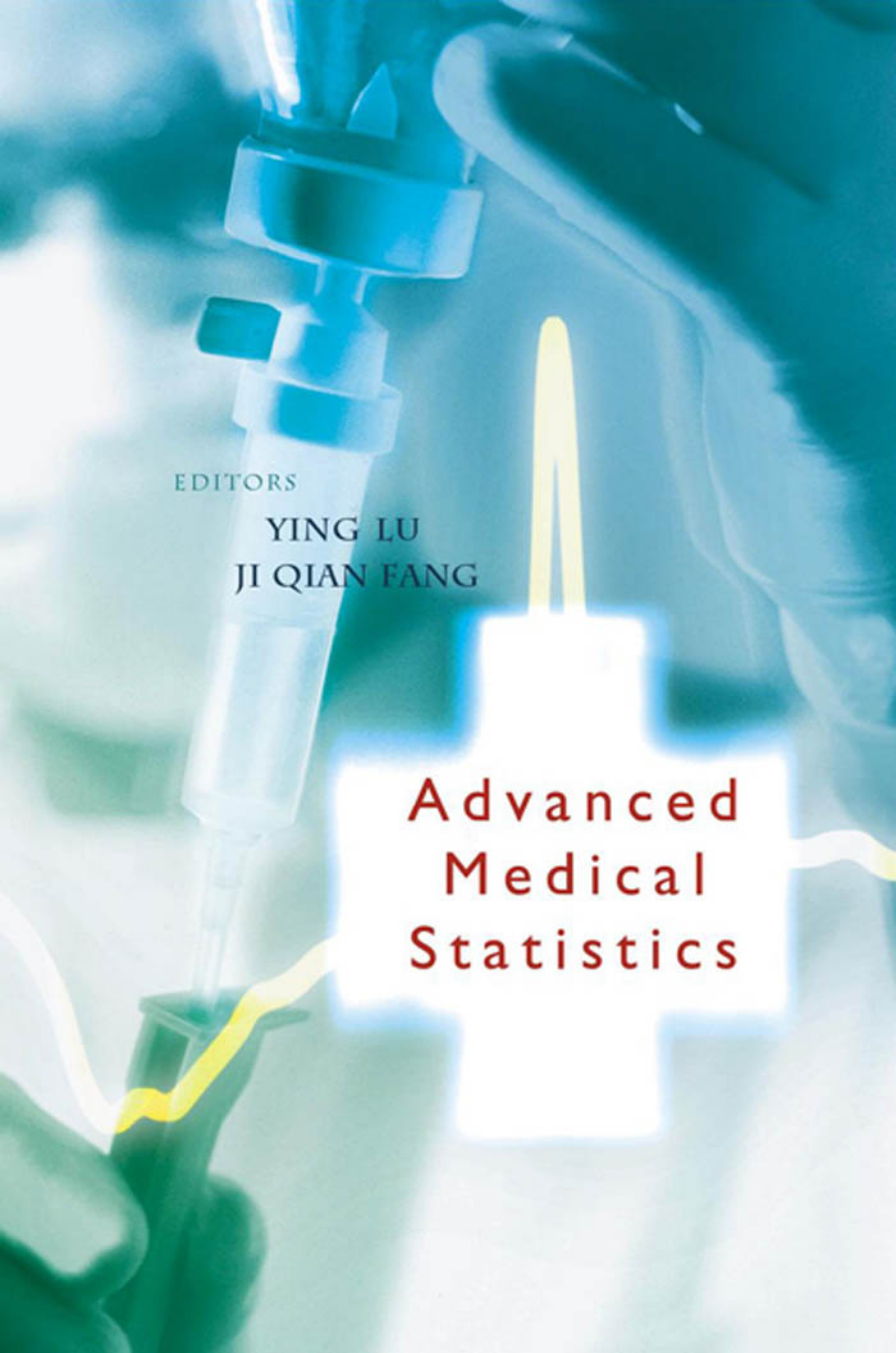

EDITORS
YING LU
JI QIAN FANG
Advanced
Medical
Statistics

Advanced Medical Statistics

Advanced Medical Statistics

EDITORS

YING LU
University of California, San Francisco, USA

JI-QIAN FANG
Sun Yat-Sen University, Guangzhou, China

World Scientific
New Jersey • London •• Singapore •• Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: Suite 202, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ADVANCED MEDICAL STATISTICS

ISBN 981-02-4799-0
ISBN 981-02-4800-8 (pbk)

Printed in Singapore.

This page intentionally left blank

PREFACE

Since the early last century, many scholars from China have studied statistics in Western countries. Some of the early pioneers, including P.L. Hsu, C.L. Chiang, C.C. Lee, K.L. Chung, and G. Tiao, etc., achieved international recognition for their significant contributions to advanced statistics. Since the 1960s, many students from Taiwan, Hong Kong, and Mainland China have received their advanced degrees from universities in North America and Europe. Some have remained, becoming professors in academia or scientists in government or industry and making significant contributions to the fields of statistics and biostatistics. Many have been elected as fellows of the American Statistical Association and/or senior members of International Biometric Society. Others have become editors or associate editors for important journals, including the *Annals of Statistics*, the *Annals of Probability*, the *Journals of the Royal Statistical Society*, the *Journal of American Statistical Association*, *Biometrika*, *Biometrics*, and *Statistica Sinnica*, etc. Several Chinese statisticians have been honored with the COPSS award, among whom Professor T.L. Lai and J. Fan have participated in the creation of this book. Meanwhile, many young statisticians have trained in Mainland China. They have accumulated a rich store of experience in teaching biostatistics and applying its theory and methods to medical research in their home country. Many overseas Chinese statisticians as well as statisticians in Mainland China, Taiwan and Hong Kong participated in publishing a book in Chinese about advances in medical statistics, which was published in 2000 by The People's Health Press, Beijing. Now, with the help of World Scientific Publishing Co, we are pleased to present the English version of this book — "*Advanced Medical Statistics*" — with a much larger professional community of English readers.

The book consists of four sections and 29 chapters. The first section is about statistical methods in biomedical research, including their history and statistical thinking in medical research, medical diagnoses, dependent

data, quality control and quality assurance in medical measurements, cost-effective and evidence-based medicine, quality of life, meta analysis, descriptive statistics, medical image processing, and time series. Many of these statistical methods were developed specifically for specific medica issues. The second section covers the most important statistical issues in pharmaceutical research and development, including pharmacology and pre-clinical studies, biopharmaceutical research, toxicological study, and confirmative clinical trials. Some of the theory and methods are published here for the first time. The third section is concerned with statistical methods in epidemiology, including statistics in genetic studies, risk assessment, infectious diseases, disease surveys, capture-recapture models for monitoring epidemics, cancer screening, and causal inferences. Most of the methods have been newly developed within the past decades. The last section is dedicated to advanced statistical theory and methods, including survival analysis, longitudinal data analysis, non-parametric curve estimation, Bayes statistics, stochastic processes, tree structured methods, EM algorithms, and artificial neural networks. These last chapters not only summarize the current status of research, future research topics and applications in medical research, but also provide some necessary theory and background for the statistical methods discussed in the first three sections.

All the chapters in the book are independent of each other; each is dedicated to a specific issue. To meet the needs of different readers, all chapters have a similar structure. The first subsection introduces the general concepts and the medical questions discussed in the chapter; examples are usually given in this section. The following sections present more specific details of concepts, methods and algorithms with the emphasis on application and significance. Derivations of proofs are generally not included, but citations in the literature are provided for interested readers.

This book is targeted to a broad readership. We hope that regardless of your background whether as a physician, a researcher in bioscience, a professional statistician, or a graduate student, you will find the book appropriate to your needs. As statistical thinking and methods are essential tools in modern medicine and biomedical research, medical researchers, leaving aside the statistical derivations and mathematical arguments, will learn what statistical tools are available to them, how to prepare the necessary information to use these methods, and how to interpret statistical results and their limitations. For professional medical statisticians, this book provides a broad perspective on medical statistics, their possible applications and interactions between special subjects, and suggestions

about future research topics, which will be helpful to their research as well as in consultation work with clients. For theoretical statisticians or applied statisticians working in other areas, the book provides many examples of statistical applications and challenges facing medical statistics, and which should help theoretical statisticians to identify new frontiers and possible application areas of their new methods. Last but not least, this book is a good reference for graduate students, providing a broad overview of medical statistics that will help them to select their research topics and guide them into the heart of the issue.

All the authors are experts in their specific areas. Each chapter reflects their own research experience, results and achievements. They have given much under the tremendous pressures of their many other obligations. As editors, we greatly appreciate their support, dedications and friendship.

Many thanks to our colleagues in the School of Public Health, Sun Yat-Sen University, who provided assistance in the preparation of the book, especially Dr. Yu Chuanhua, Dr. Yan Jie, Dr. Wang Xianhong, Dr. Ling Li, Dr. Xu Zongli, Mr. Shuming Zhu, Ms. Shaomin Wu and Ms. Fangfang Zeng. We thank the People's Health Press, Beijing, for kindly permitting us to freely publish versions other than the Chinese ones. We are most appreciative to the editors of World Scientific Publishing Co, Singapore, for their work in bringing this book to publication.

Ying Lu
Jiqian Fang
Editors

This page intentionally left blank

ABOUT THE EDITORS

Ying Lu is an associate professor of Radiology at the Department of Radiology and the director of the Biostatistics Core, UCSF Comprehensive Cancer Center, and faculty of Bioengineering Graduate Program, University of California, San Francisco. He received his BS in mathematics from Fudan University (1982) and MS in applied mathematics from Shanghai Jiao Tong University (1984), and PhD in biostatistics from the University of California, Berkeley (1990). At Berkeley, he received university fellowships (1985–1988), and Public Health Alumni Association Scholarship (1989). In 1990, he received Evelyn Fix Memorial Medal for excellent statistical dissertation on animal carcinogenicity experiments under guidance of Professors Manali and Chiang, followed by being an assistant professor of epidemiology and public health at the University of Miami School of Medicine (1990–1993). Then, he moved to the Department of Radiology at the University of California, San Francisco in 1994. He was the director of the Biostatistical Laboratory in the Osteoporosis Research Group specialized in statistical applications in quality control, clinical trial and diagnosis of osteoporosis; a member of the International Committee for Standards in Bone Measurement (1996–1998), Vice President (1995–1997) and President (1999) of the San Francisco Bay Area Chapter, American Statistical Association.

Dr. Lu has supervised two post-doctor fellows in biostatistics and more than 20 fellows in radiology and bioengineering. He has authored or co-authored more than 80 peer-reviewed articles and 4 book chapters in statistical methods for animal carcinogenicity experiments, medical diagnostic tests, and outcome prediction, as well as clinical research areas of radiology, osteoporosis, and cancer clinical trials. His papers have been published in various journals, such as *Biometrics*, *Statistics in Medicine*, *Mathematical Biosciences*, *Medical Decision Making*, *Radiology*, *Journal of Bone and Mineral Research*, *Cancer*, etc.

Ying Lu
Professor, PhD

1) Department of Radiology, Box 0629, University of California
San Francisco, CA 94143-0629, USA
ying.lu@radiology.ucsf.edu
2) Chapter 4. Statistics in Quality Control, Quality Assurance, and Quality Improvement in Radiological Studies

Ji-Qian Fang, born in Shanghai 1939, earned his BS in 1961 from the Department of Mathematics, Fudan University and PhD in 1985 from the Program of Biostatistics, the University of California at Berkeley. His PhD thesis studied multi-state survival analysis for life phenomena under the guidance of Professor Chin Long Chiang. During 1985 to 1990, Dr. Fang was a Professor and Director, the Department of Biostatistics and Biomathematics, Beijing Medical University; Since 1991, he has been the Director and Chair Professor, Department of Medical Statistics, School of Public Health, Sun Yat-Sen University. Professor Fang was a visiting professor of University of Kent, UK in 1987 and Australian National University in 1990, as well as an adjunct professor of Chinese University of Hong Kong (since 1993). He is the secretary for the Group China of the International Biometric Society and vice president of the Chinese Association of Health Statistics.

Professor Fang has published more than 100 peer-reviewed articles, monographs and text books, including "Methods of Mathematical Statistics", "Advanced mathematics", "Computer and Its Applications in Medical Field" and "Medical Statistics and Computerized Experiment."

Professor Fang has supervised 25 master students, 17 PhD students and 2 post-doctoral fellows in Biostatistics. His own and his joint research projects worked with his students cover a wide variety of fields, including "Stochastic Models of Life Phenomena", "Gating Dynamics of Ion Channels", "Biostatistical Theory and Methods for Research on Cancer Prevention", "Bootstrap Studies on Multi-state Models", "Statistical Methods for Data on Quality of Life", "Health and Air Pollution", "Analysis of DNA Finger Printing", and "Linkage Analyses between Complex Trait and Multiple Genes", etc. These projects were sponsored by either the National Foundations of China or by international organizations, such as the World Health Organization and the European Commission. Several research projects directed by Professor Fang have received awards

from the Government of Beijing Municipal Government or Ministry of Public Health of China for their significant advances in the biostatistics fields, including the projects on "Sequential Discriminant Analysis", "Multi-state Survival Analysis", "Measurement of Quality of Life in China" and "Biostatistical Theory and Methods for Research on Cancer Prevention."

Ji-Qian Fang
Professor, PhD

1) Department of Medical Statistics, School of Public Health, Sun Yat-Sen University, 74 Zhongshan Road II, Guangzhou 510080, Guangdong, PR China
fangjq@gzsums.edu.cn
2) Chapter 6. Quality of Life: Issues Concerning Assessment and Analysis
Chapter 26. Stochastic Process and Their Application in Medicine

This page intentionally left blank

Contents

This page intentionally left blank

Section 1
Statistical Methods in Biomedical Research

This page intentionally left blank

CHAPTER 1

HISTORY OF STATISTICAL THINKING IN MEDICINE

TAR TIMOTHY CHEN

Timothy Statistical Consulting, 2807 Marquis Circle East, Arlington TX 76016, USA

1. Introduction

Biostatistics is a very hot discipline today. Biostatisticians are in demand in the United States. Medical researchers appreciate statistical thinking and applications. In laboratory science, clinical research and epidemiological investigation, statisticians' collaborations are sought after. In many medical journals, statisticians are asked to serve as reviewers. In NIH (National Institutes of Health) grant applications, statisticians are required to be collaborators and statistical considerations have to be incorporated. In pharmaceutical development, drug companies recruit statisticians to guide study design, to analyze data, and to prepare reports for submission to FDA (Food and Drug Administration). All in all, statistical thinking permeates medical research and health policy. But it was not this way in the beginning. This article describes the history of application of statistical thinking in the medicine.

2. Laplace and His Vision

Near the time of American independence and the French Revolution, French mathematician Pierre-Simon Laplace (1749–1827) worked on probability theory. He published many papers on different aspects of mathematical probability including theoretical issues and applications to demography and vital statistics. He was convinced that probability theory could be applied to the entire system of human knowledge, because the principal means of finding truth were based on probabilities. Viewing medical therapy as a domain for application of probability, he said that the preferred method of

treatment would manifest itself increasingly in the measure as the number of observations was increased.[1,2]

Laplace's view that the summary of therapeutic successes and failures from a group of patients could guide the future therapy was hotly debated within the medical community. Many famous physicians like Pieere-Jean-Georges Cabanis (1757–1808) claimed that the specificity of each patient demanded a kind of informed-professional judgment rather than guidance from quantitative analysis. According to their view, the proper professional behavior for physicians in diagnosing and treating disease was to match the special characteristics of each patient with the knowledge acquired through the course of medical practice. Physicians were able to judge individual cases in all of their uniqueness, rather than on the basis of quantitative knowledge. Cabanis rejected quantitative reasoning as an intellectual distraction and viewed medicine as an "art" rather than as a "science."[3]

On the other hand, other prominent physicians like Philippe Pinel (1745–1826) said that physicians could determine the effectiveness of various therapies by counting the number of times a treatment produced a favorable response. He considered a treatment effective if it had a high success rate. He even claimed that medical therapy could achieve the status of a true science if it applied the calculus of probabilities. His understanding of this calculation, however, was restricted to counting; he did not understand the detailed nature of the probability theory being developed by Laplace.[4]

3. Louis and Numerical Method

Later another prominent clinician, Pierre-Charles-Alexandre Louis (1787–1872), considered that enumeration was synonymous with scientific reasoning. He followed Laplace's proposal that analytical methods derived from probability theory help to reach a good judgment and to avoid confusing illusions. His method consisted of careful observation, systematic record keeping, rigorous analysis of multiple cases, cautious generalizations, verification through autopsies, and therapy based on the curative power of nature. He said that the introduction of statistics into diagnosis and therapy would ensure that all medical practitioners arrive at identical results.[5]

In his study of typhoid fever, which collected patient data between 1822 and 1827, Louis observed the age difference between the groups who died (50 patients with mean age 23) and who survived (88 patients with mean

age 21). He also compared the length of residency in Paris and concluded that the group which survived lived in Paris longer. More importantly, Louis studied the efficacy of bloodletting as a therapy for typhoid fever. Among the 52 fatal cases, 39 patients (75%) had been bled. The mean survival time for the bled cases was 25.5 days contrasted to 28 days for those who were not bled. Of the 88 recovery cases, 62 patients (70%) were bled, with the mean duration of disease being 32 days as opposed to only 31 days for those not bled.[6]

Louis also studied the efficacy of bloodletting in treating pneumonitis and angina tonsillaris, and found it not useful. At that time, the method of venesection was defended by Francois Joseph Victor Broussais (1772–1838), the chief physician at the Parisian military hospital and medical school. Broussais claimed that diseases could be identified by observing the lesions of organs. Then patients could be treated by bleeding the diseased organ and by low fat, since most diseases were the result of inflammation. Louis, in contrast with Broussais, emphasized quantitative results from a population of sick individuals rather than using pathological anatomy to observe disease in a particular patient. He contended that the difference between numerical results and words, such as "more or less" and "rarely or frequently," was "the difference of truth and error; of a thing clear and truly scientific on the one hand, and of something vague and worthless on the other." He also proposed the basic concept of controlled clinical trial.[7]

Louis's work created more debates before the Parisian Academies of Sciences and Medicine in the late 1830s. The triggering issue was the question of the proper surgical procedure for removing bladder stones. A new bloodless method for removing bladder stones (lithotrity) was investigated by the surgeon and urologist Jean Civiale (1792–1867). He argued that, given the fallacy of human memory, surgeons tend to remember their successful cases more than their unsuccessful ones; errors result from inexact records. He published the relative rates of death from the traditional surgical procedure and the lithotrity. The death rate of the old procedure was 21.6% (1,237/5,715); the death rate for lithotrity was 2.3% (6/257).[3]

In response to Civiale's statistical results, the Academy of Sciences established a commission in 1835 including the mathematician Simeon-Denis Poisson (1781–1840) and the physician Francois Double (1776–1842). Rejecting the attempt to turn the clinician into a scientist through the statistical method, Double believed that the physician's proper concern should remain the individual patient. He claimed it was inappropriate to elevate

the human spirit to that mathematical certainty found only in astronomy; the eminently proper method in the progress of medicine was logical not numerical analysis.[8]

During that time, Lambert Adolphe Jacques Quetelet (1796–1874) proposed a new concept of the "average man," defined as the average of all human attributes in a country. It would serve as a "type" of the nation similar to the idea of a center of gravity in physics. He formulated this idea by combining his training in astronomy and mathematics with a passion for social statistics. He analyzed the first census of Belgium (1829) and was instrumental in the formation of the Royal Statistical Society. He maintained that the concept of statistical norms could be useful to medical practice as it had been to medical research.[9] At the same time, Poisson applied probability theory to the voting patterns of judicial tribunals. He used the "law of large numbers" to devise a 99.5% confidence interval for binomial probability.[10]

In 1837, in a lecture delivered before the French Academy of Medicine, physician Risueno d'Amador (1802–1849) used the example of maritime insurance to illustrate why the probability was not applicable to medicine. If 100 vessels perish for every 1,000 that set sail, one still could not know which particular ships would be destroyed. It depended on other prognostic variables such as the age of the vessel, the experience of the captain, or the condition of the weather and the seas. Statistics could not predict the outcome of particular patients because of the uniqueness of each individual involved. For d'Amador, the results of observation in medicine were often more variable than in other sciences like astronomy.[11]

In the ensuing debates, Double commented that a Queteletian average man would reduce the physician to "a shoemaker who after having measured the feet of a thousand persisted in fitting everyone on the basis of the imaginary model." He also claimed that Poisson's attempts to mathematize human decision-making were useless because of the pressing and immediate concerns of medical practice.

Louis-Denis-Jules Gavarret (1809–1890), trained in both engineering and medicine, addressed the criticism of d'Amador in 1840. He maintained that the probability theory merely expressed the statistical results of inductive reasoning in a more formal and exact manner. He emphasized that statistical results were useful only if certain conditions prevailed — namely, the cases must be similar or comparable, and there must be large enough observations. He followed Poisson's example in requiring a precision of 99.5% or 212:1. He commented on the insufficient sample size in Louis' study of typhoid fever.[12]

In responding to the work of Gavarret, Elisha Bartlett (1804–1855), a professor of medicine at the University of Maryland and a student of Louis, said that the value of the numerical method was exhibited by Louis, and its true principles were developed and demonstrated by Gavarret.[13] However, the British statistician William Augustus Guy (1810–1885) in his Croonian lecture before the Royal College of Physicians in 1860, said that Gavarret's confidence interval could only be applied in rare occasions, and the results obtained from averaging a small number of cases could generally be assumed to be accurate.[14] In Germany, an ophthalmologist Julius Hirschberg (1843–1925), concerning about the number of observations required by Gavarret's assumption of 212:1 odds, he modified the formula by using a lower standard of confidence of 11:1 or 91.6%.[15]

4. Statistical Analysis Versus Laboratory Investigation

In articles published in 1878 and 1881, German physician Friedrich Martius (1850–1923) commented that the dreams of Louis and Gavarret about a new era of scientific medicine had not been fulfilled due to the general "mathematical unfitness" of the medical profession as a whole. As one trained in laboratory methods, he said that the basis for science lay in laboratory experimentation rather than mere observation and the collection of numerical data.[3]

The legacy of Louis was in his claim that the clinical physician should aspire to become a scientist. But after Louis's retirement from the medical scene by the mid 1850s, some medical researchers began to argue that the compilation of numerical results might provide some useful insights about therapy; however, these results should not posses the authoritative status as "science." Friedrich Oesterlen (1812–1877) said that "scientific" results should be the discovery of knowledge which determined the causal connections, not just the discovery of the correlation.[16]

When Joseph Lister (1827–1912) published his pioneering work with antiseptic surgery in 1870, he noted that the average mortality rate was 45.7% (16/35) for all surgical procedures performed at the University of Edinburgh in the years 1864–1866 (before antiseptic methods were introduced). And it was 15% (6/40) for all surgical procedures performed in the three-year period 1867–1869 (after the introduction of antiseptic methods). Although he used this statistical result to show the efficacy of the new antiseptic method, he claimed that the science behind this was the germ theory of disease as proposed by Louis Pasteur (1822–1895).[17] Pasteur developed the

germ theory and the concept of immunity. He carried out a clinical trial in 1881 to test his new vaccine against anthrax.

The founder of 19th century scientific positivism, Auguste Comte (1798–1857), believed that mere empiricism (as practiced by Louis) was not really useful for medicine.[18] Claude Bernard (1813–1878) proposed that the science of medicine resided in experimental physiology, rather than observational statistics. As a result of his laboratory-based orientation, he claimed that the experimental investigation of each individual patient could provide an "objective" scientific result. He agreed with Louis's vision of medicine as a science but saw the science of medicine as focused on the physiological measurements of individual patients.[19]

Other prominent clinicians at that time, like German Carl Wunderlich (1815–1877), tried to steer a middle ground between Louis and Bernard and synthesized both approaches. They collected a mass of quantifiable physiological data and tried to analyze it using numerical method. However, this approach was not accepted by the medical community in general, and many still opposed the process of quantification and remained focused on the individual patient.[20]

5. The Beginning of Modern Statistics

The founders of the Statistical Society in London in 1834 chose the motto "Let others thrash it out," thus set the general aim of statistics as data collection. Near the end of the 19th century, scientists began to collect large amounts of data in the biological world. Now they faced obstacles because their data had so much variation. Biological systems were so complex that a particular outcome had many causal factors. There was already a body of probability theory, but it was only mathematics. Prevailing scientific wisdom said that probability theory and actual data were separate entities and should not be mixed. Due to the work of the British biometrical school associated with Sir Francis Galton (1822–1911) and Karl Pearson (1857–1936), this attitude was changed, and statistics was transformed from an empirical social science into a mathematical applied science.

Galton, a half-cousin of Charles Darwin (1809–1882), studied medicine at Cambridge, explored Africa during the period 1850–1852, and received the gold medal from the Royal Geographical Society in 1853 in recognition of his achievement. After reading Charles Darwin's 1859 work *On the Origin of Species*, Galton turned to study heredity and developed a new vision for the role of science in society.[21] The late Victorian intellectual movement of

scientific naturalism gave rise to the belief that scientifically trained persons must become leaders of British intellectual culture.

Galton accepted the evolutionary doctrine that the condition of the human species could be improved most effectively through a scientifically directed process of controlled breeding. His interest in eugenics led him to the method of correlation. He applied the Gaussian law of error to the intelligence of human beings and, unlike Quetelet, was more interested in the distribution and deviations from the mean than in the average value itself.

As a disciple of Galton, Karl Pearson, the founding father of modern statistics, created the statistical methodology and sold it to the world. Pearson changed statistics from a descriptive to an inferential discipline. He majored in mathematics at King's College, Cambridge. After Cambridge, he studied German literature, read law and was admitted to bar. He became professor of mathematics at King's College, London in 1881 and at University College, London in 1883. In June 1884 at age 27 he was appointed to Goldsmid Professor of Applied Mathematics at University College, London. Biologists at that time were interested in genetics, inheritance, and eugenics. In 1892 Pearson began to collaborate with zoologist WFR Weldon, Jodrell Chair of biology at University College, and developed a methodology for the exploration of life. Two years later Pearson offered his first advanced course in statistical theory, making University College the sole place for instruction of modern statistical methods before the 1920s.[22]

Following Galton, Pearson maintained that empirically determined "facts" obtained by the methods of science were the sole arbiters of truth. He argued for the almost universal application of statistical method, that mathematics could be applied to biological problems and that analysis of statistical data could answer many questions about the life of plants, animals, and men.[23] After a paper was rejected by the Royal Society, he together with Galton and Weldon founded the journal *Biometrika* in 1901 to provide an outlet for the works he and his biometrical school generated. Under Galton's generous financial support, Pearson transformed his relatively informal group of followers into an established research institute. Although he was interested in eugenics, he tried to do objective research using statistical methods and separated his institute from the social concerns of the Eugenics Education Society.

Pearson's emphasis on the statistical relevancy to the problems of biology had very few audiences. Mathematicians despised new endeavor to develop statistical methodology, and biologists thought mathematicians

had no business meddling with such things. In 1903 Pearson wrote Galton that there were only two subscribers of *Biometrika* in Cambridge, one a personal friend of Pearson and one of Weldon. Even though his major contributions were correlational methods and chi-square goodness-of-fit test, in 1906 the *Journal of the Royal Society* refused to publish a paper because they failed to see the biological significance of a correlation coefficient. In 1911 after Galton's death, Pearson became the first Galton Professor of Eugenics at University College, London.

Pearson also attempted to build an intellectual bridge to medicine by applying the statistical methods he developed. During his lifetime, the medical profession was divided about their opinion of the usefulness of statistical reasoning. Clinicians who continued to emphasize the "art" of medicine thought that statistics added little information beyond that supplied by experience. Those who argued for the existence of a "clinical science," basing diagnosis on physiological instruments or bacteriological observation, saw statistics as a way to make observation more objective, but that did not consider that as "scientific" evidence.

6. The Beginning of Medical Statistics

Major Greenwood (1880–1949) was first to respond to Pearson's "crying need" for the medical profession to appreciate the importance of new statistical methods. At the age of 18, he entered medical school and read Pearson's *Grammar of Science*. He wrote to Pearson and applied statistical analyses to his research data while a student at London Hospital. During the academic year 1904–1905, after obtaining his license to practice medicine and publishing an article in *Biometrika*, he chose to study under Pearson. Despite Pearson's warning about the difficulty of earning a living as a biometrician, Greenwood decided to stake his professional career on the application of mathematical statistical methods to medical problems.

In debating with the bacteriologist Sir Almroth Wright (1861–1947) about the efficacy of vaccine therapy and a statistical measure called "opsonic index," Greenwood invoked the distinction between functional and mathematical error.[24] The former concerned errors in techniques of measurement, while the latter concerned inferential errors derived from the fact that data were a sample of population. When he pointed out that Wright had committed mathematical error, he got the attention of the medical community.[25] Consequently the Lister Institute for Preventive Medicine in 1903 created the first department of statistics and named him

its head. Greenwood characterized his department as dealing with problems of epidemiology and pathology, in contrast to Pearson's department at the University College, which dealt with heredity, eugenics and pure mathematical statistics. By training Greenwood, Pearson had helped to create the role of medical statistician, who as a researcher, understood both medical results and statistical methods.

Greenwood left the Lister Institute in 1920 for a position at the Ministry of Health and became affiliated with the newly created Medical Research Council (MRC). He saw his position at the medical establishment as instrumental in furthering the impact of statistical methods. Raymond Pearl (1879–1940) was Greenwood's American counterpart. He went to London to study under Pearson after finishing his PhD in biology at the University of Michigan. In 1918 Pearl began a long-standing relationship with The Johns Hopkins University as professor of biometry and vital statistics in the School of Hygiene and Public Health and as statistician at The Johns Hopkins Hospital.

By the early 1920's, Greenwood was not alone in arguing for application of modern statistics in medicine. One writer said in *the Journal of the American Medical Association* in 1920 that statistics was of great practical significance and should be required in the premedical curriculum.[26] Pearl in a 1921 article in the *Johns Hopkins hospital Bulletin* said that quantitative data generated by the modern hospital should be analyzed in cooperation with expert statistician. The arguments for using statistics in medicine were framed in terms of ensuring that medical research become "scientifically" grounded.[27]

7. Randomization in Experimentation

Besides Pearson, another founder of modern statistics was Sir Ronald A. Fisher (1890–1962). He also majored in mathematics at Cambridge and studied the theory of errors, statistical mechanics, and quantum theory.[28] By the age of 22, he published his first paper in statistics introducing the method of maximum likelihood, and three years later he wrote another paper deriving the exact sampling distribution of the Pearson correlation coefficient. He was also interested in applying mathematics to biological problems. Beginning in 1919, he spent many years at Rothamsted Experimental Station and collaborated with other researchers. He developed statistical methods for design and analysis of experiments, which were collected in his books *Statistical Methods for Research Workers*[29] and

The Design of Experiments.[30] He proposed three main principles — the essentiality of replication and randomization, and the possibility of reducing errors by appropriate organization of the experiment.

Fisher's major contribution to science was using randomization to do experiments so that the variation in the data could be accounted for in the statistical analysis, and the bias of treatment assignment could be eliminated. Greenwood characterized Fisher's ideas as "epoch-making" in an article published in 1948, the year before Greenwood's death. For Fisher, statistical analysis and experimental design were only two aspects of the same whole, and they comprised all the logical requirements of the complete process of adding to natural knowledge by experimentation.[30] In other words, in order to draw inference, statisticians had to be involved in the design stage of experiments. Fisher, when addressing the Indian Statistical Congress in 1938, said, "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of".

In addition to the new developments in statistical theory brought about by Fisher's work, changes within the organization of the MRC also facilitated the emergence of the modern clinical trial. Sir Austin Bradford Hill (1897–1991), one of Greenwood's proteges, was the prime motivator behind these Medical Research Council trials. He learned statistical methods from Pearson at University College and in 1933 became Reader in Epidemiology and Vital Statistics at the London School of Hygiene and Tropical Medicine, where Greenwood became the first professor of Epidemiology and Public Health in 1927. In 1937 the editors of *The Lancet*, recognizing the necessity of explaining statistical techniques to physicians, asked Hill to write a series of articles on the proper use of statistics in medicine. These articles were later published in book form as *Principles of Medical Statistics*.[31] Upon Greenwood's retirement in 1945, Hill took his place both as honorary director of MRC's Statistical Research Unit and as professor of medical statistics at the University of London.[32]

8. First Randomized Controlled Clinical Trial

The British Medical Research Council in 1946 began the first clinical trial with a properly randomized control group trial on the use of streptomycin in the treatment of pulmonary tuberculosis. This trial was remarkable for the degree of care exercised in its planning, execution and reporting. The trial involved patient accrual from several centers, and patients were randomized

to two treatments — either streptomycin plus bed-rest, or bed-rest alone. Evaluation of patient X-ray films was made independently by two radiologists and a clinician. This blinded and replicated evaluation of a difficult disease end-point added considerably to the final agreed patient evaluation. Both patient survival and radiological improvement were significantly better on streptomycin.[33]

Hill's work set the trend for future clinical trials where both the insight of physicians and the statistical design of professional statisticians were combined. The convergence of these two separate disciplines constituted the *sine qua non* for the emergence of the probabilistically informed clinical trials. The Laplacian vision of the determination of medical therapy on the basis of the calculus of probability had finally found fulfillment.

Hill, a non-physician, acknowledged that the medical profession was responsible for curing the sick and preventing disease, but he emphasized that experimental medicine had the third responsibility of advancing human knowledge, and the statistically guided therapeutic trial was a useful way to discharge that responsibility. Unlike earlier advocates of statistical application in medicine, Hill's work became a rallying cry for supporters of therapeutic reform on both sides of Atlantic. Among many factors that contributed to this groundswell of support, one was the proliferation of new and potent industrially produced drugs in the postwar era. Supporters argued that randomized controlled clinical trials would permit the doctors to select the good treatment and prevent undue enthusiasm for newer treatments.

To those critics who believed in the uniqueness of the individual, whether patient or doctor, LJ. Witts, Nuffield Professor of Clinical Medicine of Oxford University, said in a conference in 1959, that neither patients nor doctors were as unique as they might have wanted to believe. Witts conceded that there was a conflict of loyalties between the research for truth and the treatment of the individual. However, he pointed out that similar conflict existed between the teaching of clinical students and the treatment of the patient.[34] At the same conference, Sir George Pickering, Regius Professor of Medicine at Oxford, praised the randomized controlled clinical trials and declared that, in contrast, clinical experience was unplanned and haphazard, and physicians were victims of the freaks of chance.[35]

Americans were not slow in following the British lead in applying statistics to controlled clinical trials. Americans carried out the largest and most expensive medical experiment in human history. The trial was done in 1954 to assess the effectiveness of the Salk vaccine as a protection against paralysis or death from poliomyelitis. Close to two million children

participated, and the immediate direct cost was over 5 million dollars. The reason for such a large trial was that the annual incidence rate of polio was about 1 per 2000. In order to show that vaccine could improve upon this small incidence, a huge trial was needed. Originally, there was some resistance to the randomization, but finally about one quarter of the participants did get randomized. This randomized placebo controlled double-blind trial finally established the effectiveness of the Salk vaccine.[36]

9. Government Regulation and Statistics

Later in the early 1960s, the drug Thalidomide caused an outbreak of infantile deformity. The US FDA subsequently discovered that over two and a half million tablets had been distributed to 1,267 doctors who had prescribed the drugs to 19,822 patients, including 3,760 women of childbearing age. This evidence raised the question whether the "professional judgement" of the medical community could still be trusted. The outcry from the public led the US Congress to pass the Kefauver–Harris Bill, known as the Drug Amendments of 1962 and signed by President Kennedy on October 10, 1962. This law fundamentally altered the character of research both for the drug industry and for academic medicine. It transformed the FDA into the final arbiter of what constituted successful achievement in the realm of medical therapeutics. The FDA institutionalized clinical trials as the standard method for determining drug efficacy. By the late 1960s the double-blind methodology had become mandatory for FDA approval in the US, and the procedure had become standard in most of the other Western countries by the late 1970s.

The application of statistics in medicine has scientific authority and is seen as rising above individual opinions and possessing "objectivity" and "truth." The emergence of the randomized controlled clinical trials could be seen as a special case of a more general trend — the belief that "quantification is science." This also coincided with the change of definition about statistics as a discipline. In a book written by Stanford professors Chernoff and Moses in 1959, they said, "Years ago a statistician might have claimed that statistics deals with the processing of data. Today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty."[37]

Through the work of Hill, the father of the modern clinical trial, statistical methods slowly were adapted in medical research. The reason that clinical trials gained legitimacy was because that public at large

realized that the decisions of the medical profession had to be regulated. Only when the issue of "medical decision making" was removed from the confines of professional medical expertise into the open arena of political debate could the statistical methods gain such wide acceptance. This ascendancy of the clinical trial method reflected the close connection between procedural objectivity and democratic political culture.

Above is the evolutionary history of statistical thinking in medicine. Medical research is much more than therapeutic research, but all medical research must lead to improvement of therapeutics or prevention. From this history one can see how the application of numerical methods in medicine has been debated throughout the past two hundred years. It shows that it took a long time for good concepts and procedures to prevail in science. The debates described could be applicable to the current problems about therapeutic research in alternative and complimentary medicine. Only through learning from past experience non-orthodox medicine can be modernized quickly.

10. Epilogue

Early landmarks in clinical investigation anticipated the current methodology.[38] For example, James Lind (1716–1794) in 1753 planned a comparative trial of the most promising treatment for scurvy. However, most pre-twentieth century medical experimenters had no appreciation of the scientific method. Trial usually had no concurrent control, and the claims were totally subjective and extravagant. The publication by Benjamin Rush (1745–1813) in 1794 about the success of treatment of yellow fever by bleeding was one example.

Statistics was very influential in the development of population genetics. Johann Gregor Mendel (1822–1884), a monk in the Augustinian order, studied botany and mathematics at the University of Vienna. He carried out experiments on peas to establish the three laws of genetics — uniformity, segregation and independence. After Darwin advanced the theory of evolution, there was a great debate between the evolutionists (biometricians) and those believing in the fixation of species (Mendelians). Pearson in his series of papers, *Contributions to the Mathematical Theory of Evolution, I to XVI*, gave mathematical form to the problems of genetics and evolution. However, he held the view of continuous change and never accepted Mendelism.[39]

After reading Pearson's papers while a student at Cambridge, RA Fisher made major contributions to the field of genetics, especially he synthesized and reconciled the fixed inheritance theory of Mendel and the gradual evolution theory of Darwin.[40] He was considered as one of three founders of the population genetics, together with Sewall Wright and JBS Haldane, and he occupied an endowed chair of genetics at Cambridge University. Fisher's major contributions were the theoretical foundation of statistics including estimation and the testing of hypotheses, exact distributions of various statistics, and statistical models of natural phenomena.[41]

As mentioned in the debates between the numerical methods school and the physiological school, physiological measurement data were collected using precise instruments during the later half of the nineteenth century in conjunction with the creation of research universities. Statistical methods were developed to analyze the data coming from the laboratories. Later, the controversy between the biometrical school and the bacteriologists/immunologists in the laboratory led to the further developments of correct statistical methods to analyze laboratory data.

Before the development of modern epidemiology, John Graunt (1620–1674) started to collect data on mortality, derived the life table based on survival, and thus created the discipline of demographic statistics. William Farr (1807–1883) further improved the method of the life table and created the best official vital statistics system in the world for the Great Britain.[38]

In 1848, John Snow (1813–1858) carried out the first detailed investigation of the cholera epidemic of London. Development of the discipline of bacteriology was associated with the investigation of epidemics due to infectious agents. Mathematics and statistics were used in modeling and analysis of infectious epidemic data. Modern statistical methods were developed to investigate the epidemics of non-infectious diseases in the last half of the 20th century. Epidemiological research has become another field of statistical application. It has merged with statistical survey methods to carry out surveillance and disease monitoring, and it is called population science, in contrast to clinical and laboratory sciences.

In every field of medical research, statistical thinking and methods are used to provide insight to the data and to verify the hypotheses. The generation of new data and new hypotheses also propel developments of new statistical methodology. In the twentieth century, modern statistics as created by Pearson and Fisher has made a huge impact on the advancement of human knowledge, and its application to medicine richly demonstrates the importance of statistics.

Acknowledgment

The author would like to thank Dr. James Spivey for his input to this paper.

References

1. Laplace, P. S. (1951). *A Philosophical Essay on Probabilities*, 6th ed., trans. Frederick Wilson Truscott and Frederick Lincoln Emory. Dover, New York.
2. Todhunter, I. (1865). *A History of the Mathematical Theory of Probability*, Macmillan and Co, London.
3. Matthews, J. R. (1995). *Quantification and the Quest for Medical Certainty*, Princeton University Press, Princeton, New Jersey.
4. Pinel, P. (1809). *Traite medico-philosophique sur lalienation mentale*, 2nd ed., Paris.
5. Louis, P. C. A. (1836). *Pathological Researches on Phthisis*, trans. Charles Cowan. Hilliard, Gray, Boston.
6. Louis, P. C. A. (1836). *Anatomical, Pathological and Therapeutic Researches upon the Disease Known under the Name of Gastro-Enterite Putrid, Adynamic, Ataxic, or Typhoid Fever, etc., Compared with the Most Common Acute Diseases*, Vols. 1 and 2, trans. Henry I. Bowditch. Issac R. Butts, Boston.
7. Louis, P. C. A. (1836). *Researches on the Effects of Bloodletting in Some Inflammatory Diseases, and on the Influence on Tartarized Antimony and Vesication in Pneumonitis*, trans. C. G. Putnam. Hilliard, Gray, Boston.
8. Double, F. J. (1835). Statistique appliquee a la medecine. *Comptes rendus de lAcademie des Sciences* **1**: 281.
9. Quetelet, L. A. J. (1962). *A Treatise on Man and the Development of His Faculties*, trans. R. Knox. Research Works Series #247. Burt Franklin, New York.
10. Poisson, S. D. (1837). *Recherches sur la probabilite des jugements en matiere criminelle et en matiere civile*, Bachelier, Paris.
11. D'Amador, R. (1837). *Memoire sue le calcul des probabilites applique a la medecine*, Paris.
12. Gavarret, J. (1840). *Principes generaux de statistique medicale*. Libraries de la Faculte de Medecine de Paris.
13. Bartlett, E. (1844). *An Essay on the Philosophy of Medical Science*. Lea and Blanchard, Philadelphia.
14. Guy, W. A. (1860). The numerical method, and its application to the science and art of medicine. *British Medical Journal* **469**: 553.
15. Hirschberg, J. (1874). *Die mathematischen Grundlagen der Medicinischen Statistik, elementar Dargestellt*, Veit, Leipzig.
16. Oesterlen, F. (1852). *Medical Logic*, trans. G. Whitley. Sydenham Society, London.
17. Lister, J. (1870). Effects of the antiseptic system of treatment upon the salubrity of a surgical hospital, *The Lancet* **i**: 40.
18. Comte, A. (1864). *Cours de philosophie positive*, 2nd edn., Vol. 3, JB Bailliere, Paris.

19. Bernard, C. (1957). *An Introduction to the Study of Experimental Medicine*, trans. Henry Copley Greene. Dover, New York.
20. Wunderlich, C. A. (1871). *On the Temperature in Diseases*: *A Manual of Medical Thermometry*, trans. W. Bathurst Woodman. New Sydenham Society, London.
21. Stigler, S. M. (1986). *The History of Statistics*: *The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press, Cambridge.
22. Pearson, E. S. (1938). *Karl Pearson*, Cambridge University Press, London.
23. Pearson, K. (1911). *The Grammar of Science*, 3rd edn., Macmillan, New York.
24. Cope, Z. (1966). *Almroth Wright*: *Founder of Modern Vaccine-Therapy*, Thomas Nelson, London.
25. Greenwood, M. (1909). A statistical view of the opsonic index. *Proc. Royal Soc. Med.* **2**: 146.
26. Kilgore, E. S. (1920). Relation of quantitative methods to the advance of medical science. *J. Am. Med. Assoc.* **88**, July 10.
27. Pearl, R. (1921). Modern methods in handling hospital statistics. *The Johns Hopkins Hospital Bulletin* **32**: 185.
28. Box, J. E. (1979). *R. A. Fisher*: *The Life of a Scientist*, John Wiley and Sons, New York.
29. Fisher, R. A. (1958). *Statistical Methods for Research Workers*, 13th edn., Hafner, New York.
30. Fisher, R. A. (1960). *The Design of Experiments*, 7th edn., Hafner, New York.
31. Hill, A. B. (1991). Principles of Medical Statistics. 12th edn., Lancet Ltd., London.
32. Himsworth, Sir Harold. (1982). "Bradford Hill and Statistics in Medicine," *Statistics in Medicine* **1**: 301–302.
33. MRC. (1948). Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council Investigation, *Br. Med. J.* **769**.
34. Witts, L. J. (1960). The ethics of controlled clinical trials. In *Controlled Clinical Trials*, Blackwell Scientific Publications, Oxford.
35. Pickering, Sir George. (1960). Conclusion: The Physician. In *Controlled Clinical Trials*, Blackwell Scientific Publications, Oxford.
36. Francis, T. Jr. *et al.* (1955). An evaluation of the 1954 poliomyelitis vaccines trials — Summary Report, *American Journal of Public Health* **45**(5): 1–63.
37. Chernoff, H. and Moses, L. E. (1957). *Elementary Decision Theory*, John Wiley and Sons, New York.
38. Gehan, E. A. and Lemak, N. A. (1994). *Statistics in Medical Research*: *Developments in Clinical Trials*, Plenum Publishing Co, New York.
39. Lancaster, H. O. (1994). *Quantitative Methods in Biological and Medical Sciences*: *A Historical Essay*, Springer-Verlag, New York.
40. Fisher, R. A. (1958). *The Genetical Theory of Natural Selection*, 2nd edn., Dover, New York.
41. Fisher, R. A. (1950). *Contributions to Mathematical Statistics*, ed. WA Shewhart, John Wiley and Sons, New York.

About the Author

Tar Timothy Chen is currently President, Timothy Statistical Consulting. He was Head of Biostatistics Section and Professor of Biostatistics at University of Maryland Greenebaum Cancer Center, 1998–2001; Mathematical Statistician, National Cancer Institute (1989–1998). He received BS in Mathematics (1966) from National Taiwan University; MS (1969), PhD in Statistics (1972) from the University of Chicago. His research interests include categorical data analysis, epidemiological methods, and clinical trial methodology. He has authored or coauthored 102 research papers published in *Biometrics*, *JASA*, *Statistica Sinica*, *Statistics in Medicine*, *Controlled Clinical Trials*, *New England Journal of Medicine*, *Journal of Clinical Oncology*, *Surgery*, *Ophthalmology*, *Journal of National Cancer Institute*, etc. He is an elected fellow of American Statistical Association and American Scientific Affiliation. He was the president of International Chinese Statistical Association (1999). His biosketch appeared in *Who's Who in America* (1999, 2000, 2001, 2002). *American Men and Women of Science* (1989–1998), and *Marquis Who's Who in Cancer* (1985).

This page intentionally left blank

CHAPTER 2

EVALUATION OF DIAGNOSTIC TEST'S ACCURACY IN THE PRESENCE OF VERIFICATION BIAS

XIAO-HUA ZHOU

Division of Biostatistics,
Health Services Research and Development Center of Excellence,
University of Washington, Veterans Affairs Puget Sound Health Care System,
Building 1, Room 424 (152), 1660 S. Columbian Way, Seattle, WA 98108, USA
Tel: 206-277-3588; azhou@u.washington.edu

1. Introduction

In a rapidly changing world of advancing technology, it is very important to evaluate the relative accuracies of different diagnostic tests for both quality of care and cost containment. For example, transrectal ultrasound imaging costs $150 to $400 per examination and conventional body coil magnetic resonance imaging (MRI) costs $700 to $1200 per examination. Both MRI and ultrasound could be used to detect advanced stage prostate cancer. Rifkin *et al.*[1] have shown that the accuracy of transrectal ultrasound imaging in detecting advanced stage prostate cancer was not statistically different from that of conventional body coil MRI imaging. Thus, choosing ultrasound over MRI could save $300 to $1050 without compromising quality of care.

To evaluate the accuracy of a diagnostic test, we need to determine the disease status for each patient (present or absent) independent of the patient's test result. The procedure that establishes the patient's disease status is referred to as a gold standard. The gold standard may be based on surgery, autopsy, or clinical assessments. However, some patients who underwent the test might not have had their condition status verified by the gold standard. Usually the patients who did not have their condition status verified are not a random sample but rather are a selected group. For example, if the gold standard is based on invasive surgery, then patients with negative test results are less likely to receive the gold

standard evaluation than patients with positive test results. Although this approach may be sensible and cost-effective in clinical practice, when it occurs in studies designed to evaluate the accuracy of diagnostic tests, the estimated accuracy of the tests may be biased. This type of bias is called verification bias.[2,3] For example, in a study of the accuracy of the lactose breath hydrogen test in the diagnosis of enteropathy in children, patients with negative test results had rarely undergone jejunal biopsy, the gold standard.[4] Therefore, the estimated sensitivity and specificity based on the verified cases are subject to verification bias.

Selective disease verification can lead to serious bias in estimating the accuracy of a diagnostic test. To illustrate how verification bias operates and affects the estimated accuracy of a test, we consider a hypothetical example where we want to estimate the sensitivity of a certain stress radiographic procedure in the diagnosis of coronary artery disease.[5] We use angiography as the gold standard for coronary artery disease. Assume the actual sensitivity of the radiographic procedure (which we need to estimate) is 80%. Thus, 20% of all diseased patients will have false-negative test results. Suppose 500 patients with coronary artery disease undergo the stress test; 400 respond positively and 100 respond negatively. Since angiography is a risky and expensive procedure, instead of verifying all tested patients by angiography, only 75% of patients with a positive test undergo angiography, and 10% of patients with a negative test undergo angiography. Thus, among 400 patients who tested positive, 300 have angiography, and among 100 who tested negative, only 10 have angiography. Analysis using only those patients who have angiography would lead to the mistaken conclusion that the sensitivity of the stress test is 97% (300/310), a gross overestimation of the true sensitivity. Similarly, we can show an estimator of specificity using only verified cases can also be biased.

The magnitude of verification bias depends on the association between selection for verification and the test result. The stronger the association is, the larger the bias. For example, Drum and Christacopoulos[6] studied the accuracy of hepatic scintigraph to detect liver disease. The liver disease verification procedure was either liver biopsy, exploratory laparotomy, or autopsy. In their study, they performed 650 scans (429 positive and 221 negative). Among the 429 patients with positive test results, 61% received the disease verification procedure, and of the 221 patients with negative test results, only 37% received the disease verification procedure. Using only disease verified cases, Drum and Christacopoulos reported that the hepatic scintigraph had a sensitivity of 90% and a specificity of 63%.

However, using Begg and Greenes's verification bias correction procedure (to be discussed in more detail in Sec. 2.2)[2] on all tested patients, the corrected sensitivity and specificity are 84% and 74%, respectively. Thus, the reported sensitivity in Drum and Christacopoulos's paper is inflated by 6%. A more extreme example is found in a study by Marshall *et al.*[8] Their study assessed the accuracy of diaphanography in detecting breast cancer. A total of 833 patients were tested for breast cancer using diaphanography (67 positive and 766 negative results). The verification procedure on breast cancer was biopsy. The proportion receiving the disease verification procedure was 55% for test positive patients and 7% for test negative patients. Using only verified cases, Marshall *et al.*[8] reported a sensitivity of 79% for diaphanography in detecting breast cancer. Using Begg and Greenes' correction procedure, the estimated sensitivity becomes 28%. Thus, the reported sensitivity in the paper is grossly inflated. Therefore, ignoring verification bias could grossly overestimate the accuracy of a test, and result in the misuse of a test, leading to possible mismanagement of patient care.

Although verification bias can distort the estimated accuracy of a diagnostic test, many published studies on the accuracy of diagnostic tests fail to recognize verification bias. For example, Greenes and Begg[9] reviewed 145 studies published between 1976 and 1980 and found that at least 26% of the articles had verification bias, but failed to recognize it; Bates *et al.*[10] reviewed 54 pediatric studies and found more than one third had verification bias; and Philbrick *et al.*[11] reviewed 33 studies on the accuracy of exercise tests for coronary disease and found that 31 might have had verification bias. Finally Reid *et al.*[41] looked at 112 studies published in NEJM, JAMA, BJM and Lancet between 1978 and 1993 and found 54% had verification bias.

Since it is often unethical or impractical to verify all study patients, retrospective adjustments are needed to provide correct inferences about the accuracy of tests. Assuming a gold standard exists, in this chapter, we review available statistical methods that may be used to correct for verification bias in evaluating the accuracy of diagnostic tests. In Sec. 2, we describe bias-correction methods for making inferences about the accuracy of a single diagnostic test when its response is binary, and in Sec. 3, we discuss bias-correction methods for comparing the relative accuracy of two correlated binary tests. In Sec. 4, we discuss bias-correction methods for making inferences about the accuracy of a single diagnostic test when its response is ordinal, and in Sec. 5, we present bias-correction methods for comparing the relative accuracy of two ordinal-scale diagnostic tests. We

start each section by presenting an overview of the methods, then follow this with a more detailed discussion.

2. A Single Binary Test

2.1. *An overview*

When the response of a test is binary, its accuracy is usually measured by sensitivity and specificity or positive and negative predictive values. The sensitivity measures how good the test is at providing a positive result in diseased patients, and the specificity measures how good the test is at ruling out non-diseased patients. While sensitivity and specificity are intrinsic properties of a diagnostic test, positive and negative predictive values represent the accuracy of a diagnostic test when it is applied to a particular patient.[12] Several approaches have been developed to make inferences on the accuracy of a single binary test in the presence of verification bias.[2,7,13] Begg and Greenes[2] developed a bias-correction procedure for estimating sensitivity and specificity under the conditional independence assumption, which requires that selection for verification does not depend on the true disease status directly. Zhou[7] extended their method to allow a general model for verification process and derived the maximum likelihood estimators for sensitivity and specificity of a diagnostic test and their corresponding variances.

Even though the estimated sensitivity and specificity may be biased using only verified cases, Zhou[13] showed that under the conditional independence assumption, the naive estimators of predictive values, based on only verified cases, are unbiased. However, If the conditional independence assumption does not hold, Zhou showed that the naive estimators are still biased and derived the ML estimators under a model for the verification process.

2.2. *Estimation of a single test*

To develop a bias-correction procedure for estimating sensitivity and specificity, we define the random variables, V, T, and D, to describe the verification indicator, the value of the diagnostic test result and the true disease status of a patient, respectively. Let $V = 1$ indicate a verified patient and $V = 0$ a non-verified patient; let $T = 1$ indicate a positive test result and $T = 0$ a negative test result; and let $D = 1$ indicate a diseased patient, and $D = 0$ non-diseased. Furthermore, we assume that the probability of

Table 1. Cross-classification of test results by disease status and verification status $\mathbf{X} = \mathbf{x_i}$.

		Diagnostic results	
		$T = 1$	$T = 0$
Verified	$D = 1$	s_{1i}	s_{0i}
	$D = 0$	r_{1i}	r_{0i}
Unverified		u_{1i}	u_{0i}
Total		n_{1i}	n_{0i}

verifying a patient may be influenced by not only the test results but also the discrete covariates $\mathbf{X}$, which have I different covariate patterns. Let $\mathbf{x_i}$ denote the ith covariate pattern of observed covariates, where $i = 1, \ldots, I$. Also, assume that $\mathbf{X}$ is a random sample from a discrete space $(\mathbf{x}_1, \ldots, \mathbf{x}_I)$ with probabilities $\xi = (\xi_1, \ldots, \xi_I)$. The observed data with verification bias may be displayed as in Table 1. Under the assumption that

$$k_{1i} = \frac{P(V = 1 \mid D = 1, T = 1, \mathbf{X} = \mathbf{x_i})}{P(V = 1 \mid D = 0, T = 1, \mathbf{X} = \mathbf{x_i})} \quad \text{and}$$

$$k_{0i} = \frac{P(V = 1 \mid D = 1, T = 0, \mathbf{X} = \mathbf{x_i})}{P(V = 1 \mid D = 0, T = 0, \mathbf{X} = \mathbf{x_i})}$$

are known, Zhou[7] showed that the maximum likelihood (ML) estimators for sensitivity and specificity are

$$\hat{sens} = \frac{\sum_{i=1}^{I} (\hat{sens}_i)\hat{p}_i n_i / n}{\sum_{i=1}^{I} \hat{p}_i n_i / n}$$

and

$$\hat{spec} = \frac{\sum_{i=1}^{I} (\hat{spec}_i)(1 - \hat{p}_i) n_i / n}{\sum_{i=1}^{I} (1 - \hat{p}_i) n_i / n},$$

respectively, where

$$\hat{sens}_i = \frac{s_{1i} n_{1i} / (s_{1i} + k_{1i} r_{1i})}{s_{1i} n_{1i} / (s_{1i} + k_{1i} r_{1i}) + s_{0i} n_{0i} / (s_{0i} + k_{0i} r_{0i})},$$

$$\hat{spec}_i = \frac{k_{0i} r_{0i} n_{0i} / (s_{0i} + k_{0i} r_{0i})}{k_{1i} r_{1i} n_{1i} / (s_{1i} + k_{1i} r_{1i}) + k_{0i} r_{0i} n_{0i} / (s_{0i} + k_{0i} r_{0i})},$$

which are the ML estimators for sensitivity and specificity of the test in the subpopulation with $\mathbf{X} = \mathbf{x_i}$, respectively,

$$\hat{p}_i = \frac{n_{1i}}{n_i} \frac{s_{1i}}{s_{1i} + k_{1i} r_{1i}} + \frac{n_{0i}}{n_i} \frac{s_{0i}}{s_{0i} + k_{0i} r_{0i}},$$

Table 2. Hepatic scintigraph data.

		Diagnostic results	
		$T = 1$	$T = 0$
$V = 1$	$D = 1$	231	27
	$D = 0$	32	54
$V = 0$		166	140
Total		429	221

$n_i = n_{1i} + n_{0i}$, and $n = \sum_{i=1}^{I} n_i$.[7] The corresponding variances may be computed from the inverse of the Fisher information matrix. If $k_{1i} = k_{0i} = 1$ for $i = 1, \ldots, I$, the conditional independence assumption holds, and our ML estimators reduce to the ones given by Begg and Greenes.[2]

2.3. *An hepatic scintigraph example*

Hepatic scintigraph is an imaging scan used in detecting liver disease. Drum and Christacopoulos[6] conducted an experiment to determine the sensitivity and specificity of the hepatic scintigraph in detecting liver disease. There were 650 patients who participated in the study. Of the 429 patients who had positive hepatic scintigraph results, 263 (61%) were referred to undergo a disease verification procedure, which is liver pathology. Of the 221 patients with negative hepatic scintigraph results, only 81 (37%) were referred to undergo the disease verification procedure. The data are presented in Table 2. If only patients with verified condition statuses are used in the calculation, the biased estimate of sensitivity is 0.90 with a 95% confidence interval of (0.86, 0.93); and the biased estimate of specificity is 0.63 with the 95% confidence interval of (0.53, 0.73). If the probability of verifying a patient depends only on the test results of the hepatic imaging scan, the verification process is MAR. Using the correction method described in Proposition 1, the estimated sensitivity is 0.84 with a 95% confidence interval of (0.79, 0.88), and the estimated specificity is 0.74 with a 95% confidence interval of (0.66, 0.81).

3. Comparison of Two Correlated Binary Tests

3.1. *An overview*

To compare the relative accuracies of two binary tests, several approaches have been developed to correct for verification bias.[14–16] Schartzkin *et al.*[14] considered the comparison of sensitivities and specificities of two tests in

an extreme case of verification bias where only those patients who tested positive on either test proceeded to have their true disease status verified, and they found that McNemar's test could still be used for such a comparison. Baker[15] proposed a parametric maximum likelihood procedure for estimating sensitivities and specificities of multiple tests, and Zhou[16] provided a nonparametric ML approach for comparing the relative accuracies of two correlated binary tests. While both Baker's method and Zhou's method treat the problem of verification bias as a special type of missing-data and use likelihood-based approaches for missing-data to correct for verification bias, Baker's approach primarily focused on estimation of sensitivities and specificities of multiple tests, and Zhou's approach focused on hypothesis testing for the equality of sensitivities or specificities of two diagnostic tests. Although Baker's approach allows the disease verification process to depend on the disease status, its validity still depends on the assumption that one can correctly model the disease verification process by logistic regression using the test results and the disease status. While the validity of Zhou's approach relies on the assumption that the disease verification process depends on only the test results and other observed covariates, but not on the disease status, its validity does not require modeling the diseased verification process. However, his approach assumes that the effects of covariates on disease follow a logistic regression model. Baker implemented his approach by first starting with an EM algorithm and then switching to the Newton–Raphson algorithm after a few iterations,[15] and Zhou implemented his method using the Newton–Raphson algorithm. The advantage of Zhou's approach is that the computation may be done in an existing software, such as SAS;[17] and the advantage of Baker's approach is that it may be less sensitive to starting values, but its disadvantage is that it needs a special program to carry out its computation.

3.2. *The ML approach*

In this subsection, we discuss Zhou's approach[18] for comparing the relative accuracies of two correlated binary tests. Let T_1 and T_2 be binary test results of two diagnostic tests. Let the definitions of random variables D, V, $\mathbf{X}$ be the same as those in Sec. 2.2. Then, the observed data may be summarized as in Table 3. To derive the bias-correction procedure, we need additional notation. Define

$$\theta_{ijl} = P(D = 1 \mid T_1 = j, T_2 = l, \mathbf{X} = \mathbf{x_i}), \quad \text{and}$$

$$\eta_{ijl} = P(T_1 = j, T_2 = l \mid \mathbf{X} = \mathbf{x_i}).$$

Table 3. Cross-classification of test results by disease status and verification status with $\mathbf{X} = \mathbf{x_i}$.

		$T_1 = 1$		$T_1 = 0$	
		$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$
$V = 1$	$D = 1$	s_{i11}	s_{i10}	s_{i01}	s_{i00}
	$D = 0$	r_{i11}	r_{i10}	r_{i01}	r_{i00}
$V = 0$		u_{i11}	u_{i10}	u_{i01}	u_{i00}
Total		n_{i11}	n_{i10}	n_{i01}	n_{i00}

Because the number of free parameters could grow uncontrollably as the number of covariates grows, we need to model the joint probability $P(T_1, T_2, D \mid \mathbf{X})$. We model $P(D \mid T_1, T_2, \mathbf{X})$ by a logistic regression model and $P(T_1, T_2 \mid \mathbf{X})$ by a multinomial logit model.[19] Specifically, these models are defined by the following equations:

$$P(D = 1 \mid T_1, T_2, \mathbf{X} = x_i) = \frac{\exp(\beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3' x_i)}{1 + \exp(\beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3' x_i)}$$

and

$$P(T_1 = j, T_2 = l \mid \mathbf{X} = x_i) = \frac{\exp(\alpha_{0jl} + \alpha_{1jl}' x_i)}{\sum_{h_1, h_2 = 0}^{1} \exp(\alpha_{0h_1h_2} + \alpha_{1h_1h_2}' x_i)},$$

for $j, l = 0, 1$, where $\alpha_{011} = 0$ and $\alpha_{111} = 0$. Let

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3')', \quad \alpha = (\alpha_{000}, \alpha_{001}, \alpha_{010}, \alpha_{100}', \alpha_{101}', \alpha_{110}')',$$

$\xi_i = P(\mathbf{X} = x_i)$, $\xi = (\xi_1, \ldots, \xi_{I-1})$. Let s_{ijl}, r_{ijl} and u_{ijl} be the numbers of subjects with $(V = 1, D = 1, T_1 = j, T_2 = l)$, $(V = 1, D = 0, T_1 = j, T_2 = l)$, and $(V = 0, T_1 = j, T_2 = l)$, respectively. Then, the log-likelihood function based on the observed data is

$$l(\alpha, \beta, \xi) = \sum_{i=1}^{I} \sum_{j,l=0}^{1} \{n_{ijl} \log \eta_{ijl} + s_{ijl} \log \theta_{ijl} + r_{ijl} \log(1 - \theta_{ijl})\}$$

$$+ \sum_{i=1}^{I} n_i \log \xi_i \,, \tag{1}$$

where $n_i = \sum_{j,l=0}^{1} n_{ijl}$, and $\xi_I = 1 - \xi_1 - \cdots - \xi_{I-1}$. After obtaining ML estimates of α, β, and ξ by maximizing Eq. (1) with respect to these

parameters, we can estimate the sensitivities of the two tests, π_1 and π_2, by the following formulas:

$$\hat{\pi}_1 = \left(\sum_{i=1}^{I} \sum_{l=0}^{1} \hat{\theta}_{i1l} \hat{\eta}_{i1l} \hat{\xi}_i \right) \Big/ \hat{p} \quad \text{and} \quad \hat{\pi}_2 = \left(\sum_{i=1}^{I} \sum_{j=0}^{1} \hat{\theta}_{ij1} \hat{\eta}_{ij1} \hat{\xi}_i \right) \Big/ \hat{p} \,,$$

respectively, and their specificities, ν_1 and ν_2, by

$$\hat{\nu}_1 = \left(\sum_{i=1}^{I} \sum_{l=0}^{1} (1 - \hat{\theta}_{i0l}) \hat{\eta}_{i0l} \hat{\xi}_i \right) \Big/ (1 - \hat{p}) \quad \text{and}$$

$$\hat{\nu}_2 = \left(\sum_{i=1}^{I} \sum_{j=0}^{1} (1 - \hat{\theta}_{ij0}) \hat{\eta}_{ij0} \hat{\xi}_i \right) \Big/ (1 - \hat{p}) \,,$$

where

$$\hat{p} = \sum_{i=1}^{I} \sum_{j,l=0}^{1} \hat{\theta}_{ijl} \hat{\eta}_{ijl} \hat{\xi}_i \,.$$

We may use the delta method to estimate the corresponding covariance matrix of $\hat{\pi}_1$ and $\hat{\pi}_2$ and that of $\hat{\nu}_1$ and $\hat{\nu}_2$.

4. A Single Ordinal-Scale Test

4.1. *An overview*

When the response of a diagnostic test is ordinal, there is more than one way to define a positive test result. Hence, the use of one pair of sensitivity and specificity values confounded with the chosen confidence threshold for a positive result. To overcome this limitation, a receiver operating characteristic (ROC) curve was proposed to present the accuracy of an ordinal-scale test.[20,21] An ROC curve is a plot of 1-*specificity* versus *sensitivity* as one varies the confidence threshold from the most liberal to the most conservative views on the presence of disease, and it shows the trade-off between sensitivity and specificity of a diagnostic test that can arise when one uses different confidence thresholds.[22] For estimating a single ROC curve, several bias-correction methods have been proposed.[18,23–26] Gray *et al.*[23] proposed a parametric maximum likelihood (ML) approach for estimating an ROC curve, adjusting for verification bias, under the conditional independence assumption that the selection probability of verification depends only on the test results. Their approach has three limitations: (1) it cannot be applied to the situation where some observed covariates (e.g. sex or age) may

influence the decision to verify a patient and/or affect the ROC curve itself; (2) the validity of the approach relies on the normality assumption of the latent decision variables; and (3) computation of the ML estimates requires a modified iterative scoring algorithm. Hunink *et al.*[24] proposed an ad-hoc method for estimating an ROC curve adjusting for the effects of covariates on the decision to verify and on the ROC curve. However, their approach does not necessarily provide maximum likelihood estimates nor consistent variance estimates, as shown by Rodenberg.[26] Rodenberg and Zhou[25,26] proposed a likelihood based approach for estimating an ROC curve when some observed covariates affect both the verification process and the test's accuracy. Their approach first modeled effects of covariates on the accuracy of a test by an ordinal regression model, then treated the verification bias problem as a missing-data problem, and finally used the EM algorithm[27] to compute the ML estimates under the MAR assumption for the verification process. To overcome the second and third limitations of the Gray *et al.*'s approach, Zhou[18] proposed a non-parametric maximum likelihood approach to correct for verification bias in estimating the area under an ROC curve. The main idea behind this approach was to treat the verification bias problem as a missing data problem. Under the missing data framework, he first derived an explicit expression for a ML estimator of the ROC curve area without the normality assumption. Then, he presented two approaches for estimating the corresponding variance. The first approach was based on the observed Fisher information,[28] called the information method, and the second approach was based on the jackknife method.[29] A simulation study suggests that the estimator obtained using the jackknife method outperforms the estimator obtained by the information method. The proposed approach does not require an iterative algorithm to compute the ML estimates, nor the normality assumption of the latent decision variable. The proposed approach can also apply to the setting whether some observed discrete covariates of a patient might influence the decision to verify the patient. However, this approach can only apply to the area under the ROC curve, not the ROC curve itself.

4.2. *Estimation of a single ROC curve without covariates*

Let T be the ordinal-scale test results, and the definitions of random variables D and V are the same as those in Sec. 2. Then, the observed data may be summarized as in Table 4. The rating data above may be considered as a categorization of an unobserved latent random variable T^*, representing

Table 4. Cross-classification of an ordinal-scale test by disease status and verification indicator.

		Diagnostic results		
		$T=1$	$\cdots$	$T=K$
Verified	$D=1$	z_{11}	$\cdots$	z_{K1}
	$D=0$	z_{10}	$\cdots$	z_{K0}
Unverified		u_1	$\cdots$	u_K
Total		n_1	$\cdots$	n_K

degree of suspicion on the presence of disease for a patient. By postulating a relationship between the observed T and the unobserved T^* and a parametric distribution for T^*, one can build a parametric model for the ROC curve of a diagnostic test. Several ROC models have been proposed in the literature.[30–33] The most commonly used binormal model is proposed by Dorfman and Alf,[30] and this model can be summarized in the following result.

Result 1: *Assume that* $K-1$ *cut-off points,* $\theta_1, \ldots, \theta_{K-1}$, *exist such that for each patient, if* $\theta_{k-1} < T^* \leq \theta_k$, $T = k$, *where* $k = 1, \ldots, K$, $\theta_0 = -\infty$, *and* $\theta_K = \infty$. *Further assume that given that a patient is diseased,* T^* *is normally distributed with mean* μ_1 *and variance* σ_1^2, *and that given that a patient is non-diseased,* T^* *is normally distributed with mean* μ_0 *and variance* σ_0^2. *Under these assumptions, the ROC curve of the test is a plot of* $1-\Phi(t)$ *versus* $1-\Phi(bt-a)$, *where* $\Phi(.)$ *is the cumulative distribution function of the standard normal random variable,* $a = (\mu_1 - \mu_0)/\sigma_1$, *and* $b = \sigma_0/\sigma_1$.

Hence, under the binormal model, an ROC curve is determined by two parameters, a and b, which may be estimated using the maximum likelihood method. To write down the likelihood function, one first defines the parameters for the observed data: $p_d = P(D = d)$ and $\pi_{kd} = P(T = k \mid D = d)$, and then one may write π_{kd} as functions of the parameters of an ROC curve:

$$\pi_{k0} = \Phi(\theta'_{k-1}) - \Phi(\theta'_k) \quad \text{and} \quad \pi_{k1} = \Phi(b\theta'_{k-1} - a) - \Phi(b\theta'_k - a)\,,$$

where $\theta'_k = (\theta_k - \mu_0)/\sigma_0$, $k = 1, \ldots, K$. Notice that the probability of having $T = k$ and $D = d$ for a verified patient is $p_d\pi_{kd}$ and that the probability of having T=k for an unverified patient is $\sum_{d=0}^{1} p_d\pi_{kd}$, a mixture of two

distributions. Hence, under the MAR assumption that $P(V = 0 \mid T, D) = P(V = 0 \mid T)$, the log-likelihood for the observed data may be written as

$$\sum_{k=1}^{K}\sum_{d=0}^{1} z_{kd} \log p_d \pi_{kd} + \sum_{k=1}^{K} u_k \log(p_1\pi_{k1} + p_0\pi_{k0}) \,.$$

Gray *et al.*[23] employed a modified scoring algorithm to maximize this log-likelihood function with respect to a, b, and θ'_k to obtain their ML estimates and the corresponding variance estimates.

4.3. *Estimation of ROC curves with covariates*

Let $\mathbf{X}$ be the vector of observed covariates that may affect the verification process and the accuracy of the test. Assume that $\mathbf{X}$ can be cross-classified into I distinct combinations, and $\mathbf{x_i}$ represents the values of the covariates for the ith combination. The observed data with $\mathbf{X} = \mathbf{x_i}$ form a contingency table, displayed in Table 5.

Using the Rodenberg and Zhou's approach,[25,26] we model the effects of the covariates $\mathbf{X} = \mathbf{x_i}$ on the distribution of the response of a diagnostic test by ordinal regression with a probit link[34,35]:

$$\sum_{j\le k} \pi_{jdi} = \Phi\left(\frac{\theta_k - (\alpha_D d + \alpha'_X \mathbf{x_i})}{\exp(\beta_D d + \beta'_X \mathbf{x_i})}\right), \quad \text{for } k = 1, \ldots, K-1 \,,$$

where $\pi_{jdi} = P(T = j \mid D = d, \mathbf{X} = \mathbf{x_i})$, and θ_k's are cut-off points of a latent continuous variable T^*, defined in Proposition 1. Denote $\alpha = (\alpha_D, \alpha_X)$, $\beta = (\beta_D, \beta_X)$, and $\theta = (\theta_1, \ldots, \theta_{K-1})$. To emphasize the dependence of π_{jdi} on α, β, and θ, we write $\pi_{jdi} = \pi_{jdi}(\alpha, \beta, \theta)$. Hence, under the MAR assumption, that $P(V = 0 \mid D, T, X) = P(V = 0 \mid T, X)$, the

Table 5. Observed data for the verification bias problem when $\mathbf{X} = \mathbf{x_i}$.

Verification Status V:	Disease Status D:	Diagnostic Test Result T: 1	2	$\cdots$	K
	1	z_{11i}	z_{21i}	$\cdots$	z_{K1i}
	0	z_{10i}	z_{20i}	$\cdots$	z_{K0i}
	missing	u_{1i}	u_{2i}	$\cdots$	u_{Ki}
		n_{1i}	n_{2i}	$\cdots$	n_{Ki}

log-likelihood, based on the observed data, may be written as

$$l = \sum_{k=1}^{K}\sum_{d=0}^{1}\sum_{i=1}^{I} z_{kdi}\log(p_{di}\pi_{kdi}(\gamma,\alpha,\theta)) + \sum_{k=1}^{K}\sum_{i=1}^{I} u_{ki}\log(p_{1i}\pi_{k1i}(\gamma,\alpha,\theta) + p_{0i}\pi_{k0i}(\gamma,\alpha,\theta))\,, \tag{2}$$

where $p_{di} = P(D = d \mid \mathbf{X} = \mathbf{x_i})$, the prevalence rate of disease specific to the subgroup with $\mathbf{X} = \mathbf{x_i}$. The log-likelihood, based on the observed data, has a complicated form, involving mixture distributions.

Let w_{kdi} be the number of unverified patients with $T = k$ and $\mathbf{X} = \mathbf{x_i}$ whose disease status is $d(D = d)$. Because of selective verification, one does not observe w_{kdi}, but instead one observes $u_{ki} = w_{k0i} + w_{k1i}$. If all subjects had been verified, we would have observed w_{kdi}, and a much simpler complete-data log-likelihood could be written as

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{d=0}^{1}\{z_{kdi} + w_{kdi}\}\log(p_{di}) + \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{d=0}^{1}\{z_{kdi} + w_{kdi}\}\log(\pi_{kdi}(\alpha,\beta,\theta))\,.$$

These two separate sums suggest that p_{di} and π_{kdi} can be maximized separately. They also suggest the use of the EM algorithm with a maximization step for an ordinal regression model of $\pi_{kdi}(\alpha,\beta,\theta)$ with w_{kdi} assumed known, which can be done fusing an existing computer program PLUM developed by McCullagh.[34] Here, the expectation step finds new estimates of w_{kdi} given the current values of α, β, θ, and p, $\alpha^{(m)}$, $\beta^{(m)}$, $\theta^{(m)}$, and $p^{(m)}$, using

$$u_{ki}\frac{p_{di}^{(m)}\,\pi_{kdi}(\alpha^{(m)},\beta^{(m)},\theta^{(m)})}{\sum_{d=0}^{1} p_{di}^{(m)}\,\pi_{kdi}(\alpha^{(m)},\beta^{(m)},\theta^{(m)})}\,.$$

This iterative process is continued until the relative change in successive ML estimates is small. The convergent values are the ML estimates of the parameters. Their asymptotic variance-covariance matrix is given by the inverse of the expected information matrix, defined by Eq. (2).

4.4. *Estimation of the area under an ROC curve*

If one is interested in the area under the ROC curve, a simple bias-correction procedure is available.[18] Define

$$\phi_{1ki} = P(T = k \mid \mathbf{X} = \mathbf{x_i}), \quad \text{and} \quad \phi_{2ki} = P(D = 1 \mid T = k, \mathbf{X} = \mathbf{x_i}),$$

The log-likelihood, defined in (2), may be re-written as

$$\sum_{i=1}^{I}\sum_{k=1}^{K} n_i \log(\phi_{1ki}) + \sum_{i=1}^{I}\sum_{k=1}^{K}(z_{k1i}\log(\phi_{2ki}) + z_{k0i}\log(1-\phi_{2ki})).$$

Maximizing the above log-likelihood yields the following ML estimators for ϕ_1 and ϕ_2:

$$\hat{\phi}_{1ki} = \frac{n_{ki}}{n_i} \quad \text{and} \quad \hat{\phi}_{2ki} = \frac{z_{k1i}}{z_{k1i} + z_{k0i}},$$

where $n_i = \sum_{k=1}^{K} n_{ki}$.

Notice that the area under an ROC curve A is a function of ϕ and ξ and can be written as

$$A = \frac{\begin{array}{r}\sum_{k=1}^{K-1}\sum_{j=k+1}^{K}\sum_{i=1}^{I}(1-\phi_{2ki})\phi_{1ki}\xi_i \sum_{i=1}^{I}\phi_{2ji}\phi_{1ji}\xi_i \\ + \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{I}(1-\phi_{2ki})\phi_{1ki}\xi_i \sum_{i=1}^{I}\phi_{2ki}\phi_{1ki}\xi_i\end{array}}{\sum_{k=1}^{K}\sum_{i=1}^{I}(1-\phi_{2ki})\phi_{1ki}\xi_i \sum_{j=1}^{K}\sum_{i=1}^{I}\phi_{2ji}\phi_{1ji}\xi_i}$$

Substituting unknown parameters in the equation above by their ML estimates gives the following ML estimator for A:

$$A = \frac{\begin{array}{r}\sum_{k=1}^{K-1}\sum_{j=k+1}^{K}\sum_{i=1}^{I}(1-\hat{\phi}_{2ki})\hat{\phi}_{1ki}\hat{\xi}_i \sum_{i=1}^{I}\hat{\phi}_{2ji}\hat{\phi}_{1ji}\hat{\xi}_i \\ + \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{I}(1-\hat{\phi}_{2ki})\hat{\phi}_{1ki}\hat{\xi}_i \sum_{i=1}^{I}\hat{\phi}_{2ki}\hat{\phi}_{1ki}\hat{\xi}_i\end{array}}{\sum_{k=1}^{K}\sum_{i=1}^{I}(1-\hat{\phi}_{2ki})\hat{\phi}_{1ki}\hat{\xi}_i \sum_{j=1}^{K}\sum_{i=1}^{I}\hat{\phi}_{2ji}\hat{\phi}_{1ji}\hat{\xi}_i},$$

where $\hat{\xi}_i = n_i/n$.

The corresponding variance estimator can be obtained by either the jackknife method or the information method.[18]

4.5. *A real example with fever of uncertain origin*

Gray *et al.*[23] reported data from a study on the accuracy of computed tomography in differentiating focal from nonfocal sources of sepsis among patients with fever of uncertain origin. In this study only some patients were verified, depending on their CT results. Hence, this study had verification bias. Table 6 displays the data.

Table 6. Observed Data.

		$T=1$	$T=2$	$T=3$	$T=4$	$T=5$
$V=1$	$D=1$	7	7	2	3	37
	$D=0$	8	0	1	1	4
$V=0$		40	11	3	5	12
Total		55	18	6	9	53

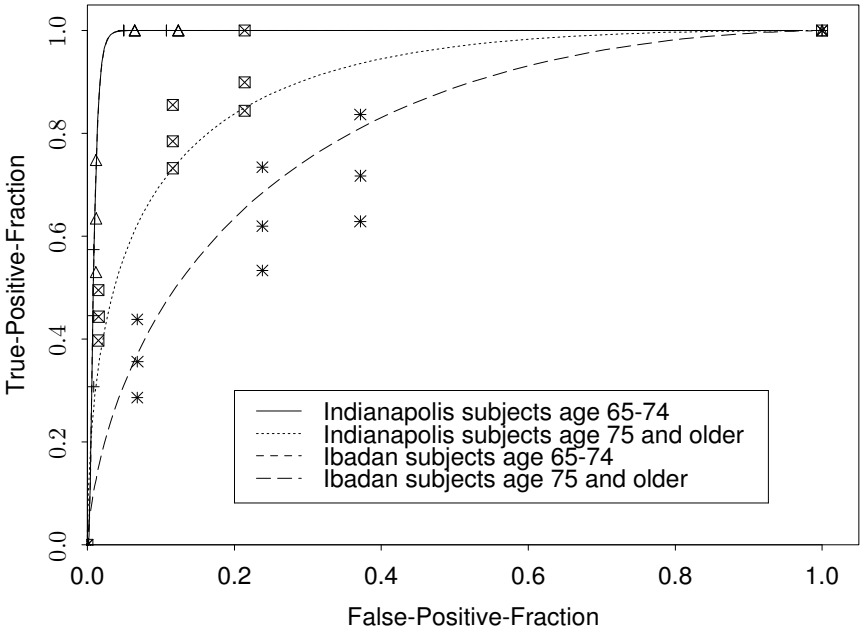


Fig. 1. ROC curves and empirical (FPF, TPF) estimates for dementia screening test by site and age group under the best model.

If we use only the verified cases, the estimated empirical and smooth ROC curves are displayed in Figure 2. The area under the smooth ROC curve is 0.75 with the standard deviation of 0.108.

If we assume that the probability of verification depends only on the result of CT, using all cases, the ML estimates of a and b are 1.80 and 1.75, respectively. We display the corrected empirical and smooth ROC curves in Fig. 2. The area under the corrected smooth ROC curve is 0.81 with standard deviation of 0.07.

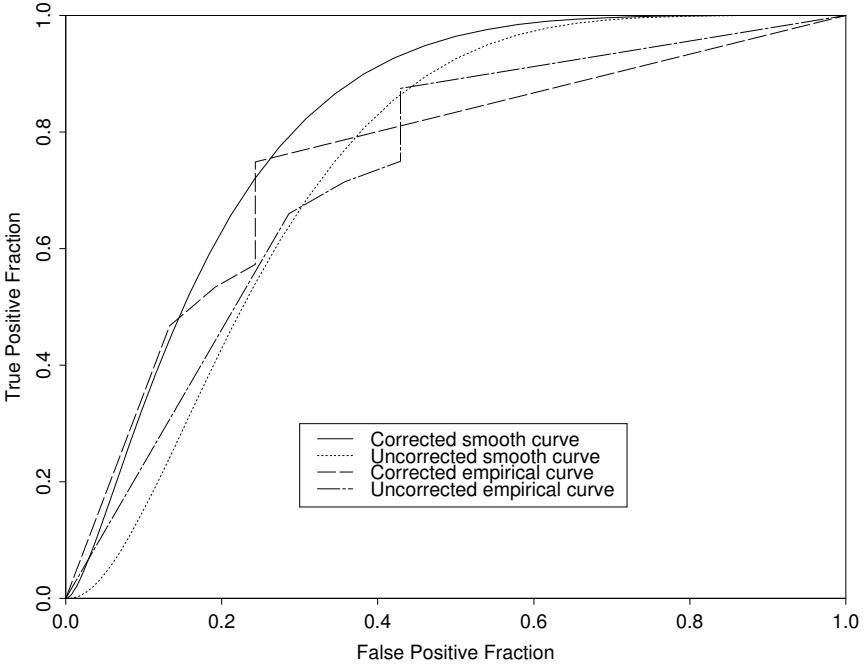


Fig. 2. Corrected and uncorrected ROC curves.

5. Comparison of Two Correlated Ordinal-Scale Tests

5.1. *An overview*

To estimate ROC curves of multiple tests, Toledano[36] adapted the idea of the weighted generalized estimation equations (GEE)[37] to correct for verification bias under the MAR assumption for the verification process. The proposed approach first modeled the verification process and then estimated the probability of verifying a patient given the patient's observed covariates, and finally, weighted verified data inversely to this estimated probability of verification. One advantage of this approach is that it permits estimation of ROC curves when some observed covariates affect the accuracy of a test and the verification process without modeling the joint probability of diagnostic test results. Two disadvantages of the approach may be: (1) its validity relies on correct modeling of the verification process; and (2) it may not be as efficient as a likelihood-based approach because it discards the unverified cases and weights the verified cases inversely to the probability of having verification.[38] For comparing two correlated ROC curve areas,

Zhou[39] extended his previous approach[18] to two correlated tests. The proposed approach first derived explicit nonparametric ML estimators for the areas under the ROC curves of two correlated tests and their corresponding variance-covariance matrix when the verification process depends on only the test results. If some categorical covariates affect the verification process, the proposed approach incorporated these covariates into the estimates of the areas by using both a logistic regression and a multinomial regression models. One strength of the proposed approach is that it does not require one to model the verification process under the MAR assumption, and its weakness is that it can only be used to estimate the areas under ROC curves, but not ROC curves themselves.

5.2. *A weighted GEE approach for ROC curves*

Let T_1 and T_2 be the responses of two diagnostic tests of a patient, ranging from 1 to K_1 and from 1 to K_2, respectively. Let the definitions of random variables D, V, and $\mathbf{X}$ be the same as those in Sec. 2.2. Then, the observed data with $\mathbf{X} = \mathbf{x_i}$ form a contingency table and are displayed in Table 7. Let T_{lj} be the result of the lth test on jth patient and Y_{ljk} be a cumulative indicator of T_{lj}. That is, $Y_{ljk} = 1$ if $T_{lj} \leq k$ and 0 otherwise. Let μ_{jlkdi} be the conditional expected value of Y_{ljk} given $D_j = d$ and $\mathbf{X} = \mathbf{x_i}$ ($\mu_{jlkdi} = E(Y_{ljk} \mid D_j = d, \mathbf{X} = \mathbf{x_i})$). One may model effects of covariates $\mathbf{X} = \mathbf{x_i}$ on μ_{jlkdi} by ordinal regression with a probit link:

$$\Phi^{-1}(\mu_{ljkdi}) = \frac{\theta_{lk} - \alpha_{Dl}d - \alpha'_{Xl}\mathbf{x_i}}{\exp(\beta_{Dl}d + \beta'_{Xl}\mathbf{x_i})}\,.$$

Let $\mathbf{B}$ be the vector of all unknown parameters, including α_{Dl}, α_{Xl}, β_{Dl}, and β_{Xl}. Using Toledano's approach[36], one estimates $\mathbf{B}$ using a two-stage procedure. First, given the test results and the observed covariates $\mathbf{X}_j$, the

Table 7. Cross-classification of ordinal-scale tests by disease and verification indicators when $\mathbf{X} = \mathbf{x_i}$.

		$T_1 = 1$			...	$T_1 = K_1$		
		$T_2 = 1$	...	$T_2 = K_2$	...	$T_2 = 1$	...	$T_2 = K_2$
$V = 1$	$D = 1$	z_{111i}	...	z_{1K_21i}	...	z_{K_11i}	...	$z_{K_1K_21i}$
	$D = 0$	z_{110i}	...	z_{1K_20i}	...	z_{K_110i}	...	$z_{K_1K_20i}$
$V = 0$		u_{11}	...	u_{1K_2}	...	u_{K_11}	...	$u_{K_1K_2}$
Total		n_{11}	...	n_{1K_2}	...	n_{K_11}	...	$n_{K_1K_2}$

probability of verifying a patient is modeled by a logistic regression model:

$$\log \frac{P(V_j = 1 \mid T_{1j}, T_{2j}, \mathbf{X_j} = \mathbf{x_i})}{P(V_j = 0 \mid T_{1j}, T_{2j}, \mathbf{X_j} = \mathbf{x_i})} = \omega'(T_{1j}, T_{2j}, \mathbf{x_i'})' .$$

The unknown parameters ω may be estimated by the method of generalized estimating equation (GEE)[40]. Denote $\hat{\omega}$ to be the resulting estimates of ω, and denote the probability of verifying the jth patient by

$$\nu_j = P(V_j = 1 \mid T_{1j},\, T_{2j},\, \mathbf{X_j} = \mathbf{x_i}) .$$

Then, the following weighted generalized estimating equation is used to estimate $\mathbf{B}$:

$$\sum_{j=1}^{n} \left(\left(\frac{V_j}{\nu_j} \frac{\partial \mu_j}{\partial \mathbf{B}} \Sigma_j(\eta)^{-1} (\mathbf{Y}_j - \boldsymbol{\mu_j}) \right) \middle| \omega = \hat{\omega},\, \eta = \hat{\eta} \right) = 0 , \tag{3}$$

where the notation $(. \,|\, \omega = \hat{\omega}, \eta = \hat{\eta})$ denotes a function of ω and η evaluated at $\hat{\omega}$ and $\hat{\eta}$; $\mathbf{Y}_j = (Y_{1j1}, \ldots, Y_{1j(K_1-1)}, Y_{2j1}, \ldots, Y_{2j(K_2-1)})'$; $\boldsymbol{\mu}_j = (\mu_{1j1}, \ldots, \mu_{1j(K_1-1)}, \mu_{2j1}, \ldots, \mu_{2j(K_2-1)})'$; $\Sigma_j(\eta) = \mathrm{cov}(\mathbf{Y}_j)$ is an assumed covariance matrix of $\mathbf{Y}_j$; and $\hat{\eta}$ is a consistent estimator of η. Toledano and Gatsonis[42] have discussed several ways of choosing the covariance matrix $\Sigma_j(\eta)$, and Toledano[36] has shown that the solution $\hat{\mathbf{B}}$ to the weighted GEE (3) is a consistent estimator of $\mathbf{B}$ and has an asymptotically normal distribution.

5.3. *A likelihood-based approach for ROC areas*

If one is interested in comparing the areas under the ROC curves, a simpler approach than the weighted GEE approach is available.[39] To derive this approach, one needs a different notation. For the observed data given as in Table 7, the following parameters are defined:

$$\phi_{2ijl} = P(D = 1 \mid T_1 = j,\, T_2 = l,\, \mathbf{X} = \mathbf{x_i}) ,$$

$$\phi_{1ijl} = P(T_1 = j,\, T_2 = l \mid \mathbf{X} = \mathbf{x_i}) , \quad \text{and} \quad \xi_i = P(\mathbf{X} = \mathbf{x_i}) .$$

Since the problem of verification bias may be considered as a missing-data problem, the likelihood approach for missing data is used to estimate ϕ_{1ijl}, ϕ_{2ijl}, and ξ_i. Assume that the missing-data mechanism is MAR, that is,

$$P(V = 0 \mid T_1,\, T_2,\, D,\, \mathbf{X} = \mathbf{x_i}) = P(V = 0 \mid T_1,\, T_2,\, \mathbf{X} = \mathbf{x_i}) .$$

Then, the contribution of a verified patient to the likelihood is $P(T_1, T_2, D, \mathbf{X}) = P(D \mid T_1, T_2, \mathbf{X})P(T_1, T_2 \mid \mathbf{X})P(\mathbf{X})$, and the likelihood contribution of an unverified case is $P(T_1, T_2, \mathbf{X}) = P(T_1, T_2 \mid \mathbf{X})P(X)$. Furthermore, one may model effects of T_1, T_2, and $\mathbf{X}$ on D by a logistic regression model:

$$\phi_{2ijl} = \frac{\exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})}{1 + \exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})}, \tag{4}$$

where $\beta_{1K_1} = 0$ and $\beta_{2K_2} = 0$, and effects of $\mathbf{X}$ on T_1 and T_2 by a multinomial logit model,

$$\phi_{1ijl} = P(T_1 = j, T_2 = l \mid \mathbf{X} = \mathbf{x_i}) = \frac{\exp(\alpha_{jl}' \mathbf{x_i})}{\sum_{h_1=1}^{K_1} \sum_{h_2=1}^{K_2} \exp(\alpha_{h_1 h_2}' \mathbf{x_i})}, \tag{5}$$

where $\alpha_{K_1 K_2} = 0$. Denote $\beta = (\beta_{11}, \ldots, \beta_{1(K_1-1)}, \beta_{21}, \ldots, \beta_{2(K_2-1)}, \beta_3')$, $\alpha = (\alpha_{11}, \ldots, \alpha_{K_1(K_2-1)})'$, $\xi_i = P(\mathbf{X} = \mathbf{x_i})$, $\xi = (\xi_1, \ldots, \xi_{I-1})$, and $n = \sum_{i=1}^{I} n_i$.

To emphasize the dependence of ϕ_{1ijl} on α and the dependence of ϕ_{2ijl} on β, one may write $\phi_{1ijl} = \phi_{1ijl}(\alpha)$ and $\phi_{2ijl} = \phi_{2ijl}(\beta)$. Under the MAR assumption, a valid log-likelihood function is

$$\begin{aligned} l(\alpha, \beta, \xi) = & \sum_{i=1}^{I} \sum_{j=1}^{K_1} \sum_{l=1}^{K_2} n_{ijl} \log \frac{\exp(\alpha_{jl}' \mathbf{x_i})}{\sum_{h_1,h_2=1}^{K} \exp(\alpha_{h_1 h_2}' \mathbf{x_i})} + \sum_{i=1}^{I} n_i \log \xi_i \\ & + \sum_{i=1}^{I} \sum_{j=1}^{K_1} \sum_{l=1}^{K_2} s_{ijl} \log \frac{\exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})}{1 + \exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})} \\ & + r_{ijl} \log \frac{1}{1 + \exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})}, \end{aligned} \tag{6}$$

where $n_{ijl} = s_{ijl} + r_{ijl} + u_{ijl}$ and $\xi_I = 1 - \xi_1 - \cdots - \xi_{I-1}$.

Let

$$l_1(\alpha) = \sum_{i=1}^{I} \sum_{j=1}^{K_1} \sum_{l=1}^{K_2} n_{ijl} \log \frac{\exp(\alpha_{jl}' \mathbf{x_i})}{\sum_{h_1,h_2=1}^{K} \exp(\alpha_{h_1 h_2}' \mathbf{x_i})},$$

$$\begin{aligned} l_2(\beta) = & \sum_{i=1}^{I} \sum_{j=1}^{K_1} \sum_{l=1}^{K_2} s_{ijl} \log \frac{\exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})}{1 + \exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})} \\ & + r_{ijl} \log \frac{1}{1 + \exp(\beta_{1j} + \beta_{2l} + \beta_3' \mathbf{x_i})}, \end{aligned}$$

and

$$l_3(\xi) = \sum_{i=1}^{I} n_i \log \xi_i \,.$$

Then, $l(\alpha, \beta, \xi)$ may be written as the sum of $l_1(\alpha)$, $l_2(\beta)$, and $l_3(\xi)$. Here, $l_1(\alpha)$, $l_2(\beta)$, and $l_3(\xi)$ may be considered as the log likelihood function of all cases modeled by the multinomial logit model defined by Eq. (5), the log likelihood function of verified cases modeled by the logistic regression model defined by Eq. (4), and the log likelihood function for a multinomial distribution based on all cases, respectively. Since the parameters α, β, and ξ are distinct, their ML estimators, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\xi}$, may be obtained by maximizing l_1, l_2, and l_3 with respect to α, β, and ξ, separately. The observed Fisher information for (α, β, ξ) is

$$\text{diag}(I_1(\alpha),\, I_2(\beta),\, I_3(\xi)) \tag{7}$$

where I_1, I_2, and I_3 are the observed Fisher information matrices on the log-likelihood functions $l_1(\alpha)$, $l_2(\beta)$, and $l_3(\xi)$, respectively.

Maximizing $l_1(\alpha)$ with respect to α and $l_2(\beta)$ with respect to β yields ML estimators $\hat{\alpha}$ and $\hat{\beta}$, respectively. Since $\xi_I = 1 - \cdots - \xi_{I-1}$, maximizing $l_3(\xi)$ with respect to ξ_i yields ML estimators of ξ_i:

$$\hat{\xi}_i = \frac{n_i}{n} \,,$$

$i = 1, \ldots, I-1$.

Note that $\gamma = P(D=1) = \sum_{i=1}^{I} \sum_{j=1}^{K_1} \sum_{l=1}^{K_2} \phi_{2ijl}\phi_{1ijl}\xi_i$ and that one may write the area under the ROC curve of a diagnostic test A_i as

$$A_i = \frac{1}{\gamma(1-\gamma)} \left[\sum_{j=1}^{K_i - 1} \phi_{i1}^*(j) \sum_{l=j+1}^{K_i} \phi_{i2}^*(l) + \frac{1}{2} \sum_{j=1}^{K_i} \phi_{1i}^*(j)\phi_{2i}^*(j) \right] ,$$

where

$$\phi_{11}^*(j) = \sum_{i=1}^{I} \sum_{k=1}^{K_2} (1 - \phi_{2ijk})\phi_{1ijk}\xi_i \,, \quad \phi_{12}^*(j) = \sum_{i=1}^{I} \sum_{k=1}^{K_2} \phi_{2ijk}\phi_{1ijk}\xi_i \,,$$

$$\phi_{21}^*(j) = \sum_{i=1}^{I} \sum_{k=1}^{K_1} (1 - \phi_{2ikj})\phi_{1ikj}\xi_i \,, \quad \phi_{22}^*(j) = \sum_{i=1}^{I} \sum_{k=1}^{K_1} \phi_{2ikj}\phi_{1ikj}\xi_i \,.$$

The delta method may be used to obtain an estimate of the covariance matrix for $\hat{A}_1$ and $\hat{A}_2$. Assuming the normality of $\hat{A}_1 - \hat{A}_2$, one can then perform the hypothesis tests and construct confidence intervals about $A_1 - A_2$.

5.4. *Availability of computer software*

For analysis of ROC data, several computer programs have been developed for carrying out computations of the bias-correction methods discussed in Secs. 4 and 5. Gray *et al.*[23] developed a program called ROCBIAS to estimate the ROC curve of a single test when the probability of verifying a patient depends on only the test results. Rodenberg and Zhou[26] developed a program called EMPLUM to estimate ROC curves when the probability of verifying a patient depends not only on the test results but also on other observed covariates. Toledano[36] developed special software written in Fortran to implement the weighted GEE approach for analyzing correlated ROC curves in the presence of verification bias. Zhou and Higgs[43] implemented the likelihood-based approach for comparing the areas under the ROC curves in SAS,[17] which can be down-loaded from *http://www.biostat.iupui.edu/~zhou*.

6. Discussion

Statistical methods in diagnostic medicine have recently received a lot of attention.[45] In this chapter we have discussed the problem of verification bias in evaluating the accuracy of diagnostic tests and some available bias-correction methods. The problem of verification bias is just one of many problems encountered in diagnostic medicine. For a completely treatment on statistical methods in diagnostic medicine, we refer readers to textbook by Zhou *et al.*[44]

References

1. Rifkin, M. D., Zerhouni, E. A., Gatsonis, C. A., Quint, L. E., Paushter, D. M., Epstein, J. I., Walsh, P. C. and McNeil, B. J. (1990). Comparison of magnetic resonance imaging and ultrasonography in staging early prostate cancer. *New England Journal of Medicine* **323**: 621–626.
2. Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease is subject to selection bias. *Biometrics* **39**: 207–216.
3. Ransohoff, D. F. and Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic Tests. *New England Journal of Medicine* **299**: 926–930.
4. Levine, J. J., Seidman, E. and Walker W. A. (1987). Screening tests for enteropathy in children. *American Journal of Diseases of Children* **141**: 435–438.
5. Tavel, M. E., Enas, N. H. and Woods, J. R. (1987). Screening tests for enteropathy in children. *American Journal of Cardiology* **60**: 1167–1169.

6. Drum, D. E. and Christacopoulos, J. S. (1972). Hepatic scintigraphy in clinical decision making. *Journal of Nuclear Medicine* **13**: 908–915.
7. Zhou, X. H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communication in Statistics — Theory and Methods* **22**: 3177–3198.
8. Marshall, V., Williams, D. C. and Smith, K. D. (1984). Diaphanography as a means of detecting breast cancer. *Radiology* **150**: 339–343.
9. Greenes, R. A. and Begg, C. B. (1985). Assessment of diagnostic technologies: Methodology for unbiased estimation from samples of selective verified patients. *Investigative Radiology* **20**: 751–756.
10. Bates, A. S., Margolis, P. A. and Evans, A. T. (1993). Verification bias in pediatric studies evaluating diagnostic tests *Journal of Pediatrics* **122**: 585–590.
11. Philbrick, J. T., Horwitz, R. I. and Feinstein, A. R. (1980). Methodologic problems of excerise testing for coronary artery disease: Groups, analysis and bias. *American Journal of Cardiology* **46**: 807–812.
12. Feinstein, A. R. (1975). On the sensitivity, specificity and discrimination of diagnostic tests. *Clinical Pharmacology and Therapeutics* **17**: 104–116.
13. Zhou, X. H. (1994). Effect of verification bias on positive and negative predictive values. *Statistics in Medicine* **13**: 1737–1745.
14. Schatzkin, A., Connor, R. J., Taylor, P. R. and Bunnag, B. (1987). Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. *American Journal of Epidemiology* **125**: 672–678.
15. Baker, S. G. (1995). Evaluating multiple diagnostic tests with partial verification. *Biometrics* **51**: 330–337.
16. Zhou, X. H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia. *Journal of Royal Statistical Society, Series C, Applied Statistics* **47**: 135–147.
17. SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 6*, SAS Institute, Inc., Cary, NC.
18. Zhou, X. H. (1996). Nonparametric ML estimate of an ROC area corrected for verification bias. *Biometrics* **52**: 310–316.
19. Agresti, A. (1990). *Categorical Data Analysis*. Wiley and Sons, New York, USA.
20. Metz, C. E. (1986). ROC Methodology in radiologic imaging. *Investigative Radiology* **21**: 720–733.
21. Swets, J. A. (1979). ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology* **14**: 109–121.
22. Hanley, J. A. (1989). Receiver operating characteristic (ROC) curve methodology. *Critical Reviews in Diagnostic Imaging* **29**: 307–335.
23. Gray, R., Begg, C. B. and Greenes, R. A. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making* **4**: 151–164.
24. Hunink, M. G. M., Richardson, D. K., Doubilet, P. M. and Begg, C. B. (1990). Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. *Medical Decision Making* **10**: 201–211.

25. Rodenberg, C. A. and Zhou, X. H. (2000). ROC curve estimation when covariates affect the verification. *Biometrics*.
26. Rodenberg, C. A. (1996). Correcting for verification bias in ROC estimation with covariates. PhD thesis, Department of Statistics, Purdue University, West Layafette, Indiana.
27. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, NY.
28. Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley and Sons, New York, NY.
29. Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall, New York, NY.
30. Dorfman, D. D. and Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: Rating data. *Journal of Mathematical Psychology* **6**: 487–496.
31. England, W. L. (1988). An exponential model used for optimal threshold selection on ROC curves. *Medical Decision Making* **8**: 120–131.
32. Grey, D. R. and Morgan, B. J. T. (1972). Some aspects of ROC curve-fitting: Normal and logistic models. *Journal of Mathematical Psychology* **9**: 128–129.
33. Ogilvie, J. C. and Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology* **5**: 377–391.
34. McCullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society, Series B* **42**: 109–142.
35. Tosteson, A. A. N. and Begg, C. B. (1985). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**: 204–215.
36. Toledano, A. Y. (1993). Generalized estimating equations for repeated ordinal categorical data, with applications to diagnostic medicine. PhD thesis, Department of Biostatistics, Harvard School of Public Health, Boston, MA.
37. Robbins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**: 846–866.
38. Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association* **92**: 1320–1329.
39. Zhou, X. H. (1998). Comparing the correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias. *Biometrics* **54**: 349–366.
40. Liang, K. Y. and Zeger, S. L. (1988). Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
41. Reid, M. C., Lachs, M. S. and Feinstein, A. R. (1995). Use of methodologic standards in diagnostic test research. Getting better but still not good. *Journal of American Medical Association* **274**: 645–651.
42. Toledano, A. Y. and Gatsonis, C. A. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* **15**: 1807–1826.
43. Zhou, X. H. and Higgs, R. (1998). COMPROC and CHECKNORM: Computer programs for comparing accuracies of diagnostic tests using ROC curves

in the presence of verification bias. *Computer Methods and Programs in Biomedicine* **57**: 179–186.
44. Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons, New York, NY.
45. Pepe, M. S. (2000). Receiver Operating Characteristic Methodology. *Journal of American Statistical Association* **95**: 308–311.

About the Author

Xiao-Hua Zhou is Director of Biostatistics Division in Puget Sound Health Care System and with a faculty appointment in the Department of Biostatistics at University of Washington. He received his BSc in Mathematics (1984) from Sichuan University, MSc in Statistics (1987) from University of Calgary, and PhD in Biostatistics (1991) from Ohio State University. He was a post-doctoral fellow in Biostatistics at Harvard University from 1991 to 1993. He joined Indiana University as Assistant Professor in 1993 and was promoted to Associate Professor in 1997. He was a visiting associate professor at Harvard University in 2000. He moved to Seattle in 2002. He was elected to the International Statistical Institute in 1998. He is the Program Chair of Section of Statistics in Epidemiology of the American Statistical Association (ASA) in 2001 and Chair-elect in 2002. In 2001, he shared Mitchell Prize from the International Society for Bayesian Analysis and Section on Bayesian Statistical Sciences of ASA with Prof. Hirano, Imbens, and Rubin. He is also an Associate Editor for *Biometrics* and Editorial Board Member for *Statistics in Medicine*. His research interests include statistical methods in diagnostic medicine, categorical data analysis, health services research, analysis of skewed data, causal inferences, analysis of observational studies, and analysis of missing data. He has published over 79 referred papers in both statistical and medical journals.

CHAPTER 3

STATISTICAL METHODS FOR DEPENDENT DATA

FENG CHEN

Department of Medical Statistics, Nantong Medical College, Nantong, Jiangsu 226001, PR China
Tel: 0086-513-5517191-2012; chenfeng@public.nt.js.cn

1. Introduction

Most classical statistical methods require independent observations. The issue here is not independence of multiple variables, rather of the samples. There are many cases of dependent in medical research when requirement of independence cannot be hold, i.e. observations are correlated.

The existence of such correlation is not a coincident, but due to the design of the experiments. In some cases, this type of correlation can be eliminated by suitable procedure without losing any information. The simplest case is paired design where the observations within the same paired is correlated. For example, to investigate a new drug's effects on hypertension, a 2-by-2 crossover design can be used to measure the diastolic pressure before and after treatment for each subject. Although the pressures across subjects are independent, the observations of the same subject are correlated.

Unfortunately, we could not eliminate intra-unit correlations in most cases by traditional statistical methods. For example, in a toxicological study, 32 pregnant rats were randomly allocated into test and control groups. Rats in control group were fed with regular food, while rats in test group were fed with combinations of regular food and suspected teratogen. The proportion of malformation of pups of two groups was compared after rat delivery. In this study, the pregnant rats are independent with each other, but genetic factors, antepartum internal womb environments and

metabolism conditions of teratogen have effects on the rat pups. Thus, the rat pups cannot be treated as independent observations because siblings are more likely to encounter the similar proportion of malformation than pups from different litters. The litter effect must be taken into account. Special procedure must be used to deal with this type of data.

The intra-unit correlation or intra-class correlation is a measure of similarity (or non-independence) among individuals that share some characteristics. The intra-unit correlation means that observations in the same unit are not dependent. There are overlaps between the information they present. It is inappropriate to ignore the intra-unit correlation. For example, in a clinical trail, many variables, i.e. vital signs, physiological index, effects and side effects, should be observed successively in different time for each subject during the trial period to show the efficacy and safety of the tested drug. Each subject should be observed several times. We refer to this type of study as repeated measurement study. There are two classical ways to deal with this sort of data. One is, to test the significance of the difference between the test group and the control group on each occasion respectively including test for the homogeneity of two groups before treatment and to compare the difference of changes (such as absolutely increase or decrease, relatively increase or decrease, etc.) between the two groups at each time. The alternative is to take k observations of each one of n subjects as one response variable, (the sample size will be nk) to fit a model (or generalized linear model) in which time is an explanatory variable. The former one will have low statistical power because it treats the observations of each occasion independently. The latter considers the correlations between the treatment effect and time. However, it ignores the intra-subject correlation of the observations and takes the data as independent data. Thus, it will increase the type I error which may result in the approval of the inefficiency drug to the market.

The set of observations taken from the same subject tend to be correlated. They provide rather less information than the same number independent observations taken from different subjects. The larger the intra-unit correlation is, the less information will be provided. Therefore, it will increase the type I error if we use nk observations to fit general linear model.

The statistical methods for dependent data are described and illustrated in this chapter. The methods cover estimation of intra-correlation coefficient, hypotheses test, estimation of sample size, etc.

Table 1. Clotting time (min) of serum from 8 volunteers, treated by 4 methods.

subject	Treatment			
	A	B	C	D
1	8.4	9.4	9.8	12.2
2	12.8	15.2	12.9	14.4
3	9.6	9.1	11.2	9.8
4	9.8	8.8	9.9	12.0
5	8.4	8.2	8.5	8.5
6	8.6	9.9	9.8	10.9
7	8.9	9.0	9.2	10.4
8	7.9	8.1	8.2	10.0

2. Examples of Dependent Data

Dependent data is omnipresent in medical researches, such as, repeated measurement data, longitudinal data, data of cross-over design, data of multicenter clinical trial, cluster sampling survey data, and infective disease, inherited disease, etc. They share the same property, which is the dependence or intra-unit correlation of observations. We refer to this type of data as dependent data. In this section, we will illustrate some types of examples for dependent data, and discuss their common and distinguishing features.

2.1. *Example* 1. *Randomized block design*

To compare the effects on the clotting time of serum of four treatments, 8 volunteers were recruited. Four samples of serum from each subject were assigned to the four treatments in a random order. The results of the experiment were presented in Table 1.

The property of the data shown in Table 1 is that the observations of the same block are correlated, while the observations from different blocks are independent. Thus, the effects of 4 treatments of 4 serums from one person are correlated. That is to say the data from block design are dependent. Observations in the same block in split-plot design and in split-split-plot design have the same property.

2.2. *Example* 2. *Cluster sampling*[1]

A simple random sample of 30 households was drawn from a census taken in 1947. The question here is whether they had consulted a doctor in the last 12 months. Data are shown below. The denominator is the number of

persons in a household, and the nominator is the number of persons who saw a doctor.

5/5, 0/5, 2/3, 3/3, 0/2, 0/3, 0/3, 0/3, 0/4, 0/4, 0/3, 0/2, 0/7, 4/4, 1/3,
2/5, 0/4, 0/4, 1/3, 3/3, 2/4, 0/3, 0/3, 0/1, 2/2, 2/4, 0/3, 2/4, 0/2, 1/4

The property of this data is that the members of the same family tend to be similar, while persons from different families are assumed to be independent. Our purpose is to estimate the proportion of people who consulted a doctor, and to measure the similarity of the members in the same family. Similar results would be obtained for any characteristic in which the members of the same family trend to act in the same way.

2.3. *Example* 3. *Toxicological study*[2]

In a toxicological study, 32 pregnant rats were randomly allocated into 2 groups: test group and control group. Rats in control group were fed with regular food, while rats in test group were fed with combination of regular food with suspected teratogen. The proportions of malformation of pups of two groups were compared after delivery. The results are shown as follows:

Control group	13/13	12/12	9/9	8/8	8/8	12/13	11/12	9/10	9/10	8/9
	11/13	4/5	5/7	7/10	7/10	9/9				
Test group	12/12	11/11	10/10	9/9	10/11	9/10	9/10	8/9	8/9	4/5
	7/9	4/7	5/10	3/6	3/10	0/7				

The denominator is the number of offspring in a litter, and the nominator is the number of offsprings that are malformation in the litter.

In this study, the pregnant rats are independent to each other, but genetic factors, antepartum internal womb environments and metabolism conditions of teratogen have effects on the rat pups. The rat pups cannot be treated as independent observations because siblings are more alike than pups from different litters. Data in Example 2 have similar property. Similar property would be obtained from genetics studies in which the members of the same family tend to be similar.

2.4. *Example* 4. *Crossover design*

For studying the bioequivalence of domestic and imported rosiglitazone maleate tablets (RMT), 24 volunteers were recruited in a 4×4 crossover study. Four sequence groups are formed by the randomized

Table 2. Results of 4×4 cross-over trial for testing bioequivalence of domestic and imported rosiglitazone maleate tablets.

Id	sequence	Stage 1			Stage 2			Stage 3			Stage 4		
		AUC	$C_{\max}$	T_{50}	AUC	$C_{\max}$	T_{50}	AUC	$C_{\max}$	T_{50}	AUC	$C_{\max}$	T_{50}
1	DCAB	884.27	204.63	3.47	905.09	222.94	3.86	2330.77	455.14	5.50	1936.98	395.55	4.35
2	CBDA	919.50	178.27	3.92	2201.98	346.89	4.58	855.89	205.31	3.79	1939.36	327.12	4.01
3	ADBC	1738.12	326.72	3.95	901.70	130.61	4.56	1889.72	375.37	4.08	870.93	158.99	4.09
4	BACD	2000.29	382.25	3.51	2350.58	479.88	3.83	952.86	187.72	3.68	955.46	202.02	3.69
5	CBDA	823.39	158.72	3.86	1864.97	329.66	3.65	710.06	133.87	3.22	1372.77	309.07	2.99
6	ADBC	2102.11	360.38	4.47	946.33	155.29	4.68	2005.84	339.67	3.53	934.45	176.23	3.72
7	DCAB	907.86	170.88	3.95	991.65	197.54	4.06	2139.65	369.21	3.98	2408.84	368.93	4.62
8	BACD	2139.72	366.84	3.80	2012.09	411.87	3.99	1134.23	200.43	3.94	924.12	223.98	3.70
9	DCAB	787.80	163.07	2.73	905.52	172.22	3.23	1966.42	362.11	3.77	1640.15	331.95	3.15
10	BACD	1785.35	347.46	3.93	1934.66	373.82	4.38	892.89	163.78	3.78	826.27	151.87	3.78
11	ADBC	2031.55	320.43	3.39	975.70	165.38	3.56	1893.99	313.81	3.48	788.06	128.56	3.23
12	BACD	1524.61	381.50	3.12	2525.23	439.42	4.35	952.05	177.75	4.07	940.57	187.22	3.58
13	ADBC	2013.54	314.76	4.99	1005.49	168.16	4.39	2322.68	406.54	4.98	946.92	152.94	4.73
14	DCAB	990.04	163.73	4.63	1118.63	177.61	4.82	2300.18	334.58	4.51	2197.79	293.69	4.72
15	CBDA	839.94	136.99	4.02	1956.45	374.10	3.97	611.43	132.20	2.63	1707.48	273.61	4.03
16	CBDA	1159.85	167.43	4.57	2760.90	349.34	5.59	1007.45	178.00	4.59	2477.37	327.36	5.88
17	DCAB	1032.22	182.99	4.01	1039.21	173.00	3.96	2440.50	380.79	4.53	1860.15	353.41	3.77
18	BACD	1782.62	376.58	3.64	1917.01	426.42	3.44	1048.27	179.42	4.04	882.46	149.33	3.23
19	CBDA	852.84	150.20	3.87	2256.02	284.50	4.04	982.67	157.35	4.19	1924.09	360.50	3.96
20	ADBC	2178.77	436.64	4.09	1273.04	186.33	4.58	2074.44	296.02	4.08	1009.09	190.86	4.57
21	ADBC	2529.23	449.49	4.58	1365.57	190.35	5.21	1868.99	412.40	4.15	1064.39	208.41	4.95
22	CBDA	989.89	167.33	3.85	1936.16	334.66	4.03	904.53	175.76	4.10	2029.20	420.12	4.24
23	BACD	1579.55	328.00	3.72	1756.96	284.65	3.69	949.38	188.22	4.11	951.75	201.15	4.12
24	DCAB	889.20	186.69	3.95	757.79	196.20	3.10	1813.93	441.25	3.57	1523.54	327.47	3.33

Latin square below

ADBC
BACD
CBDA
DCAB

Where A, B, C, D are imported RMT 2 mg, domestic RMT 2 mg, imported RMT 4 mg and domestic 4 mg, respectively.

Twenty-four volunteers were randomly allocated in 4 treatment groups with 6 in each sequence. Each subject received different treatment on different cycles. To minimize carryover effects, a 7-day wash-out period between the two treatment occasions was made. Plasma concentration of rosiglitazone maleate was detected within 24 hours after orally taking RMT. Data in Table 2 is the area under curve (AUC), maximum concentration ($C_{\max}$) and time to half maximum concentration (T_{50}). The aim is to test whether there is difference between domestic and imported RMT.

This is a four-by-four crossover design with 3 variables. In this data set, the observations in 4 periods and the variables (AUC, $C_{\max}$ and T_{50}) are correlated.

2.5. *Example* 5. *Repeated measurement, linear regression*

In a multicenter, randomized, double-blind, three doses (high, middle, and low = placebo) controlled clinical study, the researchers evaluated the efficacy and safety of urokinase (UK) in the treatment of acute cerebral infarctions within 6 hours from the onset of stroke. One interesting variable is the European stroke scale (ESS). Data are shown in Table 3.

Repeated measurement design, also known as within-subject design, is a quite common design in medical researches. The feature of this type of data set is that individuals are measured repeatedly through time. We are interested in both treatment and temporal effects. Figure 1 displays the data graphically. Each line connect the repeated observations at different times of a subject. This simple graph reveals apparent and important patterns. First, all of 30 subjects are getting better within 8 weeks as ESS is becoming larger. Second, patients with larger ESS at the beginning of the period tend to remain larger throughout. This phenomenon is called "tracking."

There are two ways to deal with this type of data by classical methods. First, we estimate the average of ESS for each week and fit a regression model of the means of ESS over time. In fact, we aggregate the data.

Table 3. ESS of 30 acute cerebral infarctions.

Id	treat	Age	weeks								
			0	1	2	3	4	5	6	7	8
1	0	27	107	106	106	108	108	112	112	112	112
2	0	21	107	106	106	106	106	112	112	114	116
3	0	21	100	100	100	106	109	108	114	116	116
4	0	36	107	106	106	107	106	111	112	117	109
5	0	17	110	111	112	112	113	113	113	116	116
6	0	22	105	108	108	106	108	108	108	109	110
7	0	29	102	101	104	100	94	106	106	105	106
8	0	15	97	97	97	99	99	99	101	101	103
9	0	21	108	108	108	110	116	116	120	128	120
10	0	27	108	108	108	114	116	118	118	124	128
11	1	34	98	98	102	121	120	124	124	132	140
12	1	37	100	98	114	118	126	126	134	138	138
13	1	31	104	123	127	129	130	130	136	140	140
14	1	28	108	120	115	119	134	126	126	127	140
15	1	32	106	108	108	108	112	112	112	114	116
16	1	18	103	102	102	104	114	114	116	128	143
17	1	15	101	103	104	108	113	113	118	122	126
18	1	31	91	90	92	93	89	95	102	105	108
19	1	39	94	94	96	99	116	124	135	138	145
20	1	34	104	104	105	105	122	128	131	129	138
21	2	36	107	111	112	127	127	128	138	141	141
22	2	45	109	114	120	130	131	132	139	142	143
23	2	40	103	103	108	112	116	118	123	125	135
24	2	44	110	114	120	124	133	135	142	144	144
25	2	22	95	103	115	113	119	122	126	134	136
26	2	25	92	102	110	108	116	116	116	122	127
27	2	32	98	106	112	112	120	124	126	136	141
28	2	38	106	121	127	126	128	130	132	138	140
29	2	22	102	112	110	119	119	123	125	133	142
30	2	19	109	109	124	127	128	132	133	144	147

As a result, it increases the correlation and causes a spurious association between ESS and time. Second, we fit a regression model for all the data on time. These two models give the same regression coefficients but different standard errors. Both of them ignore the intra-subject correlation.

2.6. *Example* 6. *Pharmacokinetics study, repeated measurements, nonlinear regression*

A single oral dose Ciclosporin A Capsule was given to 10 healthy volunteers. Plasma concentration (ng/ml) was detected after medication. The results are shown in Table 4.

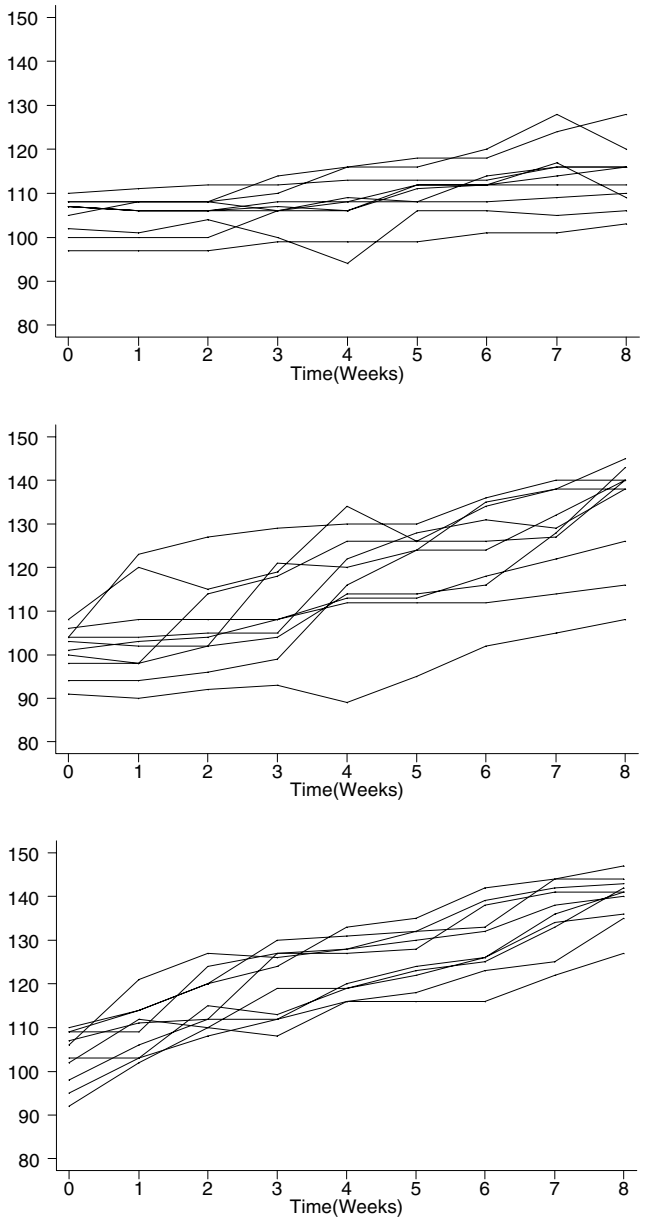


Fig. 1. ESS over time of 30 acute cerebral infarctions for three groups.

Table 4. Palsma concentrations of Ciclosporin A Capsule after medication of 10 volunteers.

Subject	Time (*hour*)									
	0.5	1	2	3	4	5	6	8	12	16
1	343.3	783.6	443.1	426.8	267.0	155.5	125.0	98.3	75.2	23.8
2	86.6	501.1	817.9	542.7	273.9	226.4	195.7	114.0	79.9	26.6
3	256.1	534.8	486.8	420.1	370.6	316.7	250.6	192.6	124.5	75.9
4	300.2	849.7	846.0	521.1	373.2	269.4	258.1	182.7	93.0	68.0
5	344.6	826.4	631.0	485.0	389.7	257.7	204.7	172.4	124.5	44.2
6	230.0	780.7	912.3	551.2	299.8	219.3	148.7	75.1	55.9	27.6
7	116.5	943.4	848.2	747.3	410.4	345.5	171.4	129.5	63.0	17.5
8	66.7	239.2	814.6	526.9	426.6	213.5	152.5	118.5	73.1	38.1
9	67.7	789.1	551.6	520.2	463.0	295.7	191.8	154.4	108.4	32.5
10	216.2	599.9	1099.5	562.9	413.9	297.5	233.2	146.6	94.8	38.7

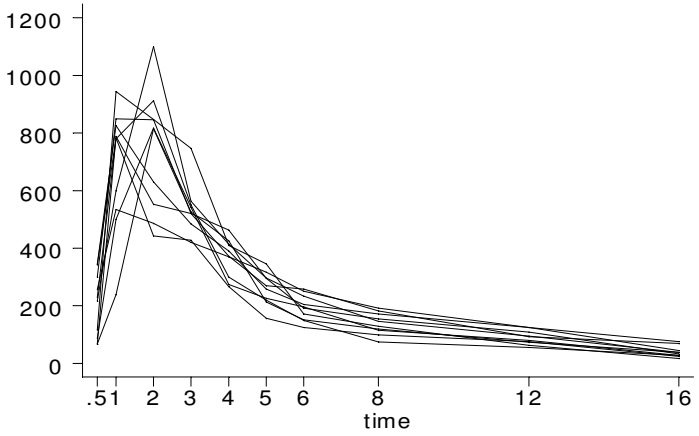


Fig. 2. Plasma concentration-time curve of Ciclosporin A capsule after a single oral dose in 10 volunteers.

This is an example of repeated measurement with nonlinear trend, which are distinct from Example 5.

In experimental or pharmacokinetical study, the sample size is relatively small and the period is usually short. Dropout seldom occurs. Furthermore both times of repeated measure and time intervals are similar to each other. However, it is not the case in clinical trial. The observed period is usually long. Compliance varies among patients and dropouts are routine. Last, but not least, the times and time intervals are different among patients.

2.7. *Example 7. Mmulti-center clinical study, ranked data*

To investigate the effect of nerve growth factor (NGF) for subjects of extra-neuritis caused by chemical products, an random, double blind, placebo controlled clinical trial was developed.

One hundred two subjects were random allocated into treatment and placebo groups. The effectiveness was observed in 8 consecutive weeks for 102 subjects. The results are shown in Table 5, in which Id represents

Table 5. Effects of 102 subjects of extra-neuritis caused by chemical products.

Id	cnt	Trt	com	sex	age	base	x1	x2	x3	x4	x5	x6	x7	x8
1	1	T	A	0	21	13	1	2	2	2	2	2	2	3
2	1	P	A	0	27	12	1	2	2	2	2	2	2	2
3	1	T	A	0	27	13	1	2	2	2	2	2	2	2
4	1	P	B	0	21	13	1	2	2	2	2	2	2	2
5	1	T	B	0	34	7	0	1	1	2	2	2	2	2
6	1	T	B	0	45	13	1	2	2	2	2	2	2	3
7	1	T	B	0	37	13	1	2	2	2	2	2	2	
8	1	P	A	0	21	13	1	2	2	2	2	2	2	2
9	1	T	A	0	31	14	1	2	2	2	3	3	3	3
10	1	T	B	0	23	14	2	2	2	2	3	3	3	3
11	1	T	A	0	22	13	1	2	2	2	2	2	2	2
12	1	P	B	1	28	14	2	2	2	2	2	2	2	2
13	2	P	A	1	24	15	2	2	2	2	3	3	m	m
14	2	T	A	0	28	13	2	2	2	2	3	3	3	m
15	2	T	B	0	29	14	2	2	2	2	3	3	m	m
16	2	T	A	0	21	10	1	2	2	2	3	3	3	3
17	2	P	B	0	31	15	2	2	2	2	m	m	m	3
18	2	T	B	1	25	10	1	1	1	1	2	2	3	3
19	2	P	A	0	29	8	0	1	1	1	2	2	2	1
20	2	T	B	0	20	10	1	1	2	2	2	3	3	3
21	2	T	A	0	32	13	1	2	2	2	2	2	2	2
22	3	T	A	0	18	14	2	2	2	2	3	3	3	3
23	3	T	A	0	31	6	0	0	0	1	1	1	1	1
24	3	T	A	0	39	10	1	1	1	2	2	2	2	3
25	3	T	B	0	34	9	1	1	1	1	1	2	2	2
26	4	P	A	0	15	12	1	1	1	1	2	2	2	2
27	5	P	B	0	36	7	1	1	1	1	1	1	1	1
28	5	T	B	0	37	5	1	1	1	1	1	1	1	2
29	6	T	A	1	15	1	0	0	0	0	0	0	1	1
30	6	T	A	1	16	1	0	0	0	0	1	1	1	1
31	6	T	B	0	15	11	1	2	2	2	2	2	2	3
32	6	P	B	1	16	8	1	1	1	1	1	1	1	1
33	6	P	A	1	17	13	1	2	2	2	2	2	2	2
34	7	T	B	0	29	13	1	2	2	2	2	2	2	3
35	7	P	B	0	19	15	2	2	2	2	3	3	3	3

Table 5. *Continued.*

Id	cnt	Trt	com	sex	age	base	x1	x2	x3	x4	x5	x6	x7	x8
36	7	T	B	0	29	13	1	2	2	2	2	2	2	2
37	7	P	A	0	29	13	2	2	2	2	m	m	m	m
38	8	T	B	0	17	13	1	2	2	2	2	1	2	1
39	8	P	A	0	18	13	1	2	2	2	2	2	2	2
40	8	T	B	0	17	12	2	2	2	2	2	2	3	3
41	8	T	A	0	18	11	1	2	2	2	2	2	2	2
42	0	P	B	0	45	7	0	1	2	2	2	2	2	2
43	9	T	A	0	30	9	0	0	1	1	1	1	2	2
44	9	P	A	0	43	12	1	2	2	2	2	2	2	2
45	9	T	A	0	36	14	2	2	2	2	3	2	2	2
46	9	T	B	0	45	14	2	2	2	2	3	3	3	3
47	9	P	B	0	32	14	2	2	2	2	3	3	3	3
48	9	T	A	0	40	11	1	1	2	2	2	2	2	2
49	1	P	A	0	36	13	1	2	2	2	2	2	2	2
50	1	T	A	0	44	13	1	2	2	2	2	3	3	3
51	10	P	B	0	35	13	1	2	2	2	2	2	2	2
52	10	T	B	0	38	13	1	2	2	2	2	2	2	2
53	10	T	A	0	33	13	1	2	2	2	2	3	3	3
54	10	T	B	0	41	13	1	2	2	2	2	2	2	3
55	11	T	A	0	22	13	2	2	2	2	2	3	m	m
56	11	P	B	0	22	14	1	1	1	1	2	2	2	2
57	11	T	B	0	28	8	1	1	1	1	1	2	2	2
58	11	P	A	0	31	14	2	2	2	2	3	3	3	3
59	11	T	A	0	37	15	2	2	2	2	3	3	3	3
60	11	T	B	0	21	14	2	2	2	2	2	3	3	3
61	11	P	B	1	24	12	1	2	2	2	2	2	2	2
62	11	T	B	0	25	7	1	1	1	2	2	2	2	2
63	11	T	A	0	32	12	1	1	2	2	2	2	2	2
64	11	T	A	0	21	15	2	2	2	2	3	3	3	3
65	11	P	A	0	20	13	2	2	2	2	3	3	3	3
66	11	T	B	1	30	14	2	2	2	2	3	m	m	m
67	11	T	A	1	38	13	1	2	2	2	2	3	3	3
68	11	T	A	0	19	16	2	2	2	2	3	3	3	3
69	11	T	B	0	21	14	2	2	2	2	2	3	3	3
70	11	P	B	0	25	11	1	m	2	2	m	m	m	m
71	11	T	B	1	22	15	2	2	2	2	m	3	m	3
72	11	P	A	1	23	10	1	1	1	1	1	1	m	m
73	11	T	A	0	21	14	2	2	2	2	3	3	3	3
74	11	T	A	0	26	14	2	m	2	m	3	m	3	m
75	11	T	B	0	20	15	2	2	2	2	3	3	3	3
76	11	P	B	0	18	13	2	2	2	m	m	m	m	m
77	11	P	A	1	23	8	1	1	1	1	1	1	1	2
78	11	T	B	0	24	15	m	2	m	2	m	3	m	3
79	11	P	B	0	26	10	1	1	1	1	1	2	2	m
80	11	T	A	0	20	15	2	2	2	2	3	3	3	3
81	11	P	A	0	18	13	2	2	m	m	m	m	m	m

Table 5. *Continued.*

Id	cnt	Trt	com	sex	age	base	x1	x2	x3	x4	x5	x6	x7	x8
82	11	T	B	0	22	14	2	2	2	2	m	m	3	m
83	11	T	B	0	25	15	2	2	2	2	3	3	3	3
84	11	T	A	0	19	15	2	2	2	2	3	3	3	3
85	11	T	B	0	20	15	2	2	2	2	3	3	m	m
86	11	T	A	0	29	15	2	2	2	2	3	3	m	m
87	11	P	A	0	25	8	m	m	1	m	m	1	m	m
88	11	T	A	0	24	15	2	2	2	m	m	3	m	m
89	11	P	B	0	21	11	m	1	1	m	m	m	m	m
90	11	T	B	0	19	15	2	2	2	2	3	3	3	m
91	11	T	A	0	21	15	2	2	2	2	3	3	3	m
92	11	T	B	0	22	14	2	2	2	2	3	3	3	m
93	11	T	A	0	20	15	2	2	2	2	3	3	3	3
94	11	P	A	0	29	4	0	0	0	0	1	1	1	1
95	11	T	B	0	19	14	2	2	2	2	3	3	3	3
96	11	P	B	0	28	12	1	2	m	2	3	2	2	m
97	11	P	B	0	21	8	1	1	1	1	1	1	m	m
98	11	T	B	0	30	15	2	2	2	2	3	3	3	3
99	11	T	B	1	27	15	2	2	2	2	3	3	3	3
100	12	P	A	0	25	7	0	1	1	1	1	2	1	2
101	12	T	B	0	35	7	1	1	1	1	1	2	2	2
102	12	T	A	0	33	8	1	1	1	1	2	2	2	2

identification of subjects; Cnt represents the center; Trt represents groups (Trt = P for placebo, Trt = T for NGF); Com represents companies; Base represents the MDNS base level before treatment; and x1–x8 are effects at time from 1 to 8 weeks after treatment. Here, x = 0 stands for invalid or worse effect of treatment on the subject, 1 for improved, 2 for notable improved, 4 for recovery, and m for missing.

Except for the intra-subject correlation, the intra-center correlation of the subjects in the same hospital should be considered for this type of data.

2.8. *Example* 8. *Repeated measurement, count data*[3]

In order to understand whether the progabide reduces the rate of epileptic seizures, 59 patients of epileptics were recruited in a clinical trial. For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks. Patients were then randomized to treatment with the anti-epileptic drug progabide, or placebo. In addition, all of the patients were treated with standard chemotherapy. The number of seizures was then recoded in 4 consecutive two-weeks for each epileptic.

Where, treatment variable is group (0 = placebo, 1 = progabide). What is different from Examples 5 and 7 is that the response variable is the seizure

counts in unit time (two-week). Poisson regression would be used here for count data.

In this study, only recurrence episodes were included, not first episode. The reason of not including the first episode was that the factors associated with recurrence of a disease are usually different from those with that disease. For example, in a model of development of breast cancer, we should not include women who already had breast cancer, because family history and late childbearing have the strongest association with development of breast cancer, whereas stage of disease, hormone receptors, and histological grade are the strongest risk factors for recurrence of breast cancer.

For those diseases for which it is sensible to speak of a second distinct episode, the risk factors for a second episode may be similar to the risk factors for a first episode. Hooton and colleagues were interested in studying urinary track infections in young women.[4] With urinary track infections, patients can have a second (or third, etc.) episode after a "crude" first episode. Repeated episodes in the same person are not independent observations because the causes of urinary track infections are likely to be more similar in repeated episodes in the same person than in separate episodes in different people. Therefore, Hooton and colleagues included repeat episodes in their analysis, which increased the power of their study.

2.9. *Other examples*

Clinical researchers in the fields of ophthalmology, orthopedics, and dentistry have a distinct advantage over cardiologists, neurologists, and hepatologists. That is while humans have only one heart, one brain, and one liver, we have two eyes, thirty-two teeth or so, and most of our joints in duplicates. In those fields with duplicate organs, it is possible to follow (or assess) a single subject and have multiple observations. For the cases with outcomes that are observed more than once in a single subject, you must use special methods to deal with outcomes that can occur in more than one body part in the same person.

In a study of complications after breast implantation most women had bilateral implants.[5] Some had multiple implants in the same breast. The investigators therefore performed follow-up of each breast implant until a complication occurred, the implant was removed, or the end of follow-up occurred. The survival times of the implants for the same woman are dependent.

In a study of the relationship of vitamin D to development of osteoarthritis of knees, the investigators used the fact that their participants had

Table 6. Four successive two-week seizure counts for each 59 patients of epileptics.

ID	y1	y2	y3	Y4	treat	baseline	Age	ID	y1	y2	y3	y4	treat	baseline	Age
1	5	3	3	3	0	11	31	31	0	4	3	0	1	19	20
2	3	5	3	3	0	11	30	32	3	6	1	3	1	10	20
3	2	4	0	5	0	6	25	33	2	6	7	4	1	19	18
4	4	4	1	4	0	8	36	34	4	3	1	3	1	24	24
5	7	18	9	21	0	66	22	35	22	17	19	16	1	31	30
6	5	2	8	7	0	27	29	36	5	4	7	4	1	14	35
7	6	4	0	2	0	12	31	37	2	4	0	4	1	11	57
8	40	20	23	12	0	52	42	38	3	7	7	7	1	67	20
9	5	6	6	5	0	23	37	39	4	18	72	5	1	41	22
10	14	13	6	0	0	10	28	40	2	1	1	0	1	7	28
11	26	12	6	22	0	52	36	41	0	2	4	0	1	22	23
12	12	6	8	5	0	33	24	42	5	4	0	3	1	13	40
13	4	4	6	2	0	18	23	43	11	14	25	15	1	46	43
14	7	9	12	14	0	42	36	44	10	5	3	8	1	36	21
15	16	24	10	9	0	87	26	45	19	7	6	7	1	38	35
16	11	0	0	5	0	50	26	46	1	1	2	4	1	7	25
17	0	0	3	3	0	18	28	47	6	10	8	8	1	36	26
18	37	29	28	29	0	111	31	48	2	1	0	0	1	11	25
19	3	5	2	5	0	18	32	49	102	65	72	63	1	151	22
20	3	0	6	7	0	20	21	50	4	3	2	4	1	22	32
21	3	4	3	4	0	12	29	51	8	6	5	7	1	42	25
22	3	4	3	4	0	9	21	52	1	3	1	5	1	32	35
23	2	3	3	5	0	17	32	53	18	11	28	13	1	56	21
24	8	12	2	8	0	28	25	54	6	3	4	0	1	24	41
25	18	24	76	25	0	55	30	55	3	5	4	3	1	16	32
26	2	1	2	1	0	9	40	56	1	23	19	8	1	22	26
27	3	1	4	2	0	10	19	57	2	3	0	1	1	25	21
28	13	15	13	12	0	47	22	58	0	0	0	0	1	13	36
29	11	14	9	8	1	76	18	59	1	4	3	2	1	12	37
30	8	17	9	4	1	38	32								

two knees to their advantage.[6] Although the Framingham's study consists of over 5000 subjects, only 556 participants had X-rays of their knees and assessments of their vitamin D intake and serum levels. Therefore, they did this by looking at both knees to maximize their statistical power.

3. Common Structures of Intra-unit Correlation for Dependent Data

The feature of dependent data is that the variance-covariance matrix of response variable is not diagonal but block diagonal.

Because the dependent data do not meet the independent requirement that is essential in classical statistical methods, special methods are needed

to deal with it. For example, random effects models and/or mixed effects models are used for repeated measurement or longitudinal data and meta-analysis is used for multicenter clinical trial.[3] Many systematic researches have been achieved in the field. In this section, we try to demonstrate the connotations of dependent data, how to judge the type of the data set, to construct reasonable covariance structure or intra-unit correlations structure, draw valid scientific inferences for the data set. In this section, we focus on the common structures of intra-unit correlation of dependent data.[7]

3.1. *A simple case*

We first consider the simplest case of a paired design. In this paired design, the subjects are independent, while two observations on the same subjects are correlated. If we assume the correlations of two observations of subjects are equal, say ρ, then the correlation matrix of $2m$ observations from m subjects could be

$$\boldsymbol{R}_Y = \begin{bmatrix} \boldsymbol{R} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{R} \end{bmatrix} \tag{1}$$

where,

$$\boldsymbol{R}_Y = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{2}$$

where $\mathbf{0}$ is 0 matrix with all elements being 0, $\boldsymbol{R}_Y$ is block diagonal matrix with $\boldsymbol{R}$ in diagonal.

For random block trial, we have a treatments and b blocks. While individuals from different blocks are independent, those from the same block tend to be similar and correlated. Because the individuals in the same block are in the same status, so we can assume that there is a positive correlation, ρ, between any two individuals from the same block. The intra-block correlation matrix is defined as

$$\boldsymbol{R}_2 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \vdots & 1 \end{bmatrix} \tag{3}$$

$a \times b$ observations form a correlation matrix which has the same structure as $\boldsymbol{R}_Y$ in (1). Matrices in diagonal block of $\boldsymbol{R}_Y$ have the same structure as $\boldsymbol{R}_2$. It is obviously that $\boldsymbol{R}_1$ is a special case of $\boldsymbol{R}_2$.

Now let's consider some types of correlation structure of longitudinal studies. The defining characteristic of a longitudinal study is that individuals are measured repeatedly through time in a follow-up study. Correlation structures vary from data set. The commonly used correlation matrices are equal correlation, neighbor correlation, autocorrelation and unstructured correlation, etc.

3.1.1. *Equal correlation*

It is similar to $\boldsymbol{R}_2$, We also refer to equal correlation as exchangeable or compound symmetry.

3.1.2. *Neighbor correlation*

Neighbor correlation is that only two closed observations are correlated, others are independent. For 5 times repeated measurement, the correlation matrix is given by

$$\boldsymbol{R}_3 = \begin{bmatrix} 1 & \rho_1 & 0 & 0 & 0 \\ \rho_1 & 1 & \rho_2 & 0 & 0 \\ 0 & \rho_2 & 1 & \rho_3 & 0 \\ 0 & 0 & \rho_3 & 1 & \rho_4 \\ 0 & 0 & 0 & \rho_4 & 0 \end{bmatrix}. \tag{4}$$

When the correlations of two closed observations are equal, the correlation is refered to as stationary 1-dependence), otherwise, nonstationary 1-dependence. Stationary 2-dependence has the structure as follows

$$\boldsymbol{R}_4 = \begin{bmatrix} 1 & \rho & \rho & 0 & 0 \\ \rho & 1 & \rho & \rho & 0 \\ \rho & \rho & 1 & \rho & \rho \\ 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & \rho & \rho & 1 \end{bmatrix}. \tag{5}$$

It is not difficult to extend to stationary k-dependence. Obviously, stationary correlation is a special case of nonstationary, and exchangeable correlation is a special case of stationary.

3.1.3. *Autocorrelation*

Autocorrelation means that correlation depends on the spacing of two measurements. The correlation between a pair of measurements on the same subject decays towards 0 as the time separation between the measurements increases. If the correlation of two observations next to each other is ρ, the correlation of two separated observations is ρ's power of number of the observations separated. For 5 times repeated measurements, the correlation matrix is given by

$$\boldsymbol{R}_5 = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}. \tag{6}$$

We refer to (6) as the first order autocorrelation or the first order autoregressive process. A natural extension of (6) is given by (7), $\boldsymbol{R}_6$, the correlation is inversed to the time interval or spacing of two measurements.

$$\boldsymbol{R}_6 = \begin{bmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} & \rho^{t_4-t_1} & \rho^{t_5-t_1} \\ \rho^{t_2-t_1} & 1 & \rho^{t_3-t_2} & \rho^{t_4-t_2} & \rho^{t_5-t_2} \\ \rho^{t_3-t_1} & \rho^{t_3-t_2} & 1 & \rho^{t_4-t_3} & \rho^{t_5-t_3} \\ \rho^{t_4-t_1} & \rho^{t_4-t_2} & \rho^{t_4-t_3} & 1 & \rho^{t_5-t_4} \\ \rho^{t_5-t_1} & \rho^{t_5-t_2} & \rho^{t_5-t_3} & \rho^{t_5-t_4} & 1 \end{bmatrix}. \tag{7}$$

3.1.4. *Unstructured or general structure*

In this case elements on nondiagonal of block matrix $\boldsymbol{R}$ are unequal.

3.1.5. *Independent, zero correlation*

Elements on nondiagonal of block matrix $\boldsymbol{R}$ are 0.

The relationships of the matrices mentioned above are as follow

$$\text{independent} \subset \text{exchangeable} \subset \text{autocorrelation} \subset \text{stationary}$$

$$\subset \text{nonstationary} \subset \text{unstructured}$$

where A $\subset$ B means A is a special case of B.

3.2. *Complicated cases*

In random cluster sample study, individuals in the same cluster (household, class in school, group in enterprise, etc.) tend to act in a similar way on healthy attitude, eating habit, and so on, and share the same environment, etc. If family is the unit in cluster sampling, genetic factor should be considered because the observations measured on the members from the same family are correlated. For example, in a cluster sampling, a simple random sample of 54 households was drawn.[8] The blood pressure observations of 209 subjects were detected. Let Y_{ij} represent a response variable, systolic pressure, for member j $(j = 1, 2, \ldots, n_i)$ in household i $(i = 1, 2, \ldots, 54)$. Where $j = 1$ stands for father, 2 for mother, 3 and more for children.

Generally speaking, if the interesting variable is affected by genetic factor or other family factors, the correlation between parents is lower than the correlations between father and children, mother and children, and children themselves. In this case, a special but common correlation structure could be defined as (for example, 4 persons in a family with parents and two children)

$$\begin{array}{c} \\ Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{array} \begin{array}{c} \begin{array}{cccc} Y_{i1} & Y_{i2} & Y_{i3} & Y_{i4} \end{array} \\ \begin{pmatrix} 1 & r_1 & r_2 & r_2 \\ r_1 & 1 & r_3 & r_3 \\ r_2 & r_3 & 1 & r_4 \\ r_2 & r_3 & r_4 & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \text{father} \\ \text{mother} \\ \text{child 1} \\ \text{child 2} \end{array} . \tag{8}$$

In fact, the correlation structure matrix of 4 members (parents and two children) in one family in the example mentioned above is

$$\begin{pmatrix} 1.0000 & 0.2056 & 0.4212 & 0.4212 \\ 0.2056 & 1.0000 & 0.4292 & 0.4292 \\ 0.4212 & 0.4292 & 1.0000 & 0.5622 \\ 0.4212 & 0.4292 & 0.5622 & 1.0000 \end{pmatrix} .$$

For stratified cluster sampling and other data with hierarchical structure, the same strategy could be used to construct the intra-cluster correlation matrices.

In the crossover design, each subject is randomized to a sequence of two or more treatments and hence acts as his own control for treatment comparisons. In the simplest paired 2×2 crossover design, two subjects are paired, the first subject in the same paired receives either of two treatments in randomized order in two successive treatment periods which often

separated by a washout period, while the other received two treatments in adverse order to the first one in two successive treatment periods. There are 3 possible correlations in this type of data: (1) correlation between two observations of the same subject in two periods; (2) correlation between two subjects in the same paired in the same period; and (3) correlation between two subjects in the same paired in different periods. The correlation structure, therefore, could be defined as:

		Subject 1		Subjects 2	
		Period 1	Period 2	Period 1	Period 2
Subject 1	Period 1	1	r_1	r_2	r_3
	Period 2	r_1	1	r_3	r_2
Subject 2	Period 1	r_2	r_3	1	r_1
	Period 2	r_3	r_2	r_1	1

.

In multicenter clinical trial, although the protocol and standard operating procedures are implemented similarly at all centers, the level and opinions of doctors and nurses, equipments, and medical conditions, etc., vary from the centers. This is so-called center-effects. Subjects in the same center are correlated. The repeated observations through time from the same subjects are also correlated. This is hierarchical structure data. If subjects from different centers are independent, the intra-center correlation structure could be defined as (3 visits for each subject):

		Subject 1			Subject 2			$\cdots$	Subject n		
		t_1	t_2	t_3	t_1	t_2	t_3	$\cdots$	t_1	t_2	t_3
Subject 1	t_1	1	r_1	r_1	r_2	r_2	r_2		r_2	r_2	r_2
	t_2	r_1	1	r_1	r_2	r_2	r_2	$\cdots$	r_2	r_2	r_2
	t_3	r_1	r_1	1	r_2	r_2	r_2	$\cdots$	r_2	r_2	r_2
Subject 2	t_1	r_2	r_1	r_2	r_2	1	r_1		r_2	r_2	r_2
	t_2	r_2	r_2	r_2	1	r_1	r_1	$\cdots$	r_2	r_2	r_2
	t_3	r_2	r_2	r_2	r_1	r_1	1		r_2	r_2	r_2
$\cdots$			$\cdots$			$\cdots$				$\cdots$	
Subject n	t_1	r_2	r_2	r_2	r_2	r_2	r_2		1	r_1	r_1
	t_2	r_2	r_2	r_2	r_2	r_2	r_2	$\cdots$	r_1	1	r_1
	t_3	r_2	r_2	r_2	r_2	r_2	r_2		r_1	r_1	1

.

Although, for a real data set, the correlation structure could be defined and selected by statistical methods, the author suggests that the biological

and medical backgrounds should be considered to get a reasonable and acceptable correlation matrix structure.

4. ANOVA Methods and Its Limitation

4.1. *Parameter estimations for dependent data*

4.1.1. *Estimation of means*

If there are n observations of variable X, denoted by $x_1, x_2, x_3, \ldots, x_n$ with mean $\bar{X}$ and variance σ^2. We assume the data are dependent.

(1) x_i is correlated with x_j with a correlate coefficient ρ

If x_i is correlated with x_j with a correlation coefficient ρ (ρ is assumed to be larger than 0 without losing general), thus the variance of $\bar{X}$ was

$$\begin{aligned}\operatorname{var}(\bar{X}) &= \frac{1}{n^2}\operatorname{cov}(x_1 + x_2 + \cdots + x_n,\, x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n^2}[n\sigma^2 + n(n-1)\rho\sigma^2] \\ &= \frac{\sigma^2}{n}[1 + (n-1)\rho]\,. \end{aligned} \tag{9}$$

Formula (9) shows that standard error of mean is larger when the data are dependent than the case when the data are independent. Moreover, it is in proportion to correlation. In this case, the confidence interval of population mean is as follows

$$\bar{X} \pm t_{n-1,\nu}\frac{\sigma}{\sqrt{n}}\sqrt{1 + (n-1)\rho}\,. \tag{10}$$

It is wider than that when the data are independent. When intra-unit correlation is 0, the confidence interval given by (10) is similar to the confidence interval when data are independent.

(2) x_i is correlated with x_j with autocorrelation

If x_i is correlated with x_j with autocorrelation

$$\operatorname{cov}(x_i, x_j) = \sigma^2\rho^{|i-j|}\,. \tag{11}$$

The variance of

$$\operatorname{var}(\bar{X}) = \frac{1}{n^2}[n + 2(n-1)\rho + 2(n-2)\rho^2 + \cdots + 2\rho^{n-1}]\sigma^2\,. \tag{12}$$

(3) Correlation between x_i and x_j is unstructured

If the correlation between x_i and x_j is unstructured

$$\mathrm{cov}(x_i, x_j) = \lfloor \rho_{ij}\sigma^2 \rfloor_{n\times n}\,. \tag{13}$$

The variance of $\bar{X}$ is

$$\mathrm{var}(\bar{X}) = \frac{1}{n^2}\left[n + 2\left(\sum_{i\neq j}\rho_{ij}\right)\right]\sigma^2\,. \tag{14}$$

Thus, the standard error of mean in this case when the data are dependent is larger than the one when the data are independent from each other. The standard error is in proportion to the correlation as well. The confidence interval is wider than that of independent data.

4.1.2. *Estimation of rate*

The independent binary data should generally be handled by the methods based on binominal distribution. Let incidence rate be π and its variance be $\pi(1-\pi)$, then the standard error is $\sqrt{\pi(1-\pi)/n}$.

If the data are correlated with each other, the variance and the standard error of rate increase. For example, in Example 2, the total incident rate is $\pi = 30/104 = 0.2885$, the variance is 0.00197 and 95% CI is 0.2038–0.3855 if we apply the methods based on binominal distribution. And its 95% CI is 0.2014–0.3756 if we apply the methods based on normal approximation.

However, the actual variance of the incident rate in each family is 0.00520, much larger than that given by the pure binominal distribution. This is because that the incidence, "visiting doctors in the last year", has a family aggregation. As a result, we underestimated the variance of dependent incidences by applying methods based on binomial distribution.

The classic way of handling dichotomous data is firstly coding the incidence that happens as 1, otherwise, as 0, and then applying Eq. (9) to the data. When it comes to a dichotomous data with equal correlation, the standard error of rate is

$$\sigma_\pi = \sqrt{\frac{\pi(1-\pi)}{n}[1 + (n-1)\rho]}\,. \tag{15}$$

Others can be handled in similar ways.

4.2. *ANOVA with random effect*

We begin with the repeated one-way testing designs, of which the block design is the simplest case. We may assume that there are a treatments and b blocks. The model of ANOVA can be represented as

$$y_{ij} = \mu + \tau_j + e_{ij} \,. \tag{16}$$

In the equation below, μ is the population's mean, τ_j is the effect of the jth treatment ($j = 1, 2, \ldots, a$), e_{ij} is the total residual error of observations receiving the jth treatment in the ith block.

When we apply a randomized block design, the units in the same block may have good homogeneity, while units in different blocks may have many differences. This is the characteristic of block design that makes the observations in every block to be homoplasy, which is called intra-block correlation. For this moment, the error term e_{ij} may be denoted as

$$e_{ij} = \nu_i + u_{ij} \,. \tag{17}$$

ν_i is the residual error of the ith block ($i = 1, \ldots, b$), u_{ij} is the residual error of observations receiving the jth treatment in the ith block. Therefore, the ANOVA model of block design should be

$$y_{ij} = \mu + \tau_j + \nu_i + u_{ij} \,. \tag{18}$$

In most cases, the treatment factors of a block design are fixed effects, while blocks are random effects. Namely, τ_j is fixed effect, μ_j is the mean of observations in the jth level, and ν_i is random effects with

$$\begin{gathered} \tau_j = \mu_j - \mu \,, \quad \Sigma\tau_j = 0 \\ \nu_i = \mu_i - \mu \,, \quad \Sigma\nu_i = 0 \,, \\ \operatorname{var}(\nu_i) = \sigma_2^2 \,, \quad \text{and } \operatorname{cov}(\nu_i, \nu_{i'}) = 0, \ i \neq i' \,, \end{gathered} \tag{19}$$

where μ_i is the mean of the ith block. u_{ij} is the random effect, and

$$\begin{gathered} u_{ij} = y_{ij} - \mu - \tau_j - \nu_i \,, \quad \Sigma u_{ij} = 0 \,, \\ \operatorname{var}(u_{ij}) = \sigma_1^2 \,, \quad \text{and } \operatorname{cov}(u_{ij}, \ u_{i'j'}) = 0, \ j \neq j' \\ \operatorname{cov}(u_{ij}, \nu_k) = 0 \quad \text{for all } i, j, k \,. \end{gathered} \tag{20}$$

In this way, the variance of y_{ij} is

$$\operatorname{var}(y_{ij}) = \sigma_2^2 + \sigma_1^2 \,, \tag{21}$$

the covariance is

$$\mathrm{cov}(y_{ij}, y_{ij'}) = \mathrm{cov}(\nu_i + u_{ij}, \nu_i + u_{ij'}) = \sigma_2^2,\ j \neq j' ,$$

and others are 0.

Expressed by matrix, the variance and covariance of y_{ij} is

$$\mathrm{cov}(e_{ij}) = \sigma^2 \begin{pmatrix} \boldsymbol{R} & 0 & \cdots & 0 \\ 0 & \boldsymbol{R} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{R} \end{pmatrix}_{ab \times ab} . \tag{22}$$

Here, $\sigma^2 = \sigma_1^2 + \sigma_2^2$,

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{a \times a} . \tag{23}$$

The intra-unit correlation coefficient is

$$\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} . \tag{24}$$

Based on the idea of ANOVA, it is obvious that

$$\begin{aligned} E(MS_{\text{treatment}}) &= b \sum_{i=1}^{a} \tau_1^2/(a-1) + \sigma_1^2 , \\ E(MS_{\text{block}}) &= b\sigma_2^2 + \sigma_1^2 , \\ E(MS_{\text{residual}}) &= \sigma_1^2 . \end{aligned} \tag{25}$$

And the variance component σ_1^2 and σ_2^2 are

$$\begin{aligned} \sigma_1^2 &= E(MS_{\text{residual}}) , \\ \sigma_2^2 &= \frac{MS_{\text{block}} - MS_{\text{residual}}}{b} . \end{aligned} \tag{26}$$

If we substitute σ_1^2 and σ_2^2 in Eq. (24) by equations above, the intra-correlation coefficient is

$$\rho = \frac{MS_{\text{block}} - MS_{\text{residual}}}{MS_{\text{block}} + (b-1)MS_{\text{residual}}} . \tag{27}$$

Table 7. ANOVA of the serum coagulation time of four methods.

Source	SS	DF	MS	F	P
Total	105.7787	31	3.4122		
Between Groups	13.0163	3	4.3388	6.62	0.0025
In Groups	92.7624	28	3.3129		
Between Block	78.9888	7	11.2841	17.20	0.0000
Residual	13.7738	21	0.6559		

If every block has a different size (e.g. missing values), the intra-unit correlation of randomized block design data can be denoted as

$$\rho = \frac{MS_{\text{block}} - MS_{\text{residual}}}{MS_{\text{block}} + (m_0 - 1)MS_{\text{residual}}}\,, \tag{28}$$

where

$$m_0 = \bar{m} - \frac{\sum (m_i - \bar{m})^2}{(a-1)\sum m_i}\,. \tag{29}$$

4.3. *Example* 9. *Analysis of randomized block design data*

The analysis of Example 1. We begin with the ANOVA Table 7.[9]

The variance component σ_1^2 and σ_2^2 are

$$\sigma_0^2 = MS_{\text{residual}} = 0.6559\,,$$

$$\sigma_1^2 = \frac{MS_{\text{block}} - MS_{\text{residual}}}{b} = \frac{11.2841 - 0.6559}{4} = 2.6571\,.$$

And the intra-correlation coefficient is

$$r = \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2} = \frac{2.6571}{0.6599 + 2.6571} = 0.8020\,.$$

Though ANOVA of correlated data is similar to that of traditional randomized block design in process and result, ANOVA of correlated data not only answers the question, "whether there is a difference between treatment groups", on which that of traditional randomized block design emphasizes, but also puts more emphasis on the further decomposition of variance and affords the intra-unit correlation. Thus, its model is more precise with richer information.

4.4. *Example* 10. 4×4 *cross-over design*

The analysis of log AUC data in Example 4. For this moment, the fixed effects that we should take into consideration is 4 treatments, A, B, C, D,

Table 8. ANOVA of log AUC.

Source	SS	DF	MS	F	P
Total	15.67277346	95			
ID(sequence)	0.96107428	20	0.04805371	4.71	< 0.0001
Sequence	0.06927165	3	0.02309055	2.26	0.0892
Period	0.13519601	3	0.04506534	4.42	0.0068
Treat	13.83396718	3	4.61132239	452.05	< 0.0001
Residual	0.67326434	66	0.01020097		

4 different periods and 4 different sequences. The 4 observations of the same subject are correlated.

And,

$$\sigma_0^2 = MS_{\text{Residual}} = 0.01020097\,,$$

$$\sigma_1^2 = \frac{MS_{ID(\text{Sequence})} - MS_{\text{Residual}}}{b} = \frac{0.04805371 - 0.01020097}{4}$$

$$= 0.009463185\,.$$

Accordingly,

$$\rho = \frac{MS_{ID(\text{Sequence})} - MS_{\text{Residual}}}{MS_{ID(\text{Sequence})} + (b-1)MS_{\text{Residual}}}$$

$$= \frac{0.04805371 - 0.01020097}{0.04805371 + (4-1) \times 0.01020097} = 0.4812\,.$$

4.5. *The condition of using ANOVA*

The ANOVA is limited to fairly balanced designs where there are tidy partitions of the total sum of squares. The model should be fairly simple so that a suitable covariance structure (symmetry) for the observations can be produced. For example, if $t = 4$ in repeated measurement data, the covariance matrix should be

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}.$$

So the symmetry means

(1) $\sigma_{ii} = \sigma_{jj} = \sigma^2$,
(2) $\sigma_{ij} = \rho\sigma^2$, $i \neq j$.

In other wards, symmetry means equal variance and equal intra-unit correlation.

When the data are not symmetry, the ANOVA would increase type I error. In 1958, Greenhouse and Geisse suggested a correction coefficient

$$\varepsilon = \frac{t^2(\bar{\sigma}_{ii} - \bar{\sigma}_{..})^2}{(t-1)(\sum \sigma_{ij}^2 + t^2\bar{\sigma}_{..}^2 - 2t\sum \bar{\sigma}_i^2)} \tag{30}$$

where t represents the times of repeated measurement, $\bar{\sigma}_{ii}$ is the average of variances in diagonal of covariance matrix, $\bar{\sigma}_{..}$ is the average of all elements in covariance matrix, and $\bar{\sigma}_i$ is the average of elements in ith row of covariance matrix.

Greenhouse and Geisse have shown that, $1/(t-1) \leq \varepsilon \leq 1$. If ε is not equal to 1, a modified $F = MS_{\text{Treatment}}/MS_{\text{Residual}}$ would not follow F distribution with degree of freedom $\nu_{\text{Treatment}}$ and ν_{Residual} but follow F distribution with degree of freedom $\varepsilon\nu_{\text{Treatment}}$ and $\varepsilon\nu_{\text{Residual}}$. Because of the cutting down of degree of freedom, the modified F test is conservative.

For the data in Example 1, the variance-covariance matrix is

$$\begin{pmatrix} 2.40286 & & & \\ 3.23143 & 5.26411 & & \\ 2.13857 & 3.00518 & 2.29125 & \\ 2.13143 & 3.41536 & 2.02036 & 3.29357 \end{pmatrix}.$$

Greenhouse–Geisser's $\varepsilon = 0.7996$. Thus, the degree of freedoms

$$\nu_{\text{Treatment}} = 0.7996 \times 3 = 2.4\,,$$

$$\nu_{\text{Residual}} = 0.7996 \times 21 = 16.8\,,$$

then $F = 6.62$, $P = 0.0056$, larger than $P = 0.0025$.

In 1970, Huynh and Feldt have proved that when $\varepsilon = 1$, the F test is valid. If covariance matrix is symmetry, then $\varepsilon = 1$ or otherwise $\varepsilon < 1$. On the other hand, $\varepsilon = 1$ does not necessary implies the covariance being symmetry. The exception is for 2×2 covariance matrix for twice repeated measurements, ε always equal to 1 even if the variances are unequal.

We should select a suitable method for dependent data according to the feature of the data set. Unfortunately, the suitable systematic methods for all types of dependent data have not been developed. Only several methods for special data set can be used now. For instance, the mixed models are employed for repeated measurements or data from randomized block design, crossover design, and some special procedures for longitudinal

data, etc. The multilevel models analysis[7] would be used if the structure of variance-covariance matrix is block diagonal. For general structure of variance-covariance matrix, which is not block diagonal, generalized least square procedure with Newton–Raphson iterations may be useful. Further research is needed.

5. GEE for Dependent Data

Generalized estimating equations (GEE) was put forward by Liang Zeger[10] which is an extension of generalized linear models that provides a unified and flexible approach to analysis of data from a longitudinal study. Of particular relevance when the repeated measurements are binary variables or counts, and a number of time dependent covariates are also measured (Qiguang Chen,[11] Lingping Xiong *et al.*[12]). GEE plays an important rule in modeling the possible correlations among the repeated observations for a given subject.[13]

5.1. *Introduction of GEE*

The key ideas are presented in terms of repeated measurements with the simplest dependent structure. Let y_{ij} be the observation of jth measurement of the ith unit, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$. $X_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{pij})$ represents the explanatory variables. The observations from the same unit are likely to be correlated, but the observations from different units are assumed in general to be independent.

If the marginal distribution of response variable y_{ij} is one of exponential family, then, by the theory of generalized linear models, the density functions would be

$$f(y_{ij}) = \exp[\{y_{ij}\mu_{ij} - a(\mu_{ij}) + b(y_{ij})\}\phi], \tag{31}$$

where ϕ is known as dispersion parameter or additional scale, $\mu_{ij} = h(\eta_{ij})$, $\eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}$. It can be proved that $E(y_{ij}) = a'(\mu_{ij})$, $\text{var}(y_{ij} = a''(\mu_{ij})/\phi$.

For random effects model, we have

$$\begin{cases} \hat{y}_{ij} = \mu_{ij} \\ g(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} \end{cases}, \tag{32}$$

where $g(\cdot) = h^{-1}(\cdot)$ as a link function. If there is correlation between the repeated observations, the correlation between n_i observations in unit i can be described by working correlation matrix $\boldsymbol{R}_i(\alpha)$. The times of repeated

measurement on subjects are different from each other, so the ranks of correlation matrices are also different from each other. $R_i(\alpha)$ depends on unknown parameter α, to which we refer as correlate parameter. For instance, for $\boldsymbol{R}_2$ in (3)

$$\rho_{st} = \begin{cases} 1 & \text{if } s = t\,, \\ \alpha & \text{if } s \neq t\,. \end{cases} \tag{33}$$

for $\boldsymbol{R}_4$ in (5)

$$\rho_{st} = \begin{cases} 1 & \text{if } s = t\,, \\ \alpha & \text{if } 0 < |s-t| \leq 2\,, \\ 0 & \text{if } |s-t| > 2\,. \end{cases} \tag{34}$$

for $\boldsymbol{R}_5$ in (6)

$$\rho_{st} = \begin{cases} 1 & \text{if } s = t\,, \\ \alpha^{|s-t|} & \text{if } s \neq t\,. \end{cases} \tag{35}$$

then the variance-covariance matrix of $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})'$ has the form

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\alpha) \boldsymbol{A}_i^{1/2} / \phi\,, \tag{36}$$

where $\boldsymbol{A}_i$ is diagonal matrix with the elements $h(\mu_{ij}) = \nu_{ij}\phi$ in diagonal, which are the function of the variance ν and the mean μ of y. Liang and Zeger[10] defined the GEE as

$$\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{E}_i = 0\,, \tag{37}$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$, $\boldsymbol{E}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$, and $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im_i})'$.

5.2. *Parameters estimations of GEE*

There are three types of parameters in GEE, covariate coefficients $\boldsymbol{\beta}$, the scale parameter ϕ, the correlation parameter α. But ϕ and α are functions of $\boldsymbol{\beta}$. We can get the estimation of $\boldsymbol{\beta}$ only if ϕ and α are known. Consequently, the estimation procedure of GEE is iterative.

The initial value of $\boldsymbol{\beta}$ will be the estimations from generalized linear model under the assumption that the observations are independent of one another, say $\boldsymbol{\beta}_i$.

The crude residuals of the model is

$$e_{ij} = y_{ij} - \mu_{ij} = y_{ij} - g^{-1}(\beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij})\,. \tag{38}$$

The Pearson residuals are

$$r_{ij} = \frac{\hat{y}_{ij} - \mu_{ij}}{\sqrt{\nu_{ij}}}\,. \tag{39}$$

Thus

$$\hat{\phi} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} r_{ij}/(N-p)\,. \tag{40}$$

The intra-unit correlation can be estimated from the current Pearson residuals. For exchangeable correlation, we have

$$\hat{\alpha} = \sum_{i=1}^{n}\left[\frac{\sum_{j=1}^{m_i}\sum_{l=1}^{m_i} r_{ij}r_{il} - \sum_{j=1}^{m_i} r_{ij}^2}{m_i(m_i-1)}\right] \Bigg/ \left[\sum_{i=1}^{n}\frac{\sum_{j=1}^{m_i} r_{ij}^2}{m_i}\right]. \tag{41}$$

For first order autocorrelation

$$\hat{\alpha} = \sum_{i=1}^{n}\frac{\sum_{j=1}^{m_i-1} r_{ij}r_{ij+1}}{m_i-1} \Bigg/ \left[\sum_{i=1}^{n}\frac{\sum_{j=1}^{m_i} r_{ij}^2}{m_i}\right]. \tag{42}$$

For stationary k-dependence

$$\hat{\alpha}=\sum_{i=1}^{n}\left[\frac{\sum_{j=1}^{m_i} r_{ij}^2}{m_i},\frac{\sum_{j=1}^{m_i-1} r_{ij}r_{ij+1}}{m_i-1},\ldots,\frac{\sum_{j=1}^{m_i-k} r_{ij}r_{i,j+k}}{m_i-k}\right] \Bigg/ \left[\sum_{i=1}^{n}\frac{\sum_{j=1}^{m_i} r_{ij}^2}{m_i}\right], \tag{43}$$

where the first element of α is 1 and the elements after k-order are 0.

At a given iteration, the scale parameter ϕ and correlation parameters α can be estimated from the current Pearson residuals. Given the estimated of ϕ and α, we can calculate an updated estimate of $\boldsymbol{\beta}$ by iteratively reweighed least squares (IRLS). These two steps are iterated until the procedure convergence.

5.3. *Analysis of examples*

5.3.1. *Example* 11. *Analyses of the data in Example* 1

The random effect model is

$$y_{ij} = \beta_0 + \beta_2 g_{2ij} + \beta_3 g_{3ij} + \beta_4 g_{4ij} + e_{ij}\,,$$

where g_1, g_2, g_3 and g_4 are dummy variables of treatment groups. The correlations between the observations of different treatments for the same subjects are assumed equal. Results are shown in Table 9.

Intra-subject correlation $\rho = 0.8020$. We obtain the same results as in Example 9.

Table 9. Estimated results of data in Example 1 by GEE.

Variables	Coefficient	SE	Z	P
g_2	0.4125	0.378783	1.09	0.276
g_3	0.6375	0.378783	1.68	0.092
g_4	1.7250	0.378783	4.55	0.000
Constant	9.3000	0.601958	15.45	0.000

Table 10. The results of fitting two GEE models for data in Example 3.

	Logistic				Probit			
	Coefficient	Std.	Z	P	Coefficient	Std.	Z	P
Group	−1.0144	0.4985	−2.03	0.042	−0.5611	0.2702	−2.08	0.038
Constant	2.1484	0.4039	5.32	0.000	1.2564	0.2086	6.02	0.000

5.3.2. *Example* 12. *Analysis of data in Example* 3

We fit both logistic regression model and probit model as follows

$$y_{ij} = \frac{e^{\alpha+\beta \text{ treat}}}{1+e^{\alpha+\beta \text{ treat}}} + e_{ij}\,,$$

$$y_{ij} = \Phi^{-1}(\alpha + \beta \text{ treat}) + e_{ij}\,,$$

where the subscripts of *treat* are omitted. Table 10 shows the results.

Two intra-litter correlation coefficients are estimated based on logistic model and probit model and they are all equal to 0.1556.

5.3.3. *Example* 13. *Analysis of data in Example* 8

Example 8 has count data, with successive two-week seizure counts for each of 59 epileptics. Poisson regression model will be used. In contrast to the examples mentioned above, beside treatment effects, the covariables, such as *age*, *ln*(*base*), and *time* effects, should also be considered. The mixed effect Poisson regression model for the data is

$$\ln(\lambda) = \alpha + \beta_1\ treat + \beta_2\ time + \beta_3\ age + \beta_4\ ln(base)\,.$$

For repeated measurement data, the intra-subject correlation structure may be exchangeable or first order auto-correlation.

For exchangeable structure, intra-subject correlation estimated from GEEs is 0.7690, and deviance = 3551.0. For autocorrelate structure the intra-subject correlation is 0.7990^t, where t is time interval between two observations (1 unit of t is 2 weeks) and deviance = 3554.79.

Table 11. GEE estimators for data in Example 8.

Parameter	Coefficient	SE	Z	P
Constant	−1.7760	0.3692	−4.81	< 0.0001
Treat	−0.2938	0.1445	−2.03	0.0420
Time	−0.0443	0.0353	−1.26	0.2092
Age	0.0231	0.0067	3.46	0.0005
ln(base)	0.9817	0.0796	12.33	< 0.0001

The working matrix of autocorrelation structure is

$$\begin{matrix} 1.0000 & 0.4533 & 0.2055 & 0.0931 \\ 0.4533 & 1.0000 & 0.4533 & 0.2055 \\ 0.2055 & 0.4533 & 1.0000 & 0.4533 \\ 0.0931 & 0.2055 & 0.4533 & 1.0000 \end{matrix}$$

The numbers of parameter of two models are equal. Therefore, the smaller the deviance is, the better the model will be. According to this, we conclude that exchangeable structure is suitable for the data. Estimated results are shown in Table 11.

The results show that the two-week seizure counts for those in test group are significantly smaller than those in placebo group. The counts are related to age and the baseline. No evidence shows that the counts change over time.

GEE can cope with data with missing values. For the numerical data in a paired design or randomized block design, the paired t-test and ANOVA require the data are balanced without missing, while the GEE does not. Furthermore, when the times of measurement are not common to all the experimental units, or when the numbers of the unit in clusters are not the same, the use of GEE will still be applicable. Liang[9] has proved that if there are not too many missing values and missing is random, the GEE estimation is robust.

GEE obtains the estimation of covariance matrix $\boldsymbol{V}$ or working correlation matrix $\boldsymbol{R}$ by using simple regression or "moment" procedures based upon functions of the actual calculated raw residuals. Theoretically, the structure of working correlation matrix can be specified arbitrarily. However, GEE focuses on modeling the fixed effects rather than exploring the structure of the random component of the model. It does not consider the case where the explanatory variables have an influence on covariance of response variable.

6. Multilevel Models for Dependent Data

Many kinds of dependent data collected in medical and biological sciences have a *hierarchical* or *clustered* structure. We refer to a hierarchy as consisting of *units* grouped at different *levels*. For example, in a clustered sampling survey where the sampling units are families, offsprings may be the level 1 units in a 2-level structure where the level 2 units are the families. Repeated measurements are the level 1 units in a 2-level structure where the level 2 units are the individuals. Repeated measurements are the level 1 unit in a 3-level structure where the level 3 units are the hospitals and level 2 units are the patients. The existence of such data hierarchies is created by experimental design. Low levels are nested in the high levels.

6.1. *Introduction of multilevel model*

Multilevel model was put forward by Harver Goldstein[14] for the data with hierarchical or clustered structure. The key ideals are to estimate variances on each level and to address how the explanatory variables affect the variances. The multilevel model, therefore, enables data analysis to obtain statistically efficient estimations of regression coefficients, and provides correct standard errors, confidence intervals and significance tests by using the clustering information.

We discuss a simple 2-level model, without lose of generalizibility, of one explanatory variable x_1.

$$y_{ij} = \beta_{0j} + \beta_{ij} x_1 + \varepsilon_{ij} \tag{44}$$

i stands for level 1 units, j for level 2 units. $i = 1, \ldots, n_j; j = 1, \ldots, m$, where, β_{0j} and β_{1j} are random variables with

$$\beta_{0j} = \beta_0 + u_{0j}\,, \quad \beta_{1j} = \beta_1 + u_{ij}\,,$$

where β_0 and β_1 are fixed parameters, u_{0j}, u_{1j} are random variables in level 2 with parameters

$$E(u_{0j}) = E(u_{1j}) = 0$$

$$\mathrm{var}(u_{0j}) = \sigma^2_{u0}\,, \quad \mathrm{var}(u_{1j}) = \sigma^2_{u1}\,, \quad \mathrm{cov}(u_{0j}, u_{1j}) = \sigma_{u01}\,.$$

ε_{ij} are random variables in level 1 with parameter

$$E(\varepsilon_{1j}) = 0\,, \quad \mathrm{var}(\varepsilon_{ij}) = \sigma^2_0\,.$$

We also assume that $\mathrm{cov}(\varepsilon_{ij}, u_{0j}) = \mathrm{cov}(\varepsilon_{ij}, u_{1j}) = 0$.

We can now write the level 2 model in the form

$$y_{ij} = \beta_0 + \beta_1 x + (u_{0j} + u_{1j} x + \varepsilon_{ij})\,. \tag{45}$$

The model consists of a fixed part and a random part. In contrast to a general mixed effect model (for example, variance component model, mixed linear model, GEE), explanatory variables can be included in random part of multilevel model with random coefficients u_{1j}. The multilevel model, therefore, is also refered to as random coefficient model.

The covariance matrices is block diagonal

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{n_1} & & & \\ & \boldsymbol{V}_{n_2} & & \\ & & \ddots & \\ & & & \boldsymbol{V}_{n_m} \end{pmatrix}. \tag{46}$$

If no covariate is included in the random part of the model, $\sigma^2_{u1} = 0$ and the model reduces to a general mixed effects model with

$$\boldsymbol{V}_{n_i} = \operatorname{cov}(y_{ij}|\boldsymbol{X}\boldsymbol{\beta}) = \begin{pmatrix} \sigma^2_{u0} + \sigma^2_0 & \sigma^2_{u0} & \cdots & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{u0} + \sigma^2_0 & \cdots & \sigma^2_{u0} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^2_{u0} & \sigma^2_{u0} & \cdots & \sigma^2_{u0} + \sigma^2_0 \end{pmatrix}_{n_i \times n_i}. \tag{47}$$

Equation (47) can be denoted as $\sigma^2_{u0}\boldsymbol{J}_{(n_i)} + \sigma^2_0 \boldsymbol{I}_{(n_i)}$. Where, $\boldsymbol{J}_{(n)}$ is $n \times 1$ vector with all elements 1, $\boldsymbol{I}_{(n)}$ is n dimension unit matrix with all elements in diagonal 1, others 0. Then the intra-unit correlation can be estimated by

$$\rho = \frac{\operatorname{cov}(u_{0j} + \varepsilon_{i_1 j} + u_{0j} + \varepsilon_{i_2 j})}{\sqrt{\operatorname{var}(u_{0j} + \varepsilon_{i_1 j}) \cdot \operatorname{var}(u_{0j} + \varepsilon_{i_2 j})}} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_0}\,. \tag{48}$$

If covariate was considered, $\sigma^2_{u1} \neq 0$ and

$$\boldsymbol{V}_{n_i} = (\sigma^2_{u0} + 2\sigma_{u01} x + \sigma^2_{u1} x^2)\boldsymbol{J}_{(n_i)} + \sigma^2_0 \boldsymbol{I}_{(n_i)}\,. \tag{49}$$

The intra-unit correlation can be estimated by

$$\rho = \frac{\sigma^2_{u0} + 2\sigma_{u01} x + \sigma^2_{u1} x^2}{\sigma^2_{u0} + 2\sigma_{u01} x + \sigma^2_{u1} x^2 + \sigma^2_0}\,. \tag{50}$$

It is thus clear that intra-unit correlation has relation to the explanatory variables.

6.2. *Estimation of parameters of multilevel model*

Parameters in multilevel model can be estimated by iterative generalized least squares (IGLS)[14] or Restricted Iterative Generalized Least Squares (RIGLS).[15]

Let $\mathrm{cov}(\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{V}$, if $\boldsymbol{V}$ is known, then according to the generalized least square estimation

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{-\boldsymbol{1}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{-\boldsymbol{1}}\boldsymbol{Y}\,, \quad \mathrm{cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{-\boldsymbol{1}}\boldsymbol{X})^{-\boldsymbol{1}}\,. \tag{51}$$

But in fact, $\boldsymbol{V}$ is usually unknown and expressed by random coefficients. For known $\boldsymbol{\beta}$ we form the residuals of y_{ij}

$$\tilde{\boldsymbol{Y}} = \{\tilde{y}_{ij}\} = \{y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}\}\,. \tag{52}$$

If we form the cross-product matrix $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^{T}$) we see that the expected value of this is simply $\boldsymbol{V}$. From the equation

$$\mathrm{vec}(\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^{T}) = \mathrm{vec}(\boldsymbol{V}) + \boldsymbol{R}\,, \tag{53}$$

we estimate parameters σ^2_{u0}, σ^2_{u1}, σ_{u01} and σ^2_0 by means of generalized least squares where $\mathrm{vec}(\cdot)$ is the vector operator.

The estimation procedure is iterative. We would usually start from "reasonable" estimates of the fixed parameters $\boldsymbol{\beta}$. Typically these will be those from an initial OLS estimation. From these we form the "raw" residuals (52), estimate random coefficients; and obtain an improved estimator of $\boldsymbol{V}$; then return to (51) to obtain new estimates of the fixed effects $\boldsymbol{\beta}$; and so on. Alternate between the random and fixed parameters estimation until the procedure convergence.

The IGLS procedure produces biased estimates in general and this can be important in small samples. Goldstein[15] shows how a simple modification leads to restricted iterative generalized least squares (RIGLS) by substituting $\boldsymbol{V} - \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{-\boldsymbol{1}}\boldsymbol{X})\boldsymbol{X}^{\boldsymbol{T}}$ for its corresponding term $\boldsymbol{V}$ in (53) to produce an unbiased estimate.

For multilevel generalized linear model, in order to work with a linearized model, we will use Taylor expansion. There are two produces to treat high-level residuals when forming Taylor expansion. One is to add current residuals to the linear component of the nonlinear function and the another does not add. The former is predictive quasi-likelihood (PQL), while the latter is marginal quasi-likelihood (MQL). In many applications, MQL procedure tends to underestimate the values of both the fixed and random parameters, especially where n_{ij} is small. So Goldstein[14] suggested that PQL be used in fitting generalized model rather than MQL. In addition, he

Table 12. MLn estimation for data in Example 1.

Variable	Coefficient	SE	Z	P
g_2	0.4125	0.4049	1.0188	0.3083
g_3	0.6375	0.4049	1.5745	0.1154
g_4	1.7250	0.4049	4.2603	0.0000
Constant	9.3000	0.6435		

also pointed out that greater accuracy is to be expected if the second-order approximation is used rather than first-order based upon the first term in the Taylor expansion.[16]

6.3. *Example* 14. *Analysis of data in Example* 1

This is the simplest case with 4 units in level 1 in a 2-level structure where the level 2 units are the subjects. The model has the form as

$$y_{ij} = \beta_0 + \beta_2 g_{2ij} + \beta_3 g_{3ij} + \beta_4 g_{4ij} + u_{0j} + e_{ij}\,.$$

To obtain IGLS estimation of the parameters, we use software MLn.[17] The results are shown in Table 12. The estimation of variance in level 1 is $\sigma_0^2 = 0.6559$, in level 2 $\sigma_{u_0}^2 = 2.6571$, with standard error $SE[\sigma_0^2] = 0.1893$, $SE[\sigma_{u_0}^2] = 1.411$, respectively.

Then

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_0^2} = \frac{2.6571}{2.6571 + 0.6559} = 0.8020\,.$$

This results are similar to those from ANOVA and GEE.

6.4. *Example* 15. *Analysis of data in Example* 7

This is a 3-level model. Subjects are level 2 units clustered within centers that are level 3 units. Repeated measurements from the same subject are level 1 units nested within level 2 unit. The results are shown in Table 13.

The multilevel model decomposes the variance into 3 levels. 0.1156 for level 1, 0.3151 for level 2 and 0.1190 for level 3. Thus, intra-subject correlation can be estimated as

$$\frac{\sigma_{u0}^2 + \sigma_{\nu 0}^2}{\sigma_0^2 + \sigma_{u0}^2 + \sigma_{\nu 0}^2} = \frac{0.3151 + 0.1190}{0.1156 + 0.3151 + 0.1190} = 0.7897\,.$$

And, intra-center correlation can be estimated as

$$\frac{\sigma_{\nu 0}^2}{\sigma_0^2 + \sigma_{u0}^2 + \sigma_{\nu 0}^2} = \frac{0.1190}{0.1156 + 0.3151 + 0.1190} = 0.2165\,.$$

Table 13. *MLn* estimation of data in Example 7.

			Coefficient	*SE*
Fixed effect	CONS		1.7230	0.2545
	TREAT		0.3273	0.1135
	COMP		0.0384	0.0900
	AGE		−0.0090	0.0076
	SEX		−0.2285	0.1483
	TIME		0.1616	0.0056
Random effect	Level 3	$\sigma^2_{\nu 0}$	0.1190	0.0657
	Level 2	σ^2_{u0}	0.3151	0.0859
	Level 1	σ^2_0	0.1156	0.0065

Theoretically, multilevel model can fit for arbitrary levels. The most powerful software *MLwin* could fit models up to 7 levels. It is sufficient in practice.

6.5. *Multilevel logistic regression*

Multilevel model can be expanded to the case where the error term in the model is non-normal distribution.[18] In the rest of this section we will focus on the multilevel models with binomial distribution, or Poisson distribution.

To make matters concrete, consider the data in Example 3. Let y_{ij} be an observation of ith pup from jth pregnant rat. If the pup is normal then $y_{ij} = 0$, else $y_{ij} = 1$. Let f_{ij} be fixed part of the model, and r_j be the random part, and π_{ij} be the expected value of the response for the ijth level 1 unit. A 2-level logistic regression model would have the form

$$y_{ij} = \pi_{ij} + \varepsilon_{ij} = \frac{\exp(f_{ij} + r_j)}{1 + \exp(f_{ij} + r_j)} + \varepsilon_{ij} \tag{54}$$

$$f_{ij} = \alpha + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij}\,,$$

$$\varepsilon_{ij} = e_{ij}\sqrt{\pi_{ij}(1 - \pi_{ij})}\,.$$

In general, ε_{ij} follows a binomial distribution, but sometimes it is extra-binomial. The variance of ε_{ij} can be written in the form of $\sigma_0^2 \pi_{ij}(1 - \pi_{ij})$. Here, σ_0^2 is refered to as extra-binomial variance (or over dispersion). When $\sigma_0^2 = 1$, it is purely binomial. We will assume $\sigma_0^2 = 1$, and $r_j \sim N(0, \sigma_0^2)$ in this section.

Let r_{ij} be the Pearson residual of the model

$$r_{ij} = \frac{y_{ij} - \pi r_{ij}}{\sqrt{\pi_{ij}(1 - \pi_{ij})/n_{ij}}}\,. \tag{55}$$

Table 14. Results of fitting 3 2-level models for data in Example 3.

Parameter	Models		
	Logit	Probit	C log-log
α	1.127(0.3380)	0.6933(0.1929)	0.3463(0.1686)
β	1.028(0.5099)	0.5655(0.2824)	0.4692(0.2383)
σ_1^2	1.212(0.5061)	0.3822(0.1578)	0.2779(0.1131)
σ_0^2	1	1	1
intra-unit correlation ρ	0.1731	0.1734	0.1739
$-2\ln(L)$	210.686	211.610	213.084

Then the intra-unit correlation can be defined as

$$\rho = \sum_{j=1}^{m} \left[\frac{\sum_{i=1}^{n_j} \sum_{k=1}^{n_j} r_{ij} r_{kj} - \sum_{i=1}^{n_j} r_{ij}^2}{n_j(n_j - 1)} \right] \Bigg/ \left[\sum_{j=1}^{m} \frac{\sum_{i=1}^{n_j} r_{ij}^2}{n_j} \right]. \tag{56}$$

In contrast to GEE, 2-level logistic model decompose the residuals into each level. The residuals in the level 1 are linear to the response, while the residues in level 2 are nonlinear.

The 2-level logistic model for Example 3 can be written as

$$y_{ij} = \frac{\exp(\alpha + \beta_{\text{Group}} + r_j)}{1 + \exp(\alpha + \beta_{\text{Group}} + r_j)} + e_{ij}\sqrt{\pi_{ij}(1 - \pi_{ij})}$$

where the subscripts of Group are omitted. The results are shown in Table 14.

6.6. *Multilevel Probit model and complementary log-log model*

The expected proportion π_{ij} in (55) is modeled using a logit link function. If we use probit link function, a 2-level probit model would have the form

$$y_{ij} = \pi_{ij} + \varepsilon_{ij} = \Phi(f_{ij} + r_j) + \varepsilon_{ij}\,. \tag{57}$$

If we use complementary log-log link function, then a 2-level complementary log-log model would have the form

$$y_{ij} = \pi_{ij} + \varepsilon_{ij} = 1 - \exp\{-\exp(f_{ij} + r_j)\} + \varepsilon_{ij}\,. \tag{58}$$

Other notations are similar to a level 2 logistic model in (54).

6.7. *Example* 16

We fitted 2-level logistic, probit, and complementary log-log models for data in Example 3 respectively. The results are shown in Table 14.

The results show that the intra-unit correlations estimated from 3 models are quite similar.

6.8. *Multilevel Poisson regression model*

For count data multilevel Poisson regression model would be fitted. For a 2-level model, it can be written as

$$y_{ij} = m_{ij} + \varepsilon_{ij} = \exp(f_{ij} + r_j) + \varepsilon_{ij}\,,$$
$$f_{ij} = \alpha + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij}\,. \tag{59}$$

We usually assume that ε_{ij} follows a Poisson distribution with $\mathrm{var}(y_{ij}|m_{ij}) = m_{ij}$. But sometimes it is extra-Poisson with conditional variance of $\mathrm{var}(y_{ij}|m_{ij}) = m_{ij} + km_{ij}^2$. When $k > 0$, it is negative binomial distribution. When $k = 0$, it is purely Poisson. Here we keep $k = 0, r_j \sim N(0, \sigma_1^2)$.

Let r_{ij} be Pearson residuals of the model as follow:

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}}\,. \tag{60}$$

The definition of intra-unit correlation is similar to (55).

6.9. *Example* 17. *Fitting a 2-level Poisson regression model for data in Example* 8

The response is two-week seizure counts for epileptics and is a count data.

The Poisson model is sufficed here. The results are shown in Table 15.

The results show that the counts are correlated with age and time. No significance can be detected in test group and placebo group. But it is

Table 15. Estimated results of random effect model for Example 8.

Parameter	Coefficient	SE	Z	P
Treat	−0.07606	0.27020	−0.28150	0.7783
Trial	0.19900	0.05859	3.39648	0.0007
Time	−0.05743	0.02026	−2.83465	0.0046
Age	−0.01685	0.01788	−0.94239	0.3460
Constant	1.88100	0.55440		

significantly different between the counts before and after medication. The intra-subject correlation estimated from the model is 0.7776.

6.10. *Multilevel logistic models for multiple response categories*

In this section we extend the multilevel logistic model for binomial response to the cases of multiple categories and ordinal categories. When the response is multiple categories without order, a multilevel polytomous logistic model will be fitted. And when the response is ordinal, a multilevel ordinal logistic model will be fitted.

For example, let's consider a 2-level model with one explanatory variable. The response is now multiple with k categories. A multilevel polytomous logistic model can be defined as

$$\pi_{ij}^{(s)} = \frac{\exp(\beta_0 + \beta_1 x_{1ij}^{(s)} + u_{0j}^{(s)})}{1 + \exp(\beta_0 + \beta_1 x_{1ij}^{(s)} + u_{0j}^{(s)})} + \varepsilon_{ij}\,, \tag{61}$$

$s = 1, 2, \ldots, k$. Under the standard assumption that the observed response proportions follow a multinomial distribution, the level 2 covariance matrix has the form

$$n_{ij}^{-1} \begin{pmatrix} \pi_{ij}^{(1)}(1-\pi_{ij}^{(1)}) & \cdots & & \\ -\pi_{ij}^{(1)}\pi_{ij}^{(2)} & \pi_{ij}^{(2)}(1-\pi_{ij}^{(2)}) & & \\ \vdots & \vdots & \ddots & \\ -\pi_{ij}^{(1)}\pi_{ij}^{(k)} & -\pi_{ij}^{(2)}\pi_{ij}^{(k)} & \cdots & \pi_{ij}^{(k)}(1-\pi_{ij}^{(k)}) \end{pmatrix}. \tag{62}$$

If k categories are ordered, we should base our model upon the cumulative response probabilities rather than the responses probabilities for each category. The multilevel ordinal logistic model can be defined as

$$\gamma_{ij}^{(s)} = \frac{\exp(\beta_0 + \beta_1 x_{1ij}^{(s)} + u_{0j}^{(s)})}{1 + \exp(\beta_0 + \beta_1 x_{1ij}^{(s)} + u_{0j}^{(s)})}\,, \tag{63}$$

where $\gamma_{ij}^{(s)}$ is cumulative probability for $s = 1, 2, \ldots, k$. If we assume an underlying multinomial distribution for the category probabilities, the cumulative proportions have a covariance matrix given by $\pi_{ij}^{(r)}(1-\pi_{ij}^{(s)})/n_{ij}$, $(r < s)$.

Table 16. The results of a 3-level cumulative logistic model for data in Example 7.

Parameter	Estimation	SE
Fixed parameters		
Y1	−1.482	0.8162
Y2	0.9309	0.8089
Y3	4.299	0.8284
Comp	−0.03902	0.2977
Treat	−1.187	0.3245
Sex	0.5575	0.4685
Age	0.03687	0.02515
Time	−0.5268	0.03924
Random parameters		
Level 3	1.001	0.5818
Level 2	1.586	0.3197
Level 1	1	0

6.11. *Example* 18. *Analysis of data in Example* 7

The effectiveness variable in Example 7 is an ordinal response, with 0 stands for invalid or worse effects of treatment on the subject, 1 for improved, 2 for notable improved, and 3 for recovery. A multilevel ordinal logistic model with cumulative odds was fitted. The results are shown in Table 16.

Where, the centers are level 3 units, the subjects are level 2 units and repeated observations are level 1 units.

6.12. *Relationship between intra-unit correlation and explanatory variable*

The key idea of multilevel model is to express the variance in each level by explanatory variables. In many applications, mean squared error is related to some explanatory variables. As a result, the intra-unit correlations are related to them, too. This issue would be resolved by adding the explanatory variables to the random part of multilevel models.

6.13. *Example* 19. *Analysis of the data in Example* 5

The data show that the variance of ESS is changing over time. Let G1 and G2 be the dummy variables of groups. We fit a 2-level model for the data in which the time variable is added into random part at level 1 of the model. The results are shown in Table 17.

The results show that in the middle dose group ($treat = 1$) and the high dose group ($treat = 2$), the intra-subject correlation is 0.7956, which

Table 17. The results of fitting a 2-level model for data in Example 5.

	Parameter	Estimate	SE
Fixed	Cons	98.19	4.175
	Time	1.283	0.1285
	Age	0.18	0.1531
	G1	−8.356	3.23
	G2	−2.972	3.255
	G1*Time	2.778	0.2771
	G2*Time	2.977	0.1902
Random			
Level 2	Cons/Cons	38.59	10.48
Level 1	Cons/Cons	9.914	1.44
Level 1	G1*Time/Cons	3.195	0.7093
Level 1	G2*Time/Cons	0.195	0.2165

is independent on time. But in the low dose (placebo) group ($treat = 0$), the intra-subject correlation coefficient depends upon time. The correlation of observations at $Time_1$ and $Time_2$ would be estimated by

$$\frac{38.59}{\sqrt{(38.59 + 9.914 + 3.195 \times Time_1)(38.59 + 9.914 + 3.195 \times Time_2)}}.$$

6.14. *Multivariate multilevel models*

So far, we have only considered a single response variable. In many applications, we wish simultaneously to model several responses functions of explanatory variables. In Example 4, AUC, $C_{\max}$ and T_{50} will be considered together as responses to test the bioequivalence of domestic and imported rosiglitazone maleate tablets (RMT). This goal could be achieved by fitting a multivariate multilevel model.

For the sake of convenience, we consider the multivariate multilevel model with two response, the logarithmic values of AUC (also denoted AUC) and $C_{\max}$, and treat the subject as a subject-level unit and 4 treatment effects (observations repeated measured on subjects) as period-level units which are clustered in subject-level. Besides intra-subject correlation, other properties of this model should be considered: observations of AUC are correlated between different periods of trial, and so do $C_{\max}$; and AUC and $C_{\max}$ are correlated either in the same period or in different periods.

The results shown that: the intra-subject correlation of AUC is

$$0.008301/(0.008301 + 0.011595) = 0.4172\,.$$

Table 18. The results fitting multivariate multilevel model for data in Example 4.

	Parameters	Coefficient	SE	Intra-subject correlation
Fixed effects	Cons_AUC	6.82577	0.02661	
	Cons_$C_{\max}$	5.13547	0.02550	
	Treat_AUC	0.03252	0.02198	
	Treat_$C_{\max}$	0.03563	0.02099	
	Drug_AUC	0.75844	0.02198	
	Drug_$C_{\max}$	0.73093	0.02099	
Random effects				
Subject level	Cons_AUC/Cons_AUC	0.008301	0.003274	1
	Cons_$C_{\max}$/Cons_AUC	0.003883	0.002426	0.486
	Cons_$C_{\max}$/Cons_$C_{\max}$	0.007678	0.003017	1
Period level	Cons_AUC/Cons_AUC	0.011595	0.001936	1
	Cons_$C_{\max}$/Cons_AUC	0.003444	0.001369	0.311
	Cons_$C_{\max}$/Cons_$C_{\max}$	0.010572	0.001765	1

The intra-subject correlations of $C_{\max}$ is

$$0.007678/(0.007678 + 0.010572) = 0.4207\,.$$

The Pearson correlation of AUC and $C_{\max}$ is 0.311 in level 1, and 0.486 in level 2.

7. Sampling Distribution and Confidence Interval of Intra-unit Correlation

7.1. *Confidence interval of intra-unit correlation*

The intra-unit correlation coefficient estimated by a generalized estimation equation or a multilevel model is a point estimation. But we did not estimate its estimation errors and had little ideas of its sampling distribution. The bootstrap may be applied to estimate the CI of intra-unit correlation.[19]

Bootstrap is a data-based simulation method for statistical inference, which can be used to study the variability of estimated characteristics of the probability distribution of a set of observations, and provide confidence intervals for parameters and hypothesis test in situations where these are difficult or impossible to derive closed form formulas. The basic idea of the procedure involves sampling with replacement to produce random samples of size n from original data, each of these is known as a bootstrap sample and each provides an estimate $\theta(b)$ of the interesting parameter, θ. Repeating the process a large number of times, say B = 500 or more,

provides the required information on the variability of the estimator. For example, the type of distribution, standard error of the bootstrap estimates. An approximate 95% confidence interval can be derived from mean ± 1.96 SD if the bootstrap estimates are normally distributed, and from the 2.5% and 97.5% quartiles of the replicate values if the bootstrap estimates is not normally distributed. The confidence interval derived from bootstrap sampling is known as bootstrap confidence interval.

If the population distribution is known, bootstrap samples can be randomly samped not from the original data, but from the population distribution. The former produce is known as non-parametric bootstrap, and the latter as parametric bootstrap.

Research shows that there are two particularities in applying the bootstrap estimation to data of dependent design.[20] First, it is not proper to adopt a parametric estimation because of the difficulty in making a judgment to the distribution of the data. So we suggest adopting a non-parametric estimation. Second, it is not proper to apply a random sample directly to observations because of the non-independence of observations. So we suggest that we sample high level units. And if some high level unit is sampled, all the observations in this unit will be sampled. As examples, data of Example 2 should be sampled by family; data of Example 3 should be sampled by litter; and data of Examples 1, 4 and 5 should be sampled by patient.

The estimator of intra-family correlation of Example 2 is 0.5674. If we sample the data on families, make 500 resamplings, and estimated by GEE, then the non-parametric 95% CI of intra-family correlation is 0.2875–0.8874. If we sample the data of Example 8 on patients and make the same analysis, we estimated the non-parametric 95% CI of correlation 0.5219–0.8874. Both of the two bootstrap sampling distributions of intra-unit correlation are skew. Because both of the two CIs do not include 0, we may accept that the intra-subject correlations exist.

7.2. *The sampling distribution of intra-unit correlation*

To estimate the type and characteristic of distribution of intra-unit correlation, we use the Monte Carlo method. One thousand simulations were generated from specific population with known ρ and corresponding assumed parameters. For each set of simulated data, we fit a 2-level model, estimate σ_0^2, σ_e^2, and then the intra-unit correlation. We then investigate the distribution of intra-unit correlation based on 1000 estimators of ρ.

We may assume that there are m individuals (two-level unit) and each individual have k repeatedly measured values (one-level unit), then we have $n = m \times k$ observations. In order to investigate the effect of units of level 1 and units of level 2 on the intra-unit correlation when overall sample size of observations are the same, we design the grids are $(k, m) = (4, 10)$, (4, 20), (4, 30), ..., (4, 100) and $(k, m) = (8, 5)$, (8, 10), (8, 15), ..., (8, 50) respectively, and the intra-unit correlation coefficients is 0.1–0.9 respectively.

Without losing generality, in analog investigation, we do not take fixed but random effect into account, because the intra-unit correlation is related only to random effect. Furthermore, we assume that the intra-unit correlation structure is exchangeable.

Now we may consider two situations. One is the simplest situation

$$y_{ij} = \mu_j + e_{ij} \,. \tag{64}$$

Only one random effect is considered both in levels 1 and level 2. Then the variance of y is $\sigma_0^2 + \sigma_e^2$; The variance-covariance matrix of y is $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{R}, \boldsymbol{R}, \ldots, \boldsymbol{R})$. If k is 4,

$$\boldsymbol{R} = \begin{pmatrix} \sigma_0^2 + \sigma_e^2 & & & \\ \sigma_0^2 & \sigma_0^2 + \sigma_e^2 & & \\ \sigma_0^2 & \sigma_0^2 & \sigma_0^2 + \sigma_e^2 & \\ \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 + \sigma_e^2 \end{pmatrix}, \tag{65}$$

the intra-unit correlation can be calculated by Eq. (48).

Another situation is more complex

$$y_{ij} = \mu_j + \nu_j x_{ij} + e_{ij} \,. \tag{66}$$

Level 1 has two random effect terms: one is random error. Another random effect term is related to independent variable, namely the variance of y, $\sigma_0^2 + \sigma_1^2 (x_{ij})^2 + \sigma_e^2$, and is affected by explanatory variable. Let $x_{ij} = j - 1$, the variance-covariance matrix of y is $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{R}, \boldsymbol{R}, \ldots, \boldsymbol{R})$, when $k = 4$,

$$\boldsymbol{R} = \begin{pmatrix} \sigma_0^2 + \sigma_e^2 & & & \\ \sigma_0^2 & \sigma_0^2 + \sigma_1^2 + \sigma_e^2 & & \\ \sigma_0^2 & \sigma_0^2 & \sigma_0^2 + 4\sigma_1^2 + \sigma_e^2 & \\ \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 + 9\sigma_1^2 + \sigma_e^2 \end{pmatrix}. \tag{67}$$

We may calculate intra-unit correlation by Eq. (48) after deducting the effect of explanatory variable to y.

The parameters of the model can be estimated by the Restricted Iterative Generalized Least Square (RIGLS) method. The simulation study is made by using specialized multilevel model software MLn.[17]

Table 19 lists the simulated results of six models. Models A–D are generated based on Eq. (64), and their amounts of units of levels 1 and 2 are model A with $(k, m) = (4, 10)$, model B with $(k, m) = (8, 5)$, model C with $(k, m) = (4, 100)$, model D with $(k, m) = (8, 50)$, respectively. The overall sample size of model A is equal to that of model B, while the overall sample size of model C is equal to that of model D. Models E and F are generated based on Eq. (66), and their amounts of units are respectively model E with $(k, m) = (4, 100)$ and model F with $(k, m) = (8, 50)$.

The result shows that the type of distribution is related to the value of intra-unit correlation. And the distribution of intra-unit correlation of these models indicates that when $\rho = 0.5$, its distribution is symmetrical and resembles the normal distribution; when $\rho > 0.5$, its distribution is positively skew; and when $\rho < 0.5$, negatively skew, just as Fig. 3 shows.

In one model, the estimated error is larger when ρ approaches 0.5, and becomes smaller gradually as ρ approaches 0.1 or 0.9.

The mean intra-unit correlation coefficients of model C is closer to the theoretical value than that of model A, and its standard error is smaller. Similarly, the mean intra-unit correlation coefficient of model D is closer to theoretical value than that of model B, and its standard error is also smaller. Therefore, the larger the sample size, the better the effect of estimation.

In comparisons of model A with B, model C with D and model E with F respectively, for which each pair has the same sample size, the estimation of model B is not as good as A, model D not as good as C and model F not as good as E except that $\rho = 0.1$ or 0.2. This is because the amount of two-level units is small while that of one-level units is large. In fact, because of the presence of intra-unit correlation, the amount of information is overlapped. As an example, the amount of information obtained by measuring k times repeatedly to the same individual is smaller than that obtained by measuring once to k individuals.

If we compare model E with C and model F with D respectively, their overall sample sizes, the amount of one-level units and two-level units are equal respectively. But models E and F have larger estimated error because the variance terms of models E and F are more complex.

Table 19. 500 simulated results of intra-unit correlation of 6 populations.

Theoretical values	Model A $k=4, m=10$	Model B $k=8, m=5$	Model C $k=4, m=100$	Model D $k=8, m=50$	Model E $k=4, m=100$	Model F $k=8, m=50$
0.1	0.1235 ± 0.1279	0.1057 ± 0.1146	0.0980 ± 0.0471	0.0965 ± 0.0411	0.1150 ± 0.0933	0.1121 ± 0.1178
0.2	0.2017 ± 0.1485	0.1786 ± 0.1467	0.1987 ± 0.0540	0.1981 ± 0.0522	0.2019 ± 0.1067	0.2075 ± 0.1450
0.3	0.2855 ± 0.1633	0.2666 ± 0.1722	0.3026 ± 0.0534	0.2988 ± 0.0572	0.2978 ± 0.1081	0.2963 ± 0.1635
0.4	0.3787 ± 0.1725	0.3428 ± 0.1884	0.3973 ± 0.0560	0.3936 ± 0.0616	0.3995 ± 0.1123	0.3934 ± 0.1599
0.5	0.4595 ± 0.1749	0.4288 ± 0.2054	0.4971 ± 0.0534	0.4942 ± 0.0609	0.4995 ± 0.1078	0.4981 ± 0.1563
0.6	0.5635 ± 0.1608	0.5207 ± 0.2042	0.5964 ± 0.0463	0.5931 ± 0.0570	0.5965 ± 0.1018	0.5920 ± 0.1507
0.7	0.6663 ± 0.1379	0.6232 ± 0.1928	0.6967 ± 0.0389	0.6947 ± 0.0481	0.6974 ± 0.0957	0.6891 ± 0.1435
0.8	0.7647 ± 0.1146	0.7278 ± 0.1702	0.7974 ± 0.0287	0.7935 ± 0.0375	0.7986 ± 0.0865	0.8042 ± 0.1211
0.9	0.8786 ± 0.0734	0.8492 ± 0.1216	0.8985 ± 0.0153	0.8965 ± 0.0213	0.8912 ± 0.0727	0.8891 ± 0.0957

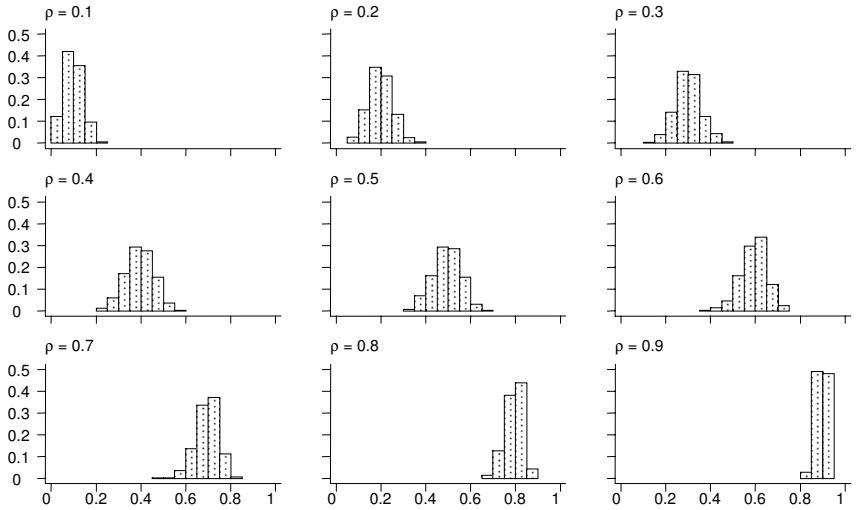


Fig. 3. Sampling distibution of the intra-unit correlations base on model C.

The statistical simulation shows that the estimator of ρ is little smaller than the theoretical value. But the larger the sample size is, the closer the estimator is to its theoretical value. And when given the same overall sample size, the larger the amount of level 2 units is (the smaller the amount of corresponding level 1 unit is), the closer the estimated value is to the theoretical value.

If ρ is close to 0.5, the sampling distribution of intra-unit correlation is approximately normal distribution. As ρ approaches 0, or approaches 1, the sampling error is becoming smaller and smaller. When ρ is close to 0, the distribution is positively skew. And when ρ is close to 1, the distribution is negatively skew.

Theoretically, as Goldstein (1998) pointed out, the estimators obtained by IGLS is biased, while that obtained by RIGLS is unbiased. But the simulated results show that estimators of ρ obtained by RIGLS are somewhat smaller than the theoretical values. And the smaller the sample size is, the further the estimated value is away from its theoretical value. When given the same overall sample size, the smaller the amount of level 2 unit is (the larger the amount of corresponding level 1 unit is, of course), the further the estimated value is biased.

The sampling error of intra-unit correlation is also related to the amount of units of every level. When given the same overall sample size, the larger

the amount of level 2 units (the smaller the amount of corresponding level 1 unit is), the smaller the estimated error. The sampling error is also related to how complex the variance of responding variable is: the larger the variance, the larger the sampling error.

This section focuses only on the situation when responding variable is numeric, that is to say that data should be distributed normally. Further investigation is needed, especially with regard to skew distributions, such as binominal and Poisson distributions.

8. Sample Size and the Cost-effect of Dependent Test

This section will take the repeated measurement (sampling) as an example to discuss sample size and power of hypothesis testing and cost-effect for dependent data. Because of the overlap of information, the dependent data tells us less than independent data given the same sample size, which leads to a low power. And the larger the relationship in groups is, the less information the data offers and the lower the power shows.

8.1. *Sample size and power of test*

Let Y_{ijg} represent the jth observation of the ith subject in the gth group ($i = 1, \ldots, m, j = 1, \ldots, k; g = 0, 1$). We also assume the individuals are independent to each other, and the intra-subject correlations are equal. If the type I error is α and the power is $1 - \beta$, the sample size of each group can be estimated by the equation below:

$$m = [1 + (k - 1)\rho]\frac{\sigma^2(Z_\alpha + Z_\beta)^2}{k\delta^2}. \tag{68}$$

Where δ is the difference of effects of the two groups ($g = 0$ and $g = 1$). It is oblivious that the number of observations m needed in this design is smaller, while the overall number of observations $n = mk$ is larger than those of the independent design. And when $\rho = 0$, it is equal to the sample size of independent design.

The table below is the result of a simulated experiment on the power of a group of repeated measurements. The intra-unit correlation is 0, 0.1, 0.2, ..., 0.9, respectively; the sample size m and the times of repeated measures k are (50, 4), (100, 2), (20, 4), (40, 2), respectively; And δ are 0, 0.2, 0.4, 0.6, 0.8, 1.0, respectively. All designs were balanced. Based on each grid, 1000 simulations were generated by using the MLn package.[17] For each set of simulated data, we fit a multilevel model. The power is then

estimated by the proportion of times of the rejections of null hypothesis in 1000 simulations.

The results corresponding to $\delta = 0$ is type I error, while that to $\rho = 0$ is the power of the independent data. From the Table 20, we can conclude that the power decreases as the intra-unit correlation within group increases. And unless δ is large enough, the extent of the decrease is large. For example, when the repeated times are the same, the power of the design with $m = 50$, $k = 4$ and $\rho = 0.9$ is only half of that with $n = 200$ and $\rho = 0$. When the repeated times are equal, the power has a tendency to increase as m increases; And when $n = mk$ are equal, the power of the design with twice measured is larger than that with 4 repeated times.

Table 20. Power of repeated measurement (times of the rejection to null hypothesis in 1000 simulations).

	$m = 50,\ k = 4$						$m = 100,\ k = 2$					
	δ						δ					
ρ	0	0.2	0.4	0.6	0.8	1.0	0	0.2	0.4	0.6	0.8	1.0
0	44	234	789	984	1000	1000	54	272	781	989	1000	1000
0.1	55	232	689	961	998	1000	43	268	757	986	999	1000
0.2	68	220	598	915	995	999	55	279	709	979	999	1000
0.3	57	169	557	854	975	999	52	226	692	963	999	1000
0.4	60	181	474	820	961	998	57	233	661	957	996	1000
0.5	56	157	447	756	936	990	43	205	645	942	999	999
0.6	66	143	393	718	931	983	62	119	600	916	989	1000
0.7	56	138	357	692	889	971	54	208	593	902	992	1000
0.8	62	151	331	647	860	970	62	190	564	883	981	999
0.9	58	117	342	595	829	952	58	168	519	875	988	1000
	$m = 20,\ k = 4$						$m = 40,\ k = 2$					
	δ						δ					
ρ	0	0.2	0.4	0.6	0.8	1.0	0	0.2	0.4	0.6	0.8	1.0
0	59	133	384	746	927	992	36	139	412	759	932	997
0.1	56	125	336	631	870	967	44	143	380	708	916	993
0.2	63	119	302	573	812	939	62	141	375	678	890	928
0.3	80	108	279	488	744	891	69	140	356	634	887	974
0.4	58	106	258	464	682	849	61	129	344	592	859	967
0.5	65	110	228	440	626	805	53	113	296	612	825	959
0.6	69	109	175	372	535	767	57	130	318	577	787	950
0.7	71	94	194	361	546	731	48	103	301	524	780	927
0.8	59	102	193	333	492	696	61	109	288	528	759	908
0.9	69	103	179	313	490	692	69	101	263	490	729	898

8.2. *Cost-effect analysis*

The estimations of design efficiency and sample size are important considerations during the experiment design. Researchers always have to balance among design efficiency, sample size and cost-benefit before making a decision. For example, a physiological experiment uses several rats' liver cells. Researchers may sample only once (single sample test, independent) or several times (repeated sample test, dependent) on each rat. The latter needs fewer rats than the former, which means the latter costs less. But data from the latter are dependent while those of the former are independent. So, the problem researchers confront is that the test should not only cost litter, but also achieve enough power of test.

When we discussed the estimation of sample size in the last section, we did not consider the cost. But the funds are limited in practice. So it is related to cost-benefit problems. On one hand, given restricted funds (the cost is constant), we should consider whether to select single sample or repeated sample to make the effect as large as possible (the variance is minimum). On the other hand, when the benefit is constant (the variance is restricted), we should consider whether to sample independently or repeatedly to make the cost the least.

8.2.1. *When the cost is constant, how to evaluate the benefits of independently sampling design and repeated sampling design?*

In a repeated sampling design, individual is independent with each other. We can assume the average elemental cost of each individual is C_1, the average direct cost of sampling once to each individual is C_2, the variation among individualities and repeated sample measures is σ_2, and the intra-subject correlation of samples from the same subject is ρ.

Let the overall cost be C, the individual numbers (or pairs) needed for repeated sampling design is m, and the times of repeated sample to each individual is k, then

$$C = m(C_1 + kC_2)\,. \tag{69}$$

It is not difficult to find:

$$\mathrm{var}(\bar{Y}) = \frac{\sigma^2}{mk}[1 + (k+1)\rho]\,. \tag{70}$$

When the restricted overall cost is C, to make $\mathrm{var}(\bar{Y})$ as little as possible (equivalent to making the power as large as possible), the optimal number

of individuals m and the times of repeated sample k to each subject are the solutions of conditional minimum of function (70) restricted by Eq. (69).

Let $f(m,k) = \sigma^2[1+(k-1)\rho]/(mk) + \lambda(mC_1 + kmC_2 - C)$, then

$$\begin{cases} \dfrac{\partial f(m,k)}{\partial k} = \dfrac{-\sigma^2(1-\rho)}{mk^2} + \lambda m C_2 = 0 \\ \dfrac{\partial f(m,k)}{\partial m} = \dfrac{-\sigma^2[1+(k-1)\rho]}{mk^2} + \lambda(C_1 + kC_2) = 0\,. \end{cases} \tag{71}$$

That is

$$\begin{cases} m = \sqrt{\dfrac{\sigma^2\rho}{\lambda C_1}} \\ k = \sqrt{\dfrac{(1-\rho)C_1}{\rho C_2}}\,. \end{cases} \tag{72}$$

Substituting m and k in (72) for their corresponding terms in (69), because C is a specific value, the optimal number of individuals is

$$m = \frac{C\sqrt{C_1\rho}[\sqrt{C_1\rho} - \sqrt{C_2(1-\rho)}]}{C_1[C_1\rho - C_2(1-\rho)]}\,. \tag{73}$$

So the minimum variance is

$$\mathrm{var}(\bar{Y}) = \sigma^2(\sqrt{\rho C_1} + \sqrt{(1-\rho)C_2})^2/C\,. \tag{74}$$

If the sample size of independent sampling design is N, the overall cost and sample error are $C = NC_1$ and $\mathrm{var}(\bar{Y}) = \sigma^2/N$ respectively.

And if the restricted overall cost is C, to make $\mathrm{var}(\bar{Y})$ as little as possible (make the power as large as possible), the optimal number of subjects m is the solution of conditional minimum of function $\mathrm{var}(\bar{Y}) = \sigma^2/N$ restricted by equation $C = NC_1$.

Let $g(N) = \sigma^2/N + \lambda NC_1$, then

$$N = C/C_1\,. \tag{75}$$

And the minimum variance of independent sample is

$$\mathrm{var}(\bar{Y}) = \sigma^2 C_1/C\,. \tag{76}$$

So given restricted overall funds C, whether to sample independently or repeatedly depends on the value of sample error, which means to work out when Eqs. (76) and (74) will have minimum values, when

$$\rho < \left(\frac{C_1 - C_2}{C_1 + C_2}\right)^2\,. \tag{77}$$

The repeated sample design can result in a minimum sample error (the power is the largest and the effect is better). Otherwise, it would be better to choose independent sampling design.

8.2.2. *When the benefit is constant, how to compare the cost of independent sample with that of repeated sample?*

We should follow the method in the last section to make the overall cost C as little as possible when $\mathrm{var}(\bar{Y}) = V$ is constant.

The optimal individual number of repeated sampling design m and the optimal sample times of each subject k can be worked out by the equations below, respectively:

$$\begin{cases} m = \dfrac{\sigma^2\sqrt{C_1\rho}[\sqrt{C_1\rho} + \sqrt{C_2(1-\rho)}]}{C_1 V}, \\ k = \sqrt{\dfrac{C_1(1-\rho)}{C_2\rho}}. \end{cases} \tag{78}$$

The minimum overall cost of repeated sampling design is

$$C = m(C_1 + kC_2) = \sigma^2[\sqrt{C_1\rho} + \sqrt{C_2(1-\rho)}]^2/V, \tag{79}$$

The optimal individual number of independent sampling design is

$$N = \sigma^2/V. \tag{80}$$

And the minimum overall cost of independent sample is

$$C = NC_1 = \sigma^2 C_1/V. \tag{81}$$

So under the condition of restricted sample error (the same benefit), should we select the independent sampling design or the repeated sampling design? This depends on the overall cost of the sample. We should compare when Eqs. (79) and (82) will have their minimum values. And only if the intra-subject correlation ρ meets the need of Eq. (77) can repeated sampling design make the sample cost as little as possible. Otherwise we'd better use an independent sampling design.

8.3. *Example* 20. *The cost benefit problems of rat's test data*

Physiology Laboratory, Nantong Medical College, had finished a test that needed four rats. Four sets of single spleen T cells turbid liquid were prepared for each rat by normal methods. Then researcher mixed ConA with

each liquid and measured OD. From pre-experiments or experiences, they estimated that

$$\sigma^2 = 0.00054844 \quad \text{and} \quad \rho = 0.52895\,.$$

Then if we restricted sampling overall cost C or sample error $\mathrm{var}(\bar{Y})$, should we select an independently sampling design or a repeated sampling design? This is related to the average elemental cost of each rat C_1, the average direct cost of repeated sampling once to each rat C_2 and the intra-subject coefficient of repeated measurement. We assume that each rat costs $C_1 = 20$ yuan, each portion (1 ml) of medium and 0.1 ml calf serum costs $C_2 = 0.12$ yuan. Because

$$\rho = 0.52895 < \left(\frac{C_1 - C_2}{C_1 + C_2}\right)^2 = \left(\frac{20 - 0.12}{20 + 0.12}\right)^2 = 0.97629\,,$$

this case meets the need of Eq. (77). Thus, it is wise to do repeated sampling instead of an independent sampling. From Eq. (72), we known that the repeated sampling times of each rat k is 13.

If the restricted overall cost $C = 110$ yuan, the repeated sampling design needs 5 rats and the minimum sampling error is $\mathrm{var}(\bar{Y}) = 0.000062$ from Eqs. (73) and (74).

If the restricted sampling error is $\mathrm{var}(\bar{Y}) = 0.000052$, the repeated sampling design needs $m = 6$ rats and the minimum sampling cost is $C = 130$ yuan from Eqs. (78) and (82).

In this section we focus on the power of the repeated sample design in one group, the estimation of sample size and some problems about cost-effect. The principles of analysis can also be applied to repeated measurement data of grouped design and longitudinal data, etc.

When estimating the power and sample size of the repeated sampling, we should take full advantages of the prior information to specify the values of variation among individuals and repeated samples, the value of intra-subject correlation coefficient and the values of acceptable error because of the affection these values have on the estimation of sample size. If there is not enough prior information, it is better to obtain it through pilot studies. And the importance of types I and II errors should be determined according to damages caused by the respectively wrong decisions.

There are two other design methods similar to repeated sampling design. One of them is the multiple repeated measures, which can improve the precision of measurements, and reflect whether the measured results have stability, namely reliablity. And the degree of reliability can be represented

by constructed validity. The intra-subject correlation among repeated measures of these data is always low and always has nothing to do with the covariates. The another one is the regular or irregular follow-up in a longitudinal study, such as the follow up studies of kid's growth and development and the metabolism of some kind of drug, etc., in which we are interested in the occurring, developing, or law of variation of an event. The intra-subject correlation of these data is always related with the interval of the follow up. However, repeated sampling is sampling from the same subject. These samples always have a low intra-subject correlation and are related to some covariates. Though in several literatures they are all refered to as repeated measurement and have similar methods of processing and analyzing, they have their own particular emphases. So the structures of covariances matrix of response variables are different. But to applied researchers, more emphases should be laid on the distinctions of different designs.

References

1. Cochran, W. G. (1977). Sampling Techniques. John Wiley and Son. Inc. New York.
2. Willams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teralogenicity. *Biometrics* **31**: 941–952.
3. Diggle, P. J., Liang, K. Y. and Zeger, S. T. (1995). Analysis of longitudinal data. Clarendon Press. Oxford.
4. Hooton, T. M., Scholes, D., Hughes, J. P. *et al.* (1996). A prospective study of risk factors for symptomatic urinary tract infection in young women. *NEJM* **335**: 468–474.
5. Gabriel, S. E., Woods, J. E. and O'Fallon, W. M. *et al.* (1997). Complications leads to surgery after breast implantation. *NEJM* **336**: 677–682.
6. McAlindon, T. E., Felson, D. T. and Zhang, Y. *et al.* (1996). Relation of dietary intake and serum levels of vitamin D to progression of osteoarthritis of the knee among participants in the Framingham study. *Ann. Int. Med.* **125**: 353–359.
7. Chen, F., Ren, S.-Q. and Lu, S. Z. (1998). On intra-correlations of dependent data. *Modern Prevent. Med.* **25**: 269–271.
8. Ren, S.-Q. and Chenfeng (1998). Expression for covariance structure for dependent data. *Chinese Health Statist.* **15**(4): 4–8.
9. Prescott, R. and Brown, H. (1995). *Mixed Models Analysis of Clinical Trials Using SAS*: *Proc Mixed and Beyond*. Course Presenters.
10. Liang, K. Y. and Zeger, S. T. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1): 13.
11. Chen, Q.-G. (1995). GEE analysis for repeated measurement in longitudinal study. *Chinese Health Statist.* **12**(1): 22.

12. Xiong, L.-P., Cao, X.-T. and Xu, Y.-Y. *et al.* (1999). Log linear model for longitudinal data. *Chinese Health Statist.* **16**(2): 68.
13. Chen, F., Ren, S.-Q. and Lu, S.-Z. (1999). Intra-unit correlation of dependent data and GEE. *Acta Academia Med. Nantong* **19**(6): 359–362.
14. Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least square. *Biometrika* **73**: 43–56.
15. Goldstein, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika* **76**: 622–623.
16. Goldstein, H. (1995) *Multilevel Statistics Models*, 2nd edn., Chapman, London.
17. Rasbash, J. and Woodhouse, G. (1995). *MLn Command Reference*, Institute of Education. London.
18. Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78**: 45–51.
19. Chen, F., Lu, S.-Z. and Yang, M. (1997). Bootstrap estimation and its apolication. *Chinese Health Statist.* **14**(5): 5–8.
20. Ren, S.-Q. (1999). The statistics methods for non-independent data. PhD dissertation. West China University of Medical Sciences.

About the Author

Feng Chen obtained his BS in Mathematics (1983) from Sun Yat-Sen University; MS in Biostatistics (1989) from Shanghai Second Medical University, and PhD in Medical Statistics (1994) from West Chinese Medical University. Dr. Chen was promoted to full professor in Nantong Medical University in 2000. He is the chairman of the project "Study on statistics methods for non-independent data" which is supported by the National Funds of Natural Sciences of China. His study visit in London University in 1996 was supported by the Royal Society. He is a member of review committee for medicines by the State Drug Administration (SDA) of China, Associate Editor of *Chinese Health Statistics*, and Vice Chairman of Statistical Theory and Methods Subcommittee of the Chinese Health Statistics Association. His main research interests are biostatistical models and statistical methods for dependent data.

This page intentionally left blank

CHAPTER 4

STATISTICS USED IN QUALITY CONTROL, QUALITY ASSURANCE, AND QUALITY IMPROVEMENT IN RADIOLOGICAL STUDIES

YING LU and SHOUJUN ZHAO

Department of Radiology, University of California, San Francisco, 3333 California Street, Suite 375, San Francisco, CA 94118, USA
Tel: 415-502-4596; ying.lu@radiology.ucsf.edu

1. Introduction

Quality control, quality assurance, and quality improvement in medical studies are active and large topics. From 1995–2000, there were more than 40,000 articles in MEDLINE database that had key words of at least one of these three terms. Quality has many connotations. The term "total quality management" (TQM) is given to an approach that related to the daily functioning of medical practices or medical research processes. All participated personnel and operational aspects are involved. Quality control is a very limited function that "controls" the product, primarily by testing, while quality assurance regulates the systems and methods for "assuring" the quality of the product.[1]

Every aspect of medical practice and research requires quality control and quality assurance. Although the statistical principles presented here can apply to other fields such as laboratory medicine, etc. this chapter is limited to quality control and quality assurance specifically in radiology. There are several reasons for this focus. First, this is the field in which the authors have the most experience. Second, radiology evaluation relies on radiological equipment, whether X-ray, ultrasound, CT, or MRI machines. As with all machinery, products of different manufacturers vary in quality. Over time, machine may draft and age can affect performance. Furthermore, precision errors are always to be expected in any radiological equipments or technique; even when the same patient is scanned under identical conditions the results will be different. Last but not least, many

radiological assessments are based on the experience of reader and are relatively subjective. It is common to have different readers to give different interpretations of the same image. Therefore, many factors will affect the results of radiological assessments. The statistical principles discussed here can resolve the conflicts among results from different devices and improve interpretation of the results.

Radiology has been used to help decisions in disease diagnosis and management of patients. Its use as tools for population screening and for drug development is increasing. The newly developed response evaluation criteria in solid tumors (RECIST) uses changes in unidimensional CT measurement of tumor lesions to define the treatment response rates.[2] Osteoporosis is defined by bone mineral density (BMD) measured by dual X-ray absorptiometry (DXA) scans[3] and osteoporosis prevention drugs are assessed according to their effect on BMD.[4] In fact, medical imaging has been used as surrogate endpoint or biomarkers in many therapeutic and diagnostic clinical trials, and radiologists are increasingly involved in these clinical trials.

Good Clinical Practice (GCP) is an international quality standard for the design, conduct, recording, and reporting of clinical trials with human subjects. GCP guidelines not only provide a framework for protecting the rights of participating patients or volunteers, they also set standards to safeguard the integrity of data that are used to evaluate treatment efficacy and submitted to regulatory agencies.[5] In radiology, GCP includes training documents and standard operating procedures, imaging device quality control, image acquisition protocols, software validation, record keeping, and reporting, etc.[6] Obviously, this is not only a statistical process. Successful quality control and quality assurance require good leadership from department chairs or principal investigators and, importantly, a team of multi-disciplinary experts. The expert team should always include a statistician. Statisticians are important in planning quality control, including determining appropriate sampling to avoid bias in selecting test samples, calculating the sample sizes, analyzing results to identify deficiencies, planning the processing control charts for monitoring machine performance, reassessing the results of quality improvement, and in reporting data and study results.

There are many aspects of quality control and quality assurance that are not directly related to statistics.[7–9] This chapter presents some statistical tools used in radiological or osteoporosis research based on the experience of the authors. It is beyond the scope of this book to

present a complete picture of quality control and quality assurance for all radiological studies.

This chapter is organized into 5 sections. In the next section, we introduce definitions of different measurement errors for continuous radiological results and different ways to evaluate these errors. In Sec. 3, we present applications of process control-charts to monitoring measurement errors over time. In Sec. 4, we review the statistics of measurement agreement. In Sec. 5, we discuss the calibration problem.

2. Measurement Errors

Radiological techniques are used to measure physical or mechanical properties that relate to disease status or progression. We use statistical techniques or procedures to transform our observations of a variable of interest into a particular category or number. This is the measurement process. For a categorical variable, we try to assign a subject into a particular, unambiguous category, as in the assessment of treatment response of solid tumors[2] or evaluation of spine fracture severity.[10] In other cases, we derive a numerical value that reflects the underlying physical quantity, such as tumor volume, bone mineral content or density, etc.

Measurement errors describe the limits of a quantitative or qualitative assessment of a disease using a particular technique or procedure. Measurement errors have many sources. This section focuses on 2 types of measurement errors — precision and accuracy — and their applications to the diagnosis of osteoporosis and monitoring changes in bone status. The implications of precision on monitoring changes are emphasized, including the concepts of standardized precision, longitudinal sensitivity, and their applications to patient measurements and quality assurance, i.e. the monitoring of machine performance.

2.1. *Measurement errors in radiological instruments*

Many sources of errors can affect the measurement and cause varying results, even when they are from the same region of interest in the same subject. Some of these variations can be controlled to minimize their impact. Some of the error sources are — in part — uncontrollable. Controllable variations are called fixed factors. Our interests, however, are usually on the uncontrollable random variations.

Errors of measurement are the differences between observed values recorded under identical conditions and a fixed true value. In osteoporosis

studies, we always assume there are true quantities for densitometry parameters for each measured subject, even though we don't always know their values. Measurement errors should be random in nature and can be attributed to two different sources: accuracy errors and precision errors.[11]

2.1.1. *Accuracy errors*

Accuracy errors here are used as equivalent to the term bias. They reflect the degree to which the measured results deviate from the true values. To evaluate accuracy errors, we need to know the true values of the measured parameters. It is not always possible, however, to measure the accuracy errors because sometimes the true values of the measured parameters cannot be verified. For example, quantitative ultrasound (QUS) bone measurements are affected by a number of quantitative and qualitative factors, and there is no single correlate for any QUS measurement. Therefore, we cannot define a single accuracy error for QUS.[12]

For clinical applications only the part of the accuracy error that varies from patient to patient in an unknown fashion is relevant. The other part, i.e. the one that is constant, can be averaged across subjects e.g. the average underestimation of bone density due to the average fat content of bone marrow in Quantitative Computed Tomography (QCT), can be ignored. There are two reasons: First, for diagnostic uses, the reference data will be affected by the same error so the difference between healthy and diseased subjects is constant. Second, the error is present at both baseline and follow-up measurements, and does not contribute to measured changes. Therefore, when discussing the impact of accuracy errors only that part of the error that changes from patient to patient in an unknown and uncontrollable fashion is of interest.[13] For this reason, small accuracy errors are of little clinical significance provided they remain constant.[14] In general they are more relevant to diagnosis and risk assessment than to monitoring.

2.1.2. *Precision errors*

They reflect the reproducibility of the technique. They measure the ability of a method to reproducibly measure a parameter for the purpose of reliably monitoring changes in bone status over time. Precision errors can be further separated into *short-term* and *long-term* precision errors. Short-term precision errors characterize the reproducibility of a technique and are useful for describing the limitations of measuring changes in skeletal status. If they are large they may affect the diagnostic sensitivity of a

technique. Long-term precision errors are used to evaluate instrument stability. Because long-term precision errors include additional sources of random variation attributable to small drifts in instrumental calibration, variations in patient characteristics, and other technical changes related to time, they provide a better measure of a technique's ability to monitor parameter changes than the short-term precision errors do. For patient measurements, estimates of long-term precision usually also include true longitudinal variability of skeletal status. For both of these reasons long-term precision errors normally are larger than short-term errors. While precision errors are easy to define, there are many ways to describe them depending on the purpose at hand, and there is no universal consensus on which definition is most appropriate.

Mathematically, let θ be the theoretical true value in which we are interested, and let X be the observed value. The difference of $\xi = X - \theta$ is the measurement error. Furthermore, if X follows a normal distribution

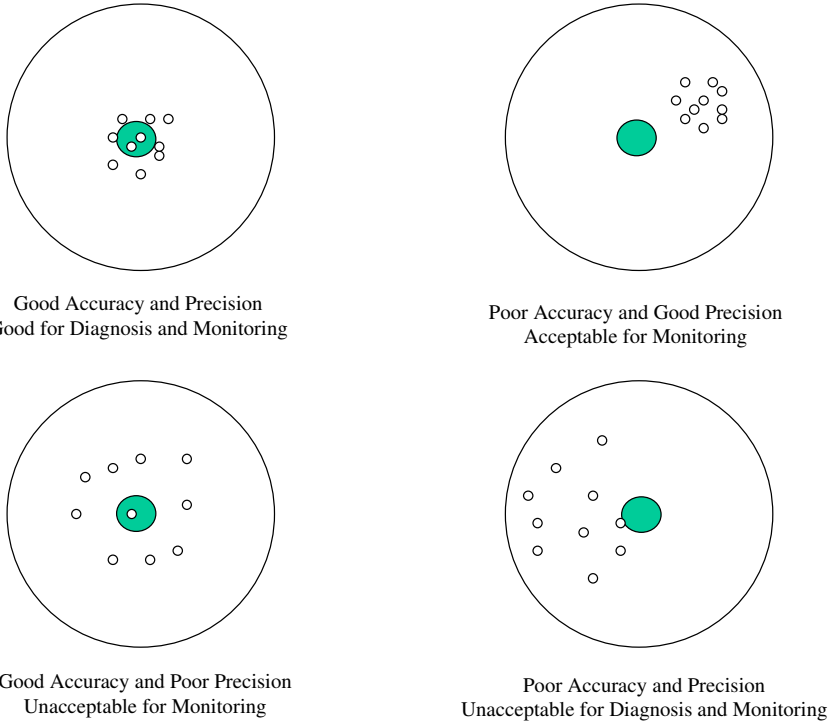


Fig. 1. Precision and accuracy.

$N(\mu, \sigma^2)$, the accuracy error is $\mu - \theta$ and precision error is σ. Here, θ is considered a gold standard.

Figure 1 illustrates the differences between precision and accuracy errors. If an archer consistently hits the target board close to the bull's-eye, but with the arrows spread out around it, it is good accuracy but poor precision. If the archer consistently hits the board far off the bull's-eye, but with all of the arrows in approximately the same location, it is poor accuracy but good precision.

2.2. *Absolute precision errors*

Although there are many different ways to describe precision errors, they can be classified as absolute or relative. For the following descriptions of precision errors, we introduce some notations. Let $X_{i,j}$ be the quantitative results (such as BMD) of the jth measurement for the ith individual, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. Because individual subjects have different underlying true values due to biological variation, it is necessary to measure individual subjects repeatedly to evaluate precision errors. We use n_i to denote the total number of measurements for the ith individual. The standard deviation (SD) of bone densitometry parameters from an individual subject i as a measure of short-term reproducibility is defined as the average distance of individual $X_{i,j}$ to the mean value for that subject, $\bar{X}_i$. Mathematically, it is the sample standard deviation:

$$\mathrm{SD}_i = \sqrt{\sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2 / (n_i - 1)}\,. \tag{1}$$

Individual precision may vary. To estimate the reproducibility of a parameter in clinical use, we need to measure a representative set of individuals and combine their individual precision errors using the root-mean-square average of individual SD values (RMS SD) or in other words, within the mean squared errors in Analysis of Variance terms. Mathematically,

$$\text{RMS SD} = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2}{\sum_{i=1}^{m} (n_i - 1)}} = \sqrt{\frac{\sum_{i=1}^{m} (n_i - 1)\mathrm{SD}_j^2}{\sum_{i=1}^{m} (n_i - 1)}}\,, \tag{2}$$

where m is the number of subjects measured for precision evaluation. When each subject has the same number of measurements, the RMS $\mathrm{SD} = \sqrt{\sum_{i=1}^{m} \mathrm{SD}_i^2 / m}$.

With long-term precision, the underlying parameter can change for individual subjects over time. Therefore, instead of measuring the distance from the observed individual values to the mean of the individual subject, we use the distances from the observed individual values to the expected value of the parameter at the time of measurement. In many situations we assume that the change of the parameter over time is linear for mathematical convenience. Thus, we can fit a regression line for observed individual measurements over time, i.e. $\hat{X}_{i,j} = \hat{a}_i + \hat{b}_i t_{i,j}$ with $t_{i,j}$ as the time of the jth measurement for the ith subject. The variation around the regression line is the standard error of the estimate (SEE):

$$\text{SEE}_i = \sqrt{\frac{\sum_{j=1}^{n_i}(X_{i,j} - \hat{X}_{i,j})^2}{n_i - 2}}\,. \tag{3}$$

In this case, SEE rather than SD should be taken as the estimate of the long-term precision error for an individual subject. For precision errors of a group of subjects, we use the root-mean-square SEE (RMS SEE) to evaluate the long-term precision error for clinical use.

$$\text{RMS SEE}_i = \sqrt{\frac{\sum_{i=1}^{m}(n_i - 2)\text{SEE}_i}{\sum_{i=1}^{m}(n_i - 2)}}\,. \tag{4}$$

The confidence intervals of RMS SD and RMS SEE can be derived using transformation of a Chi-squared distribution. The generic formula of $(1-\alpha) \bullet 100\%$ confidence interval is

$$\left(\sqrt{\frac{df}{\chi^2_{1-\frac{\alpha}{2},df}}} \cdot \text{Absolute Precision},\ \sqrt{\frac{df}{\chi^2_{\frac{\alpha}{2},df}}} \cdot \text{Absolute Precision}\right). \tag{5}$$

Thus, for short-term precision, $df = \sum_{i=1}^{m}(n_i - 1)$ and the absolute precision error is RMS SD. For long-term precision, $df = \sum_{i=1}^{m}(n_i - 2)$ and the absolute precision error is RMS SEE. The values of $\chi^2_{1-\frac{\alpha}{2},df}$ and $\chi^2_{\frac{\alpha}{2},df}$ can be obtained from most software and tables from statistics text books.

The absolute precision error depends on the unit of measurement. While it gives important information on measurement errors, it is inadequate for comparing precision errors across several techniques or measurements. For diagnosis or for monitoring longitudinal changes, we are usually more interested in the relative precision of a technique than in the absolute minimum measurement errors.

2.3. *Relative short-term precision errors*

2.3.1. *Short-term coefficient of variation*

The most commonly used measure of relative precision error is the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean measurement. It is usually given on a percentage basis. CV is unit free and therefore can be used with different techniques and instruments.

CV has a long history as a measure of reproducibility. It was first proposed by Karl Pearson in 1895 to measure the variability of a distribution. The distribution of CV is complicated. The simplest case is one individual with repeated measurements. Assuming that $X_{i,j}$ obtained from the ith individual are independent identical samples from a normal distribution $N(\mu_i, \sigma_i^2)$, the density functions for CV_i is[15]:

$$f_{\mathrm{CV}_i}(x; n_i, \lambda_i) = \begin{cases} \dfrac{e^{-n_i\lambda_i^2/2}}{\sqrt{\pi}\,\Gamma(\frac{n_i-1}{2})} \displaystyle\sum_{k=0}^{\infty} \frac{(\sqrt{2n_i}\lambda_i)^k}{k!} \Gamma\left(\frac{n_i+k}{2}\right) \frac{x^{n_i-2}}{(1+x^2)^{\frac{n_i+k}{2}}}, & x \geq 0\,, \\ \dfrac{e^{-n_i\lambda_i^2/2}}{\sqrt{\pi}\,\Gamma(\frac{n_i-1}{2})} \displaystyle\sum_{k=0}^{\infty} \frac{(-\sqrt{2n_i}\lambda_i)^k}{k!} \Gamma\left(\frac{n_i+k}{2}\right) \frac{|x|^{n_i-2}}{(1+x^2)^{\frac{n_i+k}{2}}}, & x < 0\,, \end{cases} \tag{6}$$

with $\lambda_i = \sigma_i/\mu_i$. Asymptotically, the variance of CV_i is $\lambda_i\sqrt{\frac{1}{n}(\frac{1}{2}+\lambda_i^2)}$.[16]

This individual CV is only meaningful if the subject has multiple measurements. When all individuals in a study have only one measurement, a population CV can be defined similarly to the ratio of population standard deviation and population mean. Such a CV is no longer related solely to measurement errors but to a combination of measurement errors and population variations. Feltz and Miller[17] gave an asymptotic χ^2-test (DAD test) to compare the CV from k-populations. Fung and Tsang[18] compared the DAD test with the likelihood ratio test (LRT), and the squared ranks test (SRT) in a simulation study. They concluded that the DAD test is a very good test for CVs from k-populations of normal distributions, although it is not robust, for a symmetric distribution with heavy tails. The LRT does not control type I errors correctly, although it is very powerful. The SRT is slightly liberal, but rather robust. In radiological studies, the population CV is rarely of interest, and it will not be discussed in detail here.

An alternative CV for non-normal distributions is the non-parametric CV, defined as the ratio of inter-quartile range over the median of the population.[19] The confidence interval and hypothesis testing for the

non-parametric CV can be derived using bootstrap or jackknife resampling techniques.[20,21]

In radiology, we are more interested measurement errors in a random effects model. Here, we assume that

$$X_{i,j} = \theta_i + e_{i,j}\,, \tag{7}$$

where θ_i is the unobserved true (expected) value for the ith subject that follows a $N(\mu, \tau^2)$, and $e_{i,j}$ are independent measurement errors that follow $N(0, \sigma^2)$. As in Sec. 2.2, the RMS SD in (2) is the best estimate of σ. Thus, for short-term precision, CV is defined as

$$\text{CV} = 100 \times \frac{\text{RMS SD}}{\bar{X}}\%\,, \tag{8}$$

Where $\bar{X}$ is the mean of $X_{i,j}$. This is also called within-batch CV in laboratory medicine.[22] The distribution of this short-term precision is much more complicated because means of subjects θ_i's also follow a normal distribution. Quan and Shih[23] derived the asymptotic sample variances for short-term CVs. The derivation requires two assumptions: (1) the number of repeated measurements of a patient n_i will not be more than a positive number C; (2) the proportion of subjects with $n_i = l$ converges to a constant $0 \le p_l \le 1$, as $m \to \infty$. Under these two concditions, the asymptotic standard deviation of moment estimator of short term CV defined in formula (8) is

$$\sqrt{\frac{\sigma^2}{\mu^4}\frac{(\sum_{i=1}^m n_i)\sigma^2 + (\sum_{i=1}^m n_i^2)\tau^2}{(\sum_{i=1}^m n_i)^2} + \frac{\sigma^2}{2\mu^2\sum_{i=1}^m (n_i - 1)}}\,, \tag{9}$$

for $m \to \infty$. The sample variation when $X_{i,j}$'s follow log-normal distribution can also be found in Quan and Shih.[23]

It is often useful to compare the CVs of different techniques, or of the same techniques at different research centers. When comparing the same technique at different centers, the measured subjects in different centers are independent so it is appropriate to use the DAD test similar to Feltz and Miller.[17] When comparing the CVs of different techniques, however, it is preferable to apply the techniques to the same set of subjects to control for confounding factors. This resulted correlated estimated CV and testing can be complicated. A two-step bootstrap algorithm can be used to compare two or more CVs:

Step 1. Draw m random samples with replacement from the study subjects.

Step 2. For each selected subject (possibly selected multiple times but treating each measurement as an independent sample) in Step 1, draw n_i random samples with replacement from his/her corresponding measurements.

Step 3. Calculate the difference of the two CV's based on data in Step 2.

Step 4. Repeat Steps 1 to 3 many times (1,000–2,000 times).

Step 5. Calculate the 95% bootstrap confidence intervals of the differences. If the 95% bootstrap confidence interval excludes 0, the null hypothesis that the two CVs are equal is rejected.

2.3.2. *Alternative forms of short-term coefficient of variation*

Intuitively, the larger the CV, the larger the precision errors and the poorer the technique's ability to monitor changes. However, this is not always true. To use CV, the value 0 of a measurement should have some physical meaning. For example, 0 bone mineral content and density have clear physical meanings. On the other hand, 0 value in speed of sound (SOS) in quantitative ultrasound has no physical meaning — the lower limit for speed of sound in water is around 1500 m/s. When the value 0 has no physical meaning, the origin of the parameters can be moved up or down so that CV has no physical meaning. Secondly, using CV to characterize the precision error of a technique implies that the precision error is proportional to the quantity of measurements. This is not true for many bone densitometry measurements. Normally, we see that the lower the bone density, the higher the relative precision errors (actually, even the absolute precision error increases with decreasing BMD). Thus, CV is not always a robust parameter for evaluating precision, at least for bone densitometry in osteoporosis research. Third, the mean value of the measured quantity, in many cases, is not the primary interest. We are more interested in discriminating between patients and normal controls, monitoring changes in bone status, or evaluating treatment responses, and CV is inadequate for these purposes. A major limitation of CV is that it does not take into account the impact of the technique's responsiveness to changes caused by disease or disease progression. When a technique has a very low precision error (i.e. a very "good" precision) but an even lower responsiveness (e.g. differences between healthy and diseased subjects or changes as a result of disease progression or treatment) it will not have a good longitudinal sensitivity to detect changes caused by disease over short time periods. Therefore, several approaches to adjust for differences in responsiveness have been proposed.

Miller *et al.*[24] proposed a standardized coefficient of variation (SCV) as the ratio of absolute precision over the range (5th to 95th percentiles) of parameters. The range can be obtained from manufacturer's normative data or from the observed study subjects when the sample size is large enough and sampling procedures are appropriate. Mathematically,

$$\begin{aligned} \text{SCV} &= \frac{\text{Absolute Precision}}{\text{Range}} \bullet 100\% \\ &= \frac{\text{Absolute Precision}}{95\%\ tile - 5\%\ tile} \bullet 100\%\,. \end{aligned} \tag{10}$$

Alternatively, Blake *et al.*[25] proposed using the population standard deviations as the measure for the range of the measure. Thus, the precision error is measured by the ratio of standard deviation of measurement errors over measured population standard deviation (including both measurement errors and population variations), which we call it SCV2. SCV2 relates to the attenuation parameter in the measurement error models[26] that measures the bias caused by measurement errors in linear and non-linear regression analysis. Because the width of the 90th percentile range in SCV is about 3.3 times the population standard deviation, SCV is approximately a third of SCV2.

Machado *et al.*[27] proposed a similar standardized precision measurement by replacing the range in the above formula with the differences in mean values of parameters for diseased and normal subjects, which we call SCV3. It is important to note that all these standardized CVs are also unit free.

In osteoporosis research, the population range or standard deviations of BMD change across different age groups. To adjust for the age effects on precision errors, Langton[28] proposed a precision parameter, ZSD. A ZSD is the standard deviation of an individual's Z-scores, $z_{i,j}$'s, a transformation of the observed measurement $X_{i,j}$'s. This Z-score is different from Z-statistics in statistical literature. Here, Z-score is defined as $z_{i,j} = \frac{X_{i,j} - \mu(age_i)}{\sigma(age_i)}$ where $\mu(age_i)$ and $\sigma(age_i)$ are the BMD mean and standard deviation of the age group for the ith subject. Therefore, a Z-score is the number of population standard deviations by which a subject's value varies from the population age-matched mean. It is unit free. A RMS ZSD will be a measurement for a technique.

The standard deviation of $z_{i,j}$ is $\text{ZSD}_i = \frac{\text{SD}_i(X_{i,j})}{\sigma(age_i)} \times 100\%$. Thus, ZSD_i for the ith subject is actually an age matched SCV2. A RMS ZSD is a RMS average of individual SCV2'S.

The SCV proposed by Miller *et al.*[24] is an important step in recognizing the limitations of a traditional CV. SCV often provides different information than CV. For example, PA spine BMD measured by a DXA scanner such as the Hologic QDR-1000 has a higher short term CV (1%) than speed of sound (SOS) (0.3%) measured by quantitative ultrasound machines like the Hologic Sahara. However, defining the SCV as the ratio of RMS SD over the young adult population SD gives the opposite result: the SCV of PA spine BMD is 8% and SOS is 20%.[25] Rather than using the population standard deviation, ZSD uses the age specific population standard deviation. ZSD has advantages when the population variance varies for different age groups, and the purpose of the technique is to determine the differences of individual subjects from their corresponding age group means.

An important limitation of SCV and SCV2 is their dependence on the normative data. In most cases, normative data from different equipment manufacturers are not comparable. Different manufacturers have different normative data based on different selection criteria. The procedures for collecting data may not always follow appropriate statistical sampling procedures and thus may not represent the true population distribution of the parameters. Comparing two SCVs based on two different normative data sets can be like comparing apples to oranges. Many precision studies have small sample sizes and subjects are recruited from convenient samples. The study sample may not be compatible with normative populations. These logistic difficulties severely limit the scientific validity of SCVs.

Statistical properties and hypothesis testing procedures for all the SCVs are complicated and have not been fully studied. In all these cases, the bootstrap method can be applied to resolve the real application needs.

2.3.3. *Sample size for short-term precision studies*

When planning for a short-term precision study, there are always trade-offs between the number of study subjects and the number of measurements. In most cases, we plan to have the same number of measurements n for all the m study subjects. Sample size calculations can be based on the width of the confidence intervals or on the null hypothesis. In both cases, one should have some idea of the ratio between population standard deviation τ and population mean μ.

For a given n, the asymptotic $(1 - \alpha) \bullet 100\%$ confidence width for estimated CV λ is

$$2z_{1-\alpha/2}\frac{\lambda}{\sqrt{m}}\sqrt{\frac{\lambda^2}{n}+\frac{\tau^2}{\mu^2}+\frac{1}{2(n-1)}}\,. \tag{11}$$

This is obtained by rearranging formula (9). A similar argument for the sample size to test the hypothesis $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$ is given as

$$m = \frac{\left(z_{1-\alpha/2}\lambda_0\sqrt{\frac{\lambda_0^2}{n}+\frac{\tau_0^2}{\mu_0^2}+\frac{1}{2(n-1)}}+z_{1-\beta}\lambda_1\sqrt{\frac{\lambda_1^2}{n}+\frac{\tau_1^2}{\mu_1^2}+\frac{1}{2(n-1)}}\right)^2}{(\lambda_1-\lambda_0)^2}\,. \tag{12}$$

Here, α and β are the types I and II errors; λ_1 is the alternative CV; and τ_i and μ_i are population standard deviation and means under the null ($i = 0$) and alternative ($i = 1$) hypotheses.

Equation (12) shows that the sample size m decreases as number of measurements n increases. In practice, recruiting subjects is more difficult and costly than repeating measurements. However, many factors can influence precision errors and selecting a small number of patients can either over- or under-state the true precision of the technique in clinical use. For example, measuring only healthy young women to evaluate DXA scanner precision will give smaller precision errors and will overstate the precision of the scanner. Measuring only elderly osteoporotic women will give larger precision errors and will understate the precision of the scanner. Some balance of confounding factors for precision errors must be achieved to represent the clinical population to which the machine or technique will be applied.[11] Within the given cost constraints, one should try to reach as many subjects as possible.

2.4. *Relative long-term precision errors and sensitivity of monitoring changes*

Short-term precision is useful for evaluating the utility of a diagnostic technique. The smaller the precision error, the easier to separate diseased and normal subjects. This is particularly true for standardized precision errors. They cannot, however, describe the ability of a technique to monitor changes.

2.4.1. *Longitudinal CV*

Like the limitation of short-term absolute precision errors, RMS SEE depends on the measurement unit and is not appropriate to compare across

techniques. Correspondingly, we can define a longitudinal CV as

$$\mathrm{CV} = 100 \times \frac{\text{RMS SE}}{\bar{X}}\%^{E}\,. \tag{13}$$

If we assume that the changes of measurements for individual subjects over time follows a linear model, that is

$$X_{i,j} = a_i + b_i t_{i,j} + e_{i,j}\,, \tag{14}$$

with $t_{i,j}$ the measurement time for the jth measurement of the ith subject, the longitudinal CV is

$$\mathrm{CV} = \frac{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{i,j} - \hat{a}_i - \hat{b}_i t_{i,j})^2 / \sum_{i=1}^{m}(n_i - 2)}}{\sum_{i=1}^{m}\bar{X}_i/m}\,. \tag{15}$$

Here, $\hat{a}_i$ and $\hat{b}_i$ are the estimated intercept and slop and $\bar{X}_i$ is the average for the ith subject. Derivation of asymptotic standard deviations of Eq. (15) has not yet been reported in the literature.

Although, it is inexplicitly, the longitudinal CV depends on the length of time that the measurement performed. If the length of time and frequency of measurements are different for the same technique and same subjects, the CV may be different. This is because that $\bar{X}_i \cong a + b\bar{t}_i$, which is not the case for the absolute precision. Therefore, to compare the same technique on different machines, the absolute longitudinal precision in RMS SEE is more appropriate. When comparing different techniques, the measurement times should be identical. The best plan is to measure the same subjects at the same time. Otherwise, their longitudinal CVs will not be comparable.

2.4.2. *The least significant change*

For clinical decision making it is important to know the minimum magnitude of measured change that is not caused by measurement errors. The least significant change (LSC) is defined as 2.8 times the longitudinal absolute precision,[29] i.e.

$$\mathrm{LSC} = 2.8 \times \text{RMS SEE}\,. \tag{16}$$

More specifically, if we observe a change of a subject more than LSC, we will have 95% confidence that the change is beyond measurement errors.

The derivation of the LSC is based on the following argument. Let X_1 and X_2 be two successive measurements of a subject. If there is no change in the two measurements, the difference between them is the result

of longitudinal measurement errors. If we assume the longitudinal measurement variation is σ, as estimated by RMS SEE in Eq. (4), $\Pr(|X_1 - X_2| > z_{1-\alpha/2}\sqrt{2}\sigma) = \alpha$. The least significant change is also called the "biologically significant change" in laboratory medicine.[22]

The longitudinal precision error must be used to evaluate the LSC rather than the short-term precision error, which is normally smaller than the longitudinal precision error.

The significance level of 5% has no clinical meaning. Therefore, there is no need to insist on 95% confidence when evaluating the LSC. To treat patients early, before the disease progresses, lower confidence levels can be chosen. Another parameter trend assessment margin (TAM) was proposed as $1.8 \times$ RMS SEE, which was calculated as corresponding to an 80% confidence level.[30] The LSC and TAM can also be approximately calculated in percentages based on longitudinal CV's.

2.4.3. *Follow-up time interval*

Radiological variables are often used as monitoring tools for individual patients. To assess the sensitivity of a technique for monitoring patients, Gluer[30] introduced the concept of "monitoring time interval" (MTI). The MTI for assessment of disease progression or treatment response is an estimate of the time period after which a patient will have a 50% chance of showing changes that exceed the LSC. Thus,

$$\text{MTI} = \text{LSC}/\text{Median Changes Per-Annual}\,. \tag{17}$$

The changes here can be caused by age, disease progression or treatment efficacy depending on the purpose of the study. The change also should be consistent with the units of the LSC. That is, if LSC is expressed as absolute precision, the change should be expressed as absolute changes. If the LSC is expressed as a percentage, a percentage change should be used. It is important to note that the unit of MTI is a year.

Similarly to TAM, Gluer[30] also suggested the "trend assessment interval" (TAI) an estimate of the follow-up time after which a subject will have 50% chance of changes exceeding TAM.

The determination of appropriate monitoring time intervals always represents a tradeoff between frequent visits with patient discomfort and additional costs, and fewer visits with the risk of substantial disease progress in the interval. MTI requires the usual 95% confidence level, which means the corresponding monitoring time interval would be almost double the

TAM. This shows that MTI and TAI, as applications of longitudinal precision, when defined in this fashion, have a direct and very intuitive meaning closely related to recommended monitoring time intervals. However, one should note that there is no single MTI (TAI) for each technique. They will differ substantially depending on the expected response of the patients. For their purpose, this is not a disadvantage, since it directly reflects that the frequency of follow-up measurements will depend on the type of patient examined. In osteoporosis clinics, for example, fast bone losers should have MTIs shorter than average postmenopausal women.

2.5. *Examples of applications of precision errors*

In this subsection, we give some examples of calculating absolute and relative precision as described in the previous subsections.

2.5.1. *Example* 1

The short-term precision errors of two quantitative ultrasound scanners for osteoporosis from two different manufacturers were compared. Twenty

Table 1. SOS (m/sec) at calcaneus of 20 volunteers.

Subject ID	Manufacturer 1		Manufacturer 2	
	Measure 1	Measure 2	Measure 1	Measure 2
1	1499	1505	1579	1586
2	1487	1488	1594	1590
3	1471	1465	1543	1556
4	1468	1467	1536	1545
5	1501	1504	1587	1588
6	1516	1517	1618	1605
7	1490	1491	1580	1587
8	1569	1565	1670	1683
9	1534	1543	1641	1641
10	1464	1468	1547	1558
11	1509	1510	1591	1593
12	1567	1541	1621	1647
13	1514	1509	1605	1625
14	1539	1540	1619	1614
15	1540	1537	1632	1648
16	1532	1535	1616	1617
17	1544	1531	1629	1636
18	1578	1574	1637	1644
19	1484	1482	1574	1576
20	1518	1522	1606	1610

Table 2. Short-term precisions and related parameters.

Manufacturer	Statistics (and equation number)				
	RMS SD (2)	CV (8)	SD for CV (9)	SCV (10)	SCV2
1	5.30	0.35%	0.06%	5.33%	16.28%
2	7.59	0.47%	0.07%	8.29%	21.62%

healthy elderly volunteers participated in the study. Speed of sound (SOS) at the calcaneus was measured twice on the same day for each subject. The data is given in Table 1.

Therefore, we have $m = 20$ and $n_1 = \cdots = n_{20} = 2$. The results are summarized in the following Table 2.

SCV2 was defined in Sec. 2.3.2, immediately after Eq. (10). We did not calculate SCV3 and ZSD here because SCV3 requires information from individual disease status and ZSD requires manufacturer's normative data, and neither were available. It is worth-noting that classical CV for SOS is very low compared to BMD measured by DXA (CV range from 1% to 6%). However, this does not mean that SOS is more precise in clinical use. The clinically useful range of SOS does not begin with zero and, in fact, zero is not defined here. That is why SCV and SCV2 are more meaningful in this example. The reported SCV2 for BMD measured by DXA ranged from 8% to 11%,[25] far less than SOS on a quantitative ultrasound scanner.

2.5.2. *Example* 2

Five normal volunteers participated in a longitudinal quality evaluation study for two new quantitative ultrasound (QUS) devices from different manufacturers with in one year. Table 3 lists their SOS measurements.

Table 4 displays the longitudinal precision. Although not all subjects demonstrated linear changes over time — Subject 3 in particular had some non-linear changes in Machine 1 — we applied only linear trends to all individuals. Also, as pointed out in Example 1, CV is not an appropriate measurement for SOS in QUS. CV is included in Table 4 only for demonstration.

Thus, although Machine 2 has higher precision errors, it is more sensitive to changes in age and may be a better choice for longitudinal follow-up. Of course, the sample size in this study is too small to reliably determine the monitoring time intervals.

Table 3. Longitudinal QC data for 5 normal volunteers.

		SOS (m/sec)				SOS (m/sec)	
Subject	Date	Machine 1	Machine 2	Subject	Date	Machine 1	Machine 2
1	09/21/97	1554	1636	3	05/07/98	1588	1698
1	10/04/97	1563	1642	3	06/01/98	1586	1717
1	11/05/97	1546	1634	3	07/24/98	1587	1708
1	11/18/97	1554	1635	3	09/23/98	1588	1708
1	12/29/97	1560	1656	4	09/22/97	1598	1709
1	01/09/98	1551	1626	4	10/04/97	1595	1694
1	02/04/98	1556	1648	4	10/29/97	1601	1698
1	02/24/98	1548	1642	4	11/17/97	1585	1677
1	03/22/98	1552	1658	4	12/12/97	1590	1696
1	04/11/98	1562	1665	4	12/28/97	1608	1720
1	05/07/98	1544	1637	4	01/25/98	1593	1691
1	07/11/98	1548	1653	4	02/20/98	1595	1692
1	08/13/98	1567	1672	4	03/11/98	1586	1688
1	08/26/98	1563	1658	4	03/23/98	1593	1718
1	09/21/98	1554	1646	4	04/24/98	1594	1722
2	09/22/97	1560	1654	4	06/13/98	1602	1727
2	10/05/97	1565	1660	4	06/26/98	1600	1733
2	11/06/97	1563	1643	4	07/27/98	1598	1708
2	11/19/97	1562	1652	4	08/09/98	1591	1719
2	12/22/97	1558	1663	4	09/21/98	1594	1714
2	01/04/98	1572	1680	5	09/22/97	1591	1664
2	02/05/98	1567	1674	5	11/14/97	1586	1678
2	02/19/98	1566	1667	5	12/17/97	1587	1677
2	03/23/98	1568	1677	5	12/29/97	1605	1703
2	04/12/98	1572	1663	5	02/01/98	1587	1682
2	05/08/98	1569	1661	5	02/15/98	1586	1681
2	07/12/98	1573	1650	5	03/20/98	1588	1688
2	09/21/98	1576	1695	5	04/02/98	1594	1693
3	09/22/97	1579	1654	5	05/05/98	1594	1682
3	11/17/97	1575	1666	5	05/19/98	1593	1684
3	12/15/97	1576	1667	5	06/27/98	1594	1695
3	01/13/98	1571	1669	5	07/11/98	1596	1677
3	01/29/98	1573	1670	5	08/13/98	1594	1675
3	03/03/98	1579	1676	5	08/26/98	1596	1701
3	03/27/98	1580	1690	5	09/21/98	1594	1689

Table 4. Longitudinal precision for 2 QUS machines.

		Machine 1			Machine 2		
Subject	d.f	SEE	Mean	CV	SEE	Mean	CV
1	13	7.11	1555	0.46%	10.87	1647	0.66%
2	11	3.25	1567	0.21%	12.74	1665	0.77%
3	9	4.07	1580	0.26%	8.17	1684	0.49%
4	14	6.14	1595	0.39%	13.67	1707	0.80%
5	13	4.91	1592	0.31%	10.00	1685	0.59%
Total	60	5.43	1578	0.34%	11.49	1678	0.69%
LSC (m/sec)			15.21			32.18	
MTI (yr)			2.5			1.3	

3. Statistical Process Control Charts

In Sec. 2, we introduced the concept of measurement errors and the statistics to evaluate them. Precision errors are usually evaluated whenever new techniques or new devices are developed. Precision errors are also evaluated immediately after a device is installed in clinical sites to assure that the equipment is performing according to the manufacturer's specifications at baseline. Precision errors also are always assessed before the beginning of clinical trials or longitudinal studies.[7,31] Although the manufacturer's service personnel can set up the device so that precision errors are within appropriate limits at baseline, it is very important to monitor the equipment to assure that imprecision remains within acceptable limits. Despite the remarkable accuracy and reproducibility of radiological equipment, measurements can still vary because of changes in equipment, software upgrades, machine recalibration, X-ray source decay, hardware aging and/or failure, or operator errors.

In an ideal setting, a well maintained equipment produce values that are randomly spread around a reference value. A change point is defined as the point in time at which the measured values start to deviate from the reference value. To evaluate measurement stability and identify change points, radiologists develop phantoms that simulate human measurements but, unlike humans, do not change over time.[7,32,33] Variations in phantom measurements should reflect variations in human measurements. Phantoms are measured regularly to detect one or more of the following events: (1) The mean values before and after the change point are statistically significantly different; (2) The standard deviations of measurements before and after the

Table 5. AP spine BMD of a hologic phantom in a QC study (13 March 1989 to 15 May 1989).

i	Date	BMD (X_i)	μ_0	$\sigma(=\mu\times 0.5\%)$
41	03/13/89	1.039	1.033	0.00517
42	03/15/89	1.039	1.033	0.00517
43	03/21/89	1.029	1.033	0.00517
44	03/22/89	1.036	1.033	0.00517
45	03/23/89	1.030	1.033	0.00517
46	03/27/89	1.033	1.033	0.00517
47	03/28/89	1.036	1.033	0.00517
48	03/29/89	1.038	1.033	0.00517
49	03/30/89	1.036	1.033	0.00517
50	04/03/89	1.033	1.033	0.00517
51	04/04/89	1.036	1.033	0.00517
52	04/05/89	1.034	1.033	0.00517
53	04/06/89	1.029	1.033	0.00517
54	04/07/89	1.033	1.033	0.00517
55	04/10/89	1.037	1.033	0.00517
56	04/14/89	1.042	1.033	0.00517
57	04/17/89	1.044	1.033	0.00517
58	04/18/89	1.041	1.033	0.00517
59	04/19/89	1.040	1.033	0.00517
60	04/20/89	1.036	1.033	0.00517
61	04/28/89	1.039	1.033	0.00517
62	05/01/89	1.035	1.033	0.00517
63	05/02/89	1.047	1.033	0.00517
64	05/03/89	1.028	1.033	0.00517
65	05/04/89	1.035	1.033	0.00517
66	05/05/89	1.038	1.033	0.00517
67	05/09/89	1.031	1.033	0.00517
68	05/10/89	1.041	1.033	0.00517
69	05/12/89	1.043	1.033	0.00517
70	05/15/89	1.034	1.033	0.00517

change point are statistically significantly different; (3) The measurements after the change point show a gradual but significant departure from the reference value.

In Table 5, we introduce our third example, which is roughly two months of quality control data from a DXA scanner. In this example, a Hologic spine phantom was scanned about three times a week. The purpose of the study was to monitor the stability of the DXA scanner. If the scanner is functioning acceptably, the coefficient of variation should be less than 0.5%

in the total AP spine BMD values. (Information on this data set can be found in Lu *et al.*[34]). In Table 5, i is an indicator of the observation number; date is the date the scan was performed; BMD is the ith measurement; μ_0 is the reference value based on historical QC data; and σ is the standard deviation based on 0.5% CV. We will use this data to illustrate statistical process control charts.

Statistical process control (SPC) is a powerful collection of problem solving tools for achieving process stability and improving capacity through reduction of variability.[35] There are several statistical methods for identifying change points. One is to visually check the retrospective data to determine the change points and then to verify these changes by a t-test for means and an F-test for variances. An alternative is to use statistical process control charts.[34,36] In this section, we introduce these methods and provide examples of their application in monitoring BMD measured by DXA scanners in osteoporosis studies.

3.1. *Visual inspection*

Potential change points in the data can be determined after careful visual inspection. This can be done by plotting longitudinal phantom data over time and using visual judgment to identify the potential change points created by drifts or sudden jumps. Statistical tests, such as the t-test, can be used to confirm the significance of the changes. It is important to note that there can be multiple potential change points observed for a given period of time. Careful control for type one errors for repeated tests is recommended for the t-tests.

Only experienced medical physicists or radiologists should perform periodic visual inspections. The role of primary evaluator should always be taken by the same individual to avoid subjective variations. The selection of the change points is based on the scatter plot in the most recent data. Once a change point has been identified, its cause should be investigated to determine if the change is machine related.

Visual inspection is not recommended because its efficiency depends on the experience of the reviewer and may not be reproducible.

3.2. *Shewhart control chart*

A Shewhart chart is a graphic display of a quality that has been measured over time. The chart contains a central horizontal line that represents the mean reference value. Three horizontal lines above and three below the

central line indicate 1, 2, and 3 standard deviations from the reference value. By plotting the observed quality control measurements on the chart, we can determine if the machine is operating within acceptable limits.

The reference values can be derived from theoretical values for the phantom, or from the first 25 observations measured at baseline. The reference value changes whenever the Shewhart chart indicates an out of control signal and the machine is recalibrated. The new reference value will then be the mean of the first 25 observations after recalibration. The number of observations needed to calculate the reference value may vary; the number 25 was chosen based on practical experience to balance the stability of the reference value with the length of time needed to establish it.

The standard deviation varies among individual devices, and manufacturers should be selected accordingly. For example, in one osteoporosis study, we sometimes use the BMD of a Hologic phantom to monitor DXA scanner performance. We usually assume the coefficient of variation for Hologic machines to be 0.5% and Lunar to be 0.6%, based on reported data on long-term phantom precision.[37] Therefore, the standard deviation for the scanner was calculated as 0.005 and 0.006 times the reference value for Hologic and Lunar machines respectively.

The original Shewhart chart will signal that there is a problem if the observed measurement is more than 3 standard deviations from the reference value. Although intuitive and easy to apply, the chart is not very sensitive to small but significant changes.[35] Therefore, a set of sensitizing tests for assignable causes has been developed to improve the sensitivity of Shewhart charts. Eight of the tests are available in the statistical software package SAS.[38] The tests are listed in Table 6.

Table 6. Definition of tests for assignable causes for Shewhart charts.

Tests	Pattern Description
1	One point is more than 3 standard deviation from the central line.
2	Nine points in a row on one side of the central line.
3	Six points in a row steadily increasing or steadily decreasing.
4	Fourteen points in a row alternating up and down.
5	Two out of 3 points in a row more than 2 standard deviation from the central line.
6	Four out of 5 points in a row more than 1 standard deviation from the central line.
7	Fifteen points in a row all within 1 standard deviation from the central line on either or both sides of the line.
8	Eight points in a row all beyond 1 standard deviation from the central line on either or both sides of the line.

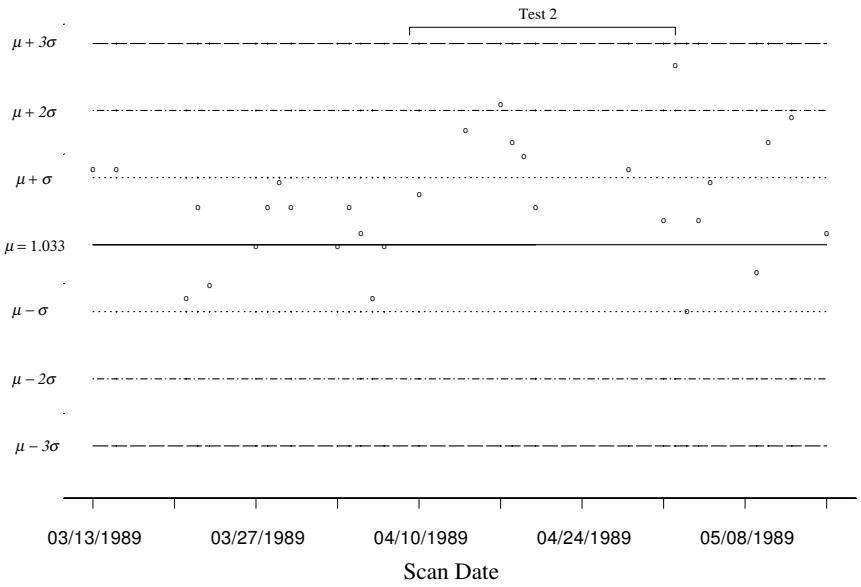


Fig. 2. Shewhart chart for QC data in Example 3.

The sensitizing rules can be used in toto or in part depending on the underlying processes of interest. For example, for quality control of DXA machines, we used four tests — 1, 2, 5 and 6.[34] Once a change point has been identified by any one of the tests, the manufacturer's repair service should be called to examine the causes and to recalibrate the machine. We then use the next 25 observations to generate new reference values and apply the tests to the subsequent data according to the new reference value.

Figure 2 shows the application of a Shewhart chart for Example 3. In this chart, the dots are the observed BMD. The six lines are the control limits 1, 2 and 3 standard deviations away from the central reference line. There is a problem with Test 2 from April 10, 1989.

The sensitizing rules increase the sensitivity of the Shewhart chart, but also increase the number of clinically insignificant alarms, which is not desirable. To overcome this problem, a threshold based on the magnitude of the mean shift can also be implemented. For example, we can select ten consecutive scans from after the possible change point identified on the Shewhart chart, and then calculate their mean values. If the mean differs by more than one standard deviation (which equals 0.5% times the reference value, in our example) from the reference value, the change point is confirmed as a true change point. Otherwise, the signal from the Shewhart

chart is ignored and the reference value is unchanged. This approach filters out small and clinically insignificant changes. However, the true difference must be more than one standard deviation for this approach to be effective, and this approach can delay the recognition of true change points.

3.3. *Moving average chart*

An alternative method is to determine the means and standard deviations of 25 consecutive measurements and then plot them over time. Control limits can be based on the assumption of a constant coefficient of variation during the process (0.5% times the reference mean) and a type one error rate comparable to the original Shewhart method (0.27%).[35] More specifically, we use X_i, for $i = 1, 2, \ldots, n$, the measured QC values of n longitudinal phantom scans from a machine. We define the moving average mean and standard deviation based on 25 scans as:

$$M_i = \sum_{j=i-24}^{i} X_j \Big/ 25\,, \quad i = 25, 26, \ldots, n \tag{18}$$

as the moving average of 25 scans to the date when the ith scan was collected, and

$$S_i = \sqrt{\sum_{j=i-24}^{i} (X_j - M_i)^2 \Big/ 24}\,, \quad i = 25, 26, \ldots, n \tag{19}$$

as the moving standard deviation of the 25 scans to the date when the ith scan was collected. Note that the first moving average can only be calculated after the first 25 scans have been collected.

Now if we assume that X_i's independently follow a normal distribution $N(\mu, \sigma^2)$, it can be shown that the M_i's follow a normal distribution $N(\mu, \sigma^2/25)$ and 24 S_i^2/σ^2's follow a chi-square distribution with 24 degrees of freedom denoted by χ^2_{24}. However, note that both M_i's and 24 S_i^2/σ^2's are not independent samples from the normal distribution and the chi-square distribution, respectively, for different i's.

Let μ_0 be the reference mean. If the machine is operating correctly, we should accept the null hypothesis, $H_0 : \mu = \mu_0$. If the machine is not operating correctly, we will accept the alternative hypothesis, $H_1 : \mu \neq \mu_0$. We select a type one error level of 0.0027 to be comparable to the original Shewhart method. We will reject the null hypothesis if $|M_i - \mu_0| > z_{1-\alpha/2}\frac{\sigma}{5} = 0.5991\sigma$. Thus, the control limits for the moving average are $\pm 59.91\%$ of the standard deviation from the reference mean.

We assumed that the CV for the machine is constant. Therefore, if it is functioning correctly, we can derive the standard deviation as equal to the reference mean times the CV. To check whether the precision of the machine is acceptable, we will test the null hypothesis, $H_0 : \sigma = \sigma_0$, versus the alternative that $H_1 : \sigma > \sigma_0$. With the same level of type one error rate as the mean difference, we will reject the null hypothesis if $24\ S_i^2/\sigma_0^2 > \chi^2_{24,1-\alpha}$, or equivalently, if $S_i > 1.41\sigma_0$. Thus, the control limit of the moving standard deviation is 1.41 times the standard deviation.

Note that there is only an upper limit for the moving standard deviation chart, as we are interested only in the increase in the standard deviation. In other words, we are looking for quality control but not quality improvement. Once the moving average moves out of the control limit, the value of the moving average at that point is used as the new reference value for scans performed after that date.

The number of scans used to calculate the moving average will affect performance of the method. Twenty-five scans were selected based on power analysis, so that the moving average chart has less than a 0.27% chance of a false alarm and a 98% chance of detecting an increase in the mean of one standard deviation. Also, the moving standard deviation chart has a 98% chance of picking up a 100% increase in the standard deviation.[34] Twenty-five scans is also a typical month's worth of quality control measurements.

3.4. *CUSUM chart*

CUSUM chart is short for Cumulative Sum Chart. In applications, we recommend a version of CUSUM known as Tabular CUSUM[35] because it can be presented with or without graphs. Mathematically, we define an upper one-sided tabular CUSUM $S_H(i)$ and a lower one-sided tabular CUSUM $S_L(i)$ for the ith QC measurement as the following:

$$S_H(i) = \max\left[0, \frac{X_i - \mu_0}{\sigma} - k + S_H(i-1)\right], \tag{20}$$

$$S_L(i) = \max\left[0, \frac{\mu_0 - X_i}{\sigma} - k + S_L(i-1)\right]. \tag{21}$$

Here, μ_0 is the reference mean, σ is the standard deviation, and k is a parameter to filter out insignificant variations and is usually set at 0.5. The initial values of $S_H(0)$ and $S_L(0)$ are 0. The chart sends an alarm message if $S_L(i)$ or $S_H(i)$ is greater than 5. In other words, when the standardized BMD value deviates more than k from zero, the cumulative upper bounded

sum increases by the amount of deviations above k. On the other hand, if the deviation is less than k, the cumulative sum will be reduced accordingly. When the cumulative sum is less than zero, we ignore the past data and set the cumulative sum as zero. However, a cumulative sum greater than 5 is a strong indication of a deviation from the reference mean in the data.

CUSUM also estimates when the change occurred and the magnitude of the change. We use the estimated magnitude of change to establish the new reference values.

Table 7 demonstrates the application of CUSUM chart to Example 3.

In this table, $S_H(i)$ and $S_L(i)$ are defined in Eqs. (20) and (21), and we selected $k = 0.5$ to detect a mean change of one standard deviation.[35] Along with the sequences $S_H(i)$ and $S_L(i)$, sequences $N_H(i)$ and $N_L(i)$ denote the number of scans since the last positive observation of $S_H(i)$ and $S_L(i)$, respectively. For example, from records one to four, the $S_H(i)$'s were positive, so that $N_H(i)$ goes from 41 to 44. However, $S_H(45)$ was zero. Therefore, the corresponding $N_H(45) = 0$. A similar rule applies for $N_L(i)$.

As explained, the initial reference value was obtained from the mean of the first 25 observations. However, once $S_H(i)$ or $S_L(i)$ exceeded 5, we concluded that the scanner was malfunctioning. For example, on April 20, 1989, $S_H(60) > 5$, suggesting that the BMD values were too high. We estimate that this event could have started on April 10, 1989, by noting the last date when $N_H(i) = 1$. Therefore, the investigation of assignable causes should focus around that time. The magnitude of change from the reference value can be estimated as $\sigma[k + S_H(i)/N_H(i)]$, which equals the average difference.[35]

Once we know the machine is malfunctioning, we will establish new reference values. If the manufacturer was involved in correcting the machine, the new mean should be established by the first 25 observations after the correction. However, if there is no intervention by the manufacturer or, as in our case, when performing retrospective data analysis, the new reference value can be estimated by $\mu_0 + \sigma[k + S_H(i)/N_H(i)]$, if the new BMD values are greater than the reference value, or by $\mu_0 - \sigma[k + S_H(i)/N_H(i)]$ when the new BMD values are smaller than the reference value. This results in a new μ_0 after the 60th scan of 1.040 mg/cm^2.

Graphical presentation of the CUSUM chart was shown in Fig. 3. In some senses, it is easier to review the Table 7 than the chart for identifying change points.

A separate CUSUM chart can be constructed for a one-sided change in variance. The one-sided variance chart was constructed according

Table 7. CUSUM table (from 13 March 1989 to 15 May 1989).

i	Date	X_i	μ_0	σ $(0.5\%_{\mu_0})$	$\frac{X_i-\mu_0}{\sigma}-0.5$	$S_H(i)$	$N_H(i)$	$\frac{\mu_0-X_i}{\sigma}-0.5$	$S_L(i)$	$N_L(i)$
41	03/13/89	1.039	1.033	0.00517	0.65	0.65	1	−1.65	0.00	0
42	03/15/89	1.039	1.033	0.00517	0.65	1.29	2	−1.65	0.00	0
43	03/21/89	1.029	1.033	0.00517	−1.29	0.00	3	0.29	0.29	1
44	03/22/89	1.036	1.033	0.00517	0.07	0.07	4	−1.07	0.00	0
45	03/23/89	1.030	1.033	0.00517	−1.10	0.00	0	0.10	0.10	1
46	03/27/89	1.033	1.033	0.00517	−0.52	0.00	0	−0.48	0.00	0
47	03/28/89	1.036	1.033	0.00517	0.07	0.07	1	−1.07	0.00	0
48	03/29/89	1.038	1.033	0.00517	0.45	0.52	2	−1.45	0.00	0
49	03/30/89	1.036	1.033	0.00517	0.07	0.58	3	−1.07	0.00	0
50	04/03/89	1.033	1.033	0.00517	−0.52	0.07	4	−0.48	0.00	0
51	04/04/89	1.036	1.033	0.00517	0.07	0.13	5	−1.07	0.00	0
52	04/05/89	1.034	1.033	0.00517	−0.32	0.00	0	−0.68	0.00	0
53	04/06/89	1.029	1.033	0.00517	−1.29	0.00	0	0.29	0.29	1
54	04/07/89	1.033	1.033	0.00517	−0.52	0.00	0	−0.48	0.00	0
55	04/10/89	1.037	1.033	0.00517	0.26	0.26	1	−1.26	0.00	0
56	04/14/89	1.042	1.033	0.00517	1.23	1.49	2	−2.23	0.00	0
57	04/17/89	1.044	1.033	0.00517	1.61	3.10	3	−2.61	0.00	0
58	04/18/89	1.041	1.033	0.00517	1.03	4.13	4	−2.03	0.00	0
59	04/19/89	1.040	1.033	0.00517	0.84	4.97	5	−1.84	0.00	0
60	04/20/89	1.036	1.033	0.00517	0.07	5.04	6	−1.08	0.00	0

Table 7. *Continued.*

i	Date	X_i	μ_0	σ $(0.5\%_{\mu_0})$	$\frac{X_i-\mu_0}{\sigma}-0.5$	$S_H(i)$	$N_H(i)$	$\frac{\mu_0-X_i}{\sigma}-0.5$	$S_L(i)$	$N_L(i)$
61	04/28/89	1.039	1.040	0.00520	−0.69	0.00	0	−0.31	0.00	0
62	05/01/89	1.035	1.040	0.00520	−1.46	0.00	0	0.46	0.46	1
63	05/02/89	1.047	1.040	0.00520	0.85	0.85	1	−1.85	0.00	0
64	05/03/89	1.028	1.040	0.00520	−2.81	0.00	0	1.81	1.81	1
65	05/04/89	1.035	1.040	0.00520	−1.46	0.00	0	0.46	2.27	2
66	05/05/89	1.038	1.040	0.00520	−0.88	0.00	0	−0.12	2.15	3
67	05/09/89	1.031	1.040	0.00520	−2.23	0.00	0	1.23	3.38	4
68	05/10/89	1.041	1.040	0.00520	−0.31	0.00	0	−0.69	2.69	5
69	05/12/89	1.043	1.040	0.00520	0.08	0.08	1	−1.08	1.62	6
70	05/15/89	1.034	1.040	0.00520	−1.65	0.00	0	0.65	2.27	7

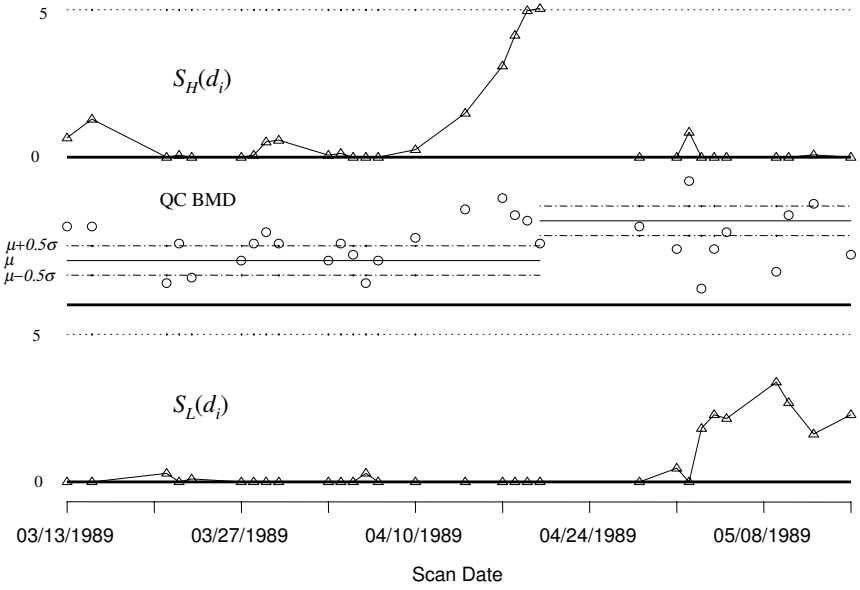


Fig. 3. CUSM chart for QC data in Example 3.

to Ryan.[39] In this approach, the observed difference of two successive scans $X_i - X_{i-1}$ was transformed to $Z_i = \{|[(X_i - X_{i-1})/\sqrt{2\sigma^2}|^{1/2} - 0.82218\}/0.34914$, which approximately follows a standard normal distribution $N(0,1)$. For the variance chart, we selected $k = 0.75$ to reduce the number of alarms due to single outliers. When an alarm for a change in variance is identified, we will investigate the causes of the alarm and may need to recalibrate the machine.

Table 8 is a variance chart for Example 3. The table has calculated values of Z_i. Since Z_i follows a standard normal distribution, the upper side CUSUM for variance is $S_H(i) = \max[0, Z_i - 0.75 + S_H(i-1)]$, which is given in the eighth column. As before, $N_H(i)$ indicates when a positive cumulative sum occurs and is useful for finding the assignable causes. The graphic presentation is similar to Fig. 3 and is not presented here.

The general procedure for deriving the algebraic boundaries of the CUSUM chart is given in Montgomery[35] and Rice[39] and theoretical comparisons of Shewhart and CUSUM can be found in both books.

The V-mask chart is another form of CUSUM chart and is essentially the same as the Tabular CUSUM.[35,40]

Table 8. CUSUM table for change of variance for Example 3.

i	Date	X_i	$X_i - X_{i-1}$	$\sigma(0.5\%_{\mu_0})$	Z_i	$Z_i - 0.75$	$S_H(i)$	$N_H(i)$
55	04/10/89	1.037	0.004	0.00517	−0.24	−0.99	0.000	0
56	04/14/89	1.042	0.005	0.00517	0.01	−0.74	0.000	0
57	04/17/89	1.044	0.002	0.00517	−0.86	−1.61	0.000	0
58	04/18/89	1.041	−0.003	0.00517	−0.52	−1.27	0.000	0
59	04/19/89	1.040	−0.001	0.00517	−1.30	−2.05	0.000	0
60	04/20/89	1.036	−0.004	0.00517	−0.24	−0.99	0.000	0
61	04/28/89	1.039	0.003	0.00520	−0.52	−1.27	0.000	0
62	05/01/89	1.035	−0.004	0.00520	−0.24	−0.99	0.000	0
63	05/02/89	1.047	0.012	0.00520	1.30	0.55	0.554	1
64	05/03/89	1.028	−0.019	0.00520	2.25	1.50	2.053	2
65	05/04/89	1.035	0.007	0.00520	0.44	−0.31	1.742	3
66	05/05/89	1.038	0.003	0.00520	−0.53	−1.28	0.467	4
67	05/09/89	1.031	−0.007	0.00520	0.44	−0.31	0.156	5
68	05/10/89	1.041	0.010	0.00520	0.99	0.24	0.391	6
69	05/12/89	1.043	0.002	0.00520	−0.86	−1.61	0.000	0
70	05/15/89	1.034	−0.009	0.00520	0.81	0.06	0.060	1

3.5. *Comparison of statistical process control charts in osteoporosis studies*

Lu *et al.*[34] compared several statistical process control procedures and their applications to monitoring DXA scanners based on daily scans of a Hologic spine phantom. The comparisons were based on their results on longitudinal quality control data from 5 clinical trial sites as well as simulation studies. They concluded that visual inspection is relatively subjective and depends on the operator's experience and alertness. The regular Shewhart chart with sensitizing rules has a high false alarm rate. The Shewhart chart with sensitizing rules and an additional filter of clinically insignificant mean changes has the lowest false alarm rate but relatively low sensitivity. This method does not require a lot of statistics and can be easily applied to clinical study sites. The CUSUM approach has the best combination of sensitivity, specificity, and identification of the time and magnitude of change. It is recommended for use in quality control centers in clinical trials, especially if patient data must be recalculated to adjust for change points.[41] Combining a moving average chart and a moving standard deviation chart comes closest to the performance of the CUSUM method as a quality control procedure for monitoring DXA scanner performance.

3.6. *Other charts*

In all the above procedures, we assumed that there is no autocorrelation between consecutive measurements. This is rarely true for longitudinal quality control for radiological equipment. The effects of such an assumption on the use of statistical process control charts and their decision structures are rather debatable. At one extreme, Wheeler[42] argues that the usual control limits are contaminated "only when the autocorrelation becomes excessive (say 0.80 or larger)." He concludes that "one need not be overly concerned about the effects of autocorrelation upon the control chart." Our personal experience with Shewhart or CUSUM charts and DXA quality control has been positive. This does not preclude autocorrelation from being a problem for other applications. Johnson and Bagshaw[43] concluded that the problem is potentially quite serious. Strike suggested "clever use" of CUSUMs in laboratory medicine, such as process control for assays.[22]

Statistical approaches for dealing with autocorrelation are to construct process charts based on residuals after removing the autocorrelation or the use of an exponentially weighted moving-average (EWMA) control chart.[35] EWMA is a flexible approach to statistical process control applications. When applied to uncorrelated data, it is a good alternative to the CUSUM chart. Applied to autocorrelated data, it can be adapted to form a control chart that eliminates the excessive false alarm problem associated with traditional control charts. Details of EWMA can be found in most books on quality control.[35,39]

While all the statistical process control charts presented here are for univariate continuous measurements, there are other types of charts for proportions and rates,[44,45] and other quality control and improvement techniques from multivariate approaches.[35,46]

4. Assessment of Agreement

In quality control for clinical trials, we must always assess the agreement of measurements. For example, during a longitudinal osteoporosis trial, a study site might upgrade its DXA machine. Because the change of BMD from baseline is the key measurement, we must be certain that the BMD values measured by the old and new machines are equivalent or in agreement. Also in clinical trials that require a radiologist's assessment of outcomes, we must be certain that readings from different radiologists are the same, and that readings at the beginning and the end of the study are similar. All these require assessment of agreement.

After a DXA scanner upgrade, multiple phantoms scans should be performed, and if possible, a group of volunteers should be scanned on both the old and new devices. If human data is available, it can be used data to assess the agreement rather than phantom data. We hope the volunteers present a range of BMD wide enough to cover the spectrum of clinical uses. Before upgrading a machine that is being used in a clinical trial, the site must first inform the trial sponsors and quality assurance centers for their approval and must rely on manufacturers to assure proper installation and calibration. The site must maintain proper documentation for machine upgrades.

Assessment of inter-reader agreement among radiologists in a clinical trial and intra-reader longitudinal consistency during a trial, normally requires group training before the trial starts. A database of representative images is assembled into a database. Potential readers for the study read the images together and discuss the grading criteria. Only trained radiologists can be readers. The group training should be documented. After training, inter-reader agreement should be assessed. If the agreement does not satisfy the requirements of the sponsors or protocols, the readers will be re-trained and a new set of test cases used to test for agreement. The trial cannot start until reader agreement reaches the pre-specified requirements. During the trial, the radiologists are required to re-read the test sets periodically to assess the agreement of their current readings with their baseline readings. This is necessary to assure longitudinal consistency. All tests for reader agreement should be documented and archived for auditing purposes.

Evaluation of agreement is also important for other purposes, such as validation of diagnostic methods or radiological devices. In these cases, a gold standard will be selected and validation is performed to assure the new measurements agree with the gold standard.

4.1. *Association versus agreement*

The concepts of agreement and association are related but different. Agreement means interchangeability of two measurements. In other words, a patient's BMD should be the same whether measured on an old DXA scanner or a new one; and the spine fracture grade of a vertebra should be the same regardless by whom or when it is read. An association, on the other hand, suggests that two machines or two readers tend to agree in the same directions. In other words, for two patients with different BMD values, both DXA machines will find the same lower and higher BMD subjects but their BMD measurements can be different.

The best example of the difference between agreement and association is the correlation coefficient of two continuous variables.[47,48] A correlation coefficient can apply to any two continuous variables regardless of their scales, such as height and weight. Even if there is a high association between height and weight, they are not interchangeable because they measure completely different things. Even when X and Y are two continuous variables that measure the same physical properties in the same units, an association still cannot indicate agreement. In fact, $cor(X, Y) = cor(a + bX, Y)$. Thus, the correlation is invariant for a shift of mean or a change of scale. Further, the estimation of the correlation depends on the range of the true quantity in the sample: the wider the range, the higher the correlation coefficient. Also, the null hypothesis in testing for a correlation coefficient is the more independent of two variables, which is not relevant to the agreement. Therefore, the use of correlation to assess agreement is inappropriate. On the other hand, a high correlation of two continuous variables in the same scale suggests that it is possible to calibrate variables so that they agree with each other.

4.2. *Assessment of agreement of two continuous variables*

As discussed above, only when two variables measure the same physical property using the same units can they be assessed for agreement. Let Y_1 and Y_2 be such continuous variables that follow normal distributions $N(\mu_{Y_1}, \sigma^2_{Y_1})$ and $N(\mu_{Y_2}, \sigma^2_{Y_2})$. They are measured from the same subjects. The correlation coefficient between Y_1 and Y_2 is ρ. Let $D = Y_1 - Y_2$ and $A = (Y_1+Y_2)/2$. We want to perform a regression analysis of $D = \alpha+\beta A+\varepsilon$. We are interested in $\alpha = \beta = 0$.[47]

It is easy to verify that

$$\beta = \mathrm{cov}(D, A)/\sigma^2_A = 0.5(\sigma^2_{Y_1} - \sigma^2_{Y_2})/(\sigma^2_{Y_1} - \sigma^2_{Y_2} + 2\rho\sigma_{Y_1}\sigma_{Y_2})\,, \qquad (22)$$

and

$$\alpha = (\mu_{Y_1} - \mu_{Y_2}) - \frac{\mu_{Y_1} + \mu_{Y_2}}{2}\beta\,. \qquad (23)$$

Therefore, $\alpha = \beta = 0$ implies that $\mu_{Y_1} = \mu_{Y_2}$ and $\sigma_{Y_1} = \sigma_{Y_2}$, i.e., the two measurements have the same distribution parameters.

Bland and Altman[47] further suggested plotting the difference D against average A and calculating the standard deviation of D (σ_D). With 95% confidence, the differences between paired data are between $\pm 2\sigma_D$. If this σ_D is less than or equal to the precision errors of Y_1 and Y_2, then these two

measurements are exchangeable and therefore, equivalent. Also, if $\sigma_D/\bar{A} \times 100\%$ is less than the CVs for Y_1 and Y_2, they should be equivalent. Here we use a bar to denote sample means.

Noting that both D and A are random variables, Bartko[49] proposed a bivariate confidence ellipse for the Bland-Altman plot. The equation of the 95% ellipse is

$$(A-\bar{A})^2/\sigma_A^2 - 2r(A-\bar{A})(D-\bar{D})^2/\sigma_A\sigma_D + (D-\bar{D})^2/\sigma_D^2 = q_{\chi^2}(0.95,2)(1-r^2)\,. \tag{24}$$

Here, $q_{\chi^2}(0.95,2) = 5.991$ is the 95% quantile of the χ^2-distribution with 2 degrees of freedom and r is the sample correlation coefficient of D and A.

The hypothesis $\alpha = \beta = 0$ can be tested using the Bradley-Blackwood procedure.[50] The test statistic is

$$F = (n-2)\frac{\left(\sum_{i=1}^n D_i^2 - \sum_{i=1}^n (D_i - \hat{\alpha} - \hat{\beta}A_i)^2\right)}{\left(2\sum_{i=1}^n (D_i - \hat{\alpha} - \hat{\beta}A_i)^2\right)} \sim F(2, n-2)\,, \tag{25}$$

which simultaneously tests for the zero intercept and slope.

Table 9 shows a dataset of AP Spine BMD (mg/cm^2) from 10 normal volunteers measured on three different DXA scanners. We are interested in the equivalence of Scanner 1 and the other two scanners.

As shown in Table 9, we can accept the null hypothesis that there is no difference in means and standard deviations between Scanners 1 and 2 by the Bradley-Blackwood test. There is, however, a significant difference between Scanners 1 and 3. Further examination of the data shows that Scanners 1 and 2 have different standard deviations. Using Bland and Altman's method, we can plot the comparison of Scanners 1 versus 2 and Scanners 1 versus 3 (Fig. 4). The dashed line shows that the 95% confidence interval is the most important measurement of these figures. Even though there is a significant non-zero intercept or slope in the Bland-Altman regression, we may still be able to treat the two measurements as interchangeable if the variation of differences is less than the *in vivo* short-term precision error. The 95% confidence ellipse of a Bland-Altman plot is useful for indicating the differences between sample variances.

A bivariate normal distribution has 5 parameters: two means, two standard deviations, and a correlation coefficient. The Bland-Altman regression compares four of the five parameters. We can have two normal random variables with the same mean and standard deviation but a negative correlation coefficient, such as Y and $-Y$, when mean Y is 0. Thus, the

Table 9. AP spine BMD of 10 patients by three DXA scanners.

	Observed BMD Data			Comparison Scanners 1 and 2		Comparison Scanners 1 and 3	
Subject	Scanner 1	Scanner 2	Scanner 3	D_1	A_1	D_2	A_2
1	1.342	1.328	1.352	0.014	1.335	−0.010	1.347
2	1.303	1.312	1.317	−0.009	1.308	−0.014	1.310
3	1.093	1.100	1.078	−0.007	1.096	0.015	1.085
4	1.092	1.116	1.087	−0.024	1.104	0.005	1.089
5	1.215	1.215	1.216	0.000	1.215	−0.001	1.216
6	1.155	1.157	1.137	−0.002	1.156	0.018	1.146
7	1.125	1.117	1.097	0.008	1.121	0.028	1.111
8	1.434	1.437	1.447	−0.003	1.436	−0.013	1.441
9	1.230	1.225	1.231	0.005	1.228	−0.001	1.231
10	1.326	1.324	1.313	0.002	1.325	0.013	1.320
	σ_D			0.0104		0.0141	
	Bradley-Blackwood Test		F	0.6733		5.3645	
			p-value	0.5367		0.0333	

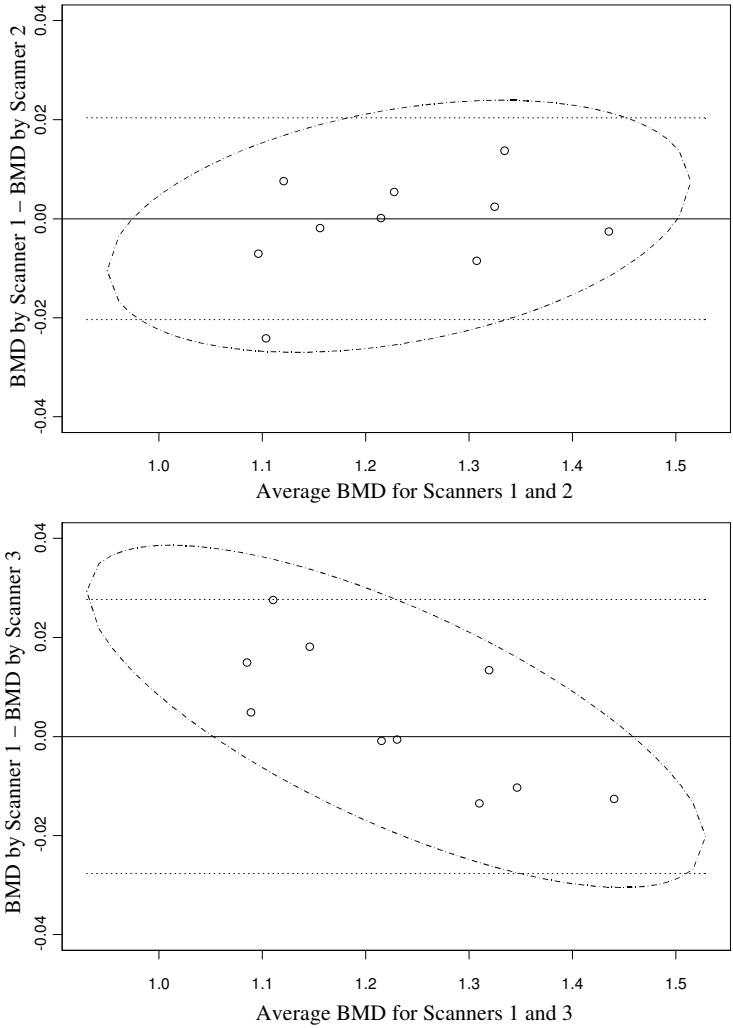


Fig. 4. Examples of Bland-Altman plots for equivalence of 3 scanners. The dashed lines are the 95% confidence intervals for the differences between two Scanners. The ellipses are the 95% bivariate confidence ellipses.

Bland-Altman regression alone is inadequate for evaluating agreement. We still need to examine the correlation coefficient between the two measurements, in addition to the Bland-Altman regression. Only a high correlation with a zero intercept and slope in the Bland-Altman regression can suggest that the two measurements are equivalent.

4.3. *Intraclass correlation coefficient*

An alternative measurement for agreement is the intraclass correlation coefficient (ICC),[51] which is simply the percentage of between readers/techniques variance in the total variance of the sum of between and within reader/technique variations. More specifically, we assume that $Y_{ij} = \mu + p_i + r_j + (pr)_{ij} + \varepsilon_{ij}$, with i representing the ith individual ($i = 1, \ldots, N$) and j representing the jth reader/devices ($j = 1, \ldots, K$). Here, Y_{ij} is the observation of the ith individual measured by jth reader/scanner/machine; μ is the overall effect common to all observations; p_i is the random patient effect; r_j is the random reader/device effect; $(pr)_{ij}$ is the interaction between patient and reader/device; and ε_{ij} is the measurement error. Here, we assume that p_i and r_j are independent and follow normal distributions $N(0, \sigma_P^2)$ and $N(0, \sigma_R^2)$, respectively, and ε_{ij} is independent of p_i and r_j and follows $N(0, \sigma_e^2)$. Without duplicate observations, the interaction term $(pr)_{ij}$ cannot be separated from measurement error and can be dropped. An intraclass correlation coefficient is defined as

$$\mathrm{ICC} = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_R^2 + \sigma_e^2}\,. \tag{26}$$

Thus, a high ICC means less difference between two readers as well as less measurement error. Lee *et al.* suggested a cut-off value of 0.75 beyond which the readers or measurement devices are considered to be in agreement.[51]

The ICC can be estimated based on the output of an ANOVA table of the two-way mixed model as the following.

$$\rho_{\mathrm{ICC}} = \frac{N(\mathrm{MSB} - \mathrm{MSE})}{N\,\mathrm{MSB} + K\,\mathrm{MSR} + (KN - K - N)\mathrm{MSE}}\,. \tag{27}$$

Here, MSB, MSR, and MSE are the mean squared between subject, between reader/device, and error respectively.

Fleiss and Shrout[52] derived an approximate formula for the confidence interval of ρ_{ICC}. Let F_U and F_L be the upper and lower $100(1 - \alpha/2)\%$ percentiles, respectively from F distribution with degrees of freedom $(N-1)$ and v, where

$$v = \frac{(K-1)(N-1)\{K\rho_{\mathrm{ICC}}\mathrm{MSR}/\mathrm{MSE} + N[1 + (K-1)\rho_{\mathrm{ICC}}] - K\rho_{\mathrm{ICC}}\}^2}{(N-1)K^2\rho_{\mathrm{ICC}}^2\mathrm{MSR}^2/\mathrm{MSE}^2 + \{N[1 + (K-1)\rho_{\mathrm{ICC}}] - K\rho_{\mathrm{ICC}}\}^2}\,. \tag{28}$$

Table 10. ANOVA Tables and ICC for Data in Example 4.

Agreement for	Scanners 1 and 2		Scanners 1 and 3		All 3 Scanners	
Source	d.f	MS	d.f	MS	d.f	MS
Between-subject	9	MSB $= \sigma_2^2 + K\sigma_P^2 = 0.02660$	9	MSB $= 0.02992$	9	MSB $= 0.04278$
Between-scanner	1	MSR $= \sigma_2^2 + N\sigma_R^2 = 0.00001$	1	MSR $= 0.00008$	2	MSR $= 0.00008$
Residual	9	MSE $= \sigma_2^2 = 0.00005$	9	MSE $= 0.00010$	18	MSE $= 0.00010$
Total	19		19		29	
v		9.3622		9.9615		19.9237
ICC and 95% C.I.		0.9963 (0.9865, 0.9991)		0.9935 (0.9755, 0.9983)		0.9931 (0.9807, 0.9981)

The approximate upper and lower bounds, ρ_U and ρ_L, respectively, for the $100(1-\alpha)\%$ confidence bounds of ρ_{ICC} are given as following.

$$\rho_U = \frac{N(\text{MSB} - F_L\text{MSE})}{F_L[K\,\text{MSR} + (NK - L - N)\text{MSE}] + N\,\text{MSB}} \tag{29}$$

and

$$\rho_L = \frac{N(\text{MSB} - F_U\text{MSE})}{F_U[K\,\text{MSR} + (NK - L - N)\text{MSE}] + N\,\text{MSB}}\,. \tag{30}$$

Table 10 shows ANOVA tables for comparison of scanners in Example 4, and the corresponding intraclass correlation coefficients.

It is clear from this example that ICC is less sensitive to agreement between two scanners. The ICC for Scanners 1 and 3 is much higher but the Bland-Altman regression shows significant disagreement. Bland and Altman[53] list other deficiencies of ICC for evaluation of agreement, including its dependence on sample variations. On the other hand, it is easier to use ICC to evaluate agreement among three or more readers or devices. Bartko[49] developed an altered version of ICC, which is simplified and has an exact formula for confidence intervals.

4.4. *Kappa statistics for agreement of categorical variables*

Like continuous measurements, agreement between two categorical variables is only meaningful when the two categorical variables have the same biological or physical meanings. Agreement of categorical variables is most commonly applied to qualitative evaluations of health or disease status by two readers or by the same reader at two different sessions, which are referred as inter-reader and intra-reader agreement respectively. In clinical studies using qualitative assessments by multiple readers, we hope that all readers will produce consistent readings, and that their assessments will remain consistent during the study period. Thus, periodic review of inter- and intra-reader agreement should be a part of quality control of clinical trials. If the readers do disagree with each other, re-training is necessary.

The simplest way to display categorical variables of two readers is a 2×2 table, displayed in Table 11. Here, X_1 and X_2 are results from two readers, with 0 indicating healthy and 1 indicating diseased, and P_{ij} representing the probability of the event. There are many ways to measure the agreement of two readers. The probability of agreement, i.e., $P(X_1 = X_2) = P_{00} + P_{11}$ is the most direct measurement. Analysis of the

Table 11. Joint distribution of outcomes of two binary variables.

		X_2		
		Health ($X_2 = 0$)	Diseased ($X_2 = 1$)	Total
X_1	Health ($X_1 = 0$)	P_{00}	P_{01}	P_{0+}
	Diseased ($X_1 = 1$)	P_{10}	P_{11}	P_{1+}
Total		P_{+0}	P_{+1}	1

probability of agreement is just like analysis of binary probability. Sample size calculations for reader agreement based on duplicated readings were presented by Freedman, Parmar, and Baker.[54]

The drawback of the probability of agreement is a positive chance of agreement even when the two readers are independent. As a result, Cohen proposed the use of Kappa statistics,[55] which offer a means of correcting measurement of agreement, defined as the following.

$$\kappa = \frac{P_{00} + P_{11} - P_{0+}P_{+0} - P_{1+}P_{+1}}{1 - P_{0+}P_{+0} - P_{1+}P_{+1}} = \frac{P_O - P_E}{1 - P_E}. \tag{31}$$

Here, $P_O = P_{00} + P_{11}$ is the observed probability of agreement and $P_E = P_{0+}P_{+0} + P_{1+}P_{+1}$ is the probability of agreement due to changes when X_1 and X_2 are independent. κ can reach 100% if there is perfect agreement and can be as low as $-P_E/(1 - P_E)$, when X_1 and X_2 are completely different.

If we use n_{ij} to denote the observed number of subjects in each category of Table 11, the maximum likelihood estimates for P_{ij}, P_{i+} and P_{+j} are $\hat{p}_{ij} = n_{ij}/n$, $\hat{p}_{+j} = n_{i+}/n$ and $\hat{p}_{+j} = n_{+j}/n$, respectively, with n as the total number of subjects. Through algebra operations, we can estimate κ by substituting the maximum likelihood estimates of the probabilities into Eq. (31).

$$\hat{\kappa} = \frac{2(n_{00}n_{11} - n_{01}n_{10})}{n_{0+}n_{+1} + n_{+0}n_{1+}}. \tag{32}$$

There are several methods for calculating the sample variations for MLE estimates in Eq. (32). Using the delta method, Fleiss *et al.*[56] derived a large sample variance of the estimator.

$$\begin{aligned} \mathrm{var}(\hat{\kappa}) = {} & \frac{1}{n(1 - P_{0+}P_{+1} - P_{1+}P_{1+})^2} \\ & \times \left\{ \sum_{i=0}^{1} P_{ii}[1 - 2(P_{i+} + P_{+i})(1 - \kappa)] \right. \end{aligned}$$

$$+(1-\kappa)^2\sum_{i=0}^{1}\sum_{j=0}^{1}P_{ij}(P_{+i}+P_{j+})^2$$

$$-[\kappa-(P_{0+}P_{+0}+P_{1+}P_{+1})(1-\kappa)]^2\Bigg\}. \tag{33}$$

Alternatively, Kraemer[57] and Fleiss and Davies[58] proposed the use of jackknife technique to calculate the variance of the estimated κ. Let $\hat{\kappa}_{ij}$ be the MLE of κ when one observation in the (i,j)th cell is excluded, and $J_{ij}(\hat{\kappa}) = n\hat{\kappa} - (n-1)\hat{\kappa}_{ij}$. The jackknife estimator of κ is given by

$$\hat{\kappa}_J = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij}J_{ij}(\hat{\kappa})/n\,, \tag{34}$$

which should be a less biased estimator than $\hat{\kappa}$. The jackknife variance can be estimated by

$$\mathrm{var}(\hat{\kappa}_J) = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij}[J_{ij}(\hat{\kappa}) - \hat{\kappa}_J]^2/[n(n-1)]\,. \tag{35}$$

Conditioned on marginal distributions of the 2×2 table in Table 11, Garner[59] proposed the following simpler formula:

$$\mathrm{var}(\hat{\kappa}) = \frac{4}{n^2(1-\hat{p}_{0+}\hat{p}_{+0}-\hat{p}_{1+}\hat{p}_{+1})^2(\sum_{i=0}^{1}\sum_{j=0}^{1}1/(n_{ij}+1))}\,. \tag{36}$$

Although all these formulas are asymptotically equivalent, there are still differences when using them for small samples. A simulation study[60] compared the different estimates for $\hat{\kappa}$ and gave guidance in methods to estimate and construct confidence intervals for Cohen's $\hat{\kappa}$ for small samples as indicated in Table 12. In this table, the "(" and ")" indicate the open-ends of an interval and "[" and "]" the closed ends of an interval. Landis and Koch[61] provided guidelines for interpreting kappa values as the level of agreement among readers. Prevalence was defined as $(2n_{11}+n_{10}+n_{01})/(2n)$.[62] The last column indicates the preferred equations for estimating the sample variance.

The use of Kappa statistics in quality control and quality assurance is mainly for estimation rather than hypothesis testing. We want to ensure that the inter-reader agreement is above an acceptable pre-specified level before we start the study. We also want to be certain that the longitudinal intra-reader Kappa statistics are beyond that given level. However, the subject of Kappa applications is very broad and goes far beyond quality

Table 12. Guidance in selecting a method for constructing confidence intervals for Cohen's $\hat{\kappa}$.[60]

Kappa ($\hat{\kappa}$)	Agreement[61]	Prevalence[62]	Sample Size	Equations
[0, 0.2)	Slight	(0.1, 0.9)	$n \geq 20$	(33)
[0.2, 0.4)	Fair	[0.1, 0.9]	$n \geq 20$	(33) or (35)
[0.4, 0.6)	Moderate	(0.2, 0.8)	$20 \leq n < 40$	(36)
		(0, 0.2] or [0.8, 1)	$n \geq 40$	(35)
[0.6, 1)	Substantial to almost perfect	(0.1, 0.9)	$n \geq 20$	(36)

assurance. The extensive literature on Kappa statistics includes agreement for ordinal or multinomial data;[63–66] for case-control studies;[67] for multiple readers or correlated samples;[68–70] and for using logistic regression models to adjust for the effects of covariates on Kappa statistics.[71] These topics are far beyond the scope of this chapter; interested readers should investigate the literature.

4.5. *Log-linear models for agreement of categorical variables*

Log-linear models can express agreement in terms of components, such as chance agreement and beyond-chance agreement. They can also display patterns of agreement among several observers, or compare patterns of agreement when subjects are stratified by values of a covariate.[72] The later is particularly useful for quality improvement to identify factors that have an affect on reader agreement.

Let $\{m_{ij} = nP_{ij}\}$ denote expected frequencies for ratings (i, j) of n subjects by two observers A and B. Chance agreement, or statistical independence of the ratings, has log-linear model representation

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B \,. \tag{37}$$

An extension of this independent model is the quasi-independent model[73]

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta_i I_{(i=j)} \,, \tag{38}$$

where the indicator $I_{(i=j)}$ equals 1 when $i = j$ and 0 otherwise. Constrains on the model parameters are $\sum_i \lambda_i^A = \sum_j \lambda_j^B = 0$. Conditional on disagreement by the observers, the rating by A is statistically independent of rating by B. When $\delta_i > 0$, more agreements regarding outcome i occur than would be expected by chance. The model is easy to fit by most statistical software.

When we assume a constant $\delta_i = \delta$, a Kappa-like index of chance-corrected agreement[74] is

$$\kappa_A = (P_{00} + P_{11})(1 - e^{-\delta}) = (P_{00} + P_{11})(1 - 1\sqrt{OR}) \,. \tag{39}$$

Graham extended above model to allow binary covariates.[75] Let X be the binary covariate with value 0 and 1. Let $\{m_{ijk} = nP_{ij}(X = k)\}$ be the frequencies of observing (i, j) by readers A and B when covariate X equals k. The extended model is

$$\log m_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^X + \lambda_{ik}^{AX} + \lambda_{jk}^{BX} + \delta^{AB} I_{(i=j)} + \delta_k^{ABX} I_{(i=j)} \,. \tag{40}$$

Here, terms with single superscripts and subscripts correspond to main effects. Terms with double superscripts and subscripts represent partial associations between the superscripted variables, controlling for the variable omitted from the superscript. As with other log-linear models, we impose the constraints of zero sums on the main effects and partial associations, respectively. In this model, δ^{AB} represents the overall agreement between two readers and δ_k^{ABX} represents the additional chance corrected agreement associated with covariate X when $X = k$. A model constraint is zero sum of δ_k^{ABX}. This model readily extends to multiple covariate situations, and estimates can be obtained using the SAS CATMOD procedure.

In model (40), δ_k^{ABX} is an interpretation of the estimates of the average of the two conditional agreement log odds ratios, $\log[(m_{iik}/m_{jik})/(m_{ii0}/m_{ji0})]$ and $\log[(m_{jjk}/m_{ijk})/(m_{jj0}/m_{ij0})]$, for any pair of distinct categories i and j. In his paper,[75] Graham applied this model to a study of the effects of age, sex, and proxy type on agreement between the primary and proxy respondents regarding the primary respondent's participation in vigorous leisure time activity.

4.6. *Latent class models*

In a latent class analysis of observer agreement, it is assumed that the ratings of observers appear related because they are, in fact, related to some latent classification of items that explains all associations in the observed agreement table. For example, we can assume that there are three types of subjects in the study population: those that all readers classify as positive or negative, and those inconclusive subjects that are rated as positives or negatives by chance by each reader.[76] Let K be the prevalence of those "agreements beyond chance" and p be the probability of conclusive items belonging to the positive category. Let π be the probability of positively

Table 13. Probability in a 2×2 table with latent classification model.

Rater A	Rater B		
	Positive	Negative	Total
Positive	$K_p + (1-K)\pi^2$	$(1-K)(1-\pi)\pi$	$K_p + (1-K)\pi$
Negative	$(1-K)(1-\pi^)\pi$	$K(1-p) + (1-K)(1-\pi)^2$	$K(1-p) + (1-K)(1-\pi)$
Total	$K_p + (1-K)\pi$	$K(1-p) + (1-K)(1-\pi)$	1

rating for inconclusive subjects. With the assumption of independent rating by the two readers for inconclusive subjects, the following Table 13 gives the probability distribution of the 2×2 table.

Thus, if $p = \pi$, K is the Cohen's Kappa statistics. If $p/(1-p) = \pi^2/(1-\pi)^2$, K equals Aickin's Kappa in Eq. (39).

Latent classification models have many uses.[76] Baker, Freedman, and Parmar[77] proposed a model with duplicate observations that allows a separation of intra- and inter-reader agreement simultaneously for binary measures.

5. Clibration and Standardization

The most important mission of quality assurance is to prevent measurement errors from exceeding a pre-specified level. For this purpose, we evaluate the performance of instruments to ensure that their precision and accuracy are acceptable for clinical diagnosis or clinical monitoring. Once we have chosen the particular devices or methods to measure study parameters, we want to be sure that they are equivalent to each other. During the study, we use the quality process control charts to monitor whether the instruments are still providing the required precision and/or whether the readers are giving consistent readings. With each step, we may still find disagreements between instruments or readers. Once we have chosen one of them as the reference standard, the process of assigning values for other instruments or readers to correct their differences from the reference standard is called calibration.

In the example of multi-center studies, we normally choose the coordinating center as the reference standard. Thus, any site/machine that produces readings or measurements that are different from the reference standards will be calibrated. This is called cross-calibration in the literature on quality control of clinical trials.[41] Although mathematically any site can be chosen as the reference standard, in practice, selection of a reference

standard should take into consideration the qualifications and quality control history of the selected site. Sometimes, multiple reference standards are needed. For example, in a clinical trial of osteoporosis that uses DXA scanners from different manufacturers, one option is to select reference standards for each manufacturer and then calibrate devices at the other study sites to the corresponding reference standards. The next step of calibration is to standardize among the reference standards.

Calibration can also occur for a single radiological machine. In the longitudinal quality control process mentioned in Sec. 3, a radiological machine was compared to a standard defined by a phantom. We normally look for the mean and variance changes in reference to the baseline value. One may also be interested in scale differences, i.e. changes in measurement unit. For DXA scanners, phantoms with different linear scaled densities can be used to serve as reference standard and calibration of a scanner may be needed if there are clinically significant deviations from that standard.

5.1. *Calibration of measurements to a standard*

To calibrate radiological equipment to the chosen standard, we need to measure the standard. One method is to measure phantoms with known theoretical measurement values.[32] Another method is to measure a set of phantoms or a group of sampled subjects to examine the differences between the reference standard device and all other study instruments, referred as cross-calibration in multi-center clinical trials.[7] In all cases, we observe pairs of data (X_i, Y_i) with X_i representing the reference standard and Y_i representing measurement of the instrument to be calibrated.

The practical question is how to assign a correct X (standard value) based on measurement Y. A naïve solution is to perform a (linear or nonlinear) regression of X_i on Y_i and use that regression model to correct future readings of Y. This solution may be adequate, but it has statistical flaws.

When we choose the standard, we assume that the standard should be accurate, that is its measurement error can be ignored. Thus, the measurement error should be associated only with Y not X. A proper linear relationship should be $Y = \alpha + \beta X + \varepsilon$, with α and β as regression parameters and ε as the random measurement error for Y. These regression parameters α and β are also referred as constant bias and relative (scale) bias.

Maximum likelihood estimates of regression parameters, denoted as $\hat{\alpha}$ and $\hat{\beta}$, and their covariance matrix as well as model RMSE are easily

available by many statistical software packages. Based on these estimates, for a given observation of y, we can calibrate it to the standard by $\hat{x} = (y - \hat{\alpha})/\hat{\beta}$.

The predicted value $\hat{x}$ is a biased estimate of true value x except when $x = \bar{X}$.

$$E(\hat{x}|y) = x + [S_e^2(x - \bar{X})]/S_{XX}\hat{\beta}^2)\,. \tag{41}$$

Here, S_e is the RMSE of the regression line and $\bar{X}$ and S_{XX} are the sample mean and sample variance of X_i's used to derive calibrations. This is because $\hat{x}$ is estimated by ratio of correlated normal variables. In most cases, such bias can be ignored for large $\hat{beta}$. More specifically, when

$$g = (t_{n-2,0.05}^2 S_e^2)/(S_{XX}\hat{\beta}^2) < 0.05\,. \tag{42}$$

When $g > 0.2$, we are not able to calibrate Y to the standard X with acceptable accuracy.[22] Details of the 95% confidence interval of calibrated $\hat{x}$ as well as simultaneous tolerance interval for it can be found in the same reference.

When we allow measurement errors for standard X, we are dealing with the calibration problem as a regression with measurement errors, and the regression and calibration problems are equivalent mathematically. Rearrangement of the linear regression gives the following relationship between X and Y:

$$X = \gamma_0 + \gamma_1 Y + \delta\,. \tag{43}$$

The difference between this calibration model and regular regression model is that Y is a random variable with $Y = U + \varepsilon$. This regression is not always identifiable unless under certain conditions.[78] When we assume that the measurement error ε and underlying true U are independent and ε has mean zero and a known variance σ_ε^2 (such as estimated through repeated measurements), the calibration formula is

$$\hat{x} = \mu_X + \hat{\gamma}_1\sigma_Y^2/(\sigma_Y^2 - \sigma_\varepsilon^2)(y - \mu_Y) = \mu_X + \hat{\gamma}_1(\sigma_U^2 - \sigma_\varepsilon^2)/\sigma_U^2(y - \mu_Y)\,. \tag{44}$$

Here, $\hat{\gamma}_1$ is the least squared estimate of slope based on observed Y with measurement errors.

5.2. *Comparative calibrations and latent structure models*

Barnett[79] first considered a model to assess "the relative calibration and relative accuracies of a set of p instruments, each designed to measure the same characteristic, on a common group of individuals." It is common for

several manufacturers to produce similar machines that measure the same physical properties. For various reasons, these machines will not produce identical measurements for the same subjects. Converting measurements from different manufacturers is important for clinical studies to reduce machine introduced variations improving study efficiency and facilitating comparisons among different studies.

For the ith subject, let a vector $\vec{Y}_i = (Y_{1i}, Y_{2i}, \ldots, Y_{pi})^T$ to denote the measurements by p instruments for the subject. Here, superscript T represents "transpose." Statistically, we assume that $\vec{Y}_i$ measures the underlying unobservable quantity X_i from an unknown normal distribution $N(\mu, \sigma_0^2)$. The relationship between $\vec{Y}_i$ and X_i is that

$$\vec{Y}_i = \vec{a} + \vec{b}\, X_i + \vec{\varepsilon}_i \tag{45}$$

with unknown regression parameters $\vec{a}$ and $\vec{a}$, and $\vec{\varepsilon}_i$ as a p-dimensional random measurement errors following $N(\vec{0}, \Sigma)$.

The difference between this model and the regular calibration model is that X_i can be observed in a regular problem, while X_i is unknown in comparative calibration problems.[80]

The number of sufficient statistics based on observations of $\vec{Y}_i$ is p means and $p(p+1)/2$ covariance matrix. The number of unknown parameters are 2 for distribution of X, $2p$ for regression coefficients, and $p(p+1)/2$ for the covariance matrix for measurement errors. Thus, for $p < 3$, comparative calibration is unidentifiable. Even for $p \geq 3$, we still need additional assumptions to make the model identifiable.

Barnett[79] assumed $a_1 = 0$ and $b_1 = 1$, and the covariant matrix of measurement errors Σ as a diagonal matrix. He used moment estimates to obtain MLE for the modal parameters. Other authors have studied similar problems,[81–84] The following EM algorithm is a shorter form of a more extended model by Lu *et al.*[85]

Like Barnett, we assume that Σ is a diagonal matrix. When we do not observe X_i, the log-likelihood of our model is pretty complicated. The log-likelihood function of observations $\vec{Y}_i$ is

$$C - \frac{p}{2}\sum_{i=1}^{n} \log(|\vec{b}\,\vec{b}^{\,T}\sigma_0^2 + \Sigma|]$$

$$- \frac{1}{2}\sum_{i=1}^{n} (\vec{Y}_i - \vec{a} - \vec{b}\,\mu_0)^T (\vec{b}\,\vec{b}^{\,T}\sigma_0^2 + \Sigma)^{-1} (\vec{Y}_i - \vec{a} - \vec{b}\,\mu_0)\,. \tag{46}$$

To make the model identifiable, we also impose linear constrains on regression parameters as $\vec{l}^{\,T}\vec{a} = c_1$ and $\vec{l}^{\,T}\vec{b} = c_2$. When $\vec{l} = (1, 0, \ldots, 0)^T$

and $c_1 = c_2 = 0$, the model is similar to Barnett.[79] When $\vec{l} = (1,1,1)^T$, $c_1 = 0$ and $c_2 = 2.912$, the model is similar to Lu *et al.*[84] While the log-likelihood function is complicated, the likelihood function for known X_i is rather simple:

$$C - \frac{p}{2}\log(|\Sigma|) - \frac{1}{2}\log\sigma_0^2 - \frac{1}{2}(\vec{Y}_i - \vec{a} - \vec{b}\,X_i)^T \times \Sigma^{-1}(\vec{Y}_i - \vec{a} - \vec{b}\,X_i) - \frac{(X_i - \mu_0)^2}{2\sigma_0^2}\,. \tag{47}$$

Thus, we can treat X_i as missing data and use the EM algorithm to derive the MLE of model parameters. The EM algorithm has the following steps:

Step 0. Set the initial values of the model parameters $\vec{a}, \vec{b}, \Sigma$, μ_0 and σ_0^2.

Step 1. E-Step: Calculate the conditional expectation of the sufficient statistics for the complete likelihood function. They are

$$V = \mathrm{var}(X_i|\,\vec{Y}_i, \vec{a}, \vec{b}, \Sigma, \mu_0, \sigma_0^2) = (\vec{b}^T\Sigma^{-1}\vec{b} + 1/\sigma_0^2)^{-1}\,, \tag{48}$$

$$E(X_i|\,\vec{Y}_i, \vec{a}, \vec{b}, \Sigma, \mu_0, \sigma_0^2) = \mu_0 + V\,\vec{b}^T\,\Sigma^{-1}(\vec{Y}_i - \vec{a} - \vec{b}\,\mu_0)\,. \tag{49}$$

Step 2. M-Step: Calculate the MLEs by replacing the conditional sufficient statistics into the following MLE formulas.

$$\hat{\vec{b}} = \frac{S_{Y,X} - (\lambda_1\bar{X} + \lambda^2)\Sigma^{-1}\vec{l}}{S_{XX}}\,, \tag{50}$$

$$\hat{\vec{a}} = \overline{\vec{Y}} - \hat{\vec{b}}\,\bar{X} - \lambda_1\Sigma\vec{l}\,, \tag{51}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}[(\vec{Y}_i - \hat{\vec{a}} - \hat{\vec{b}}\,X_i)(\vec{Y}_i - \hat{\vec{a}} - \hat{\vec{b}}\,X_i)^T]\,, \tag{52}$$

$$\hat{\mu}_0 = \bar{X}\,, \tag{53}$$

$$\hat{\sigma}_0^2 = \sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2/n\,. \tag{54}$$

Here, $\overline{\vec{Y}}$ and $\bar{X}$ are the sample means for $\vec{Y}_i$ and X_i, respectively; $S_{Y,X} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\ (\vec{Y}_i - \overline{\vec{Y}})$; and λ_1 and λ_2 are the Lagrange-coefficients for conditional maximization with $\lambda_1 = (\vec{l}^T\overline{\vec{Y}} - c_1 - c_2\bar{X})/\vec{l}^T\Sigma\vec{l}$ and $\lambda_2 = [\vec{l}^T S_{Y,X} + \vec{l}^T\overline{\vec{Y}}\,\bar{X} - (c_1 + c_2)\bar{X} - c_2 S_{X,X}]/\vec{l}^T\Sigma\vec{l}$.

Step 3. Check the convergence of the unconditional log-likelihood function and decide to stop or go back to Step 1.

Based on the MLE, we can calibrate the unobserved underlying X based on measures from any one instrument by inverse linear calibration. Moreover, this model allows us to calibrate measures from instruments k to l by the following formula:

$$\hat{Y}_{i,l} = a_l + b_l(Y_{i,k} - a_k)/b_k\,. \tag{55}$$

Here, subscript i indicates the ith subject and k, l indicate the instruments; a_k, a_l, b_k, and b_l are the kth and lth components in the vectors $\vec{a}$ and $\vec{b}$, respectively.

A much simpler model is for $p = 3$, where the closed forms of MLEs can be derived and asymptotic covariance of the MLEs can be obtained explicitly.[84] This model has been used for standardization of bone mineral densities measured by three different manufacturers.[84,86,87]

5.3. *Least square approach for comparative calibrations*

Alternatively, we define $\vec{Y}'_i = \vec{Y}_i - \overline{\vec{Y}}$ and $\vec{X}_i = G\vec{Y}'_i + k$. Here, k is a real number and G is a $p \times p$ diagonal matrix, $G = \mathrm{diag}(g_j)$ with $g_j \geq 0$. If $\vec{X}_i$ is the standard references for instruments, there should be no differences between any pairs of its components. Let H be a $p \times p$ matrix

$$H = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ -1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Hui *et al.*[88] proposed to find g_j's that minimize the differences between components in vector $\vec{X}_i$[88]:

$$\min \sum_{i=1}^{n} \vec{X}_i^T H^T H \vec{X}_i = \min \sum_{i=1}^{n} (\vec{Y}_1 - \overline{\vec{Y}})^T G^T H^T H G (\vec{Y}_i - \overline{\vec{Y}}) \tag{56}$$

under constrains $\sum_{j=1}^{p} g_j^2 = p$. Because of the quadratic constrains, the solution for minimization Eq. (56) is not in a closed form. Symbolic programming languages, such as Maple, can be used to calculate the numeric solutions.

Like the latent structure models in the previous subsection, this model needs two constraints in order to make the model identifiable. The constant parameter k can be determined by a linear constraint as demonstrated in Hui *et al.*[88]

After we derive the solutions for g_j's, we can use the following formula to calibrate values between instruments:

$$Y_{i,j} = \bar{Y}_{.,j} + g_k/g_j(Y_{i,k} - \bar{Y}_{.,k})\,. \tag{57}$$

For $p = 3$, the calibration conversion formulas between instruments are the same for the least square approach [Eq. (57)] and latent structure model [Eq. (55)] if and only if the measurement errors of instruments in latent structure model σ_j^2 are equal.[84]

6. Conclusions

Radiological instrument quality is important for both clinical diagnosis of disease and clinical monitoring of patient changes. Quality assurance and quality improvement need efforts of people who involve in the processes of manufacturing, maintaining, and operating the equipment as well as statisticians who involved in assessing the quality, monitoring the changes in quality and identifying areas for quality improvement. In this chapter, we have introduced some statistical concepts and methods that are commonly used in quality assurance of radiology studies. There are many other materials and considerations that could not be covered because of the limitation of the space. The methods discussed in this chapter have applications beyond radiological studies and are relevant to most clinical studies. Quality assurance and quality control is rather a practice than a theoretical discussion. Successful quality assurance can have visible and immediate effects. Statisticians should actively participate in quality assurance. While it is important for clinicians and biomedical researchers to realize the importance of statistics in their quality control and quality assurance practice, it is also important for biostatisticians to understand the subject issues and communicate effectively statistical principles to scientists from different backgrounds. The collaborations between statisticians and biomedical researchers in other fields will not only benefit clinical researches but also lead to new challenges for research and development of new statistical methods.

References

1. Huxsoll, J. F. (1994). Organization of quality assurance. In *Quality Assurance for Biopharmaceuticals*, ed. J.F. Huxsoll, John Wiley and Sons, Inc., New York: 2–13.
2. Therasse, P. A., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M. van Oosterom, A. T., Christian,

M. C. and Gwyther, S. G. (2000). New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute* **92**(3): 205–216.
3. WHO. (1994). Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. *Report of a WHO Study Group*. World Health Organization, Geneva.
4. Siris, E. (2000). Alendronate in the treatment of osteoporosis: A review of the clinical trials. *Journal of Womens Health and Gender-Based Medicine* **9**(6): 599–606.
5. Switula, D. (2000). Principles of good clinical practice (GCP) in clinical research. *Sciences Ethics and Engineering* **6**(1): 71–77.
6. van Kuijk, C. (1998). Good clinical practice in clinical trials: What does it mean for a radiology department? *Radiology* **209**(3): 625–627.
7. Fuerst, T., Lu, Y., Hans, D. and Genant, H. K. (1998). Quality assurance in bone densitometry. In *Bone Densitometry and Osteoporosis*, eds. H.K. Genant, G. Guglielmi and M. Jergas, Springer-Verlag, New York: 461–476.
8. Fraass, B. D. K., Hunt, M., Kutcher, G., Starkschall, G., Stern, R. and Van Dyke, J. (1998). American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: Quality assurance for clinical radiotherapy treatment planning. *Medical Physics* **25**(10): 1773–1829.
9. Laurila, JS-N. C. G., Suramo, I., Tolppanen, E. M., Tervonen, O., Korhola, O. and Brommels, M. (2001). The efficacy of a continuous quality improvement (CQI) method in a radiological department. Comparison with non-CQI control material. *Acta Radiologica* **42**(1): 96–100.
10. Genant, H., Wu, C., van Kuijk, C. and Nevitt, M. (1993). Vertebral fracture assessment using a semiquantitative technique. *Journal of Bone and Mineral Research* **8**(9): 1137–1148.
11. Gluer, C., Blake, G., Lu, Y., Blunt, B., Jergas, M. and Genant, H.(1995). Accurate assessment of precision errors: How to measure the reproducibility of bone densitometry techniques. *Osteoporosis International* **5**: 262–270.
12. Njeh, C. F., Nicholson, P. H. F. and Langton, C. M. (1999). The physics of ultrasound applied to bone. In *Quantitative Ultrasound: Assessment of Osteoporosis and Bone Status*. eds. C.F. Njeh, D. Hans, T. Fuerst, C.C. Gluer and H.K. Genant, Martin Dunitz Ltd., London: 420.
13. Gluer, C. and Genant, H. (1989). Impact of marrow fat on accuracy of quantitative CT. *Journal of Computer Assistant Tomographics* **13**(6): 1023–1035.
14. Jergas, M. and Uffmann, M. (1998). Basic considerations and definitions in bone densitometry. In *Bone Densitometry and Osteoporosis*, eds. H. Genant, G. Guglielmi and M. Jergas, Springer, New York: 269–290.
15. Liu, C.-Y. and Zheng, Z.-Y. (1989). Stabilization coefficient to random variable. *Biometrical Journal* **31**(4): 431–441.
16. Miller, G. E. (1991). Asymptotic test statistics for coefficients of variation. *Communication in Statistics — Theory and Methods* **20**(10): 3351–3363.
17. Feltz, C. J. and Miller, G. E. (1996). An asymptotic test for the equality of coefficient of variation from k populations. *Statistics in Medicine* **15**: 647–658.

18. Fung, W. K. and Tsang, T. S. (1998). A simulation study comparing tests for the equality of coefficients of variation. *Statistics in Medicine* **17**: 2003–2014.
19. Arenson, R., Lu, Y., Elliott, S., Jovais, C. and Avrin, D. (2001). Measuring the academic radiologist's clinical productivity. *Academic Radiology* **8**: 524–532.
20. Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, San Francisco.
21. Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York.
22. Strike, P. W. (1991). *Statistical Methods in Laboratory Medicine*, Butterworth-Heinemann Ltd., Oxford.
23. Quan, H. and Shih, W. J. (1996). Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics* **52**(4): 1195–1203.
24. Miller, C. G., Herd, R. J., Ramalingam. T., Fogelman, I. and Blake, G. M. (1993). Ultrasounic velocity measurements through the calcaneus: Which velocity should be measured? *Osteoporosis International* **3**(1): 31–35.
25. Blake, G. M. and Fogelman, I. (1997). Technical principles of dual X-ray absorptiometry. *Seminar in Nuclear Medicine* **27**(3): 210–228.
26. Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall, London.
27. Machado, A., Hannon, R., Henry, Y. and Estell, R. (1997). Standardized coefficient of variation for dual X-ray absorptiometry (DXA), quantitative ultrasound (QUS) and markers of bone turnover (Abstract). *Journal of Bone and Mineral Research* **12**(Suppl. 1): S258.
28. Langton, C. M. (1997). ZSD: A universal parameter for precision in the ultrasonic assessment of osteoporosis. *Physiological Measurement* **18**: 67–72.
29. Cummings, S. R. and Black, D. (1986). Should perimenopausal women be screened for osteoporosis? *Annals of Internal Medicine* **104**: 817–823.
30. Glüer C. (1999). Monitoring skeletal changes by radiological techniques. *Journal of Bone and Mineral Research* **14**(11): 1952–1962.
31. Faulkner, K. M., MR. (1995). Quality control of DXA instruments in multicenter trials. *Osteoporosis International* **5**(4): 218–227.
32. Kalender, W., Felsenberg, D., Genant, H. K., Fischer, M., Dequeker, J. and Reeve, J. (1995). The European spine phantom — A tool for standardization and quality control in spine bone mineral measurements by DXA and QCT. *European Journal of Radiology* **20**: 83–92.
33. Anderson, J. W. and Clarke, G. D. (2000). Choice of phantom material and test protocols to determine radiation exposure rates for fluoroscopy. *Radiographics* **20**(4): 1033–1042.
34. Lu, Y., Mathur, A. K., Blunt, B. A. *et al.* (1996). Dual X-ray absorptieometry quality control: Comparison of visual examination and process-control charts. *Journal of Bone and Mineral Research* **11**(5): 626–637.
35. Montgomery, D. C. (1992). *Introduction to Statistical Quality Control*, 2nd edn., Wiley, New York.

36. Orwoll, E. S., Oviatt, S. K. and Biddle, J. A. (1993). Precision of dual-energy X-ray absorptiometry: Development of quality control rules and their application in longitudinal studies. *Journal of Bone and Mineral Research* **8**(6): 693–699.
37. Jergas, M. and Genant, H. K. (1993). Current methods and recent advances in the diagnosis of osteoporosis. *Arthritis and Rheumatism* **36**(12): 1649–1662.
38. SAS/QC (2000). *User's Guide* (for SAS V8), SAS Research Institute, Cary, North Carolina.
39. Ryan ,T. P. (1989). *Statistical Methods for Quality Improvement*, Wiley, New York.
40. Pearson, D. C. and Gawte, S. A. (1997). Long-term quality control of DXA: A comparison of Shewhart rules and Cusum charts. *Osteoporosis International* **7**(4): 338–343.
41. Lu, Y., Mathur, A. K., Gluer, C. C. *et al.* (1995). Application of statistical quality control method in multicenter osteoporosis clinical trials. *International Conference on Statistical Methods and Statistical Computation for Quality and Productivity Improvement*, Seoul, Korea, 474–480.
42. Wheeler, D. J. (1991). Shewhart's Chart: Myths, Facts, and Competitors. *45th Annual Quality Congress Transactions*: *American Society for Quality Control*, 533–538.
43. Johnson, R. A. and Bagshaw, M. (1974). The effect of serial correlation on the performance of CUSUM tests. *Technometrics* **16**: 103–112.
44. Kaminsky, F. C., Maleyeff, J., Providence, S., Purinton, E. and Waryasz, M. (1997). Using SPC (statistical process control) to analyze quality indicators in a healthcare organization. *Journal of Healthcare Risk Management* **17**(4): 14–22.
45. Quesenberry, C. P. (2000). Statistical process control geometric Q-chart for nosocomial infection surveillance. *American Journal of Infection Control* **28**(4): 314–20.
46. Thompson, J. R. and Koronachi, J. (1993). *Statistical Process Control for Quality Improvement*, Chapman and Hall, New York.
47. Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **i**: 307–310.
48. Bland, J. M. and Altman, D. G. (1995). Comparing two methods of clinical measurement: A personal history. *International Journal of Epidemiology* **24**(Suppl. 1): S7–S14.
49. Bartko. J. J. (1994). General methodology II measures of agreement: A single procedure. *Statistics in Medicine* **13**: 737–745.
50. Bradley, E. L. and Blackwood, L. G. (1989). Comparing paired data: A simultaneous test of means and variances. *The American Statistician* **43**: 234–235.
51. Lee, J., Koh, D. and Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in Biology and Medicine* **19**: 61–70.

52. Fleiss, J. L. and Shrout, P. E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika* **43**: 259–262.
53. Bland, J. M. and Altman, D. G. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurements. *Computers in Biology and Medicine* **20**(5): 337–340.
54. Freedman, L. S., Parmar, M. K. B. and Baker, S. G. (1993). The design of observer agreement studies with binary assessements. *Statistics in Medicine* **12**: 165–179.
55. Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**: 37–46.
56. Fleiss, J. L, Cohen, J. and Everitt, B. S. (1969). Large-sample standard errors of kappa and wieghted kappa. *Psychological Bulletin* **72**: 323–327.
57. Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics* **36**: 207–216.
58. Fleiss, J. L. and Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology* **115**: 841–845.
59. Garner, J. B. (1991). The standard error of Cohen's Kappa. *Statistics in Medicine* **10**: 767–775.
60. Blackman, NJ.-M. and Koval, J. J. (2000). Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* **19**: 723–741.
61. Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–174.
62. Block, D. A. and Kraemer, H. C. (1989). 2×2 kappa coefficients: Measures of agreement or association. *Biometrics* **45**: 269–287.
63. Cohen, J. (1968). Weighted kappa: Nomial scale agreement with provision for scaled disagreement or partial credit. *Psycological Bulletin* **70**(4): 213–219.
64. Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edn., Wiley, New York.
65. Barlow, W., Lai, M.-Y. and Azen, S. P. (1991). A comparison of methods for calculating a stratified kappa. *Statistics in Medicine* **10**: 1465–1472.
66. Donner, A. and Eliasziw, M. (1997). A hierachical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine* **16**: 1097–1106.
67. Kraemer, H. C. and Bloch, D. A. (1990). A note on case-control sampling to estimate kappa coefficients. *Biometrics* **46**(1): 49–59.
68. Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J. and Cheney, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine* **9**: 1103–1115.
69. Oden, N. L. (1991). Estimating kappa from binocular data. *Statistics in Medicine* **10**: 1303–1311.
70. Shoukri, M. M. and Martin, S. W. (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statistics in Medicine* **14**: 83–99.
71. Shoukri, M. M. and Mian, I. U. H. (1996). Maximum likelihood estimation of the kappa coefficient from bivariate logistic regression. *Statistics in Medicine* **15**: 1409–1419.

72. Agresti, A. (1992). Modelling patters of agreement and disagreement. *Statistical Methods in Medical Research* **1**: 201–218.
73. Tanner, M. A. and Young, M. A. (1985). Modelling agreement among raters. *Journal of American Statistical Association* **80**: 175–180.
74. Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* **46**: 293–302.
75. Graham. P. (1995). Modelling covariate effects in observer agreement studies: The case of nomial scale agreement. *Statistics in Medicine* **14**: 299–310.
76. Guggenmoos-Holzmann, I. and Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class prespective. *Statistics in Medicine* **17**: 797–812.
77. Baker, S. G., Freedman, L. S. and Parmar, M. K. B. (1991). Using replicate observations in observer agreement studies with binary assessments. *Biometrics* **47**(4): 1327–1338.
78. Cheng, C.-L. and Van Ness, J. W. (1999). *Statistical Regression with Measurement Error*, Arnold, London.
79. Barnett, D. V. (1969). Simultaneous pairwise linear structural relationships. *Biometrics* **28**: 129–142.
80. Theobald, C. M. and Mallinso, J. R. (1978). Comparative calibration, linear structural relationships and congeneric measurements. *Biometrics* **34**: 39–45.
81. Fuller, W. A. (1987). *Measurement Error Models*, Wiley, New York.
82. Dunn, G. (1989). *Design and Analysis of Reliability Studies*, Oxford University Press, New York.
83. Kimura, D. K. (1992). Functional comparative calibration using EM algorithm. *Biometrics* **48**: 1263–1271.
84. Lu, Y., Ye, K., Mathur, A., Hui, S., Fuerst, T. P. and Genant, H. K. (1997). Comparative calibration without a gold standard. *Statistics in Medicine* **16**: 1889–1905.
85. Lu, Y., Ye, K., Mathur, A. K., Srivastav, S. K., Yang, S. and Genant, H. K. (1997). Application of random effects models in comparative calibration. *Proceedings of the Biometrics Section of American Statistical Association*, 170–176.
86. Hanson, J. (1997). Standardization of femur BMD [letter]. *Journal of Bone and Mineral Research* **12**(8): 1316–1317.
87. Lu, Y., Fuerst, T., Hui, S. and Genant, H. K. (2001). Standardization of bone mineral density at femoral neck, trochanter and Ward's triangle. *Osteoporosis International* **12**: 438–444.
88. Hui, S. L., Gao, S., Zhou, X.-H. *et al.* (1997). Universal standardization of bone density measurements: A method with optimal properties for calibration among several instruments. *Journal of Bone and Mineral Research* **12**(9): 1463–1470.

About the Author

Ying Lu is an associate professor of Radiology at the Department of Radiology and the director of the Biostatistics Core, UCSF Comprehensive Cancer Center, University of California, San Francisco. He was an assistant professor of the same department (1994–1998) and an assistant professor of epidemiology and public health in the University of Miami School of Medicine (1990–1993). Dr. Lu, received his BS in mathematics from Fudan University (1982) and MS in applied mathematics from Shanghai Jiao Tong University (1984), and PhD in Biostatistics from the University of California, Berkeley (1990). At Berkeley, he received university fellowships (1985–1988), the Evelyn Fix Memorial Medal (1990), and Public Health Alumni Association Scholarship (1989). Dr. Lu has authored or co-authored more than 80 peer-reviewed articles and 4 book chapters in statistical methods for animal carcinogenicity experiments, medical diagnostic tests, and outcome prediction, as well as clinical research areas of radiology, osteoporosis, and cancer clinical trials. His paper has been published in *Biometrics*, *Statistics in Medicine*, *Mathematical Biosciences*, *Medical Decision Making*, *Radiology*, *Journal of Bone and Mineral Research*, *Cancer*, etc.

CHAPTER 5

COST-EFFECTIVENESS ANALYSIS AND EVIDENCE-BASED MEDICINE

JIANLI LI

Department of Corporate Performance, St Michael's Hospital, 30 Bond Street, Toronto, Ontario, M5B 1W8, Canada
Tel: 416-864-6060 ext 6152; lij@smh.toronto.on.ca

1. Introduction

Over the past decades, as pressures to control health care spending have accelerated, the term "cost-effectiveness" has become increasingly into common parlance. It is widely used by groups as disparate as the government, the congress, the business community, managed-care organizations, the pharmaceutical industry and the press.

The central purpose of cost-effectiveness analysis (CEA) is to compare the relative value of different interventions in creating better health and/or longer life. The results of such evaluations are typically summarized in a cost-effectiveness ratio, where the denominator reflects the gain in health from a candidate intervention (measured, for example, in term of years of life gained, premature birth averted, sight years gained, symptom-free days gained) and the numerator reflects the cost of obtaining the health gain. A cost-effectiveness analysis provides information that can help decision makers sort through alternatives and decide which one best serves their programmatic and financial needs. Decision maker may be federal, state or local. They may be in the private sector or the public sector. They may control dollars or they may run programs. CEA provides a framework within which decision makers may pose a range of questions.

Cost-effectiveness analyses furnish information that can be useful in a variety of settings. For example, a managed-care organization might wish to know the cost per low- birthweight birth averted as a consequence of a prenatal outreach program. Or it might wish to take the question further

and ask the cost of this program per year of life saved for its enrolled population. Or, recognizing that programs that avert premature births may not primarily save lives but rather avert disability over the lifetime of an individual, it might want to know the cost of this intervention for each quality-adjusted life year (QALY) gained. This latter question is addressed by a particular type of CEA, some times termed "cost utility analysis," where adjustments for the value assigned to health-related quality of life are built into the calculation.

As another example, a pharmaceutical manufacturer might wish to use CEA in pricing and marketing a new cholesterol-lowering drug. It might ask the question. How much does our medication cost per year of life gained compare to a similar product manufactured by a different company? Or, if the clinical trials show clinically insignificant changes in cholesterol level between the two products but significantly decreased side effects associated with the new drug, a drug purchaser or payer might wish then to calculate the cost per quality-adjusted life year (QALY) gained in using the new drug. An industry investigator might decide to extend the considerations of the analysis and explore the cost per year of life or QALY gained when comparing pharmaceutical treatment with surgical treatment for coronary disease.

Or, an analysis of a state health department might wish to explore different strategies for control of blood lead levels in the population. It might choose to assess the cost-effectiveness of screening all children, compared to screening only those thought to be at particular risk for elevated lead levels by reason of housing or environment surrounding.

1.1. *Worked examples*

1.1.1. *Bypass angioplasty revascularization investigation*

Percutaneous transluminal coronary angioplastry was introduced in 1977 as a less invasive alternative to coronary-artery bypass surgery. Several randomized clinical trials of angioplasty and bypass surgery have compared the clinical outcomes of these procedures. The Bypass Angioplasty Revascularization Investigation (BARI) was a large trial of angioplasty and bypass surgery in US, which collected five years of follow-up data.

Mark A. Hlatky *et al.*[8] conducted a study on a total 934 of the 1829 patients enrolled in the randomized BARI. Detailed data on quality of life were collected annually, and economic data were collected quarterly. They compared quality of life, employment, and medical care costs during

five year of follow-up among patients treated with angioplasty or bypass surgery. They found that on average, functional status, which was assessed by scores on the Duke Activity Status Index, was improved more with bypass surgery than with angioplasty in the first three years ($p < 0.05$), whereas in other respects the quality of life was equivalent with either method of revascularization. Patient in the angioplasty group returned to work five weeks sooner than did patients in the surgery group ($p < 0.001$). The cost of angioplasty was initially \$11,234 lower than that of bypass surgery (a 35% saving, $p < 0.001$), but higher subsequent costs for hospitalization and medication reduced the saving to \$2,644 at five years (a 5% savings, $p = 0.047$). The five-year cost of angioplasty was significantly lower than that of surgery among patients with two-vessel disease (\$52,930 versus \$58,498, $P < 0.05$), but not among patients with three-vessel disease. After five years of follow-up, surgery had an overall cost-effectiveness ratio of \$26,177 per year of life added, but unacceptable ratios of \$100,000 or more per year of life added could not be excluded ($P = 0.13$). Surgery appeared particularly cost effective in treating patients with diabetes because of their significantly improved survival.

1.1.2. *Treatment of high blood cholesterol*

In 1985, in response to the first evidence from a randomized controlled trial that reducing cholesterol reduces the risk of death from heart disease,[10] the US National Institutes of Health created the National Cholesterol Education Program (NCEP). Three years later the NCEP published guidelines for the management of high blood cholesterol which recommended that all adults have their cholesterol checked at least every 5 years and that those with high levels (240 mg/dl) or higher), or borderline-high levels (200–239 mg/dl) plus other risk factors, be tested further. It was suggested that those whose low-density lipoproteins (LDL) levels were also high should be treated by changes in diet or with cholesterol-lowering drugs.[11] It has been estimated that more than one-third of the adult population requires dietary change and/or drugs when judged by these criteria.[15]

Cost-effectiveness analyses done in the wake of the 1988 guideline focused on the management of high blood cholesterol once detected. Both lovastatin, a frequently prescribed drug, and dietary counseling were shown to vary widely in cost-effectiveness depending on age and other risk factors for heart disease.

One study examined the use of lovastatin for people initially free of heart disease and for those who had already suffered a heart attack.[7] The authors found that, for healthy people, saving a year of life in much more costly among those with cholesterol as their only risk factor than it is for those with several risk factors, even when cholesterol is very high; the cost ranged up to $330,000 for men aged 35–44 with no other risk factor and up to $1.5 million for women in the same category. The cost was considerably lower for people with other risk factors, reflecting the widely accepted assumption that risk factors interact to make the adverse effects of any one greater when others are present. Lovastatin treatment was still more costly per life year gained for people with levels in the range 250–299 mg/dl.

By contrast, the study found that it is potentially very cost-effective to treat people with elevated cholesterol who have had heart attacks. Costs per life year gained are relatively low and for some, such as men aged 35–44, drug treatment might save money as well as extended life. Another study found similar results for a program of intensive diet therapy modeled after the one in the Multiple Risk Factor Intervention Trial (MRFIT).[18] For example, diet therapy costs more than $500,000 per year life for 20-year-old men with initial cholesterol of 240 mg/dl and no other risk factors. For men with several risk factors, the cost per life year gained in much lower.

These results suggest that management of high cholesterol in people without heart disease is often very costly per life year saved. Since they show that treatment of people whose blood cholesterol levels are not far above 240 mg/dl can be extremely costly, they suggest that the same would be true for people with levels in the borderline-high range, although the studies did not analyze this group. Taken together, cost-effectiveness results suggest that resources might better be concentrated on those with very high cholesterol levels and/or other risk factors for heart disease (and on those in whom heart disease is already present). Revised guidelines, published by NCEP in 1993,[12] were somewhat more modest in their aims, in response to studies like these as well to ongoing debate over whether reducing cholesterol lengthens life in those without heart disease.

If NCEP's 1988 guidelines were followed to the letter, it would cost, depending on the effectiveness of diet in reducing blood cholesterol levels, $20 billion to $27 billion to provide lovastatin at dose of 20 mg per day, and $47 billion to $67 billion to provide a higher, more effective, dose of 80 mg per day.[6] The saving from a more selective strategy would be substantial, freeing resources to be applied elsewhere. The CEA results suggest that

more selective treatment strategies could be designed that would lose little in health benefits.

2. Foundations of Cost-Effectiveness Analysis

2.1. *What is cost-effectiveness analysis?*

Cost-effectiveness analysis is a method designed to assess the comparative impacts of expenditures on different health interventions. As Weinstein and Stason[19] state, it is based on the premise that "for any given level of resources available, society $\cdots$ wishes to maximize the total aggregate health benefits conferred." For example, we might wish to know whether spending a certain amount of money on a public campaign to stop smoking will have greater or lesser effect on health than spending the same amount on colorectal screening. Cost-effectiveness analysis can be in decision making at different levels, such as societal level and organizational level.

2.2. *The cost-effectiveness ratio*

The central measure used in CEA is the cost-effectiveness ratio. Implicit in the cost-effectiveness ratio is a comparison between alternatives. One alternative is the intervention under study, while the other is a suitably chosen alternative — "usual care," another intervention, or no intervention. The cost-effectiveness ratio for comparing the two alternatives at the population level can be the ratio of expected costs to expected effect (CER), $E(c)/E(e)$, and ratio of incremental expected costs to incremental expected effects (ICER), $(E(c_i) - E(c_j))/(E(e_i) - E(e_j)$ or $\Delta E(c)/\Delta E(e)$.

The ratio $\Delta E(c)/\Delta E(e)$ is essentially the incremental price of obtaining a unit health effect (such as dollars per year, or per quality-adjusted year, of life expectancy) from given health intervention when compared with an alternative.

The following situations can arise:

$\Delta E(c) < 0$, $\Delta E(e) > 0$; dominance; to accept the given intervention;
$\Delta E(c) > 0$, $\Delta E(e) < 0$; dominance; to reject the given intervention;
$\Delta E(c) > 0$, $\Delta E(e) > 0$; trade-off; consider magnitude of ratio of difference in costs to difference in effectiveness;
$\Delta E(c) < 0$, $\Delta E(e) < 0$; trade-off; consider magnitude of ratio of difference in costs to difference in effectiveness.

The expected ratio of cost to effect, $E(c/e)$, can be investigated at patient level.

2.3. *The effectiveness*

The effectiveness is the extent to which medical interventions achieve health improvements in real practice settings.

2.3.1. *Individual and social well-being*

By describing CEA as a tool for improving general welfare, it can be placed squarely within the context of welfare economics. The effectiveness measures could be quantified in term of utility, such as quality-adjusted life years (QALY); and in term of health status measures, such as the number of symptom-free days.

2.3.2. *A metric of health effect: Quality-adjusted life years*

It may appear that CEA cannot even be used to compare interventions whose effects on health are qualitatively different, such as prevention of coronary artery disease and treatment of arthritis. However, such a comparison is possible if the measure of effectiveness is general enough to capture all of the important health dimensions of the effects of the interventions. Using the quality-adjusted life year (QALY) as the unit of effectiveness approaches this ideal within the framework of CEA, thus expanding considerably the range of application of CEA. The QALY is a measure of health outcome which assign to each period of time a weight, ranging from 0 to 1, corresponding to the quality of life during that period, where a weight of 1 corresponds to perfect health and a weight of 0 corresponds to a health state judged equivalent to death. The number of quality-adjusted life years, then, represents the number of healthy years of life that valued equivalently to the actual health outcome.

2.3.3. *How to obtain evidence on effectiveness?*

The foundation for economic evaluation is valid data on the effectiveness of the intervention being evaluated relative to some alternative.

The true cost and effectiveness of an intervention usually are not known but estimated. The source of estimates may be direct measurement (sampling) or indirect (non-sampling) methods such as expert opinion and published literature. There could be two types of data; sampled data where the sampling variance may or may not be known, and non-sampled data such as discount rate for which do not have sampling variation, although

the true value of the parameter may be uncertain. These data can be used in various combinations in two models of analysis: stochastic analysis where inferences are drawn using standard statistical methods based on sampling variation, and deterministic analysis where inferences are drawn from point estimates of variables but interpretation is conditional upon the range of uncertainty from sensitivity analysis. The appropriateness of methods for analyzing uncertainty in costs or effects will depend upon the mix of sampled and non-sampled data. Cost-effectiveness analysis can be wholly deterministic, partially stochastic or wholly stochastic.

2.3.4. *Deterministic cost-effectiveness analysis*

This is used where cost and effect variables are analyzed as point estimates. Sampling variation may not be available because of the source of the data (e.g. secondary data) or the variable may not have been sampled (e.g. choice of discount rate, expert opinion). Deterministic CEA models arise frequently in the early assessment of a new medical technology, where only limited data are available but some analysis is required for policy setting. For example, in their analysis of the implantable defibrillator. Kupperman *et al.*[9] constructed a cost-effectiveness model where effect data were taken from reports of patient series in the literature as point estimates of survival probabilities and cost data were derived from a Medicare claims database and expert opinion. Given these data was not possible to present cost and effect differences with 95% confidence intervals, therefore a deterministic point estimate of cost-effectiveness was subject to detailed sensitivity analysis to explore the impact of uncertainty. Therefore a point estimate based on expert opinion of resource use was used as a proxy for variables that could be sampled in the future as part of a prospective study.

2.3.5. *Partially stochastic cost-effectiveness analysis*

This is used where effectiveness has been estimated from clinical trial(s) and can be expressed as a mean effect size with an associated variance, but analysis of costs is deterministic because data are non sampled. This combination is common in decision analytic models of economic appraisal. Some studies with such data report confidence intervals for cost-effectiveness where only variation in effects has been analyzed. For example a study in ulcer maintenance theory presented 95% confidence intervals around expected one-year therapy costs including relapse management. But no primary data had been collected to determine variation between patients in

costs of managing relapse. The source of variation for the confidence interval was only the surrounding the estimated incidence of relapse on treatment and control.

2.3.6. *Wholly stochastic cost-effectiveness analysis*

This is used where both costs and effects are determined from data sampled from the same patients in a study. Although our discussion focused on the randomized controlled trials (RCT) these data might also be measured by non-experiment-design. If cost and effect data are sampled and variances are available then formal statistical tests can be performed on observed differences in costs (treatment-control) or effects.

Randomized controlled trials (RCT) are one valuable source of evidence on effectiveness, used either as single studies or combined in a meta-anlysis. There are two general ways in which RCT data can be incorporated into economic evaluation: (i) combining RCT effectiveness data retrospectively with cost data from secondary non-trial sources into a decision analysis model; or (ii) collecting effectiveness and cost data on the same patients prospectively as part of an RCT.

The growing interest in trial-based prospective cost-effective studies has raised some interesting statistical questions of study design and analysis. Given the traditional use of non-sampled secondary data (e.g. published literature, insurance claims databases, expert opinion) in cost-effectiveness models the convention for analyzing uncertainty in results has been to use sensitivity analysis, where the robustness of results is explored over a range of what if alternative values for uncertain variables. This analytical approach is marked contrast to the conventional analysis of RCT effectiveness data where standard principle s of statistical inference are used to construct tests of hypotheses and estimate intervention effect sizes, and where uncertainty is quantified by a confidence interval which has precise meaning in terms of probability.

2.4. *Sensitivity analysis and beyond*

Before considering the adaptation of stochastic methods for economic evaluation, it is necessary to review the limitation of sensitivity analysis. This method is widely recommended for assessing problems of data uncertainty in economic appraisals of health care programs and allied evaluative techniques such as clinical decision analysis. The purpose is to examine the robustness of an estimated result over a range of alternative values for

uncertain parameters. Weinstein and Stason (1977) describe the method in the following way: "The most uncertain features and assumptions... are varied one at a time over a wide range of possible values. If the basic conclusions do not change when a particular feature or assumption is varied, confidence in the conclusions is increased."

Whereas the traditional CEA model utilize sensitivity analysis, the mean-variance data on costs and effects from a prospective trial presents the opportunity to analyze cost-effectiveness using conventional inferential statistical methods.[13] The statistical approach in CEA have been discussed by many literatures.

3. Statistical Approach

3.1. *Costs and effects as point estimates*

The deterministic analysis of effectiveness is a comparison of point estimates. If we consider a treatment that is both more costly and more effective than control, then a useful way to represent incremental cost-effectiveness is illustrated in Fig. 1. In this diagram, the x axis represents the difference in effects between the experimental and control therapy (Δe) and the y axis the difference in cost between experimental and control (Δc). The slope of the line extending from origin (the control) through our study point estimate, Δe, Δc, represents the incremental cost-effectiveness of the treatment relative to control. Clearly, the steeper the slope of the line $\Delta c/\Delta e$ the

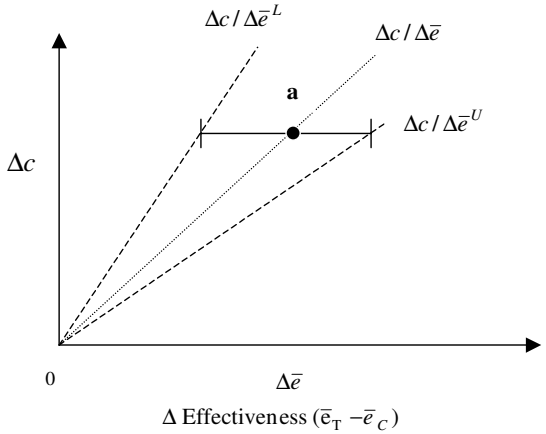


Fig. 1. Cost-effectiveness quasi-confidence interval: Deterministic analysis of cost differences and stochastic analysis of effectiveness differences.

greater is the additional cost at which additional units of effectiveness are gained by treatment relative to control, and the less attractive treatment becomes. In the absence of any data on sampling variation for costs or effects (point a) some form of sensitivity analysis would be useful to determine plausible ranges that may contain the true cost-effectiveness ratio.

3.2. *Sampled effectiveness and non-sampled costs*

In the analysis of sampled effect data (with sample variation) the null hypothesis is usually that there is no difference may come between experimental and control therapy. This is tested against either a one tailed alternative (usually that the experimental treatment more effective) or a two tailed alternative (that the experimental treatment is more or less effective than control). For a continuos clinical variable such as blood pressure we assume, by convention, that the ratio of the difference in sample means $(\bar{e}_T - \bar{e}_C)$, where the subscript T stands for the treatment group and subscript C stands for the control group, to the pooled standard error of the difference follows some known probability distribution such as Z or t. Critical values of the test statistics are determined by the analyst's judgement about the acceptable risk of making a Type 1 (false-positive) error about a difference existing, this level conventionally being set to 5%.

A problem with hypothesis testing as a form of stochastic analysis is that an overemphasis tends to be placed on the statistical significance. The advantages of the confidence interval is two-fold. First it permits hypothesis testing as described above because if a 95% confidence interval for a difference includes zero, then the treatment groups are not significantly different at 5% level. Second, in addition to statistical significance, the confidence intervals yields information on the magnitude of the observed difference (quantitative significance or clinical importance). The relationship between these two parameters is important because a difference can be highly statistically significantly but of no clinical importance, for example, a small difference (say, 0.25 mm/Hg) with $p < 0.0001$. Furthermore, the concept of a minimum clinically important difference δ to be detected is central to the design of a clinical experiment and determination of sample size.

A familiar two-tailed confidence interval for the treatment-effect size would be

$$(\bar{e}_T - \bar{e}_C) \pm t_{(n_T+n_C-2,1-\alpha/2)} \sqrt{\frac{S_{eT}^2}{n_T} + \frac{S_{eC}^2}{n_C}} \tag{1}$$

where S_{eT}^2 and S_{eC}^2 are the sample estimates of variances.

A confidence interval around (the mean effect size) has been drawn in Fig. 1. Given the confidence interval around $\Delta\bar{e}$, one approach to translating this into variation around the cost-effectiveness ratio is by creating an interval bounded by the ratio of cost difference to the lower bound of the effect interval ($\Delta_c/\Delta\bar{e}^L$) and the ratio of the cost difference to the effect upper bound of the effect interval ($\Delta_c/\Delta\bar{e}^U$). These upper and lower bounds for the cost-effectiveness ratio might be termed a quasi-confidence interval, because they are only based upon knowledge of sampling variation associated with the measurement of the denominator (effects). This reasoning can be applied analogously to a situation where we had stochastic costs but deterministic effects.

3.3. *Sampled effectiveness and sampled costs*

As we did in previous sections, we assumed that effects were measured from a trial and could be expressed as a confidence interval. However, we also assumed that resource use was measured to enable patient-specific costs to be estimated from j resources ($j = 1, \ldots, J$) in quantity Q_j at unit price P_j, then the costs for individual i can be expressed $c_i = \sum_{j=1}^{J} P_j Q_j$.

Summing over i patients ($i = 1, \ldots, n_T$) in the treatment group, mean cost per patient can be expressed as $\bar{c}_i = \frac{1}{n_T}\sum_{i=1}^{n_T} c_i$ with estimated variance

$$s^2_{EcT} = \frac{1}{n_T(n_T-1)}\sum_{i=1}^{n_T}(c_i - \bar{c}_T)^2 \,. \tag{2}$$

Therefore the difference between the mean cost associated with treatment and control can be expresses as a confidence interval:

$$(\bar{c}_T - \bar{c}_C) \pm t_{(n_T+n_C-2,1-\alpha/2)}\sqrt{\frac{S^2_{cT}}{n_T} + \frac{S^2_{cC}}{n_C}} \,. \tag{3}$$

In this situation the incremental cost-effectiveness ratio is a ratio of two random variables, both of which can be expressed as a confidence interval (around a difference in means). If we initially assume zero covariance between costs and effects then one can conceptualize this ratio in the form of a two-dimensional confidence plane.

3.4. *Joint distribution of cost and effects*

It is assumed that in an RCT (or observational study in which valid inference can made) there are J interventions where n_j patients receive

intervention j, $j = 1, 2, \ldots, J$. Costs and effects are viewed as vector random variables $\mathbf{c}_j$ and $\mathbf{e}_j - c_{ij}$ representing the costs incurred and e_{ij} the effects achieved by patient i on intervention j, $i = 1, 2, \ldots, n_j$, during a specified period. The joint probability distribution function of costs and effects on a patient level is modeled by the function $F_j(\mathbf{c}, \mathbf{e}; \mathbf{z})$. A vector of patient covariate, $\mathbf{z}$, such as diagnosis, gender and age, is introduced to cover the situation in which the cost-effect relationship of a intervention is expected to vary for different subgroups. It is assumed that $(c_{ij}(\mathbf{z}), e_{ij}(\mathbf{z}))$ are independently and identically distributed over the patients with covariates $\mathbf{z}$ receiving intervention j. The marginal distributions of F, which are the univariate distribution of cost and distribution of effect, are each associated with parameters such as expected cost $E(\mathbf{c})$, and expected effect, $E(\mathbf{e})$.

The expected cost and effect, $(E(\mathbf{c}), E(\mathbf{e}))$ could be estimated by the sample means of $\mathbf{c}$ and $\mathbf{e}$, that is, $(\bar{c}, \bar{e})$ and the covariance matrix of $(\bar{c}, \bar{e})$ could be presented as:

$$\begin{bmatrix} \dfrac{\widehat{\sigma}_c^2}{n} & \dfrac{\hat{\rho}\,\widehat{\sigma}_c\widehat{\sigma}_e}{n} \\ \dfrac{\widehat{\rho}\;\widehat{\sigma}_c\widehat{\sigma}_e}{n} & \dfrac{\widehat{\sigma}_e^2}{n} \end{bmatrix}, \tag{4}$$

where $\widehat{\sigma}_c$ and $\widehat{\sigma}_e$ are the estimated variances for cost and effect respectively and $\hat{\rho}$ is the estimated correlation coefficient between cost and effect.

The difference in expected cost and effect between two treatments/ interventions, $(\Delta E(\mathbf{c}), \Delta E(\mathbf{e}))$ could be estimated by the sample means of $\mathbf{c}$ and $\mathbf{e}$, that is, $(\bar{c}, \bar{e})$ and the covariance matrix of $(\Delta\bar{c}, \Delta\bar{e})$ could be expressed as:

$$\begin{bmatrix} \dfrac{\widehat{\sigma}_{ci}^2}{n_i} + \dfrac{\widehat{\sigma}_{cj}^2}{n_j} & \dfrac{\widehat{\rho}\,\widehat{\sigma}_{ci}\widehat{\sigma}_{ei}}{n_i} + \dfrac{\rho_j\widehat{\sigma}_{cj}\widehat{\sigma}_{ej}}{n_j} \\ \dfrac{\widehat{\rho}_i\widehat{\sigma}_{ci}\widehat{\sigma}_{ei}}{n_i} + \dfrac{\rho_j\widehat{\sigma}_{cj}\widehat{\sigma}_e}{n_j} & \dfrac{\widehat{\sigma}_{ei}^2}{n_i} + \dfrac{\widehat{\sigma}_{ej}^2}{n_j} \end{bmatrix}. \tag{5}$$

4. Statistical Inferences on Cost-Effectiveness Measures

4.1. *Parametric approaches to estimating the C-E ratio confidence interval*

4.1.1. *The confidence box approach*

A number of commentators advocated the cost-effectiveness plane (CE plane) for presenting the results of economic evaluation and for aiding

policy decision. O'Brien and colleagues[13] showed how the CE plane could be used to present the confidence limits for the estimate of incremental cost-effectiveness under the assumption of zero covariance between costs and effects. The difference in effect between two interventions is shown on the horizontal axis with mean effect difference $\Delta\bar{e}$ and upper and lower confidence limits for the effect difference $(\Delta\bar{e}^U, \Delta\bar{e}^L)$. Similarly, the difference in cost between two interventions is shown on the vertical axis with mean cost difference $\Delta\bar{c}$ and upper and lower confidence limits for the effect difference $(\Delta\bar{c}^U, \Delta\bar{c}^L)$. These "I" bars intersect at point $(\Delta\bar{e}, \Delta\bar{c})$, hence the ray that connects this point of intersection to the origin has a slope equal to the value of the ICER. Under the assumption described above, the center of the two confidence intervals intuitively can be thought of as the maximum likelihood of the two-dimensional probability density function. O'Brien and colleagues argue that combining the limits of the confidence intervals for costs and effects separately gives natural best and worst case limits on the ratio; that is, the upper limit of the cost difference over the lower limit of the effect difference $(\Delta\bar{c}^U/\Delta\bar{c}^L)$ gives the highest values of the ratio (worst case) and the lower limit of costs divided by the upper limit of effects $(\Delta\bar{e}^L/\Delta\bar{e}^U)$ gives the lowest (best) value of the ratio. Thus, in Fig. 1, the slope of the line from the origin through point a is a worst-case scenario for the incremental cost-effectiveness ratio based upon the upper 95% CI of the cost estimate and the lower 95% CI of the effect estimate. By similar reasoning, the line through point c is the best-case scenario. In contrast to Fig. 1, the slice of "pie" bounded from the origin by the best and worst cases scenarios has increased in size reflecting increased uncertainty about where the true cost-effectiveness ratio lies in this region.

There are two problems with this line of reasoning. The first is that the depiction of the two-dimensional confidence plane as being box-shaped is misleading. If costs and effects varied independently then the conditional probability of being at the lower 95% CI of both simultaneously would be less than 0.05. In principle we might expect such a bivariate probability density function to be elliptical in shape with lines of equi-probability central point-estimate (the maximum likelihood) much like an ordnance survey map of a mountain with height contours. Figure 3 illustrates how this general concept applies to the current problem. The second problem is the implicit assumption that costs and effects vary independently (i.e. have zero covariance). In principle we would expect covariance between costs and effects, and therefore we cannot assume that the numerator and denominator in the ratio are independent. This means that the bounds for the cost-effectiveness ratio depicted in Fig. 2 are still only a quasi-confidence interval

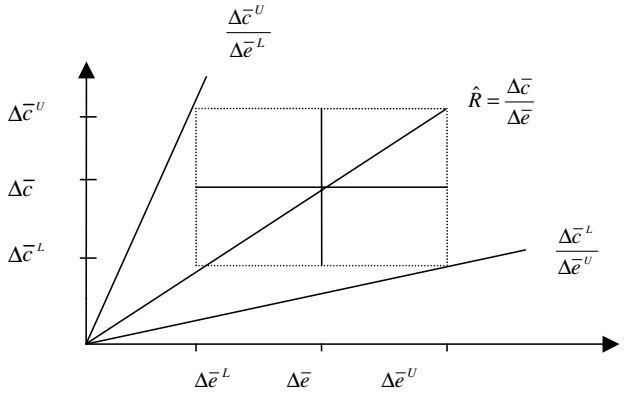


Fig. 2. Confidence limits on the cost-effectiveness plane and the "confidence box" approach to estimating confidence limits for the ICER.

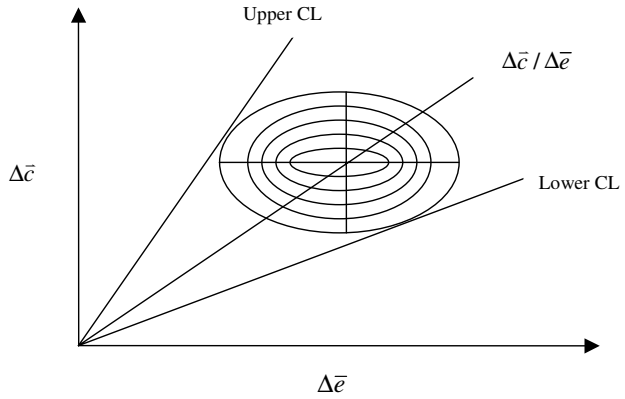


Fig. 3. Hypothetical probability density function around maximum likelihood point-estimate for cost-effectiveness.

because we have not taken account of all sampling variation. The challenge is whether a method exists for estimating the sampling distribution for the ratio of two random variables which may have nonzero covariance.

4.1.2. *The Taylor series approximation*

The Taylor approximation shows that where y is a function of two random variables x_1 and x_2, the variance of y can be expressed in term of the partial derivatives of y with respect to x_1 and x_2, weighted by the variances and

covariance of x_1 and x_2. The Taylor series formula is

$$\operatorname{var}(y) \approx \left(\frac{\partial y}{\partial x_1}\right)^2 \operatorname{var}(x_1) + \left(\frac{\partial y}{\partial x_2}\right)^2 \operatorname{var}(x_2) + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right) \operatorname{cov}(x_1, x_2)\,. \tag{6}$$

For the ICER $\Delta E(c)/\Delta E(e)$, using the sample estimates of the means and variance, the variance of the ratio estimator can be given as follows:

$$\operatorname{var}(\hat{R}) \approx \frac{1}{\Delta\bar{e}^2} \operatorname{var}(\Delta\bar{c}) + \frac{\Delta\bar{c}^2}{\Delta\bar{e}^4} \operatorname{var}(\Delta\bar{e}) - 2\frac{\Delta\bar{c}}{\Delta\bar{e}^3} \operatorname{cov}(\Delta\bar{c}, \Delta\bar{e})\,. \tag{7}$$

Since the variance of difference in mean is equal to the sum of two sampling variances for those means, then we can simplify

$$\operatorname{var}(\Delta\bar{c}) = \frac{\hat{\sigma}_{c1}^2}{n_1} + \frac{\hat{\sigma}_{c2}^2}{n_2}\,, \quad \operatorname{var}(\Delta\bar{e}) = \frac{\hat{\sigma}_{e2}^2}{n_1} + \frac{\hat{\sigma}_{e2}^2}{n_2}\,, \tag{8}$$

and the covariance term can also be simplified

$$\operatorname{cov}(\Delta\bar{c}, \Delta\bar{e}) = \frac{\widehat{\operatorname{cov}}(c_1, e_1)}{n_1} + \frac{\widehat{\operatorname{cov}}(c_2, e_2)}{n_2} = \frac{\hat{\rho}_1\hat{\sigma}_{c1}\hat{\sigma}_{e1}}{n_1} + \frac{\hat{\rho}_2\hat{\sigma}_{c2}\hat{\sigma}_{e2}}{n_2}\,. \tag{9}$$

Combining these elements gives our expression for the variance of ratio

$$\operatorname{var}(\hat{R}) \approx \left[\frac{\frac{\hat{\sigma}_{c1}^2}{n_1} + \frac{\hat{\sigma}_{c2}^2}{n_2}}{\Delta\bar{e}^2}\right] + \Delta\bar{c}^2\left[\frac{\frac{\hat{\sigma}_{e1}^2}{n_1} + \frac{\hat{\sigma}_{e2}^2}{n_2}}{\Delta\bar{e}^4}\right] - 2\Delta\bar{c}\left[\frac{\frac{\hat{\rho}_1\,\hat{\sigma}_{c1}\,\hat{\sigma}_{e1}}{n_1} + \frac{\hat{\rho}_2\,\hat{\sigma}_{c2}\,\hat{\sigma}_{e2}}{n_2}}{\Delta\bar{e}^3}\right]. \tag{10}$$

Factoring $\hat{R}^2 = \Delta\bar{c}^2/\Delta\bar{e}^2$ from the right-hand side simplifies (7) to

$$\operatorname{var}(\hat{R}) \approx \hat{R}^2[(cv(\Delta\bar{c}))^2 + (cv(\Delta\bar{e}))^2 - 2\,\widehat{\rho}\,cv(\Delta\bar{c})cv(\Delta\bar{e})]\,, \tag{11}$$

where $cv(x)$ is the coefficient of variation for the random variable x and defined as $cv(x) = \sqrt{\operatorname{var}(x)}/\bar{x}$, and ρ_{xy} is the correlation coefficient between two random variables x and y and defined as $\rho_{xy} = \operatorname{cov}(x, y)/\sqrt{\operatorname{var}(x)\operatorname{var}(y)}$. The properties of this variance are intuitively appealing: the cost-effectiveness variance will increase with a greater difference in costs or effects, with a greater population mean costs between groups and with greater negative correlation between costs and effects. Conversely the ratio variance will decrease with greater sample size, with a greater

difference in population mean effects between groups and a greater positive correlation between costs and effects.

The accuracy of the approximation in the equation above depends upon the random variables, $\Delta\bar{c}$ and $\Delta\bar{e}$, having small coefficients of variation. The coefficient of variation for each random variable is $(Z_{\alpha/2} + Z_{\beta})^{-1}$, where the two-sided level test α of significance has $1 - \beta$ power against the true difference. For even a 50% power against the true difference the coefficient of variation would be $(1.96)^{-1} = 0.51$; small enough to ensure reasonable accuracy. The accuracy of the approximation begins to fail as the difference between treatments, with respect to cost or effect, approaches zero so that the power falls well below 50%.

Similarly, for the ratio $E(c)/E(e)$, we have

$$\mathrm{var}(\hat{R}) \approx \frac{1}{\bar{e}^2}\mathrm{var}(\bar{c}) + \frac{\bar{c}^2}{\bar{e}^4}\mathrm{var}(\bar{e}) - 2\frac{\bar{c}}{\bar{e}^3}\mathrm{cov}(\bar{c},\bar{e})\,, \tag{12}$$

$$\mathrm{var}(\hat{R}) \approx \left[\frac{\widehat{\sigma}_c^2}{n}\Big/\bar{e}^2\right] + \bar{c}^2\left[\frac{\widehat{\sigma}_e^2}{n}\Big/\bar{e}^4\right] - 2\bar{c}\left[\frac{\widehat{\rho}\,\widehat{\sigma}_c\widehat{\sigma}_e}{n}\Big/\bar{e}^3\right], \tag{13}$$

$$\mathrm{var}(\hat{R}) \approx \hat{R}^2[(cv(c))^2 + (cv(e))^2 - 2\,\widehat{\rho}\,cv(c)cv(e)]\,. \tag{14}$$

Employing standard parametric assumptions gives the confidence interval as

$$\left(\hat{R} - z_{\alpha/2}\sqrt{\mathrm{var}(\hat{R})},\ \hat{R} + z_{\alpha/2}\sqrt{\mathrm{var}(\hat{R})}\right). \tag{15}$$

Knowledge of the variance of R would also enable some tests of hypotheses. For example, suppose we specified some *a priori* upper threshold for the cost-effectiveness ratio, $R_{\max}$, which was the maximum cost per unit effect that we would be willing to pay for this new treatment. Hence $R_{\max}$ would be the maximum acceptable slope of the cost-effectiveness line through the origin in Fig. 2. We might set up a one-tailed test of the hypothesis that the true ratio, R, was less than this maximum. Thus, we have a null hypothesis, $H_0 : R = R_{\max}$ which is to be tested against an alternative $H_A : R < R_{\max}$ and using our variance we might construct a test statistic of the general form:

$$Z = \widehat{R} - R_{\max}\sqrt{\mathrm{var}(\widehat{R})}\,.$$

In illustrating the possible use of $\mathrm{var}(\widehat{R})$ in estimation and hypothesis testing we have assumed that the distribution for $\widehat{R}$ will be statistically well-behaved such that some parametric distribution (e.g. normal) might

be used in the large sample case. Although this is ultimately an empirical issue it seems a questionable assumption. For example, the distribution of a ratio of two differences may not be unimodal. While a non-parametric analogue of the approach might be developed using rank-order statistics a more practical alternative might be to generate an empirical distribution for $\widehat{R}$ by non-parametric bootstrapping.

4.1.3. *Fieller's method*

An alternative method of calculating confidence intervals around ratios has been described by Fieller.[5]

The advantage of Filler's method over the Taylor series expansion is that it takes into account the skew of the ratio estimator. The method assumes that the numerator and denominator of the ratio follow a joint normal distribution such that (in the case of the ICER) $\Delta\bar{c} - R\Delta\bar{e}$ is normally distributed. Hence, dividing through by the standard deviation equation follows the standard normal distribution:

$$\frac{\Delta\bar{c} - R\Delta\bar{e}}{\sqrt{\{\mathrm{var}(\Delta\bar{c}) + R^2\,\mathrm{var}(\Delta\bar{e}) - 2R\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}}} \sim N(0,1)\,. \tag{16}$$

Setting this expression equal to $z_{\alpha/2}$ and rearranging gives the following quadratic equation in R:

$$\begin{aligned}\widehat{R}\,[1 - z^2_{\alpha/2}(cv(\Delta\bar{e}))^2] - 2R\,\hat{R}[1 - z^2_{\alpha/2}\,\rho cv(\Delta\bar{e})cv(\Delta\bar{c})] \\ + \hat{R}^2[1 - z^2_{\alpha/2}\,cv(\Delta\bar{c})] = 0\,,\end{aligned} \tag{17}$$

$$\begin{aligned}&\hat{R}\left[\frac{1 - z^2_{\alpha/2}\,\rho cv(\Delta\bar{c})cv(\Delta\bar{e})}{1 - z^2_{\alpha/2}[cv(\Delta\bar{e})]^2}\right] \\ &\pm z^2_{\alpha/2}\hat{R}\left[\frac{\sqrt{[cv(\Delta\bar{c})]^2 + [cv(\Delta\bar{e})^2] - 2\rho cv(\Delta\bar{c})cv(\Delta\bar{e})} \\ \quad - z^2_{\alpha/2}\{[cv(\Delta\bar{c})]^2[cv(\Delta\bar{e})^2] - \rho^2[cv(\Delta\bar{c})]^2[cv(\Delta\bar{e})]^2\}}{1 - z^2_{\alpha/2}[cv(\Delta\bar{e})]^2}\right].\end{aligned} \tag{18}$$

Similarly, for the ratio $E(c)/E(e)$, we have

$$\frac{\bar{c} - R\bar{e}}{\sqrt{\{\mathrm{var}(\bar{c}) + R^2\,\mathrm{var}(\bar{e}) - 2R\,\mathrm{cov}(\bar{c}, \bar{e})\}}} \sim N(0,1)\,, \tag{19}$$

$$\widehat{R}[1 - z_{\alpha/2}^2(cv(\bar{e}))^2] - 2R\,\hat{R}[1 - z_{\alpha/2}^2\,\rho cv(\bar{e})cv(\bar{c})] + \hat{R}^2[1 - z_{\alpha/2}^2 cv(\bar{c})]\,, \tag{20}$$

$$\hat{R}\left[\frac{1 - z_{\alpha/2}^2\rho cv(\bar{c})cv(\bar{e})}{1 - z_{\alpha/2}^2[cv(\bar{e})]^2}\right]$$

$$\pm z_{\alpha/2}^2\,\hat{R}\left[\frac{\sqrt{[cv(\bar{c})]^2 + [cv(\bar{e})^2] - 2\rho cv(\bar{c})cv(\bar{e})} \\ - z_{\alpha/2}^2\{[cv(\bar{c})]^2[cv(\bar{e})^2] - \rho^2[cv(\bar{c})]^2[cv(\bar{e})]^2\}}{1 - z_{\alpha/2}^2[cv(\bar{e})]^2}\right]. \tag{21}$$

Siegel *et al.*[16] proposed that $\tau = \bar{c} - R\bar{e}$ is normally distributed with mean $E\tau = 0$ and $\mathrm{var}(\tau) = (\mathrm{var}(c) - 2R\,\mathrm{cov}(c,e) + R^2\,\mathrm{var}(e))/n$. Let $F_{1,n-1}$ denote the 95th percentile of an F distribution with 1 and $(n-1)$ degrees of freedom. The probability that

$$\tau^2/(\widehat{\mathrm{var}}\,(c) - 2R\,\widehat{\mathrm{cov}}(c,e) + R^2\,\widehat{\mathrm{var}}\,(e)) < F_{1,n-1}(n-1)^{-1}$$

is 0.05 since the random variable of the left side of the inequality is distributed as an F distribution with 1 and $(n-1)$ degree of freedom. Multiplying both sides by the denominator and subtracting the right hand side from both sides of the inequality yields

$$(\bar{c}^2 - F_{1,n-1}(n-1)^{-1}\ \widehat{\mathrm{var}}\,(c)) - 2R(\bar{c}\bar{e} - F_{1,n-1}(n-1)^{-1}\ \widehat{\mathrm{cov}}(c,e))$$
$$+ R^2(\bar{e}^2 - F_{1,n-1}(n-1)^{-1}\ \widehat{\mathrm{var}}\,(e)) \le 0\,. \tag{22}$$

The set of values of R satisfying this inequality is a 95% confidence interval for the ratio $E(c)/E(e)$.

4.1.4. *Confidence interval for the expected cost to effect ratio $E(c/e)$*

Under an assumption of asymptotic normality, the expected value of the ratio $E(c/e)$ does not exist because ratios of normal random variables follow the Cauchy distribution. Therefore, in this case neither an estimator nor a confidence interval makes sense. The approximate distribution function of the random variable c/e, $F(y) = P(c/e < y\sigma_e/\sigma_c)$ is given by

$$\Phi((wE(e)/\sigma_e - E(e)/\sigma_c)(w^2 - 2\rho w + 1)^{-1/2}) \tag{23}$$

where $\Phi(\cdot)$ is the cumulative normal distribution with mean 0 and variance 1. Here, $w = y\sigma_e/\sigma_c$ where σ_e and σ_c are the population standard deviations of e and c respectively and ρ is the correlation between them. The median of

this distribution is $E(c)/E(e)$. Thus, the ratio of expected costs to expected effects is the median of the distribution of the distribution of the patient level ratio of costs to effects. A 95% confidence interval for the median of this distribution may be obtained by applying the method based on Fieller's theorem.

For some data, rather than assuming that the distribution of F is multivariate normal, it may be more appropriate to assume that the distribution has a form for which $E(c/e)$ does exist. For example, under an assumption of asymptotic normality of the ratio, the sample mean of $\overline{c/e}$ and sample variance of the ratio $\widehat{\sigma}^2_{c/e}$ can be used to form a 95% confidence interval for the mean cost-weight ratio as follows:

$$\overline{c/e} + t_{n-1}\,\widehat{\sigma}_{c/e}\sqrt{n}\,, \tag{24}$$

where n is the number of patients.

4.2. *Bootstrap approaches to estimating the C-E ratio*

The bootstrap approach for the simple one sample case is straightforward. Suppose a particular population has a real but unobserved probability distribution F from which a random sample x of n observations is taken, and the statistic of interest $s(x)$ is calculated the concern of inferential statistics is to make statements about the population parameter θ based on the sample drawn from that population. In the "bootstrap world," the observed random sample x is treated as the empirical estimate of F by weighting observation in x by the probability $1/n$. Successive random samples of size n are then draw from x with replacement to give the bootstrap samples (re-sample from the original sample). The statistic of interest is calculated for each of these samples and these bootstrap replicates of the original statistic make up the empirical estimate of the sampling distribution for that statistic. This estimated sampling distribution can be used in a variety of ways to construct confidence intervals.

In principle, the bootstrap estimate of the ICER sampling distribution can be obtained in very similar way to that of the simple one sample case. How ever, since the ICER is estimated on the basis of four estimators from two samples care must be taken to bootstrap each sample appropriately. For data structures which are more complicated than a one sample structure. Efron and Tibshirani[4] advocate that the bootstrap mechanism for the observed data mirror the mechanism by which those original data were obtained. In the case of the ICER, where data on resource use and

outcome exists for two groups of patients of size n_i and n_j receiving treatments/interventions T_i and T_j, respectively this will involve a three-stage process:

(1) Sample with replacement n_i cost/effect pair from the sample of patients who received treatment T_i and calculate the bootstrap estimates $\bar{c}_i^*$ and $\bar{e}_i^*$ for the bootstrap sample.
(2) Sample with replacement n_j cost/effect pair from the sample of patients who received treatment T_j and calculate the bootstrap estimates $\bar{c}_j^*$ and $\bar{e}_j^*$ for the bootstrap sample.
(3) Calculate the bootstrap replicate of the ICER given by the equation

$$R^* = \frac{\bar{c}_i^* - \bar{c}_j^*}{\bar{e}_i^* - \bar{e}_j^*} = \frac{\Delta\bar{c}^*}{\Delta\bar{e}^*}\,. \tag{25}$$

Repeating this three-stage process many times gives a vector of bootstrap estimates, which is an empirical estimate of the sampling distribution of the ICER statistic.

Once the sampling distribution of the ICER has been estimated in this way, several approaches exit to estimate confidence limits using the bootstrap estimate of the sampling.

4.2.1. *Normal approximation*

One method for confidence interval estimation is to take the bootstrap estimate of standard error, given by

$$\hat{\delta}^* = \sqrt{\left\{\frac{1}{B-1}\sum_{b=1}^{B}(\bar{R}^* - \bar{R}^{*b})^2\right\}}\,, \tag{26}$$

(where B is the total number of bootstrap replications) and assume that the sampling distribution is normal. The resulting $100(1-\alpha)$ per cent confidence interval is

$$(\hat{R} - z_{\alpha/2}\hat{\delta}^*,\ \hat{R} + z_{\alpha/2}\hat{\delta}^*)\,. \tag{27}$$

4.2.2. *Percentile*

The percentile method avoids the problem by making direct use of the empirical sampling distribution. The $100(\alpha/2)$ and $100(1-\alpha/2)$ percentile values of the bootstrap sampling distribution estimate are used as the upper and lower confidence limits for the ICER. The attraction of this method

is its simplicity and its avoidance of the assumption of normality for the ICER. However, skewed estimation can cause trouble for the percentile method. In particular, in this context, the percentile method assumes that the bootstrap replicates of the ICER are unbiased, whereas it is known that ratio estimators are biased and that bootstrap replicates will magnify the bias of the sample estimate.[17]

4.2.3. *Bias-corrected and accelerated*

Efron[3] suggests a modification of the percentile method, which seeks to adjust for the bias and skew of the sampling distribution. This is the bias-corrected and accelerated (BCa) percentile method, which involves algebraic adjustments to the percentiles selected to serve as the confidence interval end points. The adjusted percentiles are given by

$$\begin{aligned} \alpha_1 &= \Phi\left(\hat{z} + \frac{\hat{z} + z_{\alpha/2}}{1 - \hat{a}(\hat{z} + z_{\alpha/2})}\right), \\ \alpha_2 &= \Phi\left(\hat{z} + \frac{\hat{z} + z_{(1-\alpha/2)}}{1 - \hat{a}(\hat{z} + z_{(1-\alpha/2)})}\right), \end{aligned} \tag{28}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and z_α is the 100α percentile point of standard normal distribution. Two adjustments to the percentiles are incorporated into Eq. (28): $\hat{z}$ adjusts the sampling distribution for the bias of the estimator, while $\hat{a}$ adjusts for the skew of the sampling distribution. Setting $\hat{a} = 0$ yields the adjustment for bias on the percentile chosen to serve as end points, and is equivalent to the bias-corrected method advocated by Chaudhary and Stearns[1]:

$$\begin{aligned} \alpha_1 &= \Phi(2\hat{z} + z_{\alpha/2}), \\ \alpha_2 &= \Phi(2\hat{z} + z_{(1-\alpha/2)}). \end{aligned} \tag{29}$$

The bias correction, $\hat{z}$, is given by $\hat{z} = \Phi^{-1}(Q)$ where Q is the proportion of bootstrap replicates which are less than the sample estimate, $\hat{R}$. Therefore, if the bootstrap sampling distribution has median $\hat{R}$, $Q = 0.5$ which gives $\hat{z} = 0$ and (in the absence of a skew adjustment) the percentiles from Eq. (29) correspond to those from the straightforward percentile method. However, where the sampling distribution is not centered on $\hat{R}$ a correction is made for this bias. Notice that the nonlinear relationship between the z-score and its probability results in the percentile end points being shifted at unequal rates. It is also worth nothing that the bias correction adjustment of BCa method, while not employing distributional assumptions

concerning the distribution of the ICER itself, does make use of parametric assumptions concerning the distribution of the observed bias. This reliance on parametric assumptions has been cited as a potential weakness of the BCa method (29).

The acceleration constant adjusts for the skew of the sampling distribution. Efron and Tibshirani[4] suggest using a jack-knife estimate for $\hat{\alpha}$:

$$\hat{\alpha}^{**} = \frac{\sum_{i=1}^{n}(\bar{R}^{**} - \hat{R}_i^{**})^3}{6[\sum_{i=1}^{n}(\bar{R}^{**} - \hat{R}_i^{**})^2]^{3/2}}, \tag{30}$$

where $\hat{R}_i^{**}$ is the jack-knife replicate of the ICER with the ith observation removed, $\bar{R}^{*}* = \sum \hat{R}_i^{**}/n$ for $i = 1$ to n and $n = n_t + n_c$. In terms of the adjustments to the percentiles given in Eq. (28). In the absence of a bias correction adjustment, the skew adjustment is given by

$$\begin{aligned} \alpha_1 &= \Phi\left(\frac{z_{\alpha/2}}{1 - \hat{a}z_{\alpha/2}}\right), \\ \alpha_2 &= \Phi\left(\frac{z_{(1-\alpha/2)}}{1 - \hat{a}z_{(1-\alpha/2)}}\right). \end{aligned} \tag{31}$$

Equation (30) shows that if the sampling distribution is symmetric, $\hat{a} = 0$ and Eq. (31) shows that no adjustment to the percentile interval end points is made.

4.2.4. *Parametric bootstrap*

Efron and Tibshirani[4] outline a simulation-based method of confidence interval estimation that they refer to as a parametric bootstrap approach. Notice that from the definition of ICER, the difference in cost on the numerator and the difference in effects on the denominator of the ICER are both simply the difference between two normally distributed. The parametric bootstrap approach involves using this property of the distribution of the numerator and denominator in combination with the observe means, variance and covariance to estimate the parameters of the sampling distribution of the cost and effect differences. Sampling from each of these two distributions, while allowing for the estimated covariance between them, gives an estimate of the ICER. Repeating this process many times generates an empirical estimate of the sampling distribution of the ICER. The $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of this estimated distribution are used as estimstes for the upper and lower limits of the confidence interval, as with the percentile method.

5. Testing Difference Among the Populations

5.1. *Under assumption of normality of distribution*

5.1.1. *Testing on ICER*

Let R_0 be a specified value of the incremental cost-effectiveness ratio (ICER) R. It may be viewed as the maximum amount society is willing to pay to gain one unit of effectiveness by adopting the test intervention over the reference. We consider three tests of hypotheses on R:

(a) $H_0 : R = R_0$ verse $H_A : R \neq R_0$;
(b) $H_0 : R \geq R_0$ verse $H_A : R < R_0$;
(c) $H_0 : \Delta E(e) \geq 0$ or $R \geq R_0$ verse $H_A : \Delta E(e) > 0$ and $R < R_0$.

In (b), rejection of the null hypothesis might be interpreted to mean that the test intervention is cost-effective, in the sense that the data supports a CER below the stipulated maximum R_0. Its two-tailed version, (a) tests whether the data are consistent with a specified value R_0 of the ICER. In (c), we test the joint hypothesis on effectiveness and cost effectiveness. If the null hypothesis is tenable, the test intervention is either not effective or not cost-effective. If the alternative is true, then the test intervention is both effective and cost-effective, relative to the referent intervention. The covariance matrix of $(\Delta\bar{c}, \Delta\bar{e})'$, Σ could be represented as follows

$$\Sigma = \begin{bmatrix} \widehat{\sigma}_c^2 & \widehat{\rho}\,\widehat{\sigma}_c\widehat{\sigma}_e \\ \widehat{\rho}\,\widehat{\sigma}_c\widehat{\sigma}\,\widehat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \dfrac{\widehat{\sigma}_{c0}^2}{n_i} + \dfrac{\widehat{\sigma}_{c1}^2}{n_j} & \dfrac{\widehat{\rho}_i\widehat{\sigma}_{ci}\widehat{\sigma}_{ei}}{n_i} + \dfrac{\widehat{\rho}_j\widehat{\sigma}_{cj}\widehat{\sigma}_{ej}}{n_j} \\ \dfrac{\widehat{\rho}_i\widehat{\sigma}_{ci}\widehat{\sigma}_{ei}}{n_i} + \dfrac{\widehat{\rho}_j\widehat{\sigma}_{cj}\widehat{\sigma}_{ei}}{n_j} & \dfrac{\widehat{\sigma}_{ei}^2}{n_i} + \dfrac{\widehat{\sigma}_{e1}^2}{n_j} \end{bmatrix}. \tag{32}$$

Test of $H_0 : R = R_0$ verse $H_A : R \neq R_0$.
We formulate our test in terms of the estimated net cost $\Delta\bar{c} - R_0\Delta\bar{e}$. Under H_0, the statistic

$$T = (\Delta\bar{c} - R_0\Delta\bar{e})/\{\mathrm{var}(\Delta\bar{c} - R_0\Delta\bar{e})\}^{1/2}$$

or

$$T = (\Delta\bar{c} - R_0\Delta\bar{e})/\{\mathrm{var}(\Delta\bar{c}) + R_0^2\ \mathrm{var}(\Delta\bar{e}) - 2R_0^2\ \mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}^{1/2}$$

has an approximate standard normal distribution. The test rejects H_0 if $|T| > z_{(1-\alpha/2)}$ where $z_{(1-\alpha/2)}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution.

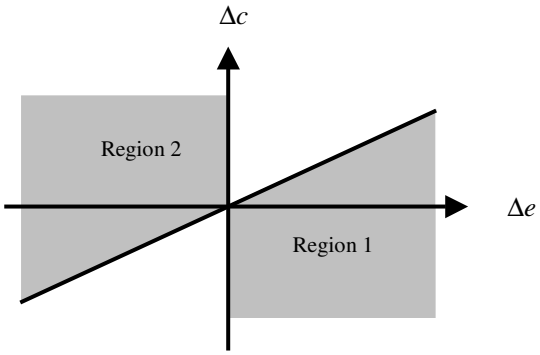


Fig. 4. Regions for one-sided test of effectiveness and cost-effectiveness. Region 1: test intervention both effective and cost-effective. Region 2: referent intervention effective and cost-effection.

Test of $H_0 : R \geq R_0$ verse $H_A : R < R_0$
We would reject $H_0 : R \geq R_0$ if $\Delta\bar{c} - R_0\Delta\bar{e} < -z_{(1-\alpha/2)}\{\text{var}(\Delta\bar{c}) + R_0^2 \text{ var}(\Delta\bar{e}) - 2R_0^2 \text{ cov}(\Delta\bar{c}, \Delta\bar{e})\}_{1/2}$
Test of $H_0 : \Delta E(e) \leq 0$ or $R \geq R_0$ verse $H_A : \Delta E(e) > 0$ and $R < R_0$.

In Fig. 4, the lower shaded region (region 1) in the C_E plane is where H_A holds. The complementary shaded region (region 2) in the second and third quadrants is where the referent intervention is both effective and cost-effective. Our one-sided test impose asymmetry between the test and the referent interventions, and region 1 is the appropriate rejection region for our test.

Based on our previous discussion, an appropriate test would reject H_0 if $\Delta\bar{e} > c_1$ and $(\Delta\bar{c} - R_0\Delta\bar{e}) < c_2$ where the constant $c_1 > 0$ and $c_2 < 0$ need to be specified. The size of the test is

$$\alpha = \sup\, P[\Delta\bar{e} > c_1,\ (\Delta\bar{c} - R_0\Delta\bar{e}) < c_2]\,,$$

where the supremum is taken over all $(\Delta\bar{e}, \Delta\bar{c})$ consistent with H_0. By normalization, we may express this in terms of the bivariate normal (Z_1, Z_2), with zero means, unit variances and correlation $-\rho^*$. Then

$$\alpha = \sup\, P\left[Z_1 < \frac{-(c_1 - \Delta\bar{e})}{\sqrt{\text{var}(\Delta\bar{e})}}\,,\right.$$

$$\left. z_2 < \frac{c_2 - (\Delta\bar{c} - R_0\Delta\bar{e})}{\{\text{var}(\Delta\bar{c}) + R_0^2 \text{ var}(\Delta\bar{e}) - 2R_0^2 \text{ cov}(\Delta\bar{c}, \Delta\bar{e})\}^{1/2}}\right]$$

$$= \max\left\{ P\left[Z_1 < \frac{-c_1}{\sqrt{\{\mathrm{var}(\Delta\bar{e})\}}} \right], \right.$$

$$\left. P\left[Z_2 < \frac{c_2}{\{\mathrm{var}(\Delta\bar{c}) + R_0^2\ \mathrm{var}(\Delta\bar{e}) - 2R_0^2\ \mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}^{1/2}} \right] \right\}. \qquad (33)$$

One solution to (33) is

$$c_1 = \sigma_1 z_{1-\alpha}\,,$$

$$c_2 = -\{\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) - 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}^{1/2} z_{1-\alpha}\,. \qquad (34)$$

5.1.2. *Testing on CER (cost-effectiveness ratio)*

If the cost-weight bivariate distributions are normal with mean vectors $(E(c_i), E(e_i))$ and common covariance matrix, the multivariate analysis of variance, MANOVA, can be used to test the hypothesis that the vectors of cost-efficiency measures are identical. If the MANOVA finds the means of the distributions of the populations to be equal and the $c - e$ measure is a function of the means, e.g. $E(c_i)/E(e_i)$, then it may be concluded that the $c - e$ measures do not differ.

A likelihood ratio test could be employed to test the hypothesis H_0 : $E(c_i)/E(e_i) = R_0$ for all i, that is, $E(c_i) - R_0 E(e_i) = 0$.

An asymptotic α-level two sided test of H_0 may be obtained by first using likelihood theory for normal variables for testing the linear hypothesis that all ratios are equal to a specific value, say, R_0. The desired likelihood ratio test is found by maximizing the previous likelihood over all possible values of R_0.

Let n_i denote the number of bivariate observations of cost and effect for treatment i and let $n = \Sigma n_i$. The available data consists of the bivariate observations (c_{ij}, e_{ij}), $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, n_i$. Let $s_{11} = \Sigma_i \Sigma_j (c_{ij} - \bar{c}_i)^2/n$, $s_{22} = \Sigma_i \Sigma_j (e_{ij} - \bar{e}_i)^2/n$, and $s_{12} = \Sigma_i \Sigma_j (c_{ij} - \bar{c}_i)(e_{ij} - \bar{e}_i)/n$. Here, s_{ij} are the elements of the pooled covariance matrix, S. The hypothesis $E(c_i)/E(e_i) = R_0$ is equivalent to the hypothesis $E(c_i) - R_0 E(e_i) = 0$ for all i. For a specific R_0 the classical test of the latter linear hypothesis is based on the Wilks' statistic, $W(R_0)$. The likelihood ratio statistic for the same linear hypothesis is given by $\Lambda(R_0) = W(R_0)^{n/2}$. Maximizing $\Lambda(R_0)$ over all possible values of R_0 yields the desired likelihood ratio test. The test rejects H_0 at the R_0 level if

$$-n \ln \chi_{\max} < \chi^2_{1-\alpha}(I - 1)\,. \qquad (35)$$

Here $\chi^2_{1-\alpha}(I-1)$ is the upper $1-\alpha$ percentage points of the chi-square distribution with $I-1$ degrees of freedom and $\chi_{\max}$ is the large of the two solutions of the following quadratic equation: $ax^2+bx+c=0$ where

$$
\begin{aligned}
a &= \Sigma_i\Sigma_j c_{ij}^2 * \Sigma_i\Sigma_j e_{ij}^2 - (\Sigma_i\Sigma_j c_{ij}e_{ij})^2\,,\\
b &= [\Sigma_i\Sigma_j c_{ij}^2 * \Sigma_i(\Sigma_j e_{ij}^2 - n_i\bar{e}_i^2) + \Sigma_i\Sigma_j e_{ij}^2 * \Sigma_i(\Sigma_j c_{ij}^2 - n_i\bar{c}_i^2)\\
&\quad - 2(\Sigma_i\Sigma_j c_{ij}e_{ij}) * (\Sigma_i(\Sigma_j c_{ij}e_{ij} - n_i\bar{c}_i\bar{e}_i)]\,,\\
c &= \Sigma_i(\Sigma_j c_{ij}^2 - n_i\bar{c}_i^2) * \Sigma_i(\Sigma_j e_{ij}^2 - n_i\bar{e}_i^2) - (\Sigma_i(\Sigma_j c_{ij}e_{ij} - n_i\bar{c}_i\bar{e}_i)^2\,.
\end{aligned}
$$

5.2. *Without assumption of normality of distribution*

The distribution of c/e is often skewed. The lifetime models can be widely applied to investigate the distributions of c/e and the difference in c/e between populations. A cost-effectiveness distribution function, or $c-e$ distribution function, could be defined as:

$$S(c_e) = \Pr(c/e > c_e)\,. \tag{36}$$

The parametric, semi-parametric and non-parametric methods are able to deal with the data, whose distributions do not meet the assumption of normality and with censored data.

The Weibull, gamma and log-normal distributions could be applied to estimate the $c-e$ distribution function and the difference in c/e between different populations.

The non-parametric approach, such as Kaplan–Meier method could be applied to estimate the $c-e$ function. The non-parametric tests such as Wilcoxon and logrank test can be used to test the equality of the different groups.

6. Power and Sample Size Assessment for Tests of Hypotheses on Cost-Effetiveness Ratios

6.1. *Test of $H_0 : R = R_0$ verse $H_A : R \neq R_0$*

The power $(=1-\beta)$ of this test at the alternative $H_A : R = R_A (\neq R_0)$ is given by

$$
\begin{aligned}
P[|\Delta\bar{c} - \bar{R}_0\Delta\bar{e}| < z_{1-\alpha/2}\{&\operatorname{var}(\Delta\bar{c}) + R_0^2\operatorname{var}(\Delta\bar{e})\\
&- 2R_0^2\operatorname{cov}(\Delta\bar{c},\Delta\bar{e})\}^{1/2}|H_A] = \beta\,.
\end{aligned}
\tag{37}
$$

Under H_A, $E(\Delta\bar{c} - \bar{R}_0\Delta\bar{e}) = \delta(R_A - R_0)$, with $\delta(\neq 0)$ denoting the true incremental effectiveness. Assuming the covariance matrix of $(\Delta\bar{c}, \Delta\bar{e})'$, Σ, is known, Eq. (37) yields

$$\begin{aligned} P[&-z_{1-\alpha/2} - \delta(R_A - R_0)\{\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) \\ &- 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e}\}^{-1/2} < Z < z_{1-\alpha/2} \\ &- \delta(R_A - R_0)\{\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) \\ &- 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e}\}^{-1/2}] = \beta\,, \end{aligned} \tag{38}$$

where Z is standard normal and $n_A = kn_0$. Depending on the sign of $\delta(R_A - R_0)$, the absolute magnitude of one of the limits on Z is usually large. In either case, we will get, approximately .

$$\begin{aligned} &|\delta(R_A - R_0)| \\ &\quad = (z_{1-\alpha/2} + z_{1-\beta})\{\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) - 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}^{1/2}\,. \end{aligned} \tag{39}$$

Routing algebraic steps gives

$$\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) - 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e}) = \widehat{\sigma}_c^{\,2}(1 - \widehat{\rho}^{\,2})(1 + v_0)\,,$$

where $v_0 = \{R_0(\widehat{\sigma}_e/\widehat{\sigma}_c) - \widehat{\rho}\}^2/(1 - \widehat{\rho}^{\,2})$. Supposing $n_1 = kn_0$, where n_0 and n_1 are the number of patients in the test intervention and referent intervention respectively, we have

$$n_0 = \frac{(\widehat{\sigma}_{c_0}^{\,2} + k^{-1}\widehat{\sigma}_{c_1}^{\,2})(z_{1-\alpha/2} + z_{1-\beta})^2(1 - \widehat{\rho}^{\,2})(1 + v_0)}{\delta^2(R_A - R_0)^2} \tag{40}$$

Under the same design set-up, the sample size n_{0b} in the referent intervention needed to guarantee power of $1 - \beta$ to detect a difference δ in the test of $H_{01} : \Delta\bar{e} = 0$ is given by

$$n_{0b} = \delta^{-2}(\widehat{\sigma}_{e0}^{\,2} + k^{-1}\widehat{\sigma}_{e1}^{\,2})(z_{1-\alpha/2} + z_{1-\beta})^2\,.$$

Therefore,

$$n_0/n_{0b} = \frac{(\widehat{\sigma}_{c0}^{\,2} + k^{-1}\widehat{\sigma}_{c1}^{\,2})(1 - \widehat{\rho}^{\,2})(1 + v_0)}{(\widehat{\sigma}_{e0}^{\,2} + k^{-1}\widehat{\sigma}_{e1}^{\,2})(R_A - R_0)^2}\,. \tag{41}$$

The parameter v_0 is a function of ρ^* between $\Delta\bar{e}$ and $(\Delta\bar{c} - \bar{R}_0\Delta\bar{e})$. In fact,

$$\rho^* = -\{R_0(\widehat{\sigma}_e/\widehat{\sigma}_c) - \widehat{\rho}\}/\sqrt{(1 - v^2)}(1 + -v_0)\,.$$

Therefore, $|\rho^*| = \{v_0/(1+v_0)\}^2$. It will be very large if ρ^* is close to one and, consequently, the sample sizes in (40) and (41) will also be large. The correlation between the incremental cost and the incremental effectiveness is related through (32) to $\widehat{\rho}$ the individual correlations $\widehat{\rho}_0, \widehat{\rho}_1$ between cost and benefit in the two interventions. As is usually the case, $R_0 \geq 0$ and both (40) and (41) are monotonically decreasing in $\widehat{\rho}$ leading to a smaller sample size n_0 and relative size n_0/n_{0b} with increasing value of $\widehat{\rho}$. Finally, these sample size formulae are dependent on both the hypothesized CER R_0 and the difference $R_A - R_0$.

6.2. *Test of* $H_0 : R \geq R_0$ *verse* $H_A : R < R_0$

Analogous sample size calculations yield the following formula, which replaces (40):

$$n_0 = \frac{(\sigma_{c0}^2 + k^{-1}\sigma_{c1}^2)(z_{1-\alpha} + z_{1-\beta})^2(1-\rho^2)(1+v_0)}{\delta^2(R_A - R_0)^2}\,. \tag{42}$$

For the one-sided test $H_{01} : \Delta E(e) = 0$, with regard to their effectiveness and, therefore, should be compared on their costs. The ratio n_0/n_{0b} compares the sample size requirement of the test $H_0 : R = R_0$ with that for $H_{01} : \Delta\bar{e} = 0$, with the latter powered to detect the difference δ.

6.3. *Test of* $H_0 : \Delta E(e) \leq 0$ *or* $R \geq R_0$ *verse* $H_A : \Delta E(e) > 0$ *and* $R < R_0$

With the solution (34), the power $(1-\beta)$ of the test can be computed from the bivariate normal distribution of (Z_1, Z_2) and is given by

$$1-\beta = P\left[Z_1 < -z_{1-\alpha} + \frac{\delta}{\sigma_1},\ Z_2 < -z_{1-\alpha} + \frac{\delta|R - R_0|}{\{\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) - 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}^{-1/2}}\right], \tag{43}$$

where $\delta > 0$ is the incremental effectiveness and $R(< R_0)$ is the true cost-effectiveness ratio. This parallels the power considerations leading to (39) for the test of cost-effectiveness only. The choice of c_1 and c_2 in (33) is optimal in order to gain maximal power for a given sample size and alternative. An implicit expression for the sample size n_0 corresponding to ( ) can be derived by making the substitutions $\sigma_e^2 = (\sigma_{e0}^2 + k^{-1}\sigma_{c1}^2)/n_0$, $\sigma_e^2 = (\sigma_{c0}^2 + k^{-1}\sigma_{c1}^2)/n_0$ and $\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) - 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e}) = \sigma_c^2(1-\rho)(1+v_0)$.

Note that the previous expression for n_0 in (42) is a lower bond for the sample size requirements for testing $H_0 : \Delta E(e) \leq 0$ or $R \geq R_0$.

6.4. *Numerical computations*

In some special cases, simplification of (40)–(43) are possible. Suppose the costs (c_0, c_1) and benefit measures (e_0, e_1) in the two interventions have the same variance $\widehat{\sigma}^2_{c0} = \widehat{\sigma}^2_{c1} (= \widehat{\sigma}^2_c)$, $\widehat{\sigma}^2_{e0} = \widehat{\sigma}^2_{e1} (= \widehat{\sigma}^2_e)$, respectively. Then, assuming equal allocation to the two interventions ($k = 1$), we have $\widehat{\rho} = (\widehat{\rho}_0 + \widehat{\rho}_1)/2$ and the sample size n_0, n_{0b} in (40) and (41) reduce to

$$\begin{aligned} n_0 &= \frac{2\widehat{\sigma}^2_e (z_{1-\alpha/2} + z_{1-\beta})^2 (R_0^2 + (\widehat{\sigma}_c/\widehat{\sigma}_e)^2 - 2\widehat{\rho} R_0 (\widehat{\sigma}_c/\widehat{\sigma}_e))}{\delta^2 (R_A - R_0)^2}, \\ n_{0b} &= \delta^{-2} 2\widehat{\sigma}^2_e (z_{1-\alpha/2} + z_{1-\beta})^2 . \end{aligned} \tag{44}$$

From (42), for one-sided testing, $z_{1-\alpha/2}$ must be replaced by $z_{1-\alpha}$. The effect size $\delta/\sigma_\varepsilon$ is the difference in effectiveness in units of standard deviation (SD). For the joint hypothesis test of $H_0 : \Delta E(e) \leq 0$ or $R \geq R_0$, the power and sample size expression (43) becomes

$$\begin{aligned} 1 - \beta = P \Bigg[& Z_1 < -z_{1-\alpha} + \sqrt{\frac{n_0}{2}} \frac{\delta}{\widehat{\sigma}_e}, \; Z_2 < -z_{1-\alpha} \\ & + \sqrt{\frac{n_0}{2}} \frac{\delta}{\widehat{\sigma}_e} \frac{|R - R_0|}{\{R_0^2 + (\widehat{\sigma}_c/\widehat{\sigma}_e)^2 - 2\widehat{\rho} \, R_0 (\widehat{\sigma}_c/\widehat{\sigma}_e)\}^{-1/2}} \Bigg]. \end{aligned} \tag{45}$$

The sample size requirement for this joint test to ensure power $(1 - \beta)$ would be greater than the sample size needed for the one-sided test for effectiveness alone. For fixed n_0, the factor

$$\zeta = \frac{R_0^2 + (\widehat{\sigma}_c/\widehat{\sigma}_e)^2 - 2\rho R_0 (\widehat{\sigma}_c/\widehat{\sigma}_e)\}^{-1/2}}{|R - R_0|}$$

would drive the power, with power decreasing with increasing ζ. This factor is the square-root of the sample size ratio n_0/n_{0b} in (41). Note that if $|\widehat{\sigma}_c/\widehat{\sigma}_e - R_0| > |R_0 - R|$, irrespective of the value $\widehat{\rho}$ we always have $\zeta > 1$. Therefore, (45) should be used to calculate power of the joint test given a sample size that might be available for testing effectiveness. On the other hand, (44) is suitable for assessing the sample size needed to establish cost-effective. It should be noted that the right-hand side of (45) is dependent on $\widehat{\rho}$ through the correlation first decrease Z_1 and Z_2. In practice, we are likely to have $R_0 > \widehat{\sigma}_c/\widehat{\sigma}_e$, in which case this correlation $\widehat{\rho}$ first decreases

with and then increase after the value $\widehat{\rho} = (R_0\sigma_e/\sigma_c)^{-1}$, therefore, in this circumstance, a strong positive correlation between cost and effectiveness would suggest a smaller sample size requirement for (45) to hold given β.

7. Examples

7.1. *Example* 1

We use summary data from Sacristan *et al.*[14] on a trial comparing two pharmacological agents in this example. Data on 150 patients using the test drug yield a mean cost of \$200,000 (SD = \$78,400). Health benefit measured in QALYs is 8 (SD = 2.1) corresponding values on 150 patients using the standard drug are \$80,000 (SD = \$27,343) for mean cost, and 5 QALYs (SD = 2.0) for mean health benefit. These values yield the following estimates: $\Delta\bar{e} = 3$, $\Delta\bar{c} = \$120,000$ and from (32) $\widehat{\sigma}_e = 0.237$ and $\widehat{\sigma}_c = 6779$. In the absence of a reported value for the correlation between cost and effectiveness, we consider values $\widehat{\rho}_0 = \widehat{\rho}_1 = 0.7$. From (32), we see that with zero correlations, the incremental cost and incremental effectiveness are uncorrelated ($\rho = 0$). For $\widehat{\rho}_0 = \widehat{\rho}_1 = 0.7$, we get $\widehat{\rho} = 0.638$ approximately.

7.1.1. *Hypothesis testing for the CER*

Suppose the hypothesized CER was $R_0 = \$50,000/\text{QALY}$. From the test of $H_0 : R = R_0$ verse $H_A : R \neq R_0$ section, the two-sided test of $H_0 : R = R_0$ based on the statistic $T = (\Delta\bar{c} - R_0\Delta\bar{e})/\{\text{var}(\Delta\bar{c} - R_0\Delta\bar{e})\}^{1/2}$ has a p-value of 0.03 if $\widehat{\rho}_0 = 0$ and approximately 0.001 if $\widehat{\rho}_0 = 0.7$. It can be shown that the p-values decrease with increasing values of $\widehat{\rho}_0$.

7.1.2. *Determining statistical power*

What power does this test have to detect an alternative CER, $R_0 = \$40,000/\text{QALY}$? We compute the power from (39) assuming an incremental effectiveness of 3 QALYs. If $\widehat{\rho}_0 = 0$ the power is about 59% and increases to 94% if $\widehat{\rho}_0 = 0.7$. A lower power may be acceptable in studies of cost-effectiveness.

7.1.3. *Testing the joint hypothesis on effectiveness and cost-effectiveness*

The power function of this one-sided test is given in (43). To test for significance of the difference in effectiveness (i.e. $H_{01} : \Delta E(e) = 0$), we would

reject if $|\Delta\bar{e}/\widehat{\sigma}_e| > z_{1-\alpha/2}$. In this example, the difference δ being highly significant makes the right-hand side of (43) essentially

$$P\left[Z_2 < -z_{1-\alpha} + \frac{\delta|R - R_0|}{\{\mathrm{var}(\Delta\bar{c}) + R_0^2\,\mathrm{var}(\Delta\bar{e}) - 2R_0^2\,\mathrm{cov}(\Delta\bar{c}, \Delta\bar{e})\}^{-1/2}}\right].$$

The power at $\delta = 3$ and $R = \$40,000$ is about 0.71 for these data.

7.2. *Example* 2

Consider the simplifications leading to (44) and (45). To ensure a power of 80% to detect an effect size $\delta/\widehat{\sigma}_e = 0.5$ with a two-sided test of $H_{01} : \Delta E(e) = 0$ with $\alpha = 0.05$, we get $n_{0b} = 63$. Suppose the hypothesized ICER is $R_0 = \$80,000$/QALY and the relative SD $\widehat{\sigma}_c/\widehat{\sigma}_e = 5,000$ (\$/QALY). Correlation between the cost and effectiveness measures is likely to be positive. Let $\widehat{\rho} = 0.7$ and assume a known effect size $\delta/\widehat{\sigma}_e = 0.5$. The sample size n_0 needed to detect an ICER of \$50,000/QALY or less with 80% power requires $n_0 \geq 6.51 n_{0b}$. For two-sided testing, this yields $n_0 \geq 410$. The sensitivity of $\rho(> 0)$ to this relative sample size is small. A zero correlation increase this ratio to 7.1.

Now consider testing the joint hypothesis $H_0 : \Delta E(e) \leq 0$ or $R \geq R_0$ under the same constraints. Suppose we want 80% power to detect an effect size 0.5 and an ICER of \$50,000/QALY. Using (45), we will get $n_0 = 323$ when $\widehat{\rho} = 0.7$. Note that the joint hypothesis is formulated as one-sided. In comparison, a one-sided test for effectiveness would need approximately 50 subjects per arm to detect an effect size of 0.5 with 80% power. As noted after (45), the power is driven by the probability involving Z_2 because $\zeta > 1$ in this case.

7.3. *Example* 3

Sample size requirements for testing $H_0 : \Delta E(e) \leq 0$ or $R \geq R$ are given in Table 1 for some values of R_0 and effective sizes $\delta/\widehat{\sigma}_e$. The test is designed with $\alpha = 0.05$ and 80% power at $R = \$30,000$/QALY. We use (45) with $\widehat{\rho} = 0.7$ and $\widehat{\sigma}_c/\widehat{\sigma}_e = 5,000$ (\$/QALY).

The last column of Table 1 gives the sample size requirement to ensure 80% power in the two-sided test of effectiveness alone. For example, to detect an effect size of 0.4 and a ICER of \$30,000/QALY, when the maximum acceptable level is \$50,000/QALY, we require a sample size of 421 for the test and referent groups. In comparison, for testing $H_{01} : \Delta E(e) = 0$, only 99 subjects are required to detect an effect size of 0.4.

Table 1. Sample size requirements for testing effectiveness and cost-effectiveness.

Effective size	Maximum ICER R_0 ($1000/QALY)					Effectiveness alone
	40	45	50	55	60	
0.3	1848	1060	748	586	490	175
0.4	1040	596	421	330	276	99
0.5	666	382	269	211	177	63
0.6	462	265	187	147	123	44

Because of the relatively large sample size needed to test the joint hypothesis of cost-effectiveness and effectiveness, in practice power could be calculated from (10) using the sample size that is needed to demonstrate a difference in effectiveness between two treatments. For example, with 175 subjects per arm, we have 80% power to detect an effect size of 0.3. with this sample size, $\widehat{\rho} = 0.7$ and $R_0 = \$50,000/\text{QALY}$ we will have 64% power to detect a ICER of $30,000/QALY at an effect size of 0.5 at an effect size of 0.3, the power is only 33%.

8. Modeling for Cost-Effectiveness Analysis

Cost-effectiveness analysis require estimation of the health effects and resource costs associated with an intervention and with the alternatives to which it will be compared. Modeling is frequently necessary since few studies provide information over sufficiently long periods or for all relevant costs, effects and population groups.

Cost-effectiveness analysis helps inform different types of decisions about health interventions. To begin, it can inform the decision to use an intervention at all by showing whether it is cost-effective enough compared to alternatives. More often decisions concern hoe to use the intervention. Should screening for hypertension be done every year, every two years, or every five years? If hypertension is diagnosed, and non-drug therapies are unsuccessful, which drugs should be used? Should folic acid supplementation be accomplished through diet, vitamin supplements, or fortification of cereal grains? If fortification, how many mg of folic acid per 100 grams of cereal grain product? Should every patient who presents at the emergency.

A model creates the framework for cost-effectiveness analysis. To serve its purpose, and enable decision makers to explore the implications of variation in the intervention, the condition, and the population, it must allow not only for substantial variation in those factors.

8.1. *Validating effectiveness estimates*

Accuracy is essential for a model. Eddy[2] described four levels of validation. First, the structure of the model should make sense to experts. Second, the model should reproduce the outcomes observed in the studies used to estimate its parameters. Third, the models predictions could be compared with results from studies not used in its construction. Fourth, the model could be used to predict outcomes for a new program and the predictions compared with the outcomes when the program is implemented. The first and second steps are essential. For the third step, randomized clinical trials (RCTs) offer a challenging, but potentially persuasive, test of a models accuracy. While trials are usually the benchmark, the model may be accurate on specific points.

It is reasonable to expect a good model to match the results of trials available at the time of its construction, but not to expect it to predict the results of future trials. Models can and should accurately reflect the state of knowledge at the time they are created.

When is a model going too far beyond the data? The medical and public health practice are the best guides. Models can appropriately be used to analyze any circumstances in which the intervention is already being applied, or in which it is being seriously considered for application. If it is appropriate to use the intervention in the real world, on real people, it is an appropriate to analyze the implications of that use of a model.

8.2. *Modeling costs*

Eddys suggestions described above should be considered for the cost estimate as well. Modelers need to pay attention to ensuring that the pathway of events described by a model represents costs as well as it does effects.

In part, the failure to validate cost estimates reflects the failure to take cost data as seriously as effectiveness data. A basic requirement for accurate predictions, often overlooked, is that both costs and effects should apply to the same population and the same circumstances. Further, data on resource use and cost need to be associated with the same care and subjected to the same sorts of consistency checks as effectiveness data — comparing one source with another, relating differences in costs to characteristics thought to be associated with those differences and so on.

In addition, the range of variation that could usefully be modeled is as wide for costs as for effects. An iterventions effectiveness differs across the

country because populations differ in incidence of the condition, risk factors and co-morbidities. Costs differ across the country because of differences in wages and other costs, in practice patterns and in suitable production technologies. While one purpose of sensitivity analyses is to determine which parameters have a major influences on cost-effectiveness, it would also be useful to explore sets of assumptions that describe, as accurately as the data allow, circumstances in another part of the country or another delivery system.

The US panel on cost-effectiveness in Health and Medicine has urged the use of micro-costing for costing events important to an analysis. Micro-costing could yield a better understanding of the factors that underlie resource use and costs for various conditions, analogous to the understanding of effectiveness built up from epidemiological and clinical research. That understanding might reveal alternatives for making interventions more cost-effective by changing the way they are delivered, not just by targeting them to population subgroup.

Models should be flexible enough to permit exploration of a range of production possibilities and cost levels for an intervention. Analysts could then examine plausible differences in costs and production technologies. It would be useful to evaluate combinations of values that occur in the real world: conditions in Michigan verse those in San Francisco, conditions in an inner city, a suburb, or a rural area.

8.3. *Modeling form*

Models are built from estimates of risk — the probability that a condition will progress to the next stage, that a test is accurate, that a treatment will be effective. In medical research, the familiar and convenient mathematical forms for fitting risk relationships are the logistic and, more recently hazard models. Both forms incorporate an assumption that the risk relationship is multiplicative, and thus that the size of the risk reduction caused by changing one risk factor differs for different levels of the other risk factors. This assumption implies, for example, that the reduction in risk caused by lowering systolic blood pressure from 160 mmHg to 140 mmHg will be larger in people who also smoke, even though they continue to smoke, than in people whose only risk factor is high blood pressure. Similarly, the reduction in risk from smoking cessation will be greater in people who are hypertensive, even if their blood pressure is unchanged, than in non-smokers.

In turn, this implies that it will be more cost-effective to apply an intervention to people with several risk factors, not because the programme achieves economies by treating several riskm factors, but because intervention against a single risk factor is more effective in these people. The point is clear in an analysis by Taylor *et al.*[18] Of a dietary programme to lower serum cholesterol modeled after the one employed in MRFIT. Effectiveness was estimated using logistic coefficients reported from the Framingham study. Results were presented separately for low-risk men, whose only risk factors for heart disease were their gender and cholesterol level, and for high-risk men, who also smoked and had high blood pressure and low HDL levels. Although the cost of the intervention was the same, cost per life-year was approximately ten times higher for low-risk men because of the multiplicative assumption incorporated in the logistic form.

Logistic and hazard models play an important role in some of the situations for which models are particularly useful — examining differences in effectiveness and cost-effectiveness among subgroups. When analysts model the implications of targeting an intervention to subgroups, or extrapolate to explore its application to less-studied groups, they need to be aware of the implications of the conventional forms. Modelers cannot supply the data to resolve this issue, but they can draw attention to it by showing how estimates change when addictive and multiplicative forms are used. The ultimate goal is to ensure that estimated differences among subgroups are not an artifact of a convenient statistical model.

References

1. Chaudhary, M. A. and Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Statistics in Medicine* **15**: 1447–1458.
2. Eddy, D. M. (1985). Technology assessment: The role of mathematical modeling in committee for evaluating medical technologies. In *Clinical Use*, *Institute of Medicine*, *Assessing Medical Technologies*, National academy Press, Washington, D.C. 144–154.
3. Efron, B. (1987). Better bootstrap confidence interval. *Journal of the American Statistical Association* **82**: 171–200.
4. Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
5. Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society*, *Series* **B16**: 175–183.
6. Garber, A. M. and Leichter, H. M. (1991). Practice guidelines and cholesterol policy. *Health Affairs* **10**(2): 52–66.

7. Goldman, L., Weinstein, M. C., Goldman, P. A. and Williams, I. W. (1991). Cost-effectiveness of HMG — CoA reductase inhibition for primary and secondary prevention of coronary heart disease. *Journal of American Medical Association* **265**: 1145–1151.
8. Hlatky, M. A., Rogers, W. J., Johnstone, I. *et al.* (1997). Medical care costs and quality of life after randomization to coronary angioplasty or coronary bypass surgery. *New England Journal Medicine* **336**: 92–99.
9. Kupperman, M., Luce, B., McGovern, B. *et al.* (1990). An analysis of the cost-effectiveness of the implantable defibrillator. *Circulation* **81**: 91.
10. Lipid Research Clinic Program (1984). Lipid research clinics coronary primary trial results, II: The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. *Journal of American Medical Association* **251**: 365–374.
11. National Cholesterol Education Program (NCEP) (1988). High blood cholesterol in adults: Report of the expert panel on detection, evaluation, and treatment. National Institutes of Health. Department of Health and Human Services, Bethesda, MD.
12. National Cholesterol Education Program (NCEP) (1994). The second report of the expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. *Circulation* **89**: 1329–1445.
13. OBrien, B. J., Drummond, M. F., Lebelle, R. J. and Willan, A. (1994). In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care* **32**: 150–163.
14. Sacristan, J. A., Day, S. J., Navarro, O., Ramos, J. and Hernandez, J. M. (1995). Use of confidence intervals and sample size calculations in health economic studies. *Annals of Pharmacotherapy* **29**: 719–725.
15. Sempos, C., Fulwood, R., Haines, C., Carroll, M. *et al.* (1989). Prevalence of high blood cholesterol levels among adults in the United States. *Journal of American Medical Association* **262**: 45–52.
16. Siegel, C., Laska, E. and Meisner, M. (1996). Statistical methods for cost-effectiveness analyses. *Controlled Clinical Trials* **17**: 387–406.
17. Stinnett, A. (1996). Adjusting for bias in C/E ratio estimates. *Health Economics* **5**: 469–472.
18. Taylor, W. C., Pass, T. M., Shepard, D. S. and Komaroff, A. F. (1990). Cost effectiveness of cholesterol reduction for the primary prevention of coronary heart disease in men. In *Preventing Disease*: *Beyond the Rhetoric*, eds. R. B. Goldbloom and R. S. Lawrence, Springer-Verlag, NY, 437–441.
19. Weinstein, M. C. and Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine* **296**: 716–721.

About the Author

Jianli Li is working in the Department of Corporate Performance, St. Michael's Hospital, University of Toronto, Canada, as biostatitician. He

worked in Ontario Joint Policy and Planning Committee, Ontario Ministry of Health and Ontario Hospital Association, as statistical consultant working on the hospital funding models and performance assessment. He has been working in healthcare management for decades and making efforts to apply the new developments in statistics, medical information science, computer science and applied mathematics into the healthcare management. He published and presented several papers and chapters, such as "An Application of Life Time Model in Estimation of Expected Length of Stay of Patient in Hospital", "Impact of the Complexity Methodology on an Ontario Teaching Hospital", "A System for Evaluating Inpatient Care Cost-Efficiency in Hospital" and "Data Mining in Health Care Organisations: A Source for Continuous Health Care Information". He was a visiting professor at the Department of Preventive Medicine and Biostatistics, University of Toronto. He was associate professor and vice chairman at the Department of Healthcare Management, Shanghai Second Medical University.

This page intentionally left blank

CHAPTER 6

QUALITY OF LIFE: ISSUES CONCERNING ASSESSMENT AND ANALYSIS

JIQIAN FANG* and YUANTAO HAO

Department of Medical Statistics, School of Public Health, Sun Yat-Sen University, 74 Zhongshan Road II, Guangzhou 510080, PR China

*Tel: 86-20-87330671; *fangjq@gzsums.edu.cn*

1. The Concept of QOL and its Components

What is quality of life? There is no universally agreed definition. Quality of life (QOL) not only means different things to different people, but it also varies according to a person's current situation. When a person falls sick he thinks QOL is good health, when he is poor, QOL is wealth. To a town planner, for example, QOL might represent access to green space and other facilities. In the context of clinical trials we are rarely interested in QOL in such a broad sense, but are concerned only with evaluating those aspects that are affected by disease or treatment of disease. This may sometimes be extended to include indirect consequences of disease such as unemployment or financial difficulties. To distinguish between QOL in its more general sense and the requirements of clinical medicine and clinical trials, the term "health-related quality of life" (HRQOL) is frequently used in order to remove ambiguity.

There are a number of reasons for developing a quality of life assessment tool. The main reason is undoubtedly that in recent years there has been a broadening of focus of the measurement of health beyond traditional health indicators such as mortality and morbidity.[1] Indeed, the measurement of health may now includes assessment of the impact of disease and impairment on daily activities and behaviour,[2] perceived health measures[3] and disability/functional status measures.[4] These measures, whilst beginning to provide an indication of the impact of disease, do not access quality of

life per se, which has been aptly described as "the missing measurement in health".[5] The increasingly mechanistic model of medicine, concerned only with the eradication of disease and symptoms, reinforces the need for the introduction of a humanistic element into health care. Health care is essentially a humanistic transaction in which the patient's well-being is the primary aim. By calling for QOL assessment in health care, attention is focused on this aspect of health, and resulting interventions will pay increased attention to the problem.

There still has not been a single, clear, universally accepted definition of HR-QOL. What domains should be included in QOL? There are five major domains of QOL which are generally referred to by most authors. These domains are physical status and functional abilities, psychological status and well being, social interactions, economic and/or vocational status and factors, and religious and/or spiritual status.

The World Health Organization (WHO) has developed an international quality of life assessment instrument (WHOQOL) which allows an enquiry into an individual's perception of own position in life in the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards and concerns. The WHOQOL measures quality of life related to health and health care. It has been developed in the framework of a collaborative project involving numerous centres in different cultural settings.[6] QOL is defined by WHO as "individuals' perceptions of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns". It is a broad ranging concept incorporating in a complex way the persons' physical health, psychological state, level of independence, social relationships, personal beliefs and their relationships to salient features of the environment.

This definition reflects the view that quality of life refers to a subjective evaluation, which is embedded in a cultural, social and environmental context. As such, quality of life cannot be equated simply with the terms "health status", "life style", "life satisfaction", "mental state" or "well-being". Because the WHOQOL focuses upon respondents' "perceived" quality of life, it is not expected to provide a means of measuring in any detailed fashion symptoms, diseases or conditions, nor disability as objectively judged, but rather the perceived effects of disease and health interventions on the individual's quality of life. The WHOQOL is, therefore, an assessment of a multi-dimensional concept incorporating the individual's perception of health status, psycho-social status and other aspects of life.

It is anticipated that the WHOQOL assessment will be used in broad-ranging ways. It will be of considerable use in clinical trials, in establishing baseline scores in a range of areas, and looking at changes in quality of life over the course of interventions. It is expected that the WHOQOL assessment will also be of value where disease prognosis is likely to involve only partial recovery or remission, and where treatment may be more palliative than curative.

For epidemiological research, the WHOQOL assessments will allow detailed quality of life data to be gathered on a particular population, facilitating the understanding of diseases, and the development of treatment methods. The international epidemiological studies that would be enabled by instruments such as the WHOQOL-100 and the WHOQOL-BREF will make it possible to carry out multi-center quality of life research, and to compare results obtained in different centers. Such research has important benefits, permitting questions to be addressed which would not be possible in single site studies. For example, a comparative study in two or more countries on the relationship between health care delivery and quality of life requires an assessment yielding cross-culturally comparable scores. Sometimes accumulation of cases in quality of life studies, particularly when studying less frequent disorders, is helped by gathering data in several settings. Multi-center collaborative studies can also provide simultaneous multiple replications of a finding, adding considerably to the confidence with which findings can be accepted.

In clinical practice the WHOQOL assessments will assist clinicians in making judgements about the areas in which a patient is most affected by disease, and in making treatment decisions. In some developing countries, where resources for health care may be limited, treatments aimed at improving quality of life through palliation, for example, can be both effective and inexpensive. Together with other measures, the WHOQOL-BREF will enable health professionals to assess changes in quality of life over the course of treatment.

It is anticipated that in the future the WHOQOL will prove useful in health policy research and will make up an important aspect of the routine auditing of health and social services. Because the instrument was developed cross-culturally, health care providers, administrators and legislators who require a valid QOL instrument for use can be confident that data yielded by work involving the WHOQOL assessment will be genuinely sensitive to their setting.

A large number of instruments have been developed for QOL assessment and we can divide them into two categories: generic instruments and

disease-specific instruments.[7] Generic instruments are intended for general use, irrespective of the illness or condition of the patient. These generic questionnaires may often be applicable to healthy people too. Some of the earliest ones were developed initially with population surveys in mind, although they were later applied in clinical trial settings.

There are many instruments that measure physical impairment, disability or handicap. Although commonly described as QOL scales, these instruments are better called measures of health status because they focus on physical symptoms. They emphasize the measurement of general health, and make the implicit assumption that poorer health indicates poorer QOL. One weakness about this form of assessment is that different patients may react differently to similar levels of impairment. Many of the earlier questionnaires such as the Sickness Impact Profile (SIP)[2] and the Nottingham Health Profile (NHP)[8] to some degree adopt this approach. Few of the earlier instruments had scales that examined the subjective non-physical aspects of QOL, such as emotional, social and existential issues. Newer instruments such as the Medical Outcomes Study 36-Item Short Form (SF-36),[9] however, emphasize these subjective aspects strongly, and also commonly include one or more questions that explicitly enquire about overall QOL. More recently, some brief instruments that place even less emphasis upon physical functioning have been developed. Two such instruments are the EuroQol,[10] which is intended to be suitable for use with cost-utility analysis, and the SEIQol,[11] which allows patients to choose those aspects of QOL that they consider most important to themselves.

Generic instruments, intended to cover a wide range of conditions, have the advantage that scores from patients with various diseases may be compared against each other and against the general population. On the other hand, these instruments fail to focus on the issues of particular concern to patient with disease, and may often lack the sensitivity required to detect differences that arise as a consequence of treatment policies that are compared in clinical trials. This has led to the development of disease-specific questionnaires, for example, the EORTC QLQ-C30 (European Organization for Research and Treatment of Cancer QLQ-C30).[12]

2. Methods of Developing QOL Measurements

The development of a new QOL instrument requires a considerable amount of detailed work, demanding patience, time and resources. Some evidence of this can be seen from the series of publications that are associated with

such QOL instruments as the SF-36, the FACT and the EORTC QLQ-C30. These and similar instruments have initial publications detailing aspects of their general design issues, followed by reports of numerous validation and field-testing studies.

Many aspects of psychometric validation depend upon collecting and analysing data from samples of patients or others. However, the statistical and psychometric techniques can only confirm that the scale is valid in so far as it performs in the manner that is expected. These quantitative techniques rely on the assumption that the scale has been carefully and sensibly designed in the first place. To that end, the scale development process should follow a specific sequence of stages, and details of the methods and the results of each stage should be documented thoroughly. Reference to this documentation will, in due course, provide much of the justification for content validity. It will also provide the foundation for the hypothetical models concerning the relationships between the items on the questionnaire and the postulated domains of QOL, which are then explored as construct validity.

Next, we will discuss the steps in instrument development in detail.

2.1. *Specifying measurement goals*

Before embarking on the development of any new instrument, the investigator should define exactly what the instrument is to measure. This initial definition will help the investigator design appropriate development and testing protocols and will enable other users of the instrument to identify its applicability to their own patients and studies. This process will include specification of the objectives in measuring QOL, a working definition of what is meant by "quality of life", identification of the intended groups of respondents, and proposals as to the aspects or main dimensions of QOL that are to be assessed. The investigator should consider at least the following criteria.

2.1.1. *Patient population*

As in a clinical trial, there should be clear inclusion and exclusion criteria that identify the precise clinical diagnosis and basic patient characteristics. A detailed definition might include age, literacy level, language ability, and presence of other illness that might have impact on QOL. An investigator may be thinking of a particular study in which the instrument is to be used, but constructing an instrument for too specific a population or function may

limit its subsequent use. One can usually choose a patient population that is narrow enough to allow focus on important impairments in that disease or function but board enough to be valid for use in other studies.

2.1.2. *Primary purpose*

The investigator needs to decide whether the primary purpose of the instrument is going to be evaluative, discriminative, or predictive. Although some instruments may be capable of all three functions, it is difficult to achieve maximum efficiency in all three.

2.1.3. *Patient function*

In most disease-specific instruments, investigators want to include all areas of dysfunction associated with that disease (physical, emotional, social, occupational). However, there are some instruments that are designed to focus on a particular function (e.g. emotional function, pain, sexual function) within a broader patient population. The investigator should decide whether all or only specific functions are to be included.

2.1.4. *Other considerations*

The investigator should also decide on the format of the instrument. Will it be interviewer and/or self-administered? Does it need to be suitable for telephone/postal interviews? Approximately how many items will the instrument contain?

Once a working definition of quality of life and study protocol are developed, a further phase of work involved operationalizing the broad domains and individual facets of quality of life. Consultants and principal investigators should draft a provisional list of domains and constituent facets of quality of life. Each facet definition should consist of a conceptual definition, a description of various dimensions along which a rating can be made for that facet, and a listing of some example situations or conditions that might significantly affect that facet at various levels of intensity. Once facets of QOL are drafted, a series of focus groups should be held with patients, well persons and health professionals to consider the facet definitions drafted by health professionals and QOL researchers. On the basis of the focus group data, a revised set of facet or domain definitions are compiled to guide subsequent item generation.

2.2. *Item generation*

The first task in instrument development is to generate a pool of all potential relevant items. For this pool, the investigator will later select items for inclusion in the final questionnaire. The most frequently used methods of item generation include unstructured interviews with patients who have insight into their condition, patient focus group discussions, a review of the disease-specific literature, discussions with health care professionals who work closely with the patients, and a review of generic QOL instruments.

A question-writing panel should be assembled. The question-writing panel should consist of the principle investigator, the main focus group moderator, at least one person with good interviewing skills and experience, and a lay person, preferably someone who participates in one of the lay focus groups, to ensure that questions are framed in a way that is easy to understand.

2.3. *Item reduction: Reducing items on the basis of their frequency and importance*

Having generated a large item pool, the investigator must select the items that will be most suitable for the final instrument. QOL instruments usually measure health status from the patients' perspective and so it is appropriate that patients themselves identify the items that are most important to them. Investigators should ensure that the patients selected represent the full spectrum of those identified in the patient population. It is important to ensure that all of the subgroups are adequately represented.

One approach to item reduction is to ask patients to identify those items that they have experienced as a result of their illness. For each positively identified item, they rate the importance using a 5-point Likert type scale ("extremely important" to "not important"). Results are expressed as *frequency* (the proportion of patients experiencing a particular item), *importance* (the mean importance score attached to each item), and the *impact*, which is the product of frequency and importance.

Very occasionally, there are items that have absolutely no potential of changing over time either as a result of an intervention or though the natural course of the disease. If one is developing an evaluative instrument, one may consider excluding such unresponsive items because they will only add to the measurement noise and the time taken to complete the questionnaire. However, if such an item is considered very important by patients

and therefore potentially a future target for therapy, exclusion because of apparent unresponsiveness to current therapies may be unwise.

A comprehensive set of items will inevitably include some redundancies. How does one decide whether to include them? One approach is to test whether the items are highly correlated. If Spearman rank order correlations are high one could consider omitting one of the items. This strategy is particularly appropriate for a discriminative instrument, for highly correlated items will, when taken together, give little information in terms of distinguishing between those with mild and severe quality of life impairment. It is somewhat riskier for evaluative instruments; just because items correlate with one another at the item reduction phase does not guarantee that they will change in parallel when measured serially over time.

Investigators can select the sample size for the item reduction process by deciding how precise they want their estimates of the impact of an item on the population. The widest confidence interval around a proportion (the frequency with which patients identify items) occurs when the proportion is 50%; any other value will yield a narrower confidence interval. If one recruits 25 subjects, and an item is identified by 50% of the population, the true prevalence of that item is somewhere between approximately 30% and 70%. If one recruit 50 subjects, the 95% CI around a proportion of 0.5 will be approximately from 0.36 to 0.64. For 100 subjects, the confidence interval will be from 0.4 to 0.6. It is recommended that researchers recruit at least 100 subjects for this part of the questionnaire development process.

There are some statistical methods we can use to determine which items should be included in the instrument. Factor analysis, cluster analysis, multiple regression, and discriminant analysis are methods often used.

2.4. *Questionnaire formatting*

2.4.1. *Selection of response options*

Response options refer to the categories or scales that are available for responding to the questionnaire items. For example, one can ask whether the subject has difficulty climbing stairs; two response options, yes and no, are available. If the questionnaire asks about the degree of difficulty, a wide variety of response options are available.

An evaluative instrument must be responsive to important changes even if they are small. To ensure and enhance this measurement property, investigators usually choose scales with a number of options, such as a 7-point scale where responses may range from 1 = no impairment to 7 = total impairment, or a continuous scale such as a 10-cm Visual Analogue Scale

(VAS). The 7-point Likert scale is often preferred, because although both yield similar data, the Likert scale has practical advantages over the VAS, being both easier to administer and easier to interpret.

Likert scale and VAS can be used as discriminative and predictive instruments, and are likely to yield optimal measurement properties. However, Likert scale and VAS are more complex than a simple yes/no response and they are very difficult to use for telephone interviews. In health surveys, investigators requiring only satisfactory discriminative or predictive measurement properties of their instrument may choose a simple response option format.

2.4.2. *Time specification*

A second feature of presentation is time specification: patients should be asked how they feeling over a well-defined period of time. Two weeks is the time frame used by most instruments on the basis of the intuitive impression that patients can accurately recall. Time specification can be modified according to the study, and other investigators may have different impressions of the limits of their population's memory.

When a new questionnaire is developed, it is necessary to test its psychometric properties including validity, reliability, responsiveness and sensitivity. Validation of instruments is the process of determining whether there are grounds for believing that the instrument measures what it intends to measure, and that it is useful for its intended purpose. Reliability concerns the random variability associated with measurements. Ideally, patients whose QOL status has not changed should make very similar, or repeatable, responses each time they are assessed. If there is considerable random variability over time, the measurements are unreliable. Sensitivity is the ability of measurement to detect differences between patients or groups of patients. Sensitivity is important in clinical trials since a measurement is of little use if it cannot detect the differences in QOL that may exist between the randomised groups. We will discuss these properties in detail in Sec. 6.

3. Linguistic Validation of QOL Instrument

3.1. *Introduction*

Most health status measures and psychological tests are used only in the setting in which they were originally developed. Some are translated into other languages and used without making any adaptations, and yet this is

necessary to ensure their usefulness in another culture or language. A very small number of instruments are produced in equivalent version in different languages, before assessing the instruments' validity and reliability that are prerequisites for the use of instrument in a new culture.

WHO has accrued considerable experience in translating health measurements. This has facilitated the development of a translation methodology which has significant advantages over the forward-translation and the translation-back-translation methodologies. We call this procedure "linguistic validation". The steps outlined below describe a sequence which has been used successfully in a number of studies. It is clear that variations of the method may well be necessary, and indeed desirable, in certain situations.

The aim of linguistic validation of a QOL questionnaire is to maintain, as far as possible, conceptual, semantic and technical equivalence between the target language and source language versions of the instrument. Conceptual equivalence refers to the same concepts underlying the questions in an instrument in both source and target languages. Semantic equivalence refers to the same denotative and connotative elements of words. Denotation refers to that which is implied by the word, and connotation refers to the emotional meaning of the word. That is to say, what the words indicate or are a sign for (denotation) or what is implied by the words in addition to their emotional meaning (connotation). Technical equivalence refers to two separate but overlapping issues: first, the equivalence of technical features of language and their relationship to the socio-cultural context; and secondly, the feasibility of the nature and mode of questioning of the instrument in both source and target culture.

The linguistic validation of a QOL questionnaire is a complex process which requires the recruitment of professional teams who are familiar with this type of work. The linguistic validation of a questionnaire is not a literal translation of the original questionnaire, but the production of a translation which is conceptually equivalent to the original, and culturally acceptable in the country in which the translation will be used.

In order to work towards an acceptable translation of an instrument in a given language the following points should be adhered to:

- The translation methodology should be adhered to and the different phases of the process should be summarised in a report
- The translated version of a questionnaire — obtained if possible in collaboration with its developer — should be recognised as the official version

in the country concerned. This will avoid the proliferation of "pirate" versions and will help to facilitate the access to translations
– Ideally, a linguistic validation of a QOL questionnaire should be complemented by a psychometric validation of the questionnaire.

3.2. *Methodology*

The original language in which the questionnaire was developed is called **source language.** The language into which the questionnaire is translated is called **target language**.

After the recruitment of a QOL specialist in each country concerned, and having explained the concepts of a linguistic validation in detail, a QOL instrument is then ideally translated according to Table 1.

Thus, in summary, the linguistic validation of a QOL questionnaire comprises 7 steps shown in the first column of Table 1.

The questionnaire should always be considered as a whole (i.e. the response choice could influence the translation of the items and vice verse).

It cannot be assumed that a questionnaire, however, extensively tested in the originating country, will be valid and reliable once it has been translated. No instrument for the assessment of psychological states of subjective

Table 1. Methodology for linguistic validation of a QOL questionnaire.

Steps	Source Questionnaire
1. "Forward" translation by two independent translators	forward version A1 and forward version A2
2. Reconciliation meeting between the 2 "forward" translators and the local project manager	forward version B
3. "Backward translation" by 1 independent translator	backward translation
4. Comparison of the source questionnaire with the "backward" translation by the local team	forward version C
5. Cognitive debriefing	forward version D
6. International harmonisation (if the original is translated into more than 1 language)	final version
7. Report	

perceptions is culture-free. In each instance the validity and other metric characteristics of the instrument must be assessed in the country of application. Important components of psychometric testing in cross-cultural quality of life studies include reliability, validity, responsiveness, and effect size interpretation.

4. Design Issues Relating to QOL Study

4.1. *Study objectives*

Clear study goals are prerequisites to developing appropriate design and analysis strategies that answer clinically relevant questions. Overly general objectives, such as "describe the QOL of..." do not adequately address aspects of study such as the comparison of the two treatment arms, whether the comparisons are limited to the period of therapy or extend across time within a treatment group. Without a focused objective, unnecessary assessments are often included in protocol designs. This increases problems of multiple comparisons and missing data, and increases the possibility that critical assessments will be omitted.

4.2. *QOL instruments*

QOL assessments should ideally be brief, using an uncomplicated and least complicated instrument or combination of instruments that adequately address primary research questions. Adding scales/instruments in order to obtain less relevant data will increase both the multiple comparisons problem and the likelihood that data will be incomplete. This will in turn potentially compromise the ability of the trial to achieve the primary objectives of the study.

4.3. *Timing of assessments*

The timing of QOL assessments must also be specified to achieve the goals of the study. Baseline measures that precede therapy allow for assessment of treatment-related changes within an individual. Depending on the goals of the study, it is also important to have a sufficiently long period of follow-up after therapy to allow for assessment of the long-term treatment effect and potential late sequelae. In the phase 3 treatment comparison setting, it is critical that QOL should be assessed regardless of treatment and disease status. Patients who have changes in status or who have discontinued

treatment should still take part in QOL assessment, as the biggest differences in QOL may be in these patients. Without these measurements it will be difficult to derive summary measures and impossible to make unbiased comparisons of the effects of different therapeutic regimens on QOL. Procedures for obtaining assessments for patients who have changed status or discontinued therapy should be explicitly stated in protocols.

The timing of assessments should be chosen to minimize missing data. It is generally recommended that the frequency of assessments be minimized for ease of patient and staff burden. However, in some cases more frequent administration linked to the clinical routine (e.g. at the beginning of every treatment cycle) may result in more complete data because the pattern of assessment is established as part of the clinical routine.

4.4. *Sample size and power*

The sample size and power to detect meaningful differences for primary QOL hypotheses is critical to any study in which QOL is an important end point. In addition to the usual estimates of variation and correlations, the sensitivity of the QOL instrument to detect clinically significant changes is the most useful information that can be provided during the validation of a QOL instrument. Specific estimates of the changes in subscales and global scales related to clinical status give the statistician and the clinician a clear and familiar reference point for defining differences that clinically relevant. This is critical for insuring an adequate sample size for the study. It should be noted that because end points may involve repeated measurements at different times and/or combinations of subscales, both test-retest correlations and among-subscale correlations are useful and should be reported for validated instruments.

If the sample size requirements for the QOL component are substantially less than for the entire study, an unbiased strategy for selection of a subset of patients in which QOL will be assessed should be identified. For example, the first 500 patients enrolled in the study might be included in the QOL substudy. This may have an additional advantage in studies with a long duration of QOL follow-up. This strategy is being used in the design of an Eastern Cooperative Oncology Group (ECOG) study, in which patient entry is expected to take 5 years, an additional follow-up of 2.5 years is planned for the survival end point, and the desired duration of QOL assessment is 5 years. By limiting the patients in which QOL is assessed to those enrolled in the first 2.5 years, the QOL study is expected

to be complete at the same time as the final analysis of the primary survival end points.

5. Characteristics of QOL Data and Statistical Issues

5.1. *Primary statistical issues*[14]

5.1.1. *Multiple comparisons*

Analysis of QOL data differs from the analysis of other clinical end points data. There are often a large number of measures resulting from both multiple dimensions of QOL (multiple instruments and/or subscales) and repeated assessments over time. Univariate tests for each subscale and time point can seriously inflate the type I error rate (false positive) for the overall trial such that the investigator is unable to distinguish between the true and false positive differences. Furthermore, it is often impossible to determine the number of tests performed at the end of analysis and adjust post hoc. Methods that allow summarization of multiple outcome both simplify the interpretation of the results and often improve the statistical power to detect clinically relevant differences, especially when small but consistent differences in QOL occur over time or across multiple domains. On the other hand, significant differences at a particular time or within a particular domain may be blurred by aggregation.

5.1.2. *Missing data*

Missing data refers to missing items in scales and missed and/or mistimed assessments. If the assessment was not completed for reasons that there are unrelated to the patient's QOL, the data are classified as "missing at random". Examples might be staff forgetting to administer the assessment, a missed appointment due to inclement weather, or the patient having moved out of the area. Data that are missing because the patient had not been on-study long enough to reach the assessment time point (i.e. the data are censored or incomplete) are also considered missing at random. Assessments may be mistimed if they are actually given but the exact timing does not correspond to the planned schedule of assessments for reasons unrelated to the patients' QOL. While these types of missing/mistimed data make analyses more complex and may reduce the power to detect differences, the estimates of QOL are unbiased even if they are based only on the observed QOL assessments.

Non-randomly missing or informatively censored data present researchers with a much more difficult problem. One example of this type of missing data is that due to death, disease progression, or toxicity where the QOL would generally be poorer in the patients who were not observed than in those who were observed. In the chronic disease setting, this relationship between QOL and missing data might manifest itself as study dropout due to lack of relief, presence of side effects, or, conversely, improvement in the condition. The difficulty occurs because analyses that inappropriately assume the data are randomly missing will result in biased estimates of QOL reflecting only the more limited population of patients who were assessed rather than the entire sample of population under study. One possibility is to limit the analysis, and thereby the inference, to patients with complete data. In most cases, however, this strategy is not acceptable to achieve the goal of comparing QOL assessment for all patients. Unless careful prospective documentation of the reasons for missing assessments is available in a clinical trial, it is generally impossible to know definitively whether the reason for the missing assessment is related to the patient's condition and/or to their QOL.

In scales based on multiple items, missing information results in a serious missing data problem. If only 0.1% of items are randomly missing for a 50-item instrument, 18% of the subjects will have one or more items missing over four assessments. If the rate is 0.5%, then only 37% of subjects will have complete data. Deletion of the entire case when there are missing items results in loss of power and potential bias if subjects with poorer QOL are more or less likely to skip an item. Individuals with a high level of non-response ($> 50\%$) should be dealt with on a case-by-case basis. Imputing missing items for an individual who has answered most questions would, in general, be preferable to deletion of the entire case or observation, although the method used for such imputing must be carefully considered. A simple method based solely on the patient's own data would use the mean of all non-missing items for the entire scale or the specific subscale. Methods based on other patients would include the mean of that item in individuals who had responded. Another method utilizing data from other participants is based on the high correlation of items within a scale or subscale and utilizes information about the individual's tendency for particular items to be scored higher or lower relative to other items. The procedure here is to regress the missing item on the non-missing items using data from individuals with complete data, and to then predict the value of

the missing items using the information gained from the items that the individual has completed.

5.1.3. *Integration of QOL and survival data*

In clinical trials with significant disease-related mortality there is need to integrate survival with QOL. This was identified by the participants in the 1990 NCI QOL workshop who "acknowledged that the use of QOL data in clinical decision-making will not routinely occur until a larger body of QOL data is available and models for integrating medical and QOL information are available". In studies where both QOL (or toxicity) and clinical end points indicate the superiority of one treatment over another, the choice of the best treatment is clear. Similarly, if either QOL or the efficacy outcome demonstrates a benefit and there is no significant difference in the other, the choice of treatment is straightforward. The dilemma occurs when there is a conflict between the QOL and efficacy outcomes. This is often the case when there is significant toxicity associated with the more effective treatment.

5.2. *Statistical methods used to analyse QOL data*

5.2.1. *Univariate methods*

One approach to the reporting of QOL data has been descriptive univariate statistics such as means and proportions at each specific point in time. These descriptive statistics may be accompanied by simple parametric or nonparametric tests such as t-tests or Wilcoxon tests. While these methods are easy to implement and often used, they do not address any of the three previously identified issues. One recommended solution to the multiple comparisons problem is to limit the number of *a priori* end points in the design of the trial to three or less. The analyses of the remaining scales and/or time points can be presented descriptively or graphically. While theoretically improving the overall type I error rate for the study, in practice investigators are reluctant to ignore the remaining data and may receive requests from reviewers to provide results from secondary analyses with the corresponding significance level.

An alternative method of addressing the multiple comparisons problem is to apply a Bonferroni correction, which adjusts the test statistics on k end points so that the overall type I error is preserved for the smallest p value. The procedure is to accept as statistically significant only those tests

with p value that are less than α/k where α is the overall type I error usually set equal to 0.05.

5.2.2. *Multivariate methods*

Multivariate analysis techniques include approaches such as repeated measures analysis of variance (ANOVA) or multivariate ANOVA (MANOVA). These techniques require complete data, which limits their use in settings where there is a low risk of mortality and very high compliance with QOL assessment. If the data are not complete, the inferences are restricted to a very select and generally non-representative group of patients. Multivariate statistics such as Hotelling's T are frequently used to control for type I error. These statistics, however, answer global questions such as "are any of the dimensions of QOL different?" or "are there differences in QOL at any point in time?" without considering whether the differences are in consistent directions. In general, the multivariate test statistics are not sensitive to differences in the same direction across the multiple end points.

The requirement for complete data can be relaxed by using repeated measures or mixed effects model with structured covariance. These methods assume that the data are missing for reasons unrelated to the patients QOL, such as staff forgetting to administer the assessment for example. If the missing assessment can reasonably be assumed to be missing at random, a likelihood-based analysis approach, such as mixed-effects models or EM (Estimation-Maximization) algorithm for repeated measures models, incorporates all patients with at least one assessment in the analysis. This approach has the additional advantages of estimation of within- and between-subject variation, inclusion of time varying variables, and of being able to test for significant changes over time.

Other methods often used to determine the risk factors related to QOL include multiple regression, stepwise discriminant analysis, canonical correlation, and Logistic regression.

5.2.3. *Other methods*

5.2.3.1. Quality-Adjusted Life Years (QALY)

An intuitive method of incorporating QOL and time would be to adjust life years by down-weighting time spent in periods of poor QOL. However, what would seem to be a simple idea has many methodological challenges. The first of these challenges is the determination of weights. Torrance[14]

describes several techniques for eliciting weights for states of health including direct ratings, time trade-offs, and standard gambles. In addition to the difficulties of administering some of these techniques in clinical trials, weights elicited by the different techniques or from different respondents may not result in equivalent measures. The choice of anchor points and content validity may mean that weights that are appropriate in one setting may be inappropriate in another. The other methodological difficulty occurs in trials with censored data. Although it might seem appropriate to undertake a standard survival analysis of individual quality-adjusted survival times, the usual product limit estimator of the survival function is biased because censoring is related by the future outcome. For example, if two groups have the same censoring time due to death, the group with the poorer QOL will be censored earlier on the QALY scale. This latter problem can be addressed by estimating the average time spent in each health state and then computing a weighted average of the time as is done in the Q-TwiST approach.

5.2.3.2. Q-TwiST

The objective of the Q-TwiST method is to evaluate therapies based on both quantity and quality of life. Q-TwiST stands for Quality-adjusted Time Without Symptoms of disease and Toxicity of treatment. It is based on the concept of quality-adjusted life years (QALYs) and represents a utility-based approach to QOL assessment in clinical trials. The starting point is to define QOL-oriented clinical health states, one of which represents relatively good health with minimal symptoms of disease or treatment associated toxicity (TWiST). Patients will progress through or skip these clinical health states, but will not back-track. The next step is to partition the area under the overall Kaplan–Meier survival curve and calculate the average time a patient spends in each clinical health state. The final step is to compare the treatment regimens using weighted sums of the mean duration of each health state, where the weights are utility based. If these utility weights are unknown, as is generally the case treatment comparisons can be made using sensitivity analyses, also called threshold utility analyses.

5.2.3.3. Markov and Semi-Markov Models

Markov and Semi-Markov models have been used to compare treatments based on estimates of the time spent in different health states and the probabilities of transitions between these states. The relevant health states

must be identified and then each is weighted to reflect the relative value of a health state compared to perfect health. The treatments are then compared in terms of the total quality-adjusted time, the weighted sum of the health state durations. In general, to calculate the transition probabilities an underlying model must be assumed. The most commonly used model is the Markov chain, which assumes that the transitions from one QOL state to another are independent and continuous and only depend on the previous state. This requires that the assessments are made at time points independent of the patients' treatment schedule or health state. Discrete-time transient semi-Markov processes are used to model the health state transition probabilities corresponding to prolonged life, while a simple recurrent Markov process is used to derive the QOL state transition probabilities. In a semi-Markov process, the state changes from an embedded Markov chain and the times spent in different health states are mutually independent, and depend only on the adjoining states.

5.3. *Conclusions*

We have identified three characteristics of QOL studies that present challenges for analysis and interpretation. The first is the occurrence of random and non-random missing data. The analysis of random missing data is generally well documented with sufficient advice and guidelines for both practical and theoretical issues. In contrast, development of methods for analysis of non-random missing data is in its infancy, and we now require an enhanced knowledge and understanding to determine which methods are most practical and appropriate.

The second issue addressed is the multivariate nature of QOL studies. Not only is QOL a multi-dimensional concept measured by multiple scales, but most studies are longitudinal. Separate analyses of each domain at multiple time points may make it difficult to communicate the results in a manner that is meaningful for clinicians and patients. Summary measures may reduce the multi-dimensionality of the problem but may not make the interpretation much easier. The issue of weights that vary by technique and study also adds to the complexity of interpretation. In general, it would be advisable to perform the analyses using various assumptions to verify that the results are not sensitive to small changes in the assumptions.

The third issue addressed is the integration of survival data with QOL measures. This can be addressed from either the perspective of QOL or from the perspective of time. From a research perspective both approaches

can be informative; however, currently time is the dimension that both clinicians and statisticians are most familiar with. Finally, interpretation of clinical trials may not always be helpful in guiding individual patient decisions. In theory, individual patients could utilize the threshold utility analysis of Q-TWiST, but this may require extensive patient education.

There are a number of statistical methodologies that can be employed in the analysis of QOL data, each of which is based on specific assumptions, yields a different summary measure, and thus emphasizes different aspects of QOL. When there is more than one analysis strategy that best anticipates the above issues should be considered. Analyses should be clearly and concisely reportable so that the relevant differences can be readily understood by those who will use the results.

6. The Validation Process: Psychometric Testing

The question of most concern relating to psychometrics is whether a measures both reliable and valid. Measurement is the process by which a concept is linked to one or more latent variables, and these are linked to observed variables. The concept can vary from one that is highly abstract, such as QOL, or intelligence, to one that is more concrete, such as age, sex, or race. One or more latent variables may be needed to represent the concept. The observed variables can be responses to questionnaire items, census figures, or any other observable characteristics.

The first step of the measurement process is to give the concept a theoretical definition. A theoretical definition explains in as simple and precise terms as possible the meaning of a concept. The second step is to identify the dimensions and latent variables that will represent it. The next step, of forming measures, depends on the theoretical definition. This is sometimes referred to as the operational definition. The operational definition describes the procedures to follow to form measures of the latent variables that represent a concept. In some situations the latent variables are operationalized as the responses to questionnaire items. The fourth step is construct the measurement model. A measurement model specifies a structural model connecting latent variables to one or more measures or observed variables. A simple measurement model for the latent variables influence on the two measures is

$$\begin{aligned} x_1 &= \lambda_{11}\xi + \delta_1\,, \\ x_2 &= \lambda_{21}\xi + \delta_2\,. \end{aligned} \tag{1}$$

where ξ represents the latent variable, x_1 and x_2 are its indicator. δ_1 and δ_2 are errors of measurement with expected values of zero and uncorrelated

with ξ and with each other. All variables are in deviation form so that intercepts terms do not enter the equations.

In sum, the four steps in measurement are to give meaning, identify dimensions and latent variables, to form measures, and to specify a model. The theoretical definition assigns meaning to a term and the concept associated with it. On the basis of this definition, we can know a concept's dimensions. Each dimension is represented by one latent variable. Guided by theoretical definitions, we form measures, and hopefully two or more measures will be formed per latent variable. Finally, we formulate the structural relation between indicators and latent variables in the measurement model. Two important properties of measures are their validity and reliability.

6.1. *Validity*

Validity[15] is concerned with whether a variable measures what it is supposed to measure. For instance, does an IQ test measure intelligence? Does the WHOQOL-100 measure people's quality of life? These are questions of validity. They can never be answered with absolute certainty. Although we can never prove validity, we can develop strong support for it. Traditionally, psychologists have distinguished four types of validity: content validity, criterion validity, construct validity, and convergent and discriminant validity. Each attempts to show whether a measure corresponds to a concept, though their means of doing so differ. Content validity is largely a "conceptual test", whereas the other three types are empirically rooted. If a measure truly corresponds to a concept, we would expect that all four types of validity would be satisfied. Unfortunately, it is possible that a valid measure will fail one or more of these tests or that an invalid measure will pass some of them.

6.1.1. *Content validity*

Content validity is a qualitative type of validity where the domain of a concept is made clear and the analyst judges whether the measures fully represent the domain. To the extent that they do, content validity is met. A key question is, how do we know a concept's domain? For the answer we must return to the first step in the measurement process. That is, to know the domain of a concept, we need a theoretical definition that explains the meaning of a concept. Ideally, the theoretical definition should reflect the meanings associated with a term in prior research so that a general rather an idiosyncratic domain results. In addition the theoretical definition should make clear the dimensions of a concept.

Does it matter if our measures lack content validity? In general, the answer is yes. Just as a nonrepresentative sample of people can lead to mistaken inferences to the population, a nonrepersentative sample of measures can distort our understanding of a concept.

The major limitation of content validity stems from its dependence on the theoretical definition. For most concepts in the social sciences, no consensus exists on theoretical definitions. The domain of content is ambiguous. In this situation the burden falls on researchers not only to provide a theoretical definition accepted by their peers but also to select indicators that fully cover its domain and dimensions. In sum, content validity is a qualitative means of ensuring that indicators tap the meaning of a concept as defined by the analyst.

6.1.2. *Criterion validity*

Criterion validity is the degree of correspondence between a measure and a criterion variable, usually measured by their correlation. To assess criterion validity, we need an objective reliable standard measure with which to compare our measure. Suppose that in a survey we ask each employee in a corporation to report his or her salary. If we had access to the actual salary records, we could assess the validity of the survey measure by correlating the two. In this case employee records represent an ideal, or nearly ideal, standard of comparison.

The absolute value of the correlation between a measure and a criterion sometimes is referred to as the validity coefficient. Does this correlation of a measure and a criterion reveal the validity of a measure? If we represent the measure as x_1 and the criterion as c_1, the validity coefficient may be represent as $\rho_{x_1c_1}$. A simple model of the relation between x_1 and c_1, and the latent variable ξ_1 that they measure appears in the following equations:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \delta_1\,, \\ c_1 &= \lambda_{21}\xi_1 + \delta_2\,, \end{aligned} \tag{2}$$

where δ_1 and δ_2 are uncorrelated with each other and with ξ_1, $E(\delta_1) = E(\delta_2) = 0$.

$$\rho_{x_1c_1} = \frac{\lambda_{11}\lambda_{21}\phi_{11}}{[\mathrm{var}(x_1)\ \mathrm{var}(c_1)]^{1/2}}\,. \tag{3}$$

As Eq. (3) reveals, the magnitude of $\rho_{x_1c_1}$ depends on factors other than the "closeness" of x_1 and ξ_1. This is made clearer if we standardize x_1, c_1,

and ξ_1 to variances of one. In this case :

$$\begin{aligned} \rho_{x_1 c_1} &= \lambda_{11}\lambda_{21}\,, \\ \mathrm{Corr}(x_1, \xi_1) &= \lambda_{11}\,, \\ \mathrm{Corr}(c_1, \xi_1) &= \lambda_{21}\,. \end{aligned} \tag{4}$$

The validity coefficient, $\rho_{x_1 c_1}$, is affected not only by $\rho_{x_1 \xi_1} (= \lambda_{11})$ but also by $\rho_{c_1 \xi_1} (= \lambda_{21})$. Even if the correlation of x_1 with ξ_1 stays at 0.5 the validity coefficient would be 0.45, 0.35, or 0.25 if the correlation of c_1 and ξ_1, is 0.9, 0.7, or 0.5. Thus, even with one change in x_1's association with ξ_1, we obtain different values of validity, depending on the criterion's relation to ξ_1.

In sum, criterion validity as measured by $\rho_{x_1 c_1}$, the validity coefficient, has several undesirable characteristics as a means to assess validity. It is not only influenced by the degree of random measurement error variance in x_1 but also by the error in the criterion. Furthermore different criteria lead to different "validity coefficient" for the same measure, leaving uncertainty as to which is an accurate reading of a measure's validity. Finally, for many measures no criterion is available.

6.1.3. *Construct validity*

Construct validity is a third type of validity. Many concepts within the social science are difficult to defined and formulated, and so content validity is difficult to apply. As mentioned earlier, appropriate criteria for some measures often do not exist. This prevents the computation of criterion validity coefficients. In these common situations construct validity is used instead.

Construct validity assesses whether a measure relates to other observed variables in a way that is consistent with theoretically derived predictions. Hypotheses may suggest positive, negative, or no significant associations between constructs. If we examine the relation between a measure of one construct to other observed variables indicating other constructs, we expect their empirical association to parallel the theoretically specified associations. To the extent that they do, construct validity exists.

The major steps in the process begin with postulating theoretical relations between constructs. Then the associations between measures of the constructs or concepts are estimated. Based on these associations, the measures, the constructs, and the postulated associations are re-examined.

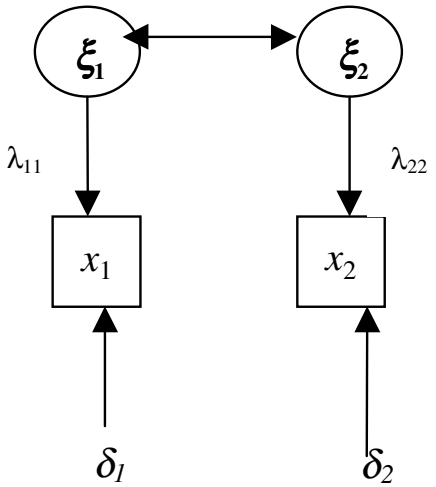


Fig. 1. Two constructs with one measure each.

Some of the difficulties with construct validity can be illustrated with a structural equation approach. As a simple example, consider Fig. 1. Assuming two constructs, ξ_1 and ξ_2. Each has one measure represented as x_1 and x_2. As usual δ_1 and δ_2 are random errors of measurement with expected values of 0, uncorrelated with each other and with ξ_1 and ξ_2. Suppose that the construct validity of x_1 is of interest. We hypothesize that the two constructs (ξ_1 and ξ_2) are positively correlated ($\phi_{12} > 0$). To test construct validity, we would compute the correlation between x_1 and x_2.

$$\rho_{x_1x_2} = (\rho_{x_1x_1}\rho_{x_2x_2})^{1/2}\rho_{\xi_1\xi_2}\,, \tag{5}$$

where $\rho_{x_ix_i}$ is the reliability of x_i. It is the squared correlation between x_i and ξ_i. The correlation of the two observed variables depends not only on the correlation of x_1 and ξ_1 but also on the correlation between the constructs ξ_1 and ξ_2 and the correlation of x_1 and x_2. Because of this, the interpretation of construct vability based on $\rho_{x_1x_2}$ is seriously complicated. For instance, if the correlation between ξ_1 and ξ_2 is relatively large and that x_1 has very high reliability but x_2 has low reliability. This would reduce $\rho_{x_1x_2}$, raising doubts about the construct validity of x_1.

In practical work, people usually use exploratory factor analysis or confirmatory factor analysis to test for construct validity. Confirmatory factor analysis is preferable than exploratory factor analysis, because its principle is similar to the definition of construct validity.

6.1.4. *Convergent and discriminant validity*

Convergent validity is another important aspect of construct validity, which is intended to show that for example, a postulated dimension of QOL correlates appreciably with all other dimensions that theory suggests should be related to it. That is, we may believe that some dimensions of QOL are related, and we therefore expect the observed measurements to be correlated. For example, one might anticipate that patients with severe pain are likely to be depressed, and that there should be a correlation between pain scores and depression ratings within group.

Many of the dimensions of QOL are interrelated. Very ill patients tend to suffer from a variety of symptoms, and have high scores on a wide range of psychological dimensions. As many dimensions of QOL are correlated with each other, assessment of convergent validity consists of predicting the strongest and weakest correlations, and confirming that subsequent observed values conform to the predictions. Analysis involves calculating all pairwise correlation coefficients between scores for different QOL scales.

Discriminant validity, or divergent validity, recognises that some dimensions of QOL are anticipated to be relatively unrelated, and that their correlations should be low. Convergent and discriminant validity represent the two extremes in a continuum of associations between dimensions of QOL. One problem when assessing discriminant validity (and to a lesser extent, convergent validity) is that two dimensions may correlate spuriously because of some third, possibly unrecognised, construct that links the two together. For example, if two dimensions are both affected by age, an apparent correlation can be introduced solely though the differing ages of the respondents. Another extraneous source of correlation could be that of social desirability, where patients may report a higher QOL on many dimensions simply to please staff or relative. When specific independent variables are suspected of introducing spurious correlations, the statistical technique of "partial correlation" should be used. This is a method of estimating the correlation between two variables, or dimensions of QOL, whilst holding other "nuisance" variables constant. In practice, there are usually many extraneous variables that contribute a little to the spurious correlations obtained.

Convergent validity and discriminant validity are commonly assessed across instruments. For convergent validity to exist, those scales from each instrument that are intended to measure similar constructs should have higher correlations with each other than with scales that measure unrelated constructs.

Table 2. Template for the multitrait-multimethod (MTMM) correlation matrix.

		Emotional function		Social function		Role function	
	Instrument	1	2	1	2	1	2
Emotional	1	R					
function	2	C	R				
Social	1	D		R			
function	2		D	C	R		
Role	1	D		D		R	
function	2		D		D	C	R

The multitrait-multimethod (MTMM) correlation matrix is a method for examining convergent and discriminant validity. The general principle of this technique is that two or more methods, such as different instruments, are each used to assess the same traits, for example QOL aspects, items or subscales as estimated by the different methods. Various layouts are used for MTMM matrices, the most common being shown in Table 2.

In Table 2, the two instruments are methods, while the functioning scales are traits. Cells marked C show the correlations of the scores when different instruments are used to assess the same trait. Convergent validity is determined by the C cells. If the correlations in these cells are high, say above 0.7, this suggests that both instruments may be measuring the same thing. If the two instruments were developed independently of each other, this would support the inference that the traits are defined in a consistent and presumably meaningful manner.

Similarly, the D cells show the scale-to-scale correlations for each instrument, and these assess discriminant validity. Lower correlation are usually expected in these cells, because otherwise scales purporting to measure different aspects of QOL are in fact more strongly related than supposedly similar scales from different instruments. The main diagonal cells, marked R, can be used to show reliability coefficients, as described later. These can be either Cronbach's α for internal reliability or, if repeated QOL assessments are conducted on patients whose condition is stable, test-retest correlations. Since repeated values of the same trait measured twice by the same method will usually be more similar than values of the same trait measured by different instruments, the R cells containing test-retest repeatability scores should usually contain the most significant correlations.

One common variation on the theme of MTMM matrices is to carry out the patient assessments on two different occasions. The upper-right

triangle of Table 2 can be used to display the correlations at time 1, and the correlations at time 2 can be shown the lower-left triangle that we have been describing above. The diagonal cells dividing the two triangles, marked R, should then show the test-retest repeatability correlations.

6.1.5. *Alternatives to classical validity measures*

Thus far we have reviewed four common types of validity: content, criterion, construct, and convergent and discriminant validity. Content validity is largely a theoretical approach to validation. Criterion validity is largely an empirical means of validating. Construct validity and convergent-discriminant validity are both theoretical and empirical. They are theoretical in the sense that theory suggests which constructs should correlate and which should not. The empirical aspect concerns the correlations of observed measures. The empirical aspect concerns the correlations between measures, although there are a number of limitations associated with this. One problem is that they rely on correlations rather than structural coefficients to test validity. Criterion validity examines the correlation between the criterion and the observed measure. Construct validity and convergent-discriminant validity are based on the correlation between measures of the same and different constructs. These correlations may have little to do with the validity of a measure. A second problem with these empirical tests is that they use only observed measures, rather than incorporating the latent variables into the analysis. The implicit assumption is that the correlation between two observed variables mirrors an association involving latent variables, so it is implicitly assumed that the correlation of the criterion and the measure adequately approximates the correlation between the latent variable and the measure. In construct and convergent-discriminant validities the correlation of observed measures is a proxy for the correlation of the latent constructs. But in fact, it can be a poor proxy under a number of conditions.

To overcome these limitations, Bollen[15] proposed an alternative definition that based on a structural equation approach. In his definition, the validity of a measure x_i of ξ_j is the magnitude of the direct structural relation between ξ_j and x_i. Therefore, for a measure to be valid, the latent and observed variable must have a direct link. Using this approach, a natural question is how to measure validity based on it? There is probably no one ideal measure of validity, but several correspond to this theoretical definition.

6.1.5.1. Unstandardized Validity Coefficient (λ)

One important gauge of validity, the direct structural relation between an and x_i and ξ_j, is λ_{ij} the unstandardized coefficient linking them. For instance,

$$x_1 = \lambda_{11}\xi_1 + \delta_1\,, \tag{6}$$

where λ_{11} is the unstandardized coefficient, it provides the expected change in x_1 for a one-unit change in ξ_1. The λ_{ij} coefficients are in the Λ_x and Λ_y matrices.
As in multiple regression x_i may have a number of explanatory variables. Consider the following measurement model:

$$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \lambda_{13}\xi_3 + \delta_1\,. \tag{7}$$

The validity of x_1 with respect to ξ_1 is indicated by λ_{11}. The λ_{11} coefficient is interpreted as the expected change in x_1 for one-unit change in ξ_1, holding constant ξ_2 and ξ_3. In addition the validity of x_1 with respect to ξ_2 and ξ_3 can be gauged by λ_{12} and λ_{13} respectively. Thus the unstandardized validity coefficient λ_{ij} is appropriate for measures that depend on one or more latent variables.

The unstandardized validity coefficient λ_{ij} is also useful for comparing samples from different populations. For example, the same observed variable may be measured in samples of males and females, samples from two different countries, or samples of some other groups. A comparison of validity could be made by comparing the corresponding $\hat{\lambda}_{ij}$ coefficients in the separate samples. They represent a better measure of the structural relation of the variables, and are less influenced by differences in population variances.

One disadvantage in comparing the unstandardized validity coefficients of measures that depend on the same latent variable is that the observed variables may be measured on very different scales. Direct comparison of the magnitude of λ's to determine the relative validity of measures generally is not appropriate.

6.1.5.2. The Standardized Validity Coefficient, λ^s

The standardized validity coefficient λ^s is defined as

$$\lambda^s_{ij} = \lambda_{ij}\left[\frac{\phi_{jj}}{\mathrm{var}(x_i)}\right]^{1/2}\,, \tag{8}$$

where ϕ_{jj} is the variance of latent variable ξ_j.

Unlike λ_{ij}, λ^s_{ij} is one means to compare the relative influence of ξ_j on several x_i variables. For example, if x_1 and x_2 depend on ξ_j and λ^s_{1j} is 0.8 and λ^s_{2j} is 0.1, this would indicate that x_1 is more responsive to ξ_j than is x_2 in standard deviation units. In addition, if x_i depends on two or more latent variables the relative influence of the latent variables can be compared. The standardized λ^s_{ij} is less useful than λ_{ij} in comparing different populations because it is greatly influenced by the varying standard deviations of the variables in different populations.

6.1.5.3. Unique Validity Variance, $U_{x_i\xi_j}$

The unique validity variance measures that part of explained variance in x_i that is uniquely attributable to ξ_j. The formular for $U_{x_i\xi_j}$ is

$$U_{x_i\xi_j} = R^2_{x_i} - R^2_{x_i(\xi_j)}\,, \tag{9}$$

where $R^2_{x_i}$ is the squared multiple correlation coefficient or proportion of variance in x_i explained by all variables in a model that have a direct effect on x_i (excluding error terms) and $R^2_{x_i(\xi_j)}$ is the proportion of explained variance in x_i by all variables with a direct effect on x_i excluding ξ_j.

$U_{x_i\xi_j}$ always varies between zero and one. If only ξ_j has a direct effect on x_i, $U_{x_i\xi_j}$ equals the squared correlation between ξ_j and ξ_i. $U_{x_i\xi_j}$ is more general than $\rho^2_{x_i\xi_j}$ since it allows the observed variable to depend on more than one latent variable and it is zero if ξ_j has no direct effect on x_i. If multiple correlated latent variables underlie x_i, $U_{x_i\xi_j}$ will generally not equal $\rho^2_{x_i\xi_j}$ unless the latent variables are uncorrelated.

6.2. *Reliability*

Reliability is the consistency of measurement. It is not the same as validity since we can have consistent but invalid measures. To illustrate reliability, suppose that I wish to measure your level of education. I narrowly define education as completed years of formal schooling. I operationalize it by asking:“ How many completed years of formal schooling have you had?” Next, I record your answer. If I had the ability to erase your memory of the question and the response you gave, I could repeat the same question and again, record your answer. Repeating this process an infinite number of times, I could determine the consistency of your response to the same question. The reliability of this education measure is the consistency in your response over the infinite trials. The greater the fluctuation across your answers, the lower the reliability of the measure.

It is possible to have a very reliable measure that is not valid. For example, repeatedly weighing yourself on a bathroom scale may provide a reliable measure of your weight but the scale is not valid if it always gives a weight that is 5 kg too light. A more extreme example would be obtaining a measure of intelligence by asking individuals their shoes size. This may provide a very reliable measure, but it lacks validity as an intelligence measure. Thus the distinction between reliability and validity is a very important one.

Much of the social science literature on reliability originates in classical measurement theory from psychology. A fundamental equation of the theory is

$$x_i = \tau_i + e_i \,, \tag{10}$$

where x_i is the ith observed variable (or "test" score), e_i is the error term and τ_i is the true score that underlies x_i. It is assumed that $\mathrm{cov}(\tau_i, e_i)$ is zero and that $E(e_i) = 0$. According to classical test theory, the errors of measurement for different items are uncorrelated. The correlation between two measures results from the association of their true scores. Thus the true scores are the systematic components that lead to the association of observed variables.

Parallel, τ-equivalent, and congeneric measures are the three major types of observed variables in test theory. They can be defined using two measures x_i and x_j as shown in the example below:

$$\begin{aligned} x_i &= \alpha_i \tau_i + e_i \,, \\ x_j &= \alpha_j \tau_j + e_j \,. \end{aligned} \tag{11}$$

The e_i and e_j are uncorrelated. Assume that the true scores are the same. If $\alpha_i = \alpha_j = 1$, $\mathrm{var}(e_i) = \mathrm{var}(e_j)$, then x_i and x_j are parallel measures. If $\alpha_i = \alpha_j = 1$, $\mathrm{var}(e_i) \neq \mathrm{var}(e_j)$, the measures are τ-equivalent. Finally, if $\alpha_i \neq \alpha_j$, $\mathrm{var}(e_i) \neq \mathrm{var}(e_j)$, then the measures are congeneric. Congeneric measures are the most general of the three types.

The reliability of a measure $\rho_{x_i x_i}$ is defined as

$$\rho_{x_i x_i} = \frac{\alpha_i^2 \, \mathrm{var}(\tau_i)}{\mathrm{var}(x_i)} \,. \tag{12}$$

For τ-equivalent or parallel measures, this simplifies to

$$\rho_{x_i x_i} = \frac{\mathrm{var}(\tau_i)}{\mathrm{var}(x_i)} \,. \tag{13}$$

Reliability is the ratio of true score's variance to the observed variable's variance. It is equals to the squared correlation of the observed variable and the true score:

$$\begin{aligned}\rho^2_{x_i\tau_i} &= \frac{[\mathrm{cov}(x_i,\tau_i)]^2}{\mathrm{var}(x_i)\ \mathrm{var}(\tau_i)}\\ &= \frac{\alpha_i^2[\mathrm{var}(\tau_i)]^2}{\mathrm{var}(x_i)\ \mathrm{var}(\tau_i)}\\ &= \frac{\alpha_i^2\ \mathrm{var}(\tau_i)}{\mathrm{var}(x_i)}\\ &= \rho_{x_ix_i}\,. \end{aligned} \tag{14}$$

Thus, $\rho_{x_ix_i}$ can be interpreted as the variance of x_i that is explained by τ_i with the remaining variance due to error.

A number of methods have been proposed for estimating the reliability of measures. Here will review the four most common: test-retest, alternative forms, split-halves, and Cronbach's α.

6.2.1. *Test-retest method*

The test-retest method is based on administering the same measure for the same observations at two points in time. The equations for the two measures are

$$\begin{aligned} x_t &= \alpha_t\tau_t + e_t\,,\\ x_{t+1} &= \alpha_{t+1}\tau_{t+1} + e_{t+1}\,, \end{aligned} \tag{15}$$

where t and $t+1$ are subscripts representing the first and second time periods for the x, α, τ and e. Here it is assumed that $E(e_t) = E(e_{t+1}) = 0$, that the true scores (τ_t, τ_{t+1}) are uncorrelated with errors (e_t, e_{t+1}), and that the errors are uncorrelated. In addition this method assumes that x_t, x_{t+1} are parallel measures and that the true scores are equal.
The reliability estimate is the correlation of x_t and x_{t+1}. Using the definition of the correlation between two variables and covariance algebra leads to

$$\rho_{x_tx_{t+1}} = \frac{\mathrm{cov}(x_t, x_{t+1})}{[\mathrm{var}(x_t)\ \mathrm{var}(x_{t+1})]^{1/2}} = \frac{\mathrm{var}(\tau_t)}{\mathrm{var}(x_t)} = \rho_{x_tx_t}\,. \tag{16}$$

In fact, the correlation of any two parallel measures equals their reliability since all parallel measures have identical reliability.

Despite the intuitive appeal of the test-retest reliability technique, it has several limitations. First, it assumes perfect stability of the true score. In many cases the true score may change over time so that this assumption is not reasonable. If lack of equivalence of true scores is the only violated assumption, then $\rho_{x_t x_{t+1}}$ is less than the reliability. Secondly, memory effects are sometimes present. People's memories of response during the first interview can influence their response in a second interview. They may have the tendency to give the same responses.

In short, the test-retest method of estimating reliability has the advantage of simplicity, but it is dependent on assumptions that are unrealistic in practice.

6.2.2. *Alternative forms*

Another method for estimating reliability is that of alternative forms. This is similar to the test-retest method, except that different measures instead of the same measure are collected at t and $t+1$. The equations for the two measures are

$$\begin{aligned} x_1 &= \tau_t + e_t\,, \\ x_2 &= \tau_{t+1} + e_{t+1}\,. \end{aligned} \tag{17}$$

The x_1 variable is a measure of τ at time t, x_2 is a different measure at $t+1$, and x_1 and x_2 are parallel measures. Like the test-retest method it is assumed that τ_t equal τ_{t+1}, that the expected value of e_t and e_{t+1} are zero, and that the errors are uncorrelated with each other and with τ_t and τ_{t+1}. With these assumptions the correlation between x_1 and $x_2(\rho_{x_1,x_2})$ equals the reliability of both measures.

The alternative form does have two advantages. One is that compared to the test-retest, the alternative form measures are less susceptible to memory effects since time t and $t+1$ have different scales. Second, the errors of measurement for one indicator are less likely to correlate with a new measure at the second time period. Compared to test-retest, correlated errors of measurement are less likely to happen. Although the alternative forms estimate of reliability overcomes some of the limitations of the test-retest approach, several unrealistic assumptions remain there. For example, it is assumed that τ_t is still equal to τ_{t+1}. The assumption that the error variances are equal is less likely since x_1 and x_2 are different measures, that are administered at different time points.

6.2.3. *Split-halves*

A third means to estimate reliability is with split-halves. The split-halves method assumes that a number of items are available to measure τ. Half of these items are combined to form a new measure, say, x_1, and the other half to form x_2. Note that in contrast to the test-retest and alternative form, x_1 and x_2 are measures of τ in the same time period. It is still assumed that $E(e_1) = E(e_2) = 0$, $\operatorname{cov}(e_1, e_2) = 0$, $\operatorname{cov}(\tau_1, e_1) = \operatorname{cov}(\tau_1, e_2) = 0$, and that x_1 and x_2 are parallel measures. The equations for x_1 and x_2 are

$$\begin{aligned} x_1 &= \tau_1 + e_1\,, \\ x_2 &= \tau_1 + e_2\,. \end{aligned} \tag{18}$$

The correlation of x_1 and x_2 equals to

$$\rho_{x_1x_2} = \frac{\operatorname{cov}(x_1, x_2)}{[\operatorname{var}(x_1)\ \operatorname{var}(x_2)]}^{1/2} = \frac{\operatorname{var}(\tau_1)}{\operatorname{var}(x_1)} = \rho_{x_1x_1} = \rho_{x_2x_2}\,. \tag{19}$$

In many cases the unweighted sum of two halves forms a composite to measure τ_1 so that the reliability of $x_1 + x_2$ may be determined. As demonstrated earlier, in general the squared correlation of τ_1 with observed score represents the reliability of a measure. Employing this notion, the squared correlation of τ_1 with $x_1 + x_2$ is

$$\begin{aligned} \rho^2_{\tau_1(x_1+x_2)} &= \frac{[\operatorname{cov}(\tau_1, x_1 + x_2)]^2}{\operatorname{var}(\tau_1)\ \operatorname{var}(x_1 + x_2)} \\ &= \frac{4[\operatorname{var}(\tau_1)]^2}{\operatorname{var}(\tau_1)[(\operatorname{var}(x_1) + \operatorname{var}(x_2) + 2\operatorname{cov}(x_1, x_2)]} \\ &= \frac{2\operatorname{var}(\tau_1)/\operatorname{var}(x_1)}{\operatorname{var}(\tau_1)/\operatorname{var}(x_1) + \operatorname{var}(x_1)/\operatorname{var}(x_1)} \\ &= \frac{2\rho_{x_1x_1}}{1 + \rho_{x_1x_1}}\,. \end{aligned} \tag{20}$$

This formula is well known as the Spearman-Brown Prophey formula for gauging the reliability of a full test based on split-halves.

The split-halves test has several aspects more desirable than the test-retest and alternative forms methods. For one, the split-halves method does not assume perfect stability of τ since τ is only gauged in one time period. Secondly, the memory effects that can occur if the same item is asked at two points in time do not operate with this approach. Third, the correlated errors of measurement that are likely in test-retest approaches are less likely for split-halves. A practical advantage is that split-halves are often cheaper and more easily obtained than overtime data.

One disadvantage is that the split-halves must be parallel measures. Often we cannot know whether the variance of the measurement errors are equal, or whether α_1 and α_2 are equal to one. Another drawback is the way that the halves are allocated is somewhat arbitrary. There are many possible ways of dividing a set of items in half, and each split could lead to a different reliability estimate.

6.2.4. *Cronbach's α coefficient*

Cronbach's α coefficient overcomes some of the disadvantages of the split-halves method. The Coefficient α is the most popular reliability coefficient in social science research. It measures the reliability of a simple sum of τ-equivalent or parallel measures. For α, the observed variables $x_1, x_2, \ldots, x_q$ are summed. The x_i's should be scored so that they are all positively or all negatively related to τ_1. I will call this index H so that $\sum_{i=1}^{q} x_i = H$. The squared correlation of τ_1 and H or the reliability of H is

$$\begin{aligned}
\rho^2_{\tau_1 H} &= \frac{[\operatorname{cov}(\tau_1, H)]^2}{\operatorname{var}(\tau_1)\ \operatorname{var}(H)} \\
&= \frac{[\operatorname{cov}(\tau_1, x_1 + x_2 + \cdots + x_q)]^2}{\operatorname{var}(\tau_1)\ \operatorname{var}(H)} \\
&= \frac{[\operatorname{cov}(\tau_1, q\tau_1 + \sum_{i=1}^{q} e_i)]^2}{\operatorname{var}(\tau_1)\ \operatorname{var}(H)} \\
&= \frac{[q\ \operatorname{var}(\tau_1)]^2}{\operatorname{var}(\tau_1)\ \operatorname{var}(H)} \\
&= \frac{q^2\ \operatorname{var}(\tau_1)}{\operatorname{var}(H)} \\
&= \rho HH\,. \qquad (21)
\end{aligned}$$

This equation provides a general formula for the reliability of the unweighted sum of q τ-equivalent or parallel measures. As the next equation shows, this can be manipulated so that it appears as the typical formula for Cronbach's α:

$$\begin{aligned}
\rho HH &= \frac{q^2\ \operatorname{var}(\tau_1)}{\operatorname{var}(H)} \\
&= \frac{q(q-1)q\ \operatorname{var}(\tau_1)}{(q-1)\ \operatorname{var}(H)}
\end{aligned}$$

$$= \left(\frac{q}{q-1}\right)\left(\frac{q^2 \text{ var}(\tau_1) - q \text{ var}(\tau_1)}{\text{var}(H)}\right)$$

$$= \left(\frac{q}{q-1}\right)\left(\frac{q^2 \text{ var}(\tau_1) + \sum_{i=1}^{q} \text{var}(e_i) - q \text{ var}_(\tau_1) - \sum_{i=1}^{q} \text{var}(e_i)}{\text{var}(H)}\right)$$

$$= \left(\frac{q}{q-1}\right)\left(\frac{\text{var}(H) - [q \text{ var}(\tau_1) + \sum_{i=1}^{q} \text{var}(e_i)]}{\text{var}(H)}\right)$$

$$= \left(\frac{q}{q-1}\right)\left(1 - \frac{\sum_{i-1}^{q} \text{var}(x_i)}{\text{var}(H)}\right). \tag{22}$$

With these features the advantages of α over the other reliability measures should be evident. There are no assumptions needed for the stability of τ_1. The measures need not be parallel. The possibility of memory effects are remote since measures for only one time period are applied. There is no problem in selecting splits of items for testing since all measures can be treated individually. In addition, computation of α is relatively easy. However, two drawbacks to α are that it underestimates reliability for congeneric measures, and it is not suited to work with single indicators.

Measurement is a broad topic in social science research. This section emphasized the issues of measurement most relevant to a structural equations approach to measurement models. Most basic is the need to begin with a clear definition of the concepts to be measured. Without such a definition, we have little hope of identifying dimensions and latent variables. Validity and reliability are two basic characteristics of measures. Validity refers to the direct correspondence between a measure and a concept. Reliability refers to the consistency of a measure, regardless of whether it is valid. Many researchers have proposed empirical techniques to estimate validity and reliability. These often are based on correlation coefficients and restrictive assumptions about the properties of measures. Several alternative means have been shown here, that are more general than the traditional procedures, and they also fit well into a structural equations approach to measurement.

References

1. World Health Organization (1991). *World Health Statistics Annual*. WHO, Geneva.
2. Bergner, M., Bobbitt, R. A. and Carter, W. B. *et al* (1981). The sickness impact profile: Development and final revision of a health status measure. *Medical Care* **19**: 787–805.

3. Hunt, S. M., McKenna, S. P. and McEwan, J. (1989). *The Nottingham Health Profile. Users Manual*, Revised edition.
4. Ware, J. E., Snow, K. K., Kosinski, M. and Gandek, B. (1993). *SF-36 Health Survey*: *Manual and Interpretation Guide*. New England Medical Center, MA, USA.
5. Fallowfield, L. (1990). *The Quality of Life*: *The Missing Measurement in Health Care*, Souvenir Press.
6. The WHOQOL Group (1998). The World Health Organization Quality of Life Assessment (WHOQOL): Development and general psychometric properties. *Social Science and Medicine* **12**: 1569–1585.
7. Fayers, P. M. and Machin, D. (2000). *Quality of Life*: *Assessment*, *Analysis and Interpretation*. New York: John Wiley and Sons.
8. Hunt, S. M., McKenna, S. P. and McEwen, J. *et al.* (1981). The Nottingham Health Profile: Subjective health status and medical consultations. *Sosial Science and Medicine* **15A**: 221–229.
9. Ware, J. E. Jr, Snow, K. K. and Kosinski, M. *et al.* (1993). *SF-36 Health Survey Manual and Interpretation Guide*, New England Medical Centre, Boston, MA.
10. Brooks, R. and with the EuroQol group (1996). EuroQol: The current state of play. *Health Policy* **37**: 53–72.
11. Hickey, A. M., Bury, G. and O'Boyle, C. A. *et al.* (1996). A new short-form individual quality of life measure (SEIQol-DW): Application in a cohort of individuals with HIV/AIDS. *British Medical Journal* **313**: 29–33.
12. Bjordal,
K., Hammerlid, E. and Ahlner-Elmqvist, M. *et al.* (1999). Quality of life in head and neck cancer patients: Validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. *Journal of Clinical Oncology* **17**: 1008–1019.
13. Spilker, B. (ed.) (1996) *Quality of Life and Pharmacoeconomics in Clinical Trials*. Lippincott Williams and Wilkins, Philadelphia.
14. Torrance G. W. (1986). Measurement of health state utilities for economic appraisal: A review. *Journal Health Econometrica* **5**: 1–30.
15. Bollen, K. A. (1989). *Structural Equations with Latent Variables*, John Wiley and Sons, NY.

About the Author

Jiqian Fang, born in Shanghai 1939, obtained BS in Mathematics (1961) from the Fudan University and PhD in Biostatistics (1985) from the University of California at Berkeley. From 1985 to 1990, he was Professor and Director, the Department of Biostatistics and Biomathematics, Beijing Medical University; Since 1991, he has been Director and Chair Professor,

Department of Medical Statistics, School of Public Health, Sun Yat-Sen University.

His research projects covers widely in various fields, including "Stochastic Models of Life Phenomena", "Gating Dynamics of Ion Channels", "Biostatistical Theory and Methods for Research on Cancer Prevention", "Bootstrap Studies on Multi-state Models", "Statistical Methods for Data on Quality of Life", "Health and Air Pollution", "Analysis of DNA Finger Printing", and "Linkage Analyses between Complex Trait and Multiple Genes", etc. Some of them have received awards from the Beijing Municipal Government or Ministry of Public Health of China for their significant advances in the biostatistics fields.

This page intentionally left blank

CHAPTER 7

META-ANALYSIS

XUYU ZHOU

Medical Information Institute, Sun Yat-Sen University,
74 Zhongshan Road II, Guangzhou 510080, PR China

JIQIAN FANG,* CHUANHUA YU and ZONGLI XU

Department of Medical Statistics, School of Public Health,
Sun Yat-Sen University, 74 Zhongshan Road II,
Guangzhou 510080, PR China
*Tel: 86-20-87330671; *fangjq@gzsums.edu.cn*

YING LU

Department of Radiology, University of California,
3333 California Street, Suite 375,
San Francisco, CA 94118, USA
Tel: 415-502-4596; ying.lu@radiology.ucsf.edu

The best possible synthesis of available information is essential for medical researchers, health policy-makers, clinicians and other decision makers. With the explosion of information in the literature, literally hundreds of studies may exist on the same topics, and the designs, participants, outcomes, sample sizes, and interventions among these studies may differ. How can information derived from those studies be combined to arrive at a general conclusion? During the past 20 years, meta-analysis, a statistical procedure for systematically combining and analyzing the results of previous research, has been applied with increasing frequency to health-related contexts, especially in the fields of clinical trials.

1. Introduction

1.1. *Definition*

The term "meta-analysis" was coined by psychologist Glass in 1976.[1] The prefix "meta" has several related meanings, including the ideas of occurring after something else, of transcending, or of being more comprehensive than the precursor. Glass' first definition of meta-analysis is the statistical analysis of a large collection of analyses results from individual studies for the purpose of integrating the findings. A useful definition was given by Huque: "...the term 'meta-analysis"' refers to a statistical analysis which combines or integrates the results of several independent clinical trials, considered by the analyst to be 'combinable'.[2]" Similar synonyms of meta-analysis include "overview", "quantitative review", "quantitative synthesis", and "pooling". But these alternative terms may be less specific or less poignant, and were not accepted broadly.

More recently, Evidence-Base Medicine (EBM) has been greatly developed. EBM, systematic review, and meta-analysis get widely used terms in medical journals. Systematic review denotes any type of review that has been prepared using strategies to avoid bias and that which includes a material and methods section. Systematic review may or may not include formal meta-analysis. The Cochrane Collaboration aims to prepare, maintain, and disseminate comprehensive and systematic reviews of the effects of health care. Systematic reviews provided by Cochrane Collaboration are regarded as the best evidence for practicing EBM.[3] Nowadays, meta-analysis is not limited to a statistical approach, and defined as a systematic approach to identifying, appraising, synthesizing, and (if appropriate) combining the results of relevant studies to arrive at conclusions about the body of research.[4]

1.2. *Historical notes*

The origins of pooling the results may be traced to statistician Karl Pearson in 1904, who was the first researcher to report the use of formal techniques to combine data from different samples. The first article which quantitatively synthesized the previous research in medicine, *The Powerful Placebo*, and written by Beecher, was published in 1955.[5] As a formal statistical technique to combine data from studies for the same topic, meta-analysis began to be applied to social sciences in the mid-1970s, particularly in educational and psychological research.

Widespread use of meta-analysis in medicine quickly followed its popularization in the social sciences, and mainly focused the research on the randomized clinical trials. In the late 1980s, there has been a rapid growth in interest and use of the method. At that time, descriptions of the method of meta-analysis and guidelines for its application appeared almost simultaneously in many general influential medical journals, such as the *New England Journal of Medicine*, *Lancet*, and *Annals of Internal Medicine*. Meta-analysis has been adopted by MEDLINE as a Medical Subject Heading (MeSH) term in 1989 and as a sort of Publication Type (PT) in 1993. Meta-analysis of observational studies has also been advocated.

Meta-analysis is now commonplace in a wide range of medical research contexts. Concurrent with the increased number of articles using meta-analysis in the last decade, there have been numerous articles relating to statistical issues or concerns. Many methods have been proposed and used, from crude "vote counting" of studies showing significant or non-significant results, through method for combination of effect size estimates based on fixed or random-effects models, to general linear mixed models and Bayesian methods. Meta-analysis has established itself as an influential branch of biostatistics.

With the sharp increasing use of meta-analysis, several unresolved issues concerning meta-analysis still remain. Incomplete or un-standardized reporting of results, and combing "apples and oranges and the occasional lemon" — failure to make allowance for varying nature and quality of the studies reviewed.[6] Therefore, both the uncritical synthesis of data from observational studies and the unconsidered synthesis of disparate results from randomized controlled trials can threaten to damage the validity and reliability of conclusions of meta-analysis. Other stubborn problems involved in meta-analysis may be biases, especially publication bias, and heterogeneity across studies.

1.3. *Objectives of meta-analysis*

Traditionally, research synthesis was done in a fairly simple way. The classic narrative reviews have several disadvantages that meta-analysis appear to overcome. The traditional review is a subjective method of summarizing research data and therefore prone to bias and error. Without guidance by formal rules, a narrative review expresses the personal opinions of their

authors and depends heavily on the perspicacity and personal experience of the reviewer. Selective inclusion of studies that support the reviewer's view is common. On the other hand, a narrative review tends to present a series of effect measures in the narrative in most situations, and reviewers potential to ignore the factors that greatly influence the results of primary study, such as research design, sample size, and effect size. Meta-analysis provides a logical framework to research a review: Similar measures from comparable studies are listed systematically and the available effect measures are combined where possible.

For example, in 1982, use of thrombolytic agents after acute myocardial infarction was controversial. Table 1 presents the data of eight randomized clinical trials at that time, which examined the effects of a loading dose of at least 250,000 international units of intravenous streptokinase on mortality given a short time after an acute myocardial infarction had occurred. As shown in Table 1, two trials showed a higher risk of mortality in treated patients, with both 95% confidence intervals covering one, which means no statistical significance; five showed a lower risk, with four of those 95% confidence intervals covering one; and one showed same mortality rate in the treated and the control patients. The trials were all fairly small, and the difference in mortality between treated and controlled patients was

Table 1. Results of randomized trials of effect on mortality of intravenous streptokinase following acute myocardial infarction published before 1982.

	N Deaths/Total		Mortality (%)		Estimated relative
Included Study	Treated	Control	Treated	Control	Risk and its 95% CI
Avery (1969)	20/83	15/84	24.1	17.9	1.35(0.74–2.45)
European Working Party (1971)	69/373	94/357	18.5	26.3	0.70(0.53–0.92)
Heikinheimo (1971)	22/219	17/207	10.0	8.2	1.22(0.67–2.24)
Dioguardia (1973)	19/164	18/157	11.6	11.5	1.01(0.55–1.85)
Breddin (1973)	13/102	29/104	12.7	27.9	0.46(0.26–0.81)
Bett (1973)	21/264	23/253	8.0	9.1	0.88(0.50–1.54)
Aber (1979)	43/302	44/293	14.2	15.0	0.95(0.64–1.40)
UCSG for Streptokinase in AMI(1979)*	18/156	30/159	11.5	18.9	0.61(0.36–1.04)
			Summary relative risk		0.80(0.68–0.95)

*European Cooperative Study Group for Streptokinase in acute myocardial infarction.

statistically significant in only one trial. These studies were interpreted as inconclusive about the benefit of early treatment with intravenous streptokinase.

In a meta-analysis based on these trials, Stampfer estimated the relative risk of mortality in patients treated with intravenous streptokinase to be 0.80 with 95% confidence limits of 0.68 and 0.95, and draw the conclusion that streptokinase reduces the mortality following acute myocardial infarction.[7] The findings were published in the famous medical journal, *New England Journal of Medicine*, and were not accepted by clinician due to poor understanding of meta-analysis in early 1980s. Until 1986, a large clinical trial of intravenous streptokinase after acute myocardial infarction involving thousands of patients (GISSI 1985) confirmed the conclusion based on the meta-analysis, and streptokinase got to be widely used in clinical practice.

The objectives of meta-analysis are:

1.3.1. *To increase statistical power*

Meta-analysis effectively provides a gain in statistical power for average estimates. In clinical trials, meta-analysis offers an opportunity to observe more events of interest in the groups followed, when incidence or mortality is rare, and combined estimates are likely to be more precise. In some cases, a single study often cannot detect or exclude a modest, albeit clinical relevant, difference in the effects of two treatments with great confidence. For example, suppose a drug could reduce the risk of death from myocardial infarction by 10%, to detect such an effect with 90% confidence (that is, with a type II error of no more than 10%) over 10,000 patients in each treatment group would be needed. However, such large samples were difficult to recruit in a single study. Clearly, if data from more than one study are available and can be combined, the "sample size" and, thus, power increase, and relatively small effects can be detected or excluded with confidence.

1.3.2. *To improve estimate of effect size*

Meta-analysis has historically been useful in summarizing prior research based on randomized trials when individual studies are too small to yield a valid conclusion. Results from studies may disagree as to the magnitude of effects or, of more concern, as to the direction of effects. By integrating the actual evidence, meta-analysis allows a more objective appraisal, which can help to resolve uncertainties when the original researches, classic

reviews, and editorial comments disagree. As an effective tool for quantitative synthesis, meta-analysis may resolve issues relating to inconsistent or conflicting results from studies, provide the pooled estimate of effect size with a more precise confidence interval, and draw an explicit conclusion.

1.3.3. *To assess the disagreement and generalizability of study results*

Studies for the same topic may use different eligibility criteria for participants, different definitions of disease, different methods of measuring or defining exposure, or different variations of treatment. It means there is heterogeneity between studies. When heterogeneity is large enough to be detected by a statistical test, it is important to explore its source. Meta-analysis also systematically assesses the biases and confounding in primary studies.

On the other hand, meta-analysis can contribute to considerations about the generalizability of study results. The findings of a particular study may be valid only for a specific population of patients with the same characteristics as those investigated in the trial. If many trials are available for different groups of patients, and show similar results, it can be concluded that the effect of the intervention under study has some generality. Furthermore, meta-analysis is also superior to individual trials when answering questions about whether an overall study result varies among subgroups — for example, among men and women, older and younger patients, or subjects with different degrees of severity of disease. These questions can be addressed in the analysis and often lead to insights beyond what is provided by the calculation of a single combined effect estimate.

1.3.4. *To answer new questions that were not previously posed in the individual studies*

Meta-analysis includes the epidemiological exploration and evaluation of results, new ideas (hypotheses) that were not posed in the individual studies can thus be developed and tested for further research and further original studies.

1.4. *The main steps involved in a meta-analysis*

Meta-analysis should be viewed as an observational study of the evidence. The steps involved are similar to any other research undertaking:

Formulation of the problem to be addressed, collection and analysis of the data, and reporting of the results.

1.4.1. *Formulating the problem*

It is as important to carefully plan a study that involve in a meta-analysis as to carefully plan a clinical trial, a cross-sectional survey, and a case-control or a cohort study. Documentation of all aspects of study design and conduct of the study is a crucial and often overlooked step in carrying out the meta-analysis.

As with any research, a meta-analysis begins with a well-formulated question and design. Meta-analysis can, in general, be motivated by a number of factors. It can be conducted in an effort to resolve conflicting evidence, to answer the questions where the answer is uncertain or to explain variations in practice.

A well-formulated question is essential for determining the structure of a meta-analysis. Specifically, it will guide much of the meta-analysis process including strategies for locating and selecting studies or data, for critically appraising their relevance and validity, and for analyzing variation among their results.

There are several key components to a well-formulated question. A clearly defined question should specify the types of people (participants), types of interventions or exposures, types of outcomes that are of interest, and types of study design. In general the more precise one is in defining components, the more focused the meta-analysis.

The first step in planning the study is to define the problem. The problem definition is a general statement of the main questions that the study addresses. For examples, does the thrombolytic therapy lower the risk of death for patients with acute myocardial infarction? A meta-analysis for randomized clinical trials. Does the passive smoking increase the risk of lung cancer for women? A meta-analysis for case-control studies. These two topics are well-formulated questions that contain the main elements for a meta-analysis.

Once the problem is defined, developing a detail study protocol is essential. A protocol is the blueprint for conduct of the meta-analysis. The protocol should clearly state the objectives, the background, the hypotheses to be tested, the subgroups of interest, the proposed methods and criteria for identifying and selecting relevant studies, and extracting

and analyzing information. The statement of objectives should be concise and specific.

1.4.2. *Searching the relevant information*

A comprehensive, unbiased information search is one of the critical differences between a meta-analysis and a traditional review.

Systematic procedures for literature searching should be described in protocol in detail. Ideally, all of the relevant information, including the published literature, unpublished literature, uncompleted research reports, and work in progress, would be searched and identified in meta-analysis. In practice, the meta-analyst begins with searches of regular medical databases of published literature. Developing a search strategy is very important, which means to present the exact search terms and the search algorithm for each computer databases. Sometimes restrictions are necessary, such as language, study objects, publication year, or publication types, and it is easy to carry out in computer database search.

Skipping over important documents available in databases in searching process may affect the validity and reliability for the results of meta-analysis. The ability of a search algorithm to identify all of the pertinent literature can be improved by consultation with a professional librarian or an expert searcher. Two useful concepts in information retrieval can be used to describe the success of the search process: Sensitivity and precision. Sensitivity of a search is its ability to identify all of the relevant material. Precision (which is the positive predictive value of the search) is the amount of relevant material among the materials retrieved by the search. The overall strategy for searching is to maximize sensitivity and precision. But with the increase of the recall, the precision may be reduced. For meta-analysis, a higher percent sensitivity may be more important than precision.

MEDLINE is the most powerful bibliographic database that is the primary source of information on publication in the biomedical literature. It contains information on publications in over 3,500 and covers the period from 1966 to the present. MEDLINE provides more than 10 search entries and is very friendly to users. The use of MeSH (Medical Subject Headings) terms allows searches of MEDLINE to be focused and specific, which gives higher sensitivity and precision. Free access to MEDLINE through the Internet (www.ncbi.nlm.nih.gov/PubMed) greatly enhances the ability to conduct searches. Other broadly used biomedical databases include EMBASE, SCI (Web of Science), Cochrane Library, and specific databases, such as CANCERLINE, TOXLINE, etc.

The citations or abstracts in databases are browsed in search process, and those obviously unrelated to the topic are eliminated. The full-text of the remaining articles is then collected. These articles are read quickly, and those clearly irrelevant ones are excluded. The remaining publications are then systematically reviewed to determine whether they are eligible for the meta-analysis based on predetermined criteria for eligibility. The reference lists of the articles that contain useful information are searched for more references, then the new publications retrieved, and the process is repeated, until all potentially articles on the topic are identified.

Medical information is also presented in professional website, especially in the medical journal's website, and some of them also provide free full-text. Handsearching is often used. Scanning new information in key journals in the area of interest is an important supplement.

Furthermore, "fugitive" literatures, such as proceedings of conferences, dissertations and master's theses, books chapters, and government reports, are not included in MEDLINE and most other databases. To ignore these material have the potential to cause bias in the meta-analysis. One of the effective ways to obtain the information about publications in the fugitive literature is to consult experts.

Unpublished studies are the ultimate example of fugitive literature. The existence of large numbers of unpublished studies may cause publication bias, which will be discussed in detail in the final section in this chapter.

1.4.3. *Selecting the studies eligible for inclusion*

Studies are chosen for meta-analysis on the basis of inclusion and exclusion criteria. Inclusion criteria are ideally delineated at the stage of the development of the meta-analysis protocol, and should depend on the specific objectives of the analysis. The process of determining whether studies are eligible for inclusion in the meta-analysis should be systematic and rigorous. Each article must be assessed to see whether the inclusion criteria for the meta-analysis are met. To ensure reproducibility and minimize bias in selecting studies, the following six aspects should be addressed in almost all meta-analyses.

1.4.3.1. Study Population

What types of people should be included in meta-analysis? This involves deciding whether one is interested in a specific population group determined on the basis of factors such as age, sex, educational status, or

presence of a particular condition such as the severity of disease and types of disease.

For example, in a meta-analysis of the effects of estrogen replacement therapy on the risk of breast cancer, the inclusion criteria for study population is limited to the women who experienced the natural menopause or who underwent premenopausal hysterectomy, with or without bilateral oophorectomy. The studies that included subjects with a previous history of breast cancer are excluded.

1.4.3.2. Study Design

In clinical trials, the effect of non-randomized controlled study is often overestimated compared with that of randomized study. The treat effect of single blind design may be different from that of double blind design, even though other aspects of the studies are the same. When both randomized and nonrandomized studies are available for a topic, estimates of effect size should be made separately for the randomized and the nonrandomized studies.

In observational studies, the results of case-control study and cohort study may be discrepant for identical problem due to the effects of confounding factors, the influence of biases, or both. The results of meta-analysis need to be reported respectively, according to the study design.

1.4.3.3. Intervention or Exposures

One of the key components about eligibility for a meta-analysis is to specify the intervention or exposure that is of interest, and what types of control groups that are acceptable also need to be defined. In other words, how similar intervention (exposure) should be to use them in the same analysis, such as studies with different doses of the same drug in clinical trials, and studies with the different intensity of exposures in observational data.

For example, a meta-analysis of low-dose aspirin for the prevention of pregnancy-induced hypertensive disease included the studies in which the intervention is aspirin in doses of less than 325 mg/day.

1.4.3.4. Outcomes

Researchers on primary studies often report more than one outcome, and may report the same outcome using different measures. When defining eligibility criteria for the meta-analysis, eligibility based on the similarity

of the outcome will enhance the homogeneity of the studies. Generally, the end-points that are comparable, quantitative and reflecting the final outcomes are appropriate to be chosen for meta-analysis. For example, the chief endpoints, which included in the meta-analysis of randomized trials of angiotensin-converting enzyme (ACE) inhibitors on mortality and morbidity in patients with congestive heart failure (CHF), are total and cause-specific mortality (i.e. progressive heart failure, myocardial infarction, and sudden or presumed arrhythmic death) and hospitalization for CHF.

1.4.3.5. Inclusive dates of publication and English-language publication

Meta-analysis should be as up-to-date as possible, the cutoff date for identification of eligible studies should be specified in the report of the meta-analysis. The inclusive date of publication should be chosen based on consideration of the likelihood of finding important and useful information during the period that is chosen, but not simply on convenience, such as availability of MEDLINE.

A meta-analysis solely based on English-language publications has been shown to have the potential to cause bias. It is not valid to conduct a meta-analysis to rely only on the publications and reports that are easily found and understood.

1.4.3.6. Restriction on sample size or length of follow-up

Most of classical the statistical methods for meta-analysis are based on asymptotic. Normal under moderately large samples. The precision of small studies may tend to be overestimated. To avoid the problem of weighting small studies inappropriately in the meta-analysis, it is reasonable to make sample size an eligibility criteria for the meta-analysis. Small studies are excluded.

Sometimes, the length of follow-up may influence the likelihood of observing a true association in clinical trials. For observational studies, there are many situations where exposure would not affect the risk of disease until after a latent period. To avoid these problems, the length of follow-up could be a criterion for eligibility for the meta-analysis.

An alternative to making study size or length of follow-up an eligibility criterion is to estimate effect with and without small studies or with and without studies with short follow-up or low-dose exposure.

For example, in a meta-analysis of the efficacy of screening mammography, one of the inclusion criteria is, the length of follow-up is least 5 years and with minimum of 10 breast cancer mortality cases in each eligible study.

Generally, highly restrictive eligibility criteria tend to give meta-analysis greater validity. But the criteria may be so restrictive and require so much homogeneity as to limit the eligible studies to only one or two studies, which is conflicted with one of the goals of meta-analysis as a method to increase statistical power. However, less restrictive criteria may lead to the accusation that the meta-analysis "mixes apples and oranges".

1.4.4. *Abstracting the data*

The process of abstraction of information for meta-analysis from eligible studies should be reliable, valid, and free of bias. In order to enhance the reliability of data collection, a standardized form should be developed to record the information. The key components of a data collect form generally include study characteristic with methods, participants, interventions, outcome measures and results.

To avoid the selection bias, the abstraction of information should be done by two abstractors separately, and experts should be consulted for disagreement. Furthermore, the abstractors should be blinded to the information of the authors, the journals, and the funding sources. It is believed that these factors possibly influence the judgment of the abstractor.

1.4.5. *Assessing study quality*

It is important to systematically complete critical appraisal of all included studies, which primarily focus on the validity of studies. If the quality of original study is poor, the results of meta-analysis will be less reliable and valid.

The validity of a study is the extent to which its design and conduct are likely to prevent systematic errors, or bias. Generally, there are four sources of systematic errors in clinical trials: Selection bias, performance bias, attrition bias and detection bias. The randomization process, the measurement of patient compliance, the blinding of patients and observers, the statistical analyses, and the handling of withdrawals in each primary study should be examined. For non-experimental studies, control for confounding, measurement of exposure and completeness of follow-up are all the main factors that need to be greatly considered in the process of study quality assessment.

Because quality assessment is a subjective process, it may potential cause error and bias. There is not a "gold standard" for study quality appraise yet. So, the reliability of the quality rating scales in published meta-analysis is often not formally evaluated.

1.4.6. *Statistical analysis*

The process of quantitative combining the data is the key step for meta-analysis, which is distinguish from the traditional narrative review. The main procedures involved in the statistical analysis are: Defining the outcome; homogeneity test for the effect size; model choice (fixed-effects model or random-effects model); pooled estimate of effect size (point estimate and confidence interval estimate); hypothesis test for overall effect size and graphic display of the results.

1.4.7. *Sensitivity analysis*

The goal of sensitivity analysis in meta-analysis is to assess the robustness of conclusion when different assumptions are made in conducting the analysis. Sensitivity analysis is usually conducted to examine the change of the pooled estimate of effect size, when both fixed- and random-effects model are used. Sensitivity analysis is also often done including and excluding certain studies, which are controversial, have large effects and thus dominate the analysis, or cannot be determined to meet the eligibility criteria but whose exclusion may be problematic. When there is more than one estimate of effect size available from a study, sensitivity analysis can be performed using one estimate and then the other.

For example, Egger did a sensitivity analysis in the meta-analysis of β-blockade in secondary prevention after myocardial infarction.[8] Firstly, the overall effect was calculated by different statistical model, the results showed that the overall effect estimates are virtually identical and that confidence intervals are only slightly wider with random-effects model. Secondly, methodological quality was assessed in terms of how patients were allocated to treatment or control groups, how outcome was assessed, and how the data were analyzed. The results showed that the three low quality studies presented more benefit than high quality trials. Exclusion of these three studies, however, leaves the overall effect and the confidence intervals practically unchanged. Third, when stratifying the analysis by study size, the results showed the trials with smallest sample sizes have the largest effect. However, exclusion of such studies has little effect on the

overall estimate. Thus, sensitivity analysis showed that the results from this meta-analysis were robust.

1.4.8. *Discussion of results*

As with any medical article, the last step in meta-analysis is discussion.

- Investigating and explaining the source of heterogeneity are critically important component of meta-analysis, when there is "statistically significant" heterogeneity across studies. Heterogeneity is easier to be observed in observational studies due to the diversity in their designs, the methods for collecting data, definitions of endpoints, and the degree of control for bias and confounding. Indeed, there are no statistical methods that can deal with the bias and confounding in the original studies. Meta-regression model and mixed model may adjust somewhat of heterogeneity by controlling the confounding, but it still cannot explain the source of heterogeneity. Sensitivity analysis and subgroup are useful for exploring the heterogeneity. It may not be appropriate with great difference.
- Subgroup analysis is necessary when treatment effect vary according to patient-level covariance or trial-level characteristics. For example, the effect of a given treatment is unlikely to be identical across different group of participant — for example, young people versus elderly people, those with mild disease versus with severe disease. A relationship between the underlying risk of patient and treatment effect may crucially affects decisions about which patients should be treated from a cost-effectiveness perspective: Patient at high risk with a small proportionate treatment benefit may be preferentially treated compared to low risk patients with a larger proportionate treatment benefit. Sometimes the treatment effect may be in the opposite direction for patients at low and high risk. Meta-analysis thus offers a sounder basis for subgroup analysis. But meta-analytic subgroup analyses are prone to bias and need to be interpreted with caution. Ideally, if individual patient data in each eligible study can be obtained, a standardized subgroup analysis can be performed.
- Meta-analysis is essentially viewed as an observational study. Bias can occur at multiple steps in the process of meta-analysis. Bias may seriously influence the validity and reliability of meta-analysis, and more attention needs to be paid to detect and assess of the bias.
- When reporting the conclusion, we should summarized the key finding, interpret the results in light of the total of available evidence, and suggest

a future research agenda. But for meta-analysis of observational studies, generalization of the conclusions must be explained in caution, because bias and confounding may distort the findings as we have shown above.

For example, the hypothesis from ecological analyses that higher intake of saturated fat could increase the risk of breast cancer generated much observational research often with contradictory results. A comprehensive meta-analysis showed an association from case-control but not from cohort studies (odds ratio was 1.36 from case-control studies versus relative rate 0.95 from cohort study), and this discrepancy was also shown in two separate large collaborative meta-analyses of case-control and cohort studies. The most likely explanation for this situation is that biases in the recall of dietary items and in the selection of study participants have produced a spurious association in the case-control comparisons.[9]

2. Statistical Methods in Meta-Analysis

2.1. *Definition of the study outcome*

The primary studies included in the meta-analysis may report several different end points. Often the meta-analyst has little control over the choice of the study outcome, and it is very important to select pooled statistic that is comparable across all studies. In some situations this task will be impossible. Here, three classes of outcome measures are discussed: Measures based on discrete outcome data, that may generally be thought of as odds ratios, relative risks, or risk differences; those based on continuous data, such as mean difference, and standardized mean difference; and a miscellaneous set of outcome measures that may be based on test statistics.[10]

2.1.1. *Odds ratios, relative risks and risk differences*

Suppose there are K studies for binary discrete measurements included in the meta-analysis, whose data are in the form of 2×2 tables (see Table 2). Let i index study, in a typical one, clinical trials, let 1 denote treatment group, and 2 control group. We denote a_i, b_i, c_i, and d_i as the number of observations in each of the cells defined by the treatment and outcome table, with n_{1i} subjects in the treatment group and n_{2i} in the control group. p_{1i} and p_{2i}, are the proportions of having the characteristic under study, such as death, relapse or some other kind of failure. In an epidemiological case-control study, the two groups would be the cases and controls and

Table 2. Arrangement of data for 2×2 table.

	Treated (Exposed)	Not Treated (Not Exposed)	Total
Death (Case)	a_i	b_i	n_{1i}
Survival (Control)	c_i	d_i	n_{2i}
Total	m_{1i}	m_{2i}	T_i

Table 3. Parameter estimation for three binary measurements.

	Parameter	Estimator	Standard Error
Risk Difference	$D = P_1 - P_1$	$d_i = \hat{p}_{1i} - \hat{p}_{2i}$	$s_{di} = \left(\frac{p_{1i}(1-p_{1i})}{n_{1i}} + \frac{p_{2i}(1-p_{2i})}{n_{2i}} \right)^{\frac{1}{2}}$
Relative Risk	$R = P_1/P_2$	$r = \hat{p}_{1i}/\hat{p}_{2i}$	$S_{\mathrm{Log}}(ri) = \left(\frac{(1-p_{1i})}{n_{1i}p_{1i}} + \frac{(1-p_{2i})}{n_{2i}p_{2i}} \right)^{\frac{1}{2}}$
Odds Ratio	$\Omega = \frac{P_1/(1-P_1)}{P_2/(1-P_2)}$	$\omega_I = \frac{\hat{p}_{1i}/(1-\hat{p}_{1i})}{\hat{p}_{2i}/(1-\hat{p}_{2i})}$	$s_{\mathrm{Log}}(\omega i) = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)^{\frac{1}{2}}$

the characteristic under study would be exposed to the hypothesized risk factor.

Table 3 gives the formula of parameter inferences in three potential study summary statistics: The ratio of the odds for the treated group to the odds for the control group (odds ratio, OR), the ratio of two probabilities (relative risk, RR), and the difference between two probabilities (risk difference, RD). OR and RR are typically analyzed on logarithmic scale with normal distribution approximation, and the confidence intervals for OR and RR are also computed on the logarithmic scale, then transformed back to the original scale. In practice, OR is widely used as an outcome measure for its convenient mathematical properties, which allow for easily combining data and testing the significance of the overall effect. The OR will be close to the RR, if the end point occurs relatively infrequently, such as less than 20%. RD or absolute risk reduction is easy to interpret and defined for boundary values (proportions of 0 or 1), and is approximately normally distributed for the modest sample sizes. RD reflects both the underlying risk without treatment and the risk reduction associated with treatment. Taking the reciprocal of the RD gives the "number needed to treat" (the number of patients needed to be treated to prevent one event), which is very useful in making a decision in clinical practice.

2.1.2. *Means differences and standardized means differences*

When the primary studies report means as outcome measure on a continuous scale, there are two situations to be considered. First, all of the eligible studies use the same measure of effect, and mean difference may be used as summary measure to estimate pooled effect in the meta-analysis. Suppose the n_{1i} and n_{2i} are the sample sizes, x_{1i} and x_{2i} are the means, for treatment and control group, respectively. $Y_i = \bar{X}_{1i} - \bar{X}_{2i}$, with standard error, s_i, calculated as with

$$s_i^2 = s_{pi}^2 \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}} \right) \quad \text{with} \quad s_{pi}^2 = \frac{(n_{1i} - 1)s_{1i}^2 + (n_{2i} - 1)s_{2i}^2}{n_{1i} + n_{2i} - 2},$$

where s_{1i}^2 and s_{2i}^2 are the treatment and control group variance, respectively, of the ith study.

Second, all of the eligible studies address the same question, but the measure of effect is made using different instruments and thus different scales. When there is no direct measure common to all the studies, it may be feasible to transform the study-specific summary to a standardized (scale-free) statistic denoted as effect size. One common estimator of effect size is

the standardized mean difference, which is calculated as the difference of means divided by the variability of the measures. If

$$Y_{ij}^1 \sim N(\mu^1, \sigma^2)\,, \quad j = 1, 2, \ldots, n_{1i}\,,$$

$$Y_{ij}^2 \sim N(\mu^2, \sigma^2)\,, \quad j = 1, 2, \ldots, n_{2i}\,,$$

then the standardized means difference is defined as

$$\delta = \frac{\mu^1 - \mu^2}{\sigma}\,,$$

which denotes the gain (or loss) as the fraction of the measurements. The estimator of δ, Hedge's g, is defined as

$$h_i = \frac{\bar{Y}_i^1 - \bar{Y}_i^2}{s_p}\,.$$

Such standardization leads to a unitless effect measure. The results from the original studies, where "success" is measured in different ways, can be standardized to unitless measures and then pooled. The estimated variance of h_i is

$$\mathrm{var}(h_i) = \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right) + \frac{h_i^2}{2(n_{1i} + n_{2i})}\,.$$

2.1.3. *Other measures*

When the summary data from the primary studies consist of test statistics, then it is sometimes possible to recover the estimated effect size if the appropriate pieces of information are also reported. For example, if the z-statistics is reported, the estimated standardized mean difference may be calculated as

$$\hat{\delta} = z\sqrt{\left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right)}\,.$$

2.2. *Model choice*

In meta-analysis, pooled effects and confidence intervals are usually obtained by using appropriate parametric statistical models. Just like ANOVA, analysis the sources of variation may be critical for the model used in meta-analysis.[11,12]

There are at least two sources of variation to consider before combining summary statistics across studies. One is the inner- or within-study variation, which is derived from sampling error. Sampling error may vary with

studies. In general, the sampling error may be relatively small for studies with large sample sizes, which means high degrees of precision and large weight would be given. The other is the inter- or between-study variation. The fixed-effects (FE) model assumes each study is measuring the same underlying parameter and there is no inter-study variation, in other words, the population from which the given studies were drawn comprises studies exactly like those in the sample, the only source of variation in the observations is due to within-study sampling. By contrast, the random effects (RE) model assumes each study is associated with a different but related parameter, which means the population believed to produce the sampled set of studies is a population of studies not exactly alike. For the RE model, each study's observed effect results from sampling variation about a random effect measure, which itself is "drawn" from a distribution of effect measures.

2.2.1. *Fixed-effects model*

A fixed-effects model assumes that each observed study effect, $Y_i (i = 1, 2, \ldots, K)$, is a realization of a population of independent studies with common parameters. Let θ be the parameter of interest, which quantifies the average treatment effect. Assume that Y_i is such that $E(Y_i) = \theta$ and let $s_i^2 = \text{var}(Y_i)$ be the estimate of variance of the effect in the ith study. For moderately large study sizes, each Y_i should be asymptotically normal distributed (by the central limit theorem) and approximately unbiased. Thus,

$$Y_i \overset{\text{indep}}{\sim} N(\theta, s_i^2) \tag{1}$$

and s_i^2 is assumed known.

2.2.2. *Random-effects model*

The random-effects model framework postulates that each observed study effect, Y_i, is a draw from a normal distribution with a study-specific mean, θ_i, and variance, s_i^2. θ_i is interpreted as the "true effect" in study i. Furthermore, θ_i is assumed to be a draw from some hyper-distributions of effects with mean θ and variance τ^2. θ is the true underlying effect of interest, represent the average treatment effect, and τ^2 is the inter-study variance, or heterogeneity parameter. Thus,

$$Y_i | \theta_i,\ s_i^2 \overset{\text{indep}}{\sim} N(\theta_i, s_i^2)\,, \tag{2}$$

$$\theta_i|\theta,\ \tau^2 \overset{\text{indep}}{\sim} N(\theta, \tau^2)\,. \tag{3}$$

Random-effects model "borrow strength" across studies when estimating study-specific effects, θ_i, as well as the population effect θ. RE model of (2) and (3) is refer to "hierarchic" model. This structure will be particularly useful in the development of the Bayesian paradigm.

2.3. *Statistical inference*

A test of homogeneity should be done before any further analysis. If no significant inter-study variation is found, a fixed-effects approach is adopted. Otherwise, the meta-analyst either adopts a random-effects approach or identifies study characteristics that stratify the studies into subsets with homogeneous effects. The test of heterogeneity is described next and followed by a description of inference for fixed-effects and random-effects models. Maximum likelihood, and restricted maximum likelihood methods are given for both types of models.

2.3.1. *Test of homogeneity*

The investigation of homogeneity is a crucial part of the meta-analysis. The fixed effects model assumes that the K study-specific summary statistics share a common mean θ. A statistical test for the homogeneity of study means is equivalent to testing

$$\mathrm{H}_0 : \theta = \theta_1 = \theta_2 = \cdots = \theta_K\,,$$

$$\mathrm{H}_1 : \text{At least two } \theta_i\, s \text{ different}\,.$$

The test statistic

$$Q_w = \sum_i^k W_i (Y_i - \hat{\theta})^2 \tag{4}$$

will asymptotically follow χ^2_{k-1} under H_0 for large sample sizes. The overall treatment effect θ, is estimated as a weighted average, that is

$$\hat{\theta} = \sum W_i Y_i \Big/ \sum W_i \quad \text{and} \quad W_i = 1/s_i^2\,.$$

If Q_w is greater than the $100(1-\alpha)$ percentile of the χ^2 distribution, the hypothesis of equal means, H_0, would be rejected at the $100(1-\alpha)$ level. If H_0 is rejected, the meta-analyst may conclude that the study means arose from two or more distinct populations and proceed by either attempting to identify covariates that stratify studies into the homogeneous populations

or adopting a random-effects model. If H_0 cannot be rejected, it would be concluded that the K studies share a common mean, θ.

Tests of homogeneity have low power against the alternative $\text{var}(\theta_i) > 0$. Note that not rejecting H_0 is equivalent to asserting that the between-study variation is small. The results of simulation by Hardy show that the power of homogeneity test depends on the number of included studies, the total information (i.e. total weight or inverse variance) available and the distributions among the different studies.[13] In practice, if the studies are homogeneous, then the choice between the fixed- and random-effects model is not important, as the models will yield similar results. The use of the random-effects model is not considered to be a defensible solution to the problem of heterogeneity. The random-effects model is generally "conservative". That is, in most situations, use of the random-effects model will lead to wider confidence inference and a low chance to call a difference "statistically significant".

2.3.2. *Parameter estimation*

For fixed-effects model, when s_i^2 is assumed known, $\log(L(\theta|y, s^2)) \propto \sum_i \left(\frac{(Y_i-\theta)^2}{s_i^2}\right)$, which leads to the maximum likelihood estimator (MLE)

$$\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i} \quad \text{with} \quad W_i = \frac{1}{s_i^2}. \tag{5}$$

Standard inferences about θ are available using the fact that

$$\hat{\theta}_{\text{MLE}} \sim N\left(\theta \left(\sum_i w_i\right)^{-1}\right).$$

For random-effects model, if τ^2 is known, the MLE of θ is given by

$$\hat{\theta}(\tau)_{\text{MLE}} = \frac{\sum_{i=1}^k w_i(\tau) Y_i}{\sum_{i=1}^k w_i(\tau)} \quad \text{with} \quad W_i(\tau) = \frac{1}{s_i^2 + \tau^2}. \tag{6}$$

However, in the more realistic case of unknown τ^2, restricted maximum likelihood (RMLE) can be employed as a method for estimating variance components in a general linear model. Using the marginal distribution for y, the log-likelihood to be maximized is

$$\log(L(\theta, \tau^2|s^2 y) \propto \sum_i \left\{\log(s_i^2 + \tau^2) + \frac{(Y_i - \hat{\theta}_R)^2}{s_i^2 + \tau^2}\right\} + \log\left(\sum (s_i^2 + \tau^2)^{-1}\right).$$

The REML of τ^2 is the solution of

$$\tau_R^2 = \frac{\sum_i w_i^2(\hat{\tau})\left(\frac{k}{k-1}(Y_i - \hat{\theta}_R)^2 - s_i^2\right)}{\sum_i w_i^2(\hat{\tau})}.$$

The estimator for the population mean is then calculated as

$$\hat{\theta}_R = \frac{\sum_i^k w_i(\hat{\tau}_R)Y_i}{\sum_i^k w_i(\hat{\tau}_R)}, \quad w_i(\hat{\tau}_R) = \frac{1}{s_i^2 + \hat{\tau}_R^2},$$

and inferences are made using $\hat{\theta}_R \sim N(\theta, (\sum_i w_i(\hat{\tau}_R))^{-1})$.

By equating the homogeneity test, Q_w, to its corresponding expected value, DerSimonian and Laird proposed a non-iterative (method of moments) estimator of τ^2 as

$$\tau^2 = \max\left\{0, \frac{Q - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}\right\}.$$

This leads to

$$\hat{\theta}_{DL} = \frac{\sum_i w_i(\hat{\tau}_{DL})Y_i}{\sum_i w_i(\hat{\tau}_{Dl})} \quad \text{with} \quad w_i(\hat{\tau}_{Dl}) = \frac{1}{s_i^2 + \hat{\tau}_{DL}^2}.$$

$\hat{\theta}_{DL}$ is also denoted Cochran's semi-weighted estimator of θ and can be easily programmed using most software packages.

A third estimator of τ^2 and θ is to adopt a fully Bayesian approach, which reflect the uncertainty in the estimates of hyperparameters.

2.4. *Classical approaches for meta-analysis*

Many methods of meta-analysis have been proposed. Here we focus on the classic approaches based on two kinds of measures, discrete outcome and continuous outcome.

2.4.1. *Measures based on a discrete outcome*

For measures based on discrete outcome, we primary discuss the methods involve the data in the form of 2×2 table, which is widely used in the meta-analysis of clinical trials, cohort studies and case-control studies. Suppose the arrangement of data and table notation is still as shown in Table 2.

2.4.2. *Mantel-Haenszel method*

The Mantel-Haenzel method is a well-known approach for pooling data across strata. Since each study included in meta-analysis could be regarded as a stratum, Mantel-Haenzel method is appropriate for analyzing data for a meta-analysis. The method is based on the assumption of fixed-effects model, and the pooled measure is expressed as a combination of stratum-specific measures. Mantel-Haenzel method can be used when the measure of effect is a ratio measure, typically an odds ratio.[14] In meta-analysis, the pooled estimate using Mantel-Haenzel method is the weighted average of the maximum-likelihood estimate of the odds ratios in each study, using the inverse of study level variances as weights.

The odds ratio for the ith study $OR_i = \frac{a_i d_i}{b_i c_i}$.

The weight for the ith study $w_i = \frac{b_i c_i}{T_i}$.

The pooled estimate of odds ratio is

$$OR_{MH} = \frac{\sum(w_i OR_i)}{\sum w_i} = \frac{\sum(a_i d_i / T_i)}{\sum(b_i c_i / T_i)}\,. \tag{7}$$

The variance of the OR_{MH} is equal to

$$\mathrm{var}(OR_{MH}) = \frac{\sum F}{2\sum R^2} + \frac{\sum G}{2\sum R \sum S} + \frac{\sum H}{2\sum S^2}\,,$$

with

$$F = \frac{a_i d_i (a_i + d_i)}{T_i^2}\,,$$

$$G = \frac{a_i d_i (b_i + c_i) + b_i c_i (a_i + d_i)}{T_i^2}\,,$$

$$H = \frac{b_i c_i (b_i + c_i)}{T_i^2}\,,$$

$$R = \frac{a_i d_i}{T_i}\,, \quad S = \frac{b_i c_i}{T_i}\,.$$

The 95% confidence interval for pooled odds ratio is equal to

$$\exp\left(\ln OR_{MH} \pm 1.96\sqrt{\mathrm{var}(OR_{MH})}\right). \tag{8}$$

The Q statistics for homogeneity test is given by

$$\begin{aligned} Q &= \sum w_i (\ln OR_{MH} - \ln OR_i)^2 \\ &= \sum w_i [\ln(OR_i)]^2 - \frac{[\sum w_i \ln(OR_i)]^2}{\sum w_i}\,. \end{aligned} \tag{9}$$

Under the null hypothesis of homogeneity, Q has an approximate χ^2_{k-1} distribution.

The test based on Mantel-Haenszel χ^2 has optimal statistical properties, being the uniformly most powerful test. But application of the method requires that data to complete a 2×2 table of outcome by treatment groups for each study are available.

2.4.1.2. Peto method

The Peto method is a modification of Mantel-Haenszel method. It is based on the fixed-effects model and the effect measure of interest is odds ratio.[15] Peto method uses a score statistics and Fisher information statistics from conditional likelihood for study-specific effects to estimate pooled effects. The computation involved in Peto method is relatively simple compared to Mantel-Haenszel method. Peto method has been extensively used, especially in clinical trials.

Let O_i and E_i be the observed and expected number of events in the treatment group for ith study, respectively, where $E_i = \frac{n_{1i}m_{1i}}{T_i}$.

The pooled estimate of odds ratio is equal to

$$OR_p = \exp\left(\frac{\sum(O_i - E_i)}{\sum V_i}\right), \tag{10}$$

where $V_i = \frac{n_{1i}m_{1i}n_{2i}m_{2i}}{T_i^2(T_i-1)}$ is the variance of the difference $O_i - E_i$.

The 95% confidence interval for pooled odds ratio is

$$\exp\left(\ln OR_p \pm \frac{1.96}{\sqrt{\sum V_i}}\right) = \exp\left(\frac{\sum(O_i - E_i) \pm 1.96\sqrt{\sum V_i}}{\sum V_i}\right). \tag{11}$$

The homogeneity test, Q, is given by

$$Q = \sum \frac{(O_i - E_i)^2}{V_i} - \frac{(\sum(O_i - E_i))^2}{\sum V_i}. \tag{12}$$

Under the null hypothesis of homogeneity, Q has an approximate χ^2_{k-1} distribution.

Although Peto method is widely used, it has been demonstrated to be potentially biased when the true common odds ratio is far from unity or when there are large unbalances between the numbers of death and survival or exposed and non-exposed. In this situation, Mantel-Haenszel may be preferred.

Example 1. Table 4 shows data from seven randomized clinical trials of the effect of aspirin in preventing death after myocardial infarction[16] The Peto

method is used to estimate a summary odds ratio and its 95% confidence interval for these data is as follows:

Table 4. Data form seven randomized trials of the effectiveness of aspirin after myocardial infarction and the results of meta-analysis (Peto method).

	Aspirin		Placebo						
Study	No. Deaths	No. patient	No. death	No. Patient	E_i	$O_i - E_i$	V_i	OR_i	$(O_i - E_i)^2/v_i$
1	49	615	67	624	5i.6	−8.6	26.3	0.720	2.8
2	44	758	64	771	53.5	−9.5	25.1	0.681	3.6
3	102	832	126	850	112.8	−10.4	49.3	0.803	2.4
4	32	317	38	309	35.4	−3.4	15.5	0.801	0.7
5	85	810	52	406	91.3	−6.3	27.1	0.798	1.5
6	246	2267	219	2257	233.0	13.0	104.3	1.133	1.6
7	1570	8587	1720	8600	1643.8	−73.8	665.1	0.895	8.2
Total						−99.4	912.7		20.8

Source: Fleiss and Gross.[16]

2.4.1.2.1. Homogeneity test

Calculate $E_i, V_i, O_i - E_i$, and $(O_i - E_i)^2/V_i$, and the results are show in Table 4.

$$Q = \sum \frac{(O_i - E_i)^2}{V_i} - \frac{(\sum(O_i - E_i))^2}{\sum V_i} = 20.8 - \frac{(-99.4)^2}{912.7} = 10.1\,.$$

Here, $df = 6$, $\chi^2_{(0.05,6)} = 12.6 > 10.1$, $P > 0.05$, the null hypothesis of homogeneous odds ratio would not be rejected at 5 percent level, so that the fixed-effects model may be appropriate to be adopted for pooling the odds ratio.

2.4.1.2.2. Calculate the pooled estimate of odds ratio and its 95% confidence interval

$$OR_p = \exp\left(\frac{\sum(O_i - E_i)}{\sum V_i}\right) = \exp\left(\frac{-99.4}{912.7}\right) = 0.09$$

$$\exp\left(\frac{\sum(O_i - E_i) \pm 1.96\sqrt{\sum V_i}}{\sum V_i}\right) = \exp\left(\frac{-99.4 \pm 1.96\sqrt{912.7}}{912.7}\right)$$

$$= (0.84, 0.96)\,.$$

2.4.1.2.3. Graphical presentation of the results

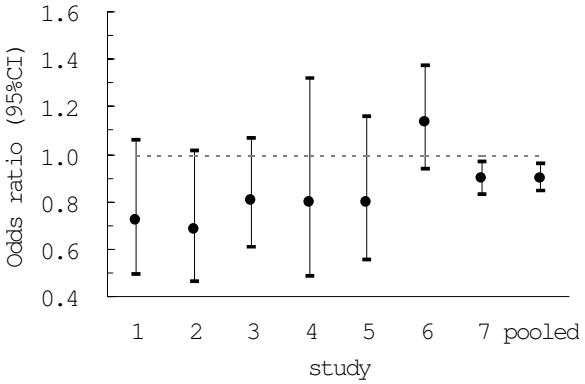


Fig. 1. The odds ratios of seven studies and their 95% confidence interval, and pooled odds ratio and its 95% confidence interval.

2.4.1.3. Fleiss method

When data to complete a 2×2 table is not available, the Peto method could not be adopted unless those studies are excluded. Sometimes the individual study may report the proportions having the characteristic under study, the Fleiss method can be used as alternative the Peto method based on fixed-effect model. For a clinical trial or cohort study, let p_{1i} and p_{2i} be the mortality rate or incidence rate for treated (exposed) and control group, respectively. For a case-control study, let p_{1i} and p_{2i} be the exposure rate for case and control group, respectively. Fleiss draws the formula of pooling the log odds ratio when p_{1i} and p_{2i} are given in the included study in meta-analysis.[16]

The effect for ith study, denoted by y_i, is the logarithm of the odds ratio:

$$y_i = \ln(OR_i) = \ln(p_{1i}(1 - p_{2i})/p_{2i}(1 - p_{1i}))\,.$$

The variance and weight of y_i are given by

$$\mathrm{var}(y_i) = \frac{1}{n_{1i}p_{1i}(1 - p_{1i})} + \frac{1}{n_{2i}p_{2i}(1 - p_{2i})}\,, \quad w_i = \frac{1}{\mathrm{var}(y_i)}\,.$$

The pooled estimate of odds ratio is equal to

$$OR_F = \exp(\bar{y}) = \exp\left(\sum w_i y_i \Big/ \sum w_i\right). \tag{13}$$

The 95% confidence interval for summary odds ratio is given by

$$\exp\left(\bar{y} \pm 1.96 \Big/ \sqrt{\sum w_i}\right). \tag{14}$$

The Q statistic for homogeneity test is

$$Q = \sum w_i (y_i - \bar{y})^2 = \sum w_i y_i^2 - \frac{(\sum w_i y_i)^2}{\sum w_i}. \tag{15}$$

Example 1 (continued). The Fleiss method is used to estimate a pooled odds ratio and its 95% confidence interval for data in Example 1 in Table 5.

First, calculate the observed effect $y_i = \ln(OR_i)$, variance v_i weight w_i and $w_i y_i$, $w_i y_i^2$ for each individual study, results shown in Table 5.

The Q statistic for homogeneity test is

$$Q = \sum w_i (y_i - \bar{y})^2 = \sum w_i y_i^2 - \frac{(\sum w_i y_i)^2}{\sum w_i}$$

$$= 20.7849 - \frac{(-99.1391)^2}{910.559} = 10.8\,.$$

$df = 6$, $\chi^2_{(0.05,6)} = 12.6 > 10.1$, $P > 0.05$, H_0 would not be rejected, so the fixed-effects model may be appropriate.

Then the pooled estimate of odds ratio is equal to

$$OR_F = \exp\left(\sum w_i y_i \Big/ \sum w_i\right)$$

$$= \exp(-99.1391/910.559) = \exp(-0.1089) = 0.90\,.$$

Table 5. Results of meta-analysis for the effectiveness of aspirin after myocardial infarction (Fleiss method).

Study	$y_i = \ln(OR_i)$	$w_i = 1/v_i$	$w_i y_i$	$w_i y_i^2$
1	−0.3285	25.710	−8.4457	2.7744
2	−0.3842	24.291	−9.3326	3.5856
3	−0.2194	48.801	−10.7069	2.3491
4	−0.2194	15.440	−3.3875	0.7432
5	−0.2332	28.409	−6.6250	1.5449
6	0.1249	103.985	12.9877	1.6222
7	−0.1109	663.923	−73.6291	8.1655
Total		910.559	−99.1391	20.7849

The 95% confidence interval for pooled odds ratio is given by

$$\exp\left(\bar{y} \pm 1.96 \Big/ \sqrt{\sum w_i}\right) = \exp(-0.1089 \pm 1.96/\sqrt{910.559})$$
$$= (0.84, 0.96)\,.$$

Note that, results of Fleiss method are the same as those of Peto method. If complete 2×2 tables are available for all included studies, Peto method is simpler than Fleiss method, but the latter can be used for those only proportions reported.

2.4.1.4. General variance-based method

When the effect size is measured as a rate difference, the general variance-based method would be applied to estimation of the pooled rate difference. The general variance-based method also used to estimate the pooled risk ratio, rate ratio and odds ratio.[12] The general variance-based method is also based on fixed-effect model.

2.4.1.4.1. Effect size is measured as a rate difference

The rate different for ith study is $RD_i = \frac{a_i}{n_{1i}} - \frac{c_i}{n_{2i}}$.

The variance and weight of rate difference are $\text{var}(RD_i) = \frac{n_{1i}n_{2i}}{m_{1i}m_{2i}T_i}$, $w_i = 1/\text{var}(RD_i)$.

The pooled estimate of rate difference is

$$RD_{GV} = \frac{\sum(w_i RD_i)}{\sum w_i}\,. \tag{16}$$

The 95% confidence interval of pooled estimate of rate difference is equal to

$$RD_{GV} \pm 1.96 \Big/ \sqrt{\sum w_i}\,. \tag{17}$$

2.4.1.4.2. Effect size is measured as an incidence density ratio or as a risk ratio

The relative risk for ith study is $RR_i = \frac{a_i}{n_{1i}} / \frac{c_i}{n_{2i}}$.

The variance and weight of relative risk are $\text{var}(RR_i) = \frac{n_{2i}T_i}{m_{1i}m_{2i}n_{1i}}$, $w_i = 1/\text{var}(RR_i)$.

The pooled estimate of relative risk is

$$RR_{GV} = \exp\left(\frac{\sum(w_i \ln(RR_i))}{\sum w_i}\right). \tag{18}$$

The 95% confidence interval of pooled estimate of relative risk is equal to

$$\exp\left(\ln(RR_{GV}) \pm 1.96 \Big/ \sqrt{\sum w_i}\right). \tag{19}$$

When each study in meta-analysis just presents the relative risk and its 95% confidence interval, whereas a complete 2×2 table is unavailable, general variance-based method also could be applied to estimate the pooled effect using Eqs. (18) and (19). The formula for estimating variance from the 95% confidence interval is

$$\mathrm{var}(RR_i) = \left(\frac{\ln(RR_i/RR_l)}{1.96}\right)^2 = \left(\frac{\ln(RR_u/RR_i)}{1.96}\right)^2, \tag{20}$$

where RR_u and RR_l are the upper and lower bound of the 95% confidence interval for ith study.

2.4.1.4.3. Effect size is measured as odds ratio

The pooled estimate of odds ratio is

$$OR_{GV} = \exp\left(\frac{\sum(w_i \ln(OR_i))}{\sum w_i}\right). \tag{21}$$

The 95% confidence interval of pooled estimate of odds ratio is equal to

$$\exp\left(\ln(OR_{GV}) \pm 1.96 \Big/ \sqrt{\sum w_i}\right), \tag{22}$$

where

$$w_i = [\mathrm{var}(\ln(OR)_i)]^{-1} = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)^{-1}. \tag{23}$$

Note that when pooling the effect, relative risk and odds ratio should be transform to logarithmic scale in order to be approximately normally distributed, whereas the rate difference could be computed directly.

2.4.1.5. DerSimonian-Laird method

The approaches we previously described are all based on the fixed-effect model. When the studies included in meta-analysis lack of homogeneity, the random-effects model may be appropriate to combine the effect size. Formulas of applying the DerSimonian-Laird method summarizing studies in the case where effects are measured as odds ratios are given as follows[17]:

$$OR_{DL} = \exp\left(\frac{\sum w_i^* \ln(OR_i)}{\sum w_i^*}\right). \tag{24}$$

The 95% confidence interval of pooled estimate of odds ratio is equal to

$$\exp\left(\ln OR_D L \pm 1.96 \Big/ \sqrt{\sum w_i^*}\right), \tag{25}$$

where w_i^* is the weighting factor for the ith study, is estimated as

$$w_i^* = \frac{1}{D + (1/w_i)}. \tag{26}$$

D is derived from the homogeneity test statistic, Q, in Eq. (4). As described previously about the moment estimate of inter-study variance τ^2 in model choice and homogeneity test, we have

$$D = \frac{Q - (K-1)\sum w_i}{(\sum w_i)^2 - \sum w_i^2} \quad \text{and} \quad D = 0 \text{ if } Q < k - 1, \tag{27}$$

where k is the number of included studies.

Example 1 (continued). In the example of meta-analysis of seven clinical trials in which aspirin was used to prevent the death after myocardial infarction, we have calculated the pooled effect sized using the approaches based on fixed-effects model. The results of homogeneity test is, $Q = 10.8$, and $df = 6$, $\chi^2_{(0.05,6)} = 12.6 > 10.8$, $P > 0.05$, the null hypothesis was not rejected. In order to evaluate the dependence of the conclusions of the analysis on the model assumption, now we calculate the pooled effect using random-effects model.

$$D = \frac{Q - (k-1)\sum w_i}{(\sum w_i)^2 - \sum w_i^2} = \frac{10.1 - (7-1) \times 910.559}{910.559 - 456284.69} = 0.00977.$$

Each w_i^* for individual study is calculated using Eq. (26), results shown in Table 6.

Table 6. Results of meta-analysis for the effectiveness of aspirin after myocardial infarction (DerSimonian-Laird method).

Study	$y_i = \ln(OR_i)$	w_i	w_i^2	w_i^*	$w_i^* y_i$
1	−0.3285	25.710	661.004	20.54	−6.747
2	−0.3842	24.291	590.053	19.63	−7.542
3	−0.2194	48.801	2381.538	33.04	−7.219
4	−0.2194	15.440	238.394	13.42	−2.944
5	−0.2332	28.409	807.071	22.24	−5.186
6	0.1249	103.985	10812.880	51.58	−6.442
7	−0.1109	663.923	440793.750	88.68	−9.835
Total		910.559	456284.69	249.13	−33.061

The pooled estimate of odds ratio and its 95% confidence interval are

$$OR_{DL} = \exp\left(\frac{\sum w_i^* \ln(OR_i)}{\sum w_i^*}\right) = \exp\left(\frac{-33.061}{249.13}\right)$$
$$= \exp(-0.1327) = 0.88\,,$$

$$\exp\left(\ln OR_{DL} \pm 1.96\Big/\sqrt{\sum w_i^*}\right)$$
$$= \exp(-0.1327 \pm 1.96/\sqrt{249.13}) = (0.77, 0.99)\,.$$

Now, if we compare the results of fixed-effects and random-effects model, the pooled point estimate of odds ratio, 0.88 and 0.90, respectively, is almost the same. The length of the 95% confidence interval based on random-effects model is 0.22 (0.99–0.77), which is greater than that based on fixed-effects model, 0.12 (0.96–0.84). So the result of random-effects model is potentially more conservative. But the two methods yield the same conclusion, that is, in general, aspirin make the risk of death after myocardial infarction decrease by nearly 10%.

2.4.3. *Measures based on a continuous scale*

When the effect size in the studies included in a meta-analysis is measured on a continuous scale, we primarily focus on the estimates of pooled mean difference and standardized mean difference.[10,12]

Suppose the n_{1i} and n_{2i} are the sample sizes, x_{1i} and x_{2i} are the means, of treatment and control group, respectively. The mean difference $y_i = \bar{x}_{1i} - \bar{x}_{2i}$, with standard error, s_i, calculated as

$$s_i^2 = s_{pi}^2\left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right), \quad \text{where} \quad s_{pi}^2 = \frac{(n_{1i}-1)s_{1i}^2 + (n_{2i}-1)s_{2i}^2}{n_{1i}+n_{2i}-2}\,.$$

2.4.2.1. Fixed-Effect model

2.4.2.1.1. Effect size is measured on the same scale

The pooled measure of size effect (mean difference) is $y_s = \frac{\sum w_i y_i}{\sum w_i}$, where $w_i = \frac{1}{s_i^2}$.

The Q statistic for homogeneity test is given by

$$Q = \sum w_i(y_s - y_i)^2 = \sum w_i y_i^2 - \frac{(\sum w_i y_i)^2}{\sum w_i}\,.$$

The 95% confidence interval of summary measure of effect size is $y_s \pm 1.96/\sqrt{\sum w_i}$.

Example 2. Table 7 presents data about the change in Kurtzke Disability Status Scale at two years in four randomized trials of the effect of azathioprine treatment in multiple sclerosis. The summary estimate of mean difference is given as follows:

Table 7. A meta-analysis for Change in Kurzke Disability Status Scale at two years in four randomized trials of the effect of azathioprine treatment in multiple sclerosis.

Study	Treated			Control			y_i	y_i^2	s_i^2	w_I	$w_i y_i$	$w_i y_i^2$
	x_{1i}	s_{1i}	n_{1i}	x_{2i}	s_{2i}	n_{2i}						
1	0.30	1.26	162	0.42	1.28	175	−0.12	0.0144	0.019	52.632	−6.316	0.758
2	0.17	0.90	15	0.83	0.98	20	−0.66	0.4356	0.105	9.524	−6.286	4.149
3	0.20	1.10	30	0.45	1.12	32	−0.25	0.0625	0.080	12.500	−3.125	0.781
4	0.17	1.38	27	0.42	1.36	25	−0.25	0.0289	0.145	6.897	−1.724	0.431
Total								0.5414		81.553	−7.451	6.119

Source: Yudkin *et al.* (1991). *Lancet* **338**: 1051–1055 and Petitti.[12]

2.4.2.1.1.1. Homogeneity test

$$Q = \sum w_i y_i^2 - \frac{(\sum w_i y_i)^2)}{\sum w_i} = 6.119 - \frac{(-17.451)^2}{81.553} = 2.385\,.$$

Here, $df = 3$, $\chi^2_{(0.05,3)} = 7.28 > 2.385$, $p > 0.05$ therefore, the null hypothesis that the studies are homogeneous is not rejected, and it is appropriate to use fixed-effects model to estimate the pooled weight mean.

2.4.2.1.1.2. Calculating the pooled effect size and its 95% confidence interval

$$y_s = \frac{\sum w_i y_i}{\sum w_i} = \frac{-17.451}{81.553} = -0.197\,,$$

$$y_s \pm 1.96 \Big/ \sqrt{\sum w_i} = -0.197 \pm (1.96/\sqrt{81.553}) = (-0.414, 0.02)\,.$$

The results of meta-analysis suggest, the pooled mean difference of the Kutzke Disability Scale for the effect of azathioprine treatment in multiple sclerosis is −0.197, but the results are statistically non-significant (95% confidence interval covers zero). Based on these results, we still cannot draw the conclusion that azathioprine is beneficial for multiple sclerosis.

2.4.2.1.2. Effect size is measured on different scale

When studies used different scales to measure effect, the standardized mean difference is calculated as the estimate of effect size. Let

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{s_{pi}},$$

then the pooled estimate of effect size is

$$d_s = \frac{\sum w_i d_i}{\sum w_i}, \tag{28}$$

where w_i is the weight assigned to each study. This weighted estimator of the effect size was shown by Hedges to be asymptotically efficient when sample sizes in the two groups are both greater than 10 and the effect sizes are less than 1.5.[18] When the sample sizes are about equal in the two groups and both greater than 10, the weight of each study can be estimated as follows:

$$w_i = \frac{2N_i}{8 + d_i^2}. \tag{29}$$

The 95% confidence interval for the pooled estimate of effect size is

$$d_s \pm 1.96/\sqrt{w_i}. \tag{30}$$

The Q statistic for homogeneity test is given by

$$Q = \sum w_i (d_s - d_i)^2 = \sum w_i d_i^2 - \frac{(\sum w_i d_i)^2}{(\sum w_i)}. \tag{31}$$

2.4.2.2. Random-effects model

If H_0 of homogeneity is rejected, which means that the between-study variance is relatively large, a random-effect model should be used.

The calculation of effect size is the same, that is, $d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{s_{pi}}$.

The pooled estimate of effect size and variance are

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i},$$

$$s_d^2 = \frac{\sum w_i (d_i - \bar{d})^2}{\sum w_i} = \frac{\sum w_i d_i^2}{\sum w_i} - \bar{d}^2, \tag{32}$$

where, $w_i = N_i = n_{1i} + n_{2i}$.

The random-effect model assumes $d_i = \delta_i + e_i$, with

$$\bar{\delta} = \bar{d}, \quad \bar{e} = 0 \quad \text{and} \quad s_i^2 = \frac{4k}{\sum w_i}\left(1 + \frac{\bar{d}^2}{8}\right). \tag{33}$$

(1) If $s_d^2 > s_e^2$, $s_\delta^2 = s_d^2 - s_e^2$, and the 95% confidence interval of pooled effect size is

$$\bar{d} \pm 1.96 s_\delta \,. \tag{34}$$

(2) If $s_d^2 \leq s_e^2$, $s_\delta^2 = 0$, and random-effects model is actually fixed-effect model, that is

$$d_i = \delta + e_i \,.$$

The standard error for $\bar{d}$ is

$$s_{\bar{d}} = \frac{s_e}{\sqrt{k}} \,. \tag{35}$$

Then the 95% confidence interval of pooled effect size is

$$\bar{d} \pm 1.96 s_{\bar{d}} \,. \tag{36}$$

In random-effect model, the statistic for homogeneity test is given by

$$x^2 = \frac{k s_d^2}{s_e^2} \,. \tag{37}$$

Under the null hypothesis of homogeneity, the statistic follows an approximate x_{k-1}^2 distribution.

Table 8. Data from meta-analysis of the effect of aminophylline treatment in severe acute asthma.

Study	$N_i(w_i)$	s_{pi}	d_i	$w_i d_i$	$w_i d_i^2$
1	20	0.76	−0.43	−8.6	3.698
2	50	320.00	−0.04	−2.00	0.08
3	48	0.65	−0.84	−40.32	33.869
4	24	0.42	−1.67	−40.08	66.934
5	29	0.22	−1.03	−29.87	30.766
6	20	17.00	−2.41	−48.2	116.162
7	23	0.62	−0.08	−1.84	0.147
8	13	110.00	0.26	3.38	0.879
9	23	2.10	2.93	67.39	197.453
10	51	6.30	0.51	26.01	13.265
11	61	0.50	0.72	43.92	31.622
12	66	0.67	0.03	1.98	0.059
13	40	0.58	−0.02	−0.8	0.016
··	468			−29.03	494.95

Source: Littenberg (1988). *JAMA*. **259**: 1678–1684 Petitti.[12]

Example 3. Table 8 presents data from a meta-analysis of the effect of aminophylline in severe acute asthma. The 13 studies included in the meta-analysis reported different measures on pulmonary function. A standardized mean difference should be used as common metric. The pooled estimate of effect size and 95% confidence interval are calculated as follows:

2.4.2.2.1. Homogeneity test

$$s_d^2 = \frac{\sum w_i(d_i - \bar{d})^2}{\sum w_i} = \frac{\sum w_i d_i^2}{\sum w_i} - \left(\frac{\sum w_i d_i}{\sum w_i}\right)^2$$

$$= \frac{494.95}{468} - \frac{(-29.03)^2}{468^2} = 1.054\,,$$

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i} = \frac{-29.03}{468} - 0.062\,,$$

$$s_e^2 = \frac{4k}{\sum w_i}\left(1 + \frac{\bar{d}^2}{8}\right) = \frac{4 \times 13}{468}\left[1 + \frac{(-0.062)^2}{8}\right] = 0.111\,,$$

$$x^2 = \frac{k s_d^2}{s_e^2} = \frac{13 \times (1.054)^2}{0.111} = 130.107\,.$$

$df = 12$, $x^2_{(0.05,12)} = 21.03$, $p < 0.05$, the null hypothesis of homogeneity is rejected, which means between-study variance is relatively large, and random-effects model should be adopted.

2.4.2.2.2. Calculating the summary effect size and its 95% confidence interval

$$s_\delta^2 = s_d^2 - s_e^2 = 1.054 - 0.111 = 0.94\,,$$

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i} = \frac{-29.03}{468} = -0.062\,,$$

$$\bar{d} \pm 1.96 s_\delta = -0.062 \pm (1.96 \times \sqrt{0.94}) = (-1.962, 1.838)\,.$$

The results of meta-analysis suggest that the effect of aminophylline treatment in severe acute asthma is statistically non-significant (95% confidence interval covers zero). In fact, the heterogeneity between studies is greatly large in the example, the smallest effect size is −0.02, whereas the

largest is 2.93. It is necessary to explore the source of heterogeneity before meta-analysis, and assess the source of biases and confounding. If the combinability of studies is poor, meta-analysis should be abandoned. The process above is just a typical example for computation.

3. Bayesian Methods in Random-Effects Models for Meta-analysis

The methods discussed above are basically frequentist procedures. There have been considerable discussions in the literature on the relative merits of fixed- and random-effects model. In practice, when combining the effect, the choice between fixed- and random-effects models is determined by the results of statistical tests of homogeneity (Q statistic). But the power of statistical tests of homogeneity is low. The results of random-effects model may be more "conservative", which leads to somewhat wider confidence intervals than the fixed-effects model. Little is known about the approach describing the random effects quantitatively. The appropriate treatment for small studies and extreme results included in meta-analysis is still unresolved in classic methods. Furthermore, the uncertainty of the parameters, such as the pooled effect size and variance, is not taken into account to use current approaches for meta-analysis.

Bayesian methods for meta-analysis give several options to deal with these problems and have been well-developed in the past decades. Under the Bayesian framework for random-effect model in meta-analysis, the parameter is an unknown random variable that has a specific distribution. The posterior distribution of parameter is derived from prior distribution and sample information available.

DuMouchel gave a fully Bayesian analysis of the hierarchical model with a complete conjugate prior structure.[19] Carlin developed and implements a fully Bayesian approach to meta-analysis for 2×2 tables, in which uncertainty about effects in comparable studies is represented by an exchangeable prior distribution.[20]

A Bayesian analysis requires integration of each of the conditional posterior distributions. Unfortunately, such integration cannot be performed in closed form in most situations. Approximate solution can be obtained through asymptotic or numerical techniques. With the great progress in Bayesian computational tools, especially the rapid development of Markov Chain Monte Carlo (MCMC) method, it is effective to deal with the problems that could not be resolved by classical meta-analysis method. Gibbs

sampling is a recently developed simulation tool for Bayesian inferences, obtaining the simulated joint posterior distribution from the full conditional distributions of parameters.[21,22]

In this section, the Bayesian approaches are introduced, especially the hierarchical model under a full Bayesian framework and the Gibbs sampling in random-effects model for meta-analysis.

3.1. *Bayesian meta-analysis for DuMouchel's model*

Supporse there are K individual studies included in the meta-analysis, and the effect for each study is $Y_1, Y_2, \ldots, Y_K$. The random-effects model is

$$Y_i = \mu_i + \varepsilon_i\,, \quad \varepsilon_i \sim N(0, \sigma_i^2)\,,$$

$$\mu_i = \mu + e_i\,, \quad e_i \sim N(0, \tau^2)\,,$$

with $\{\varepsilon_i, i = 1, \ldots, K\}$ and $\{e_i, i = 1, \ldots, K\}$ are independent. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_K)'$, $\mathbf{1} = (1, \ldots, 1)'$, $\boldsymbol{m} = (\mu_1, \ldots, \mu_k)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)'$, $\boldsymbol{e} = (e_1, \ldots, e_k)'$, $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2)$ and I the $K \times K$ identity matrix, then the random-effect model in matrix form as $Y|m \sim N(m, \sum)$, $m \sim N(1\mu, \tau^2 I)$.

Under the full Bayesian framework, we have the model

$$Y|m\,,\ \sigma^2 \sim N(m, \sigma^2 C)\,,$$

$$\sigma^2 \sim x^2(df_\sigma)\,,$$

$$m|\mu\,,\ \tau^2 \sim N(1\mu, \tau^2 H)\,,$$

$$\mu|\tau^2 \sim N(0, D \to \infty)\,,$$

$$\tau^{-2} \sim x^2(df_\tau)\,.$$

Here, σ^2, τ^2 and μ are hyperparameters, and $\mathbf{C}$ and $\mathbf{H}$ are assumed as known $K \times K$ covariance matrices with unknown scale factor σ^2 and τ^2, respectively. The degrees of freedom df_σ and df_τ for inverse-$\boldsymbol{\chi}^2$ prior distributions allow incorporation of how incorporation of how well known $\mathbf{C}$ and $\mathbf{H}$ are, respectively. The prior distribution for μ is the standard diffused and independent of τ^2. In fact, as noted by DuMouchel, these particular prior distributions are chosen for convenience, so that the posterior distribution of $\mathbf{m}$ given Y is a mixture of multivariate student-t distribution, each with degrees of freedom $df_\sigma + df_\tau + K - 1$. For computational convenience, however, he suggests using a multivariate normal approximation to the posterior, which can then be completely described through the posterior mean and covariance matrices.

Reparameterize the variance parameters as ϕ, let $\phi = \tau^2/\sigma^2$,

$$W(\phi) = (\phi\mathbf{H} + \mathbf{C})^{-1},$$

$$\beta(Y,\phi) = [1'W(\phi)1]^{-1}1'W(\phi)Y,$$

$$S(Y,\phi) = [Y - 1\beta(Y,\phi)]'W(\phi)[Y - 1\beta(Y,\phi)],$$

$$\gamma(Y,\phi) = \frac{df_\tau + df_\sigma/\phi + S(Y,\phi)}{df_\sigma + df_\tau + K - 3}.$$

The posterior estimate $E(\mu|Y)$ of μ is then given by integrating what is essentially the weighted least squares estimator of μ over the posterior density of ϕ, $f(\phi|Y)$

$$E(\mu|Y) = \int E(\mu|\phi,Y)f(\phi|Y)d\phi.$$

Similarly,

$$\mathrm{var}(\mu|Y) = \int \{\gamma(Y,\phi)[1'W(\phi)1]^{-1} + [\beta(Y,\phi) - E(\mu|Y)] \times [\beta(Y,\phi) - E(\mu|Y)]'\} f(\phi|Y)d\phi.$$

The approximate 95% credible interval for μ using $E(\mu|Y)$ and $\mathrm{var}(\mu|Y)$ and the normal distribution, will be

$$E(\mu|Y) \pm 1.96\sqrt{\mathrm{var}(\mu|Y)}.$$

Posterior mean of σ^2 are obtained using

$$E(\sigma^2|Y) = \int \gamma(Y,\phi)f(\phi|Y)d\phi.$$

3.2. *Bayesian meta-analysis for Carlin's model*

Carlin adopts a Bayesian approach to meta-analysis for 2×2 tables, in which an exchangeable prior distribution is used. A hierarchical normal model assumes that

$$Y_i|\mu_i,\ \sigma_i^2 \sim N(\mu_i, \sigma_i^2), \tag{38}$$

$$\mu_i|\mu,\ \tau^2 \sim N(\mu, \tau^2), \tag{39}$$

where σ_i represents the corresponding estimated standard error, which is assumed known without error. μ_i is interpreted as the "true effect" in ith study, which has an exchangeable normal prior, and also it means effects are independently and identically distributed conditional on the values of

unknown hyperparameters μ and τ^2. Here, τ^2 is between-study variance. Assume the prior distributions of μ and τ^2 are non-informative or locally uniform prior. Under the framework of Bayesian, the posterior distributions of quantities of interest, conditional on the variance hyperparameter, have closed form solutions. Let $B_i = \tau^2/(\tau^2 + \sigma_i^2)$, then we have

$$\hat{\mu} = E(\mu|Y, \tau^2) = \frac{\sum B_i Y_i}{\sum B_i}\,, \tag{40}$$

$$\text{var}(\mu|Y, \tau^2) = \frac{\tau^2}{\sum B_i}\,. \tag{41}$$

The posterior mean and variance for the individuals μ_i, conditional on both μ and τ^2, for each i, are

$$E(\mu_i|Y, \mu, \tau^2) = B_i Y_i + (1 - B_i)\mu\,, \tag{42}$$

$$\text{var}(\mu_i|Y, \mu, \tau^2) = B_i \sigma_i^2\,. \tag{43}$$

Note that, B_i is usually referred to as the shrinkage factor for the ith study. The larger the inter-study variation, τ^2, is the smaller the shrinkage B_i of the observed study effects. Because $0 \le B_i \le 1$, the mean is compromised between the average treatment effect μ and the observed study summary statistics, Y_i. When $\sigma_i^2 = 0$, shrinkage is maximized to $B_i = 1$ so that $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$ and the random-effects model reduces to the fixed-effects model.

Integrating Eqs. (42) and (43) over the posterior distribution of μ conditional on τ^2 we have

$$\begin{aligned} E(\mu_i|Y, \tau^2) &= \int E(\mu_i|Y, \mu, \tau^2) f(\mu|Y, \tau^2) d\mu \\ &= B_i Y_i + (1 - B_i)\hat{\mu}_\tau^2\,, \end{aligned} \tag{44}$$

$$\text{var}(\mu_i|Y, \tau^2) = B_i \sigma_i^2 + (1 - B_i)^2 \frac{\tau^2}{\sum B_i}\,. \tag{45}$$

The marginal likelihood function

$$f(Y|\tau^2) = \left(\frac{IIB_i}{(\tau^2)^{k-1} \sum B_i}\right)^{1/2} \exp\left\{-\frac{1}{2\tau^2}\left[\sum B_i Y_i^2 - \frac{(\sum B_i Y_i)^2}{\sum B_i}\right]\right\}$$

can be obtained by integrating μ out of the full likelihood. The posterior density for τ^2 is then

$$f(\tau^2|Y) = f(Y|\tau^2) f(\tau^2)\,,$$

where $f(\tau^2)$ is the prior density of τ^2. Carlin used Monte Carlo procedure to compute posterior density of estimates of interest, τ^2, τ and μ_i.

3.3. *Gibbs sampling in random-effects model for meta-analysis*

Gibbs sampling is a procedure for numerical integration of complex functions that has come from its origins in statistical mechanics, through image processing into modern statistics. It is based on a simple, although computationally demanding, idea. All unknown quantities are given some initial values. The technique then involves successively sampling from the conditional distribution of each variable in turn, given the current value of all the other variables. These "full conditional" distributions are often of fairly standard form. It can be shown that under broad conditions eventually one will be sampling from the correct posterior distributions of the unknown parameters. Recently, there are many literatures on this topic, both on methodology and applications.

The key feature of Gibbs sampling is, given a joint posterior density $\mathbf{P}(\boldsymbol{\theta}|\mathbf{X})$, K univariate full conditional densities (the distribution of each individual component of $\boldsymbol{\theta}$ conditional on known values of the data $\mathbf{X}$ and all other components) can be written down in close form.

Now we derive the full conditional distributions for parameters in Gibbs sampling based on random-effects model for meta-analysis. Consider the typical Bayesian hierarchical model as previouly described in Eqs. (38) and (39), that is

$$\begin{aligned}
&\text{Level I}: Y_i|\mu_i,\ \sigma_i^2 \sim N(\mu_i, \sigma_i^2)\,,\\
&\text{Level II}: \mu_i|\mu,\ \tau^2 \sim N(\mu, \tau^2)\,,\\
&\text{Level III}: \mu|(a,b) \sim N(a,b)\,, \quad \tau^2|(c,d) \sim IG(c,d)\,.
\end{aligned}$$

For computationally convenient, the prior distributions for hyperparameters μ and τ^2 are generally normal distribution and inverse Gamma distribution, respectively. Under the full Bayesian framework, all full conditional distributions are easily estimated using Gibbs sampling. Samples from the marginal posterior distributions of interest are simulated using the following full conditional distributions:

$$\begin{aligned}
&\mu_i|Y_1, \ldots Y_k, \mu_j \neq i, \mu\,,\\
&\qquad \tau_\mu^2 \sim N\left(Y_i\left(\frac{\tau^2}{\sigma_i^2+\tau^2}\right) + \mu\left(\frac{\sigma_i^2}{\sigma_i^2+\tau^2}\right),\ \frac{\sigma_i^2\tau^2}{\sigma_i^2+\tau^2}\right). \qquad (46)
\end{aligned}$$

$$\begin{aligned}
&\mu|Y_1, \ldots, Y_k, \ldots, \mu_k\,,\\
&\qquad \tau^2 \sim N\left(\sum_{i=1}^{K}\mu_i\left(\frac{Kb}{\tau^2+Kb}\right) + a\left(\frac{\tau^2}{\tau^2+Kb}\right), \frac{\tau^2}{\tau^2+Kb}\right), \qquad (47)
\end{aligned}$$

$$\tau^2|Y_1,\ldots,Y_k,\mu_1,\ldots,\mu_k\,,$$
$$\mu \sim IG\left(\frac{1}{2}K+c,\ \frac{1}{2}\sum_{i=1}^{K}(\mu_i-\mu)^2+d\right). \tag{48}$$

The processes involved in Gibbs sampling are: (i) μ_i, μ, and τ^2 are given some initial values; (ii) Gibbs sampling values are obtained in turn, from the conditional distributions in Eqs. (46), (47) and (48); (iii) update the Gibbs sampling values successively for t iterations. For each run a "burn-in" of m iterations is followed by a further t-m iterations during which the posterior marginal density of parameters, μ_i, μ, and τ^2 are computed; (iv) check the convergence of Gibbs sampling.

Note that, given the prior and conditional distribution, Gibbs sampling is easy to be carried out. When deriving the full conditional distributions, the prior and likelihood are conjugate in the model we discussed above. In some situations, it may be reasonable to assume that the prior of μ and τ^2 are non-informative prior, and the form of full conditional distribution is simpler.

WinBUGS is a program that carries out Bayesian inference for complex statistical analysis via MCMC simulation technique.[23] Using WinBUGS software, Gibbs sampling is easily implemented for many common models and distributions. The WinBUGS language allows the model to be specified by way of construction of a directed graphical model. The summary statistics for the variable, which calculate from the posterior distributions of parameters of interest, are given in the output. The software also produce the plots of the kernel density estimate, dynamic trace for sampling, and autocorrelation function for parameters.

3.4. *An example of Gibbs sampling for meta-analysis*

Table 9 gives the results of 16 case-control studies about the role of hepatitis B virus (HBV) infection, hepatitis C virus (HCV) infection, and dual infection in the patients with primary hepatocellular carcinoma (PHC) in Chinese.

The classic approaches for meta-analysis are not suitable for estimating quantitatively the risk of HBV, HCV and dual infection for PHC. As shown in Table 9, extreme values (zero) are observed for dual infection in the control groups in several studies, due to the quite low population-based dual infection rate. Classic approaches could not deal with the extreme values unless 0.5 is used to substitute zero or the studies containing zero

Table 9. The data of 16 case-control studies for HBV, HCV and dual infection in PHC.

	Statue of HBV, HCV Infection													
	Non-	HBV Infection				HCV Infection				Dual Infection				
Study No.	Infection ca/co	ca/co var(Y^{10})	OR^{10}		Y^{10}	ca/co	OR^{01}	Y^{01}	var(Y^{01})	ca/co var(Y^{11})	OR^{11}		Y^{11} ca/co	Total
1	42/198	77/40	9.08	2.21	0.07	6/8	3.54	1.26	0.32	15/1	70.71	4.26	1.10	140/247
2	33/101	102/10	31.22	3.44	0.15	3/3	3.06	1.12	0.71	14/1	42.85	3.76	1.11	152/115
3	34/81	43/8	12.81	2.55	0.19	4/3	3.18	1.16	0.63	11/0*	52.41	3.96	2.13	92/92
4	20/70	49/16	10.72	2.37	0.15	0/1*	1.75	0.56	3.06	8/0*	56.00	4.03	2.19	77/87
5	21/36	28/24	2.00	0.69	0.15	8/10	1.37	0.32	0.30	14/1	24.00	3.18	1.15	71/71
6	20/62	64/31	6.40	1.86	0.11	7/7	3.10	1.13	0.35	9/0*	55.80	4.02	2.18	100/100
7	9/75	50/21	19.84	2.99	0.19	11/3	30.56	3.42	0.55	30/1	250.00	5.52	1.16	100/100
8	35/122	53/20	9.24	2.22	0.11	3/1	10.46	2.35	1.37	5/1	17.43	2.86	1.24	96/144
9	9/123	51/14	49.79	3.91	0.21	4/2	27.33	3.31	0.87	6/1	82.00	4.41	1.29	70/140
10	22/278	232/73	40.16	3.69	0.07	49/8	77.40	4.35	0.19	58/2	366.45	5.90	0.57	361/361
11	5/57	87/45	22.04	3.09	0.25	6/3	22.80	3.13	0.72	11/4	31.35	3.45	0.56	109/109
12	7/109	45/16	43.79	3.78	0.24	3/1	46.71	3.84	1.49	9/2	70.07	4.25	0.76	61/128
13	11/179	80/26	50.07	3.91	0.15	3/1	48.82	3.89	1.43	10/2	81.36	4.40	0.70	104/208
14	13/105	79/105	6.08	1.80	0.11	3/4	6.06	1.80	0.67	15/6	20.19	3.01	0.32	110/220
15	15/120	100/27	29.63	3.39	0.12	23/2	92.00	4.52	0.62	12/1	96.00	4.56	1.16	150/150
16	10/138	23/10	31.74	3.46	0.25	4/4	13.80	2.62	0.61	1/0*	27.60	3.32	3.11	38/152

Source: Zhou Xuyu (1999). Postgraduate Dissertation of Sun Yat-Sen University of Medical Science.

ca: Case; co: Control.

*: The data of the included study contain extreme value, zero.

are excluded, and it may potentially lead to biasness. Furthermore, classic approaches neglect the uncertainly of the parameter of interest, especially the parameter for inter-study variance. The Bayesian approach based on random-effects model is flexible. Here, Gibbs sampling is adopted via the WinBUGS software to obtain pooled estimate of parameters by directly fitting three logistic models using the data available in 16 studies.

Arrangement of data and table notation for each individual study is shown in Table 10.

Table 10. Arrangement of data and table notation for 16 case-control studies.

	Non Infection	HBV Infection	HCV Infection	Dual Infection	Total
Case	a_i	c_i	e_i	g_i	m_i
Control	b_i	d_i	f_i	h_i	n_i

For each individual study, the odds ratio (OR), logarithm of OR, and variance for logarithm of OR, are given using following formula (here 00 denote non-infection, 10 denote HBV infection, 01 denote HCV infection, and 11 denote dual infection). The results are also shown in Table 9.

$$OR_i^{10} = \frac{c_i \times b_i}{d_i \times a_i}, \quad Y_i^{10} = \ln OR_i^{10}, \quad \mathrm{var}(Y_i^{10}) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i},$$

$$OR_i^{10} = \frac{e_i \times b_i}{f_i \times a_i}, \quad Y_i^{01} = \ln OR_i^{01}, \quad \mathrm{var}(Y_i^{01}) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{e_i} + \frac{1}{f_i},$$

$$OR_i^{11} = \frac{g_i \times b_i}{h_i \times a_i}, \quad Y_i^{11} = \ln OR_i^{11}, \quad \mathrm{var}(Y_i^{11}) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{g_i} + \frac{1}{h_i}.$$

Three logistic models are introduced for HBV, HCV, and dual infection. Take HBV infection for example. Let r_i^{00} and r_i^{10} denote the number of infection in the control and case group in ith study, arising from n_i^{00} and n_i^{10} subjects which are assumed to have probability of p_i^{00} and p_i^{10} of HBV infection, respectively. β_i^{10} is defined as

$$\beta_i^{10} = \mathrm{logit}(p_i^{10}) - \mathrm{logit}(p_i^{00}) = \ln\left(\frac{p_i^{10}}{1-p_i^{10}}\right) - \ln\left(\frac{p_i^{00}}{1-p_i^{00}}\right).$$

β_i^{10} is the true effect for ith individual study, that is, posterior mean of Y_i^{10}. The prior distribution of β_i^{10} is $N(\mu^{10}, (\sigma_\mu^{10})^2)$. μ^{10} is the pooled effect size of interest, and $(\tau^{10})^2$ is the variance of inter-study. Thus, the full model can be written as

HBV infection	HCV infection	Dual infection
$r_i^{00} \sim B(p_i^{00}, n_i^{00})$	$r_i^{00} \sim B(p_i^{00}, n_i^{00})$	$r_i^{00} \sim B(p_i^{00}, n_i^{00})$
$r_i^{10} \sim B(p_i^{10}, n_i^{10})$	$r_i^{01} \sim B(p_i^{01}, n_i^{01})$	$r_i^{11} \sim B(p_i^{11}, n_i^{11})$

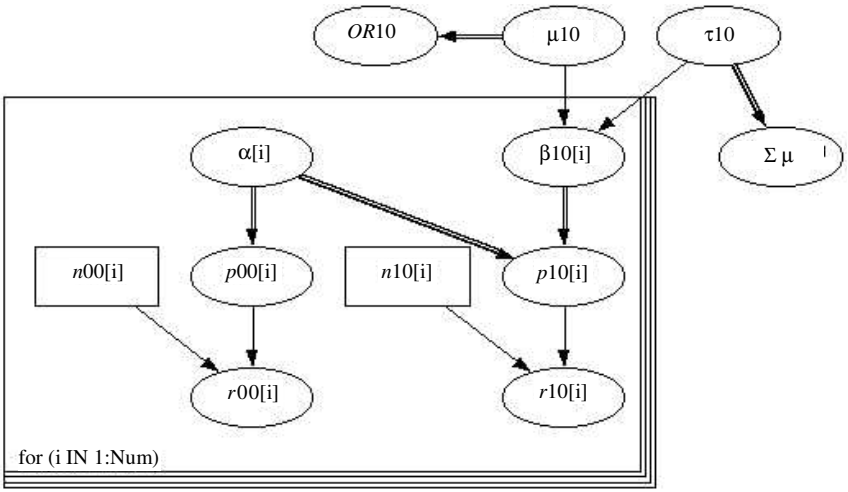


Fig. 2. Directed graphic model for HBV infection in WinBUGS.

$$\begin{array}{lll} \text{logit}(p_i^{00}) = \alpha_i^{10} & \text{logit}(p_i^{00}) = \alpha_i^{01} & \text{logit}(p_i^{00}) = \alpha_i^{11} \\ \text{logit}(p_i^{10}) = \alpha_i^{10} + \beta_i^{10} & \text{logit}(p_i^{01}) = \alpha_i^{01} + \beta_i^{01} & \text{logit}(p_i^{11}) = \alpha_i^{11} + \beta_i^{11} \\ \beta_i^{10} \sim N(\mu^{10}, (\tau^{10})^2) & \beta_i^{01} \sim N(\mu^{01}, (\tau^{01})^2) & \beta_i^{11} \sim N(\mu^{11}, (\tau^{11})^2)\,. \end{array}$$

Still take HBV infection for example. The prior distribution of hyperparameters μ^{10} and $(\tau^{10})^2$ are "non-informative", $\mu^{10} \sim N(0.0, 10^6)$, $(\sigma_\mu^{10})^2 \sim IG(10^{-3}, 10^{-3})$. The prior of parameter α_i^{10} are also "non-informative", $\alpha_i^{10} \sim N(0, 10^{-5})$.

In the WinBUGS, we can describe above models intuitively by the way of construction of directed graphical models, in which nodes in the graph represent the data and parameters of the model (See Fig. 2).

From the conditional independence conditions expressed in the graph, the joint distribution takes the form (ignoring n_i^{00}, n_i^{10}, and using the fact that p_i^{00}, p_i^{10} can be expressed in terms of α_i, β_i^{10})

$$\begin{aligned} p(r^{00}, r^{10}, \mu^{10}, \tau^{10}, \beta, \alpha) \propto \Pi_i [&p(r_i^{00}|\alpha_i, \beta_i^{10}) p(r_i^{01}|\alpha_i, \beta_i^{10}) \\ &\times p(\alpha_i) p(\beta_i^{00}|\mu^{10}, \tau^{10}) p(\mu^{10}) p(\tau^{10})\,. \end{aligned}$$

First 5000 iterations were used as a "burn in" in order to reduce the effect of initial value of parameters. Then running another 20,000 iterations and the summary statistics of posterior distribution for parameters were estimated. The main results of Gibbs sampling were seen in Table 11, which contains the means, standard deviations, and 95% confidence intervals from

Table 11. The results of Gibbs sampling for HBV, HCV and dual infection.

Status of Infection	Parameter	Mean	SD	95%CI
HBV Infection	μ^{10}	2.862	0.250	2.371–3.360
	OR^{10}	18.050	4.668	10.710–28.800
	τ^{10}	0.892	0.209	0.565–1.380
HCV Infection	μ^{01}	2.489	0.410	1.663–3.297
	OR^{01}	13.110	5.712	5.276–27.020
	τ^{01}	1.344	0.359	0.785–2.191
Dual Infection	μ^{11}	4.489	0.308	3.901–5.120
	OR^{11}	93.540	31.530	49.440–167.300
	μ^{11}	0.487	0.366	0.033–1.320

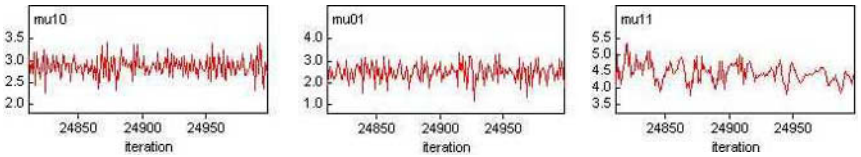


Fig. 3. The trace of Gibbs sampling for parameters μ^{10}, μ^{01}, μ^{11}.

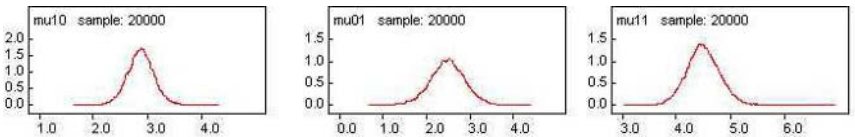


Fig. 4. The kernel density of Gibbs sampling for parameters μ^{10}, μ^{01}, μ^{11}.

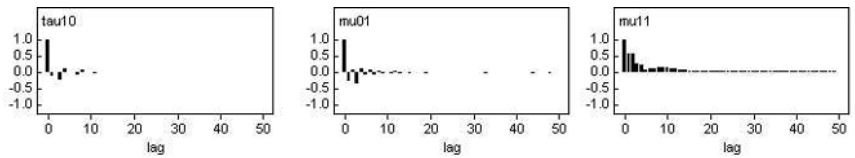


Fig. 5. The autocorrelation of Gibbs sampling for parameters μ^{10}, μ^{01}, μ^{11}.

the posterior distribution of parameters, μ^{10}, τ^2 and OR. WinBUGS also gives the posterior distributions of "true effect" for each study, β_i^{10}, β_i^{01} and β_i^{11}.

The trace, kernel density and autocorrelation plots for summary effects, μ^{10}, μ^{01}, μ^{11}, in WinBUGS were presented in Figs. 3–5. The dynamic traces

Table 12. The results of meta-analysis using DerSimonian-Laird method.

Parameter	HBV Infection	HCV Infection	Dual Infection
μ	2.815	2.423	4.026
95%CI	$2.359 \sim 3.270$	$1.639 \sim 3.208$	$3.553 \sim 4.500$
τ^2	0.711	1.816	0

showed that the Gibbs sampling tends to balance, the plots of kernel density estimate are smooth, and the autocorrelation of sampling is low.

For comparison, the classical DerSimonian-Laird random-effects model is used to estimate the pooled effect. For those studies in which the number of dual infection in control group is zero, 0.5 is substituted in order to calculate the OR. Results are shown in Table 12.

The results in Table 11 and 12 show that, for HBV, HCV infection, point estimations and 95% confidence intervals of summary effects for parameters μ^{10} and μ^{01} from Gibbs sampling and classical method are similar. But for dual infection, the number of dual infection in control group is quite small in most of 16 case-control studies, and four of them even contain zero. The pooled estimation of μ^{11} is 4.489 (95%CI is 3.901–5.120) via Gibbs sampling, and μ^{11} is 4.026 (95%CI is 3.553–4.500) using classical method, so the difference is relatively large. Moreover, the pooled estimation of between-study variance, $(\sigma_\mu^{11})^2$, is zero, when using DerSimonian-Larid method, which means the between-study variance could not be identified for dual infection and result in bias obviously.

In fact, when the data in meta-analysis contain many extreme values, the pooled estimation of true effect and variance is unreliable using classic methods, which are basically based on approximately normalization with large samples.

Gibbs sampling, almost the standard tool for Bayesian method, can be flexibly deal with a large of complex models that the classical approaches may difficult handle. The key of Gibbs sampling is to obtain the joint posterior distribution from the full conditional distributions of parameters using MCMC method, given the prior distribution and likelihood function. When the full conditional distribution is not given in a close form, Metropolis-Hastings method may be adopted.

Gibbs sampling can be effectively implemented using WinBUGS software, as demonstrated in the example. Furthermore, one can quite easily adjust for specific covariance that may influence the treatment effect by

fitting a new model under full Bayesian framework in WinBUGS. For the choice of prior distribution, *student-t* distribution as a population prior may be reasonable and proper in some situations.

4. Meta-analysis of Diagnostic Tests

Studies of the diagnostic accuracy of a test conducted at different centers often produce estimates of the sensitivity and specificity of a test that vary greatly. These differences may be due to random sampling variation and differences in the cutoff points of diagnostic test. In order to get summary results of diagnostic tests for different centers, meta-analysis of diagnostic tests is necessary.

The steps in conducting a meta-analysis of diagnostic tests are as follows:

(i) *Determine the objective and scope of meta-analysis*
In order to get the diagnostic accuracy, we must determine the test of interest, the disease of interest and reference standard by which it is measured, and the clinical question and context.

(ii) *Retrieve the relevant literatures and judge the validity of the literatures*
Extract and sort data of primary studies, and assess the eligibility and the quality of retrieved studies for inclusion in the analysis by two or more reader. Analyze the situations that come from different primary studies and get differences of diagnostic accuracy. The situations include as follows: If the reference standard is acceptable as a good representation of the true presence or absence of the disease of interest; if between the test and the reference standard are read independently each other; whether verification by the reference standard is done for all patients who had the test or a stratified random sample of them; if the design of primary studies is correct; how much the cutoff point is; whether the prevalence of population who accept the test is similar to etc.[24–26] The first author should consider the results from all readers overall.

(iii) *Estimation of a summary diagnostic accuracy of a test*
There are several statistical methods to calculate a summary diagnostic accuracy of a test. In this section, we will introduce *summary receiver operating characteristic* (SROC for short). In the last part of this section, we will introduce briefly the other methods to calculate a summary diagnostic accuracy of a test.

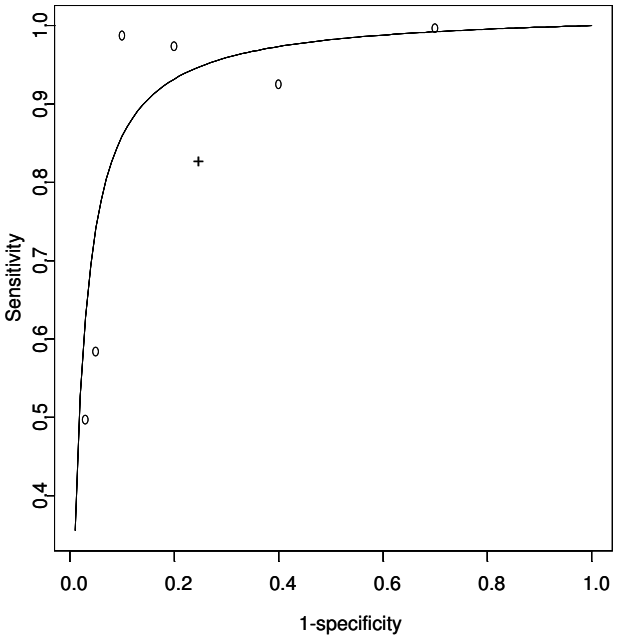


Fig. 6. Mean sensitivity and specificity cannot summarize results of diagnostic test in meta analysis.

While the goal of meta-analysis for diagnostic tests and the corresponding protocol development are similar in principals to the meta-analysis for clinical trials mentioned earlier, there are some specific issues. First, the performance of a diagnostic test is determined by the sensitivity and the specificity. Meta-analysis for diagnostic tests has two simultaneous endpoints. Secondly, because of the need to balance both sensitivity and specificity, the usual meta-analysis for rates, such as weighted average of sensitivity and specificity separately will miss the essential non-linear relationship between sensitivity and specificity. Figure 6 illustrates why the average sensitivity and specificity will not work for meta-analysis of a diagnostic test. Here, the six points are the observed means for sensitivity and specificity from six studies. The solid line is the corresponding ROC curve. When we take the average of sensitivity and specificity without considering their inter-relationship, we have the average point in "+", which is not on the ROC curve.[27–30] This figure demonstrated that using traditional meta-analysis on sensitivity and specificity separately results in the summary characteristics that do not belong to the test.

The mathematical reason for this difficulty is because of the non-linear relationship between sensitivity and specificity. Any transformation that reasonably related 1-specificity in a linear form to sensitivity will help to simplify the meta-analysis of diagnostic tests. One of these approaches is the SROC.

4.1. *SROC analysis*

In order to evaluate the diagnostic accuracy of a test, at first we must be aware of the true presence or absence of the disease of interest. The standard which identifies an individual as disease (case) or non-disease (control) is the reference standard or golden standard. Golden standards which are used in medical research include biopsy, autopsy, surgery exploration, follow-up and so on. Although a golden standard need not be perfect, it should be more credible than the diagnostic test of interest and it should be independent with the diagnostic test. For the individuals which are determined case or control by golden standard, the results which are determined by a diagnostic test are labeled as positive or negative respectively. The data can be presented as the form of fourfold table. Among them there are two true results, that is, case is diagnosed as positive (true positive, TP) and control is diagnosed as negative (true negative, TN). There are two false results, that is, case is diagnosed as negative (false negative, FN) and control is diagnosed as positive (false positive, FP) (see Table 13).

The true positive rate (TPR), i.e. sensitivity, is the probability that a test result is positive in patients with disease of interest, namely:

$$TPR = a/(a+c)\,, \tag{49}$$

$(1-TPR) = c/(a+c)$ is called false negative rate.

The false positive rate (FPR) which equals to (1-specificity), is the probability that a test result is positive in patients without the disease of

Table 13. A diagnostic test results for 2×2 table.

Test Results	Golden Standard		Total
	Case	Control	
Positive	a(TP)	b(FP)	$a+b$
Negative	c(FN)	d(TN)	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=N$

interest, namely:

$$FPR = b/(b+d)\,, \tag{50}$$

$(1 - FPR) = d/(b+d)$ is true negative rate or specificity.

4.1.1. *SROC linear regression model*

For TPR and FPR, we use logit translation, namely:

$$\text{logit}(TPR) = \ln[TPR/(1-TPR)]\,, \tag{51}$$

$$\text{logit}(FPR) = \ln[FPR/(1-FPR)]\,, \tag{52}$$

let

$$D = \text{logit}(TPR) - \text{logit}(EPR)\,, \tag{53}$$

$$S = \text{logit}(TPR) + \text{logit}(FPR)\,. \tag{54}$$

Through the formula (53), we can get:

$$\begin{aligned} D &= \ln \frac{TPR/(1-TPR)}{FPR/(1-FPR)} \\ &= \ln \frac{\text{true positive rate} \times \text{false negative rate}}{\text{false positive rate} \times \text{true negative rate}} = \ln OR\,. \end{aligned} \tag{55}$$

Through the formula (54), we can get:

$$\begin{aligned} S &= \ln \frac{TPR \times FPR}{(1-TPR)(1-FPR)} \\ &= \frac{\text{true positive rate} \times \text{false positive rate}}{\text{true negative rate} \times \text{false negative rate}}\,. \end{aligned} \tag{56}$$

Let D be dependent variable and S be independent variable. In order to make *SROC curve* into a linear in (S, D) plane, we establish an *SROC* linear regression model as:

$$\hat{D} = A + B \times S\,, \tag{57}$$

where D is a log odds ratio [see formula (55)], representing the odds of a positive test result among people with the disease relative to the odds of a positive test result among people without the disease. D value can reflect the distinguishing ability of a diagnostic test. S is a measure of threshold for classifying a test as positive, which has a value of 0 when a sensitivity equals specificity [see formula (56)]. It becomes positive, i.e. $S > 0$, when

a threshold is used that increases sensitivity (and decreases specificity) and becomes negative, i.e. $S > 0$, when a threshold is used that decreases sensitivity (and increases specificity). A is the intercept of the linear model and a log odds ratio when sensitivity equals specificity ($S = 0$). B is the regression coefficient and examines the extent to which the odds ratio (D) is dependent on the threshold (S) used. If the regression coefficient (B) is near zero and not statistically significant, test accuracy for each primary study can be summarized by a common odds ratio given by the intercept A.

4.1.2. *Solving the parameter of SROC linear regression model*

Unweighted least squares linear regression, weighted least squares linear regression, and robust method can be used to solve the parameters of SROC linear regression model (57).

4.1.2.1. Conventional least squares methods

This method can be introduced in a general statistical textbook. The parameter A and B are solved by making minimum of the square sum of the difference between observed value and fitted value (i.e. residual). The disadvantage of the method is not paying more attention to larger study, it does not consider the sample size of primary studies.

4.1.2.2. Weighted least squares method

In order to give more weight to studies of larger sample size, weighted least squares method can be used, weighting each observation using the reciprocal of the variance of log odds ratio ($\ln OR$). The parameter A and B are solved by making minimum of the square sum of weighted residual. Let a, b, c, and d be the number of true positive, false positive, false negative, and true negative respectively (see Table 13). The weight can be calculated by

$$W = [\mathrm{var}(D)]^{-1} = (1/a + 1/b + 1/c + 1/d)^{-1}. \tag{58}$$

To deal with the 0 of denominator, if a cell of cross-classification of test and golden standard value is 0 among a, b, c, d, we add 0.5 to each cell of the primary study. The observation values of the study become $(a + 0.5)$, $(b + 0.5)$, $(c + 0.5)$, and $(d + 0.5)$.

Weighted method is inappropriate if one assumes that individual primary studies are all measuring the same underlying test accuracy. So,

Moses, Shapiro and Littenberg suggested a robust modeling technique of SROC in 1993.[25]

4.1.2.3. Robust method

D plotted against S, coordinate points (S, D) of primary studies are plotted. According to the value of S value, we order the scatters (S, D) pairs and divide the points into 3 approximately equal groups. The total of studies divide 3 and round it, we can get the number of scatter points for left or right side. For example, 10 scatter points are divided, left or right side is round $(10/3) = 3$ respectively. Find the medians of S and D among the left and right side respectively and label them. Link the labeled scatter point into a line. The slope of the line is regression coefficient B. The intercept A is derived by positioning the line so that half of the points lie above and half below it. Let (S_1, D_1) and (S_2, D_2) represent two points which are on the line and far from each other (for example, the two median points of S and D among left or right side respectively). Using the follow formula, we can calculate the regression parameters A and B.

$$A = \frac{D_1 S_2 - D_2 S_1}{S_2 - S_1}, \quad B = \frac{D_2 - D_1}{S_2 - S_1}. \tag{59}$$

Solveing the parameter of SROC curve using robust method see Fig. 7. This figure is plotted using the S and D in Table 14. The regression coefficient of the line is 0.0011. The line parallel approximately the abscissa.

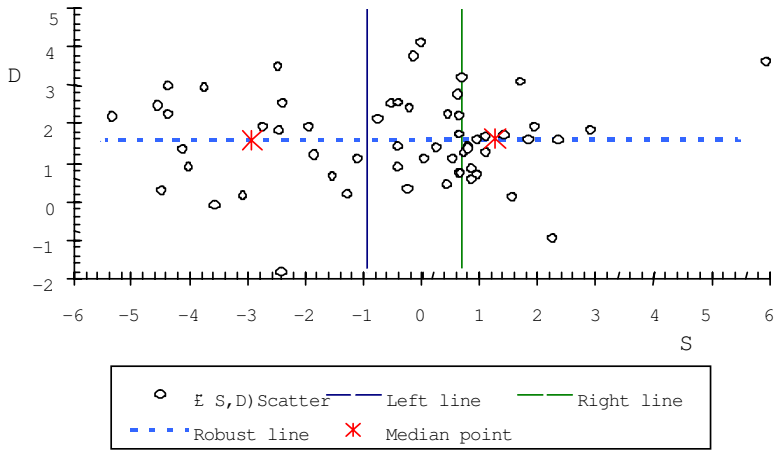


Fig. 7. Solving the parameter of SROC curve using robust.

Table 14. The data of Pap test from 59 primary studies.

Studies	TP	FP	FN	TN	Sensitivity	1-specificity	Weight		
i	a	b	c	d	TPR	FPR	W	D	S
1	8	3	23	84	0.258	0.034	1.947	2.276	−4.388
2	31	3	43	14	0.419	0.176	2.173	1.213	−1.868
3	70	12	121	25	0.66	0.324	6.855	0.187	−1.281
4	65	10	6	6	0.915	0.625	2.229	1.872	2.893
5	20	3	19	4	0.513	0.429	1.458	0.339	−0.236
6	35	92	20	156	0.636	0.371	10.433	1.088	0.032
7	39	8	111	270	0.260	0.029	6.122	2.473	−4.565
8	567	117	140	157	0.802	0.427	41.976	1.693	1.105
9	25	37	11	18	0.694	0.673	4.684	0.100	1.542
10	38	28	17	37	0.691	0.431	6.762	1.083	0.526
11	45	35	15	48	0.750	0.422	7.231	1.414	0.783
12	71	87	10	306	0.877	0.221	7.761	3.218	0.702
13	4.5	0.5	36.5	5.5	0.110	0.083	0.411	0.305	−4.491
14	2	2	3	21	0.400	0.087	0.724	1.946	−2.757
15	5	9	3	182	0.625	0.047	1.539	3.518	−2.496
16	38	21	7	62	0.844	0.253	4.293	2.774	0.609
17	4	2	16	31	0.200	0.061	1.184	1.355	−4.127
18	87	13	12	9	0.879	0.591	3.535	1.613	2.349
19	15	3	65	15	0.188	0.167	2.074	0.143	−3.076
20	41	1	61	29	0.402	0.033	0.930	2.970	−3.765
21	76	12	11	12	0.874	0.500	3.694	1.933	1.933
22	10	4	48	174	0.172	0.022	2.655	2.204	−5.341
23	28	11	28	77	0.500	0.125	5.704	1.946	−1.946
24	3.5	0.5	5.5	1.5	0.389	0.250	0.319	0.647	−1.551
25	79	26	13	182	0.859	0.125	7.489	3.750	−0.141
26	61	20	27	35	0.693	0.364	7.576	1.375	0.255
27	62	20	16	49	0.795	0.290	6.710	2.251	0.458
28	284	31	68	68	0.807	0.313	15.340	2.215	0.644
29	66	25	20	44	0.767	0.362	7.820	1.759	0.629
30	40	43	12	47	0.769	0.478	6.542	1.293	1.115
31	11	1	1	2	0.917	0.333	0.386	3.091	1.705
32	23	50	10	44	0.697	0.532	5.370	0.705	0.961
33	65	13	42	13	0.607	0.500	5.180	0.437	0.437
34	1269	928	264	1084	0.828	0.461	152.068	1.725	1.415
35	223	22	74	83	0.751	0.210	13.245	2.431	−0.225
36	154	30	20	237	0.885	0.112	10.633	4.108	−0.026
37	6	2	12	81	0.333	0.024	1.312	3.008	−4.394
38	7	4	3	4	0.700	0.500	1.024	0.847	0.847
39	12	5	11	60	0.522	0.077	2.558	2.572	−2.398
40	348	41	212	103	0.621	0.285	23.987	1.417	−0.426

Table 14. *Continued.*

Studies	TP	FP	FN	TN	Sensitivity	1-specificity	Weight		
i	a	b	c	d	TPR	FPR	W	D	S
41	8	4	11	34	0.421	0.105	2.019	1.822	−2.459
42	12.5	2.5	6.5	0.5	0.658	0.833	0.380	−0.956	2.263
43	95	9	2	1	0.979	0.900	0.617	1.664	6.058
44	40	18	20	19	0.667	0.486	5.459	0.747	0.639
45	71	13	20	18	0.780	0.419	5.087	1.592	0.942
46	1204	186	455	241	0.726	0.436	79.655	1.232	0.714
47	6	20	51	27	0.105	0.426	3.659	−1.840	−2.440
48	35	9	12	12	0.745	0.429	3.264	1.358	0.783
49	10	31	5	32	0.667	0.492	2.751	0.725	0.661
50	3	5	3	15	0.500	0.250	1.071	1.099	−1.099
51	118	40	44	183	0.728	0.179	16.216	2.507	−0.534
52	13	3	82	17	0.137	0.150	2.078	−0.107	−3.576
53	38	14	13	62	0.745	0.184	5.241	2.561	−0.415
54	14	25	67	291	0.173	0.079	7.705	0.889	−4.020
55	12	14	6	12	0.667	0.538	2.471	0.539	0.847
56	238	52	2	16	0.992	0.765	1.707	3.600	5.958
57	111	44	20	39	0.847	0.530	9.313	1.593	1.834
58	491	165	250	701	0.663	0.191	73.944	2.122	−0.772
59	48	16	38	31	0.558	0.340	7.047	0.895	−0.428

4.1.3. *Establishing SROC curve regression model*

Both regression parameters A and B are solved using above methods. We can establish SROC curve regression model as follow:

$$TPR = \left[1 + e^{-A/(1-B)} \left(\frac{1 - FPR}{FPR}\right)^{(1+B)/(1-B)}\right]^{-1}, \tag{60}$$

where TPR represents true positive rate and FPR represents false positive rate.

For a general ROC analysis, the area under ROC cure is taken as the diagnostic accuracy of a test. For SROC analysis, we can take TPR^* as the diagnostic accuracy of a test. TPR^* is the sensitivity taken by SROC curve of Eq. (60) and line equation

$$TPR + FPR = 1\,. \tag{61}$$

It reflects the extent to which SROC curve approach the top left corner. The larger the value of TPR^* is, the higher the diagnostic accuracy of a test is. $TPR+FPR = 1$ is a line through both the top left corner (1, 0) and

the bottom right corner (0, 1). For the line, sensitivity equals specificity, namely $S = 0$.

Using $S = 0$ and formula (54), we have

$$S = \text{logit}(TPR) + \text{logit}(FPR) = 0$$

or

$$\text{logit } FPR = -\text{logit } TPR\,. \tag{62}$$

Substituting formula (62) into formula (53), we have

$$D = \text{logit}(TPR) - \text{logit}(FPR) = 2 \text{ logit}(TPR) = A + B \cdot S = A\,,$$

and

$$\text{logit}(TPR) = A/2\,, \tag{63}$$

and

$$TPR = (1 + e^{-A/2})^{-1}\,. \tag{64}$$

In order not to be confused with general TPR, we take the diagnostic accuracy of a test of SROC curve as

$$TPR^* = (1 + e^{-A/2})^{-1}\,.$$

Its standard error can be calculated by

$$SE(TPR^*) = \frac{SE(\hat{A})}{8[\cosh(A/4)]^2}\,, \tag{65}$$

where $SE(\hat{A})$ is the standard error of the intercept A of linear regression model. Cosh(.) is the hyperbolic cosine function.

To compare the diagnostic accuracy between 2 independent groups, if the numbers of the primary studies is large enough (more than 10), we can use Z statistic, namely

$$Z = \frac{TPR_1^* - TPR_2^*}{\sqrt{SE^2(TPR_1^*) + SE^2(TPR_2^*)}}\,, \tag{66}$$

where Z is a quantile from the standard normal distribution. Both TPR_1^* and TPR_2^* are the diagnostic accuracy of compared SROC curves. Either $SE(TPR_1^*)$ or $SE(TPR_2^*)$ is the standard error of TPR_1^* or TPR_2^*, respectively.

If the regression coefficient of a SROC curve has $B = 0$, for the FPR of each primary study, the confidence interval of the TPR can be taken as:

$$\left(\left[1 + e^{-A_L}\left(\frac{1 - FPR}{FPR}\right)\right]^{-1}, \left[1 + e^{-A_U}\left(\frac{1 - FPR}{FPR}\right)\right]^{-1}\right), \quad (67)$$

where A_L and A_U are the lower and upper confidence interval of the intercept A respectively.

4.1.4. *Analysis using an example*

The Pap test involves the collection, preparation, and examination of exfoliated cervical cells. It is quick, noninvasive, and relatively inexpensive. These properties make the test appealing for cervical precancer. Currently some doctors use it as a screening test and as a follow-up test for women. Because the accuracy of the test is affected by a doctor understanding the natural history of cervical cancer, morbidity of cervical cancer, the number of sampling of cell, the diagnostic accuracy of test has been reported wide variation. The value of the sensitivity and the specificity ranges from 11% to 99% and from 14% to 97% respectively. The method of SROC analysis is illustrated using the data of 59 primary studies reported by Fahey, Irwig and Macaskill.[26]

Example 5. In the Data of Fahey, Irwigand and Macaskil, the number of true positive (TP, a), false positive (FP, b), false negative (FN, c), true negative (TN, d) is not presented, but the number of with disease, the number of without disease, sensitivity, (1-specificity) were given. For the method need them, according to the known data we calculate a, b, c, d (see Table 14). Because there were 0s in b of 13th and 24th and d of 42nd of primary studies, to avoid 0 of denominator, 0.5 was added to a, b, c, d of the 3 studies (see Table 14).

In the 1st study, the weight was calculated using formula (58),

$$W_1 = \left(\frac{1}{8} + \frac{1}{3} + \frac{1}{23} + \frac{1}{84}\right)^{-1} = 1.947\,.$$

The true positive rate is calculated by formula (49), i.e. $TPR = 8/(8{+}23) = 0.2581$. The false positive rate is calculated by formula (50), i.e. $FPR = 3/(3 + 84) = 0.0345$. D and S are calculated by formula (55) and (56) respectively, i.e. $D = \ln\frac{0.2581(1-0.0345)}{0.0345(1-0.2581)} = 2.276$, $S = \ln\frac{0.2581\times 0.0345}{(1-0.2581)(1-0.0345)} = -4.388$ and so on.

Using above weight, the weighted least square linear regression model is established taking D as dependent variable, S as independent variable. The residual standard deviation of the weighted model is 2.430. Intercept is $A = 1.720$ and its standard error is $SE(A) = 0.100$. The result of t test is $t = 17.227$ and $P = 0.001$. Using $A \pm t_{0.05,58}SE(A)$, the 95% confidence interval of A is 1.520–1.920. The results suggest that the difference between A and 0 has statistical significance under 0.05 test level.

The regression coefficient is $B = -0.015$ and its standard error is $SE(B) = 0.070$. The result of t test is $t = -0.215$, $P = 0.830$. The results suggest that the difference between B and 0 has no statistical significance under 0.05 test level.

The odds ratio is $\exp(A) = \exp(1.720) = 5.585$. It suggests the odds of positive test in abnormal group is larger than in the normal group.

According to formula (64) and (65), we can get the diagnostic accuracy of the test $TPR^* = 0.703$ is and its standard error is $SE(TPR^*) = 0.010$.

The general least square linear regression model is established taking D as dependent variable and S as independent variable. The residual standard deviation of the model is 1.1144. Intercept is $A = 1.590$ and its standard error is $SE(A) = 0.151$. The result of t test is $t = 10.522$ and $P = 0.001$. Using $A \pm t_{0.05,58}SE(A)$, the 95% confidence interval of A is 1.288–1.892. The results suggest that the difference between A and 0 has statistical significance under 0.05 test level.

Regression coefficient is $B = -0.020$ and its standard error is $SE(B) = 0.063$. The result of t test is $t = 0.319$, $P = 0.751$. The results suggest that the difference between B and 0 has no statistical significance under 0.05 test level.

The odds ratio is $\exp(A) = \exp(1.590) = 4.904$. It suggests the odds of positive test in abnormal group is larger than in the normal group.

According to formula (64) and (65), we can get the diagnostic accuracy of the test is $TPR^* = 0.689$ and its standard error is $SE(TPR^*) = 0.016$.

D plotted against S, coordinate points (S, D) of 59 primary studies are plotted. According to the value of S value, we order the scatters (S, D) pairs and divide the points by 3 approximately equal groups. The 59 studies were divided into 3 groups. The number of scatter points for left or right side is round $(59/3) = 20$. The medians of S and D among left side are $(S_1, D_1) = (-2.916, 1.588)$ and among right side are $(S_2, D_2) = (1.265, 1.593)$. The intercept A and regression coefficient B are $A = 1.5914$ and $B = 0.0011$ respectively obtained by formula (59). So, the linear regression model is

$$\hat{D} = 1.5914 + 0.0011S\,.$$

To make half the points lie above and half below the line, we need to move the line up and down. In this situation, the regression parameter is 0.0011 constantly and the intercept A is derived by positioning the line. In fact, this equals that the number of positive sign equals negative sign of residual which is from the difference between observed value and predicted value. Through changed the A value many times, we got the line which scatter points lie above equals below approximately and the intercept is $A = 1.5914$, the odds ratio is $\exp(A) = \exp(1.5914) = 4.9106$, and the diagnostic accuracy is $TPR^* = 0.6891$.

The results obtained from the weighted linear regression, general linear regression, robust regression are presented in Table 15.

Table 15. The diagnostic accuracy and related result from 3 methods.

Methods	A	$SE(A)$	95%CL	B	$SE(B)$	95%CL	TPR^*	SE (TPR^*)	Odds Ratio
Weighted	1.720	0.100	$1.520 \sim 1.920$	-0.015	0.070	$-0.155 \sim 0.125$	0.703	0.010	5.585
Unweighted	1.590	0.151	$1.288 \sim 1.892$	0.020	0.063	$-1.241 \sim 1.281$	0.689	0.016	4.904
Robust	1.591	–	–	0.001	–	–	0.689	–	4.911

Substituting A, B of 3 methods into formula (60), we obtained the SROC curves of weighted, unweighted and robust method respectively. They are as follows:

$$TPR_{\text{weighted}} = \left[1 + e^{-1.694}\left(\frac{1-FPR}{FPR}\right)^{0.970}\right]^{-1},$$

$$TPR_{\text{unweighted}} = \left[1 + e^{-1.623}\left(\frac{1-FPR}{FPR}\right)^{1.041}\right]^{-1},$$

$$TPR_{\text{robust}} = \left[1 + e^{-1.593}\left(\frac{1-FPR}{FPR}\right)^{1.002}\right]^{-1}.$$

To obtain the smooth SROC curve, let FPR from 0.002 to 0.998 (can also setup other value) and increase in arithmetic series 0.002. According to the above SROC curve equations TPR is calculated. 499 SROC coordinate points were obtained. Using the above coordinate points obtained and point (0, 0), (1, 1) we can plot the smooth SROC curve. Figure 8 presents smooth SROC curve and SROC coordinate points of the 59 primary studies from Table 14.

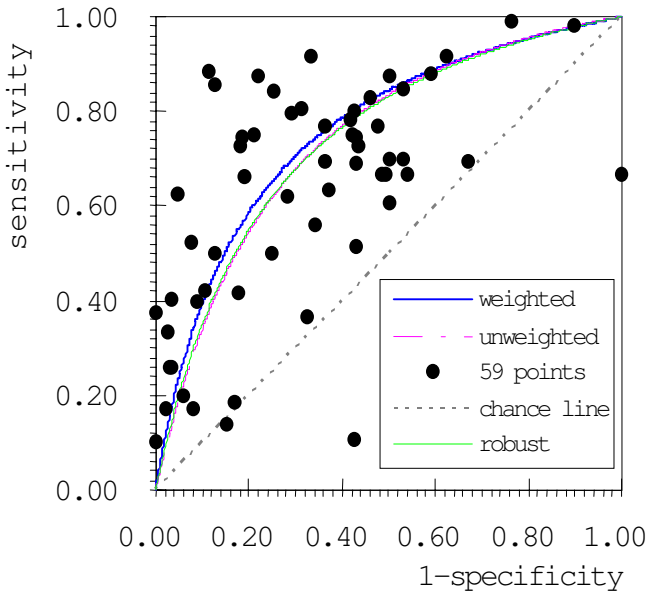


Fig. 8. SROC curves of 3 methods and the scatter points of 59 primary studies.

From Fig. 8, it is suggested that the area under curve of weighted method is larger. Of unweightd method and of robust method are similar. These results are consistent with the diagnostic accuracy TPR^* and odds ratio in Table 15.

If the association between weighted and unweighted method is ignored and assuming TPR^* obtained by 2 methods is approximately normal distribution. The formula (66) can be used to test the difference between 2 TPR^* s. The result of test is $Z = 0.7132$, $P = 0.4757$ for two-side test. This suggests that the TPR^* difference between weighted and unweighted method have not statistical significance.

4.1.5. *The SAS code of solving SROC curves parameter*

SAS code 1. SROC analysis of weighted, unweighted and robust method.[31]

Number	SAS Code
01	`OPTIONS LS=76 PS=MAX NODATE;`
02	`%LET N=59;  /*the number of primary studies N= **********/`
03	`%LET A_ROB=1.5914;  /* changed robust intercept A_ROB= **********/`
04	`DATA SROC;  RETAIN I 0;`

SAS code 1. (*Continued*).

Number	SAS Code
05	INPUT TP FN FP TN@@;
06	I+1; N_RL=ROUND(&N/3);
07	W=1/ (1/TP+1/FN+1/FP+1/TN); TPR=TP/(TP+FN);FPR=FP/(FP+TN);
08	D=LOG(TPR/(1-TPR))-LOG(FPR/(1-FPR));
09	S=LOG(TPR/(1-TPR))+LOG(FPR/(1-FPR));
10	CARDS;
11	8 23 3 84 76 11 12 12 8 11 4 34
12	31 43 3 14 10 48 4 174 12.5 6.5 2.5 0.5
13	70 121 12 25 28 28 11 77 95 2 9 1
14	65 6 10 6 3.5 5.5 0.5 1.5 40 20 18 19
15	20 19 3 4 79 13 26 182 71 20 13 18
16	35 20 92 156 61 27 20 35 1204 455 186 241
17	39 111 8 270 62 16 20 49 6 51 20 27
18	567 140 117 157 284 68 31 68 35 12 9 12
19	25 11 37 18 66 20 25 44 10 5 31 32
20	38 17 28 37 40 12 43 47 3 3 5 15
21	45 15 35 48 11 1 1 2 118 44 40 183
22	71 10 87 306 23 10 50 44 13 82 3 17
23	4.5 36.5 0.5 5.5 65 42 13 13 38 13 14 62
24	2 3 2 21 1269 264 928 1084 14 67 25 291
25	5 3 9 182 223 74 22 83 12 6 14 12
26	38 7 21 62 154 20 30 237 238 2 52 16
27	4 16 2 31 6 12 2 81 111 20 44 39
28	87 12 13 9 7 3 4 4 491 250 165 701
29	15 65 3 15 12 11 5 60 48 38 16 31
30	41 61 1 29 348 212 41 103
31	;
32	TITLE 'to calculate sensitivity, 1-specificity, weight, D, S using
33	TP,FN,FP,TN ' ;
34	PROC PRINT;RUN;
35	TITLE 'weighted regression model?W=1/(VAR(LN(OR)))';
36	PROC REG DATA=SROC OUTEST=W OUTSEB SIMPLE;
37	MODEL D=S; WEIGHT W;
38	DATA W1; SET W;
39	PROC TRANSPOSE DATA=W PREFIX=AW OUT=WW;
40	DATA XX1; SET WW; OR_SROC=EXP(AW1);
41	A_L=AW1-AW2*TINV(1-0.05/2,&N-1); A_U=AW1+AW2*TINV(1-0.05/2,&N-1);
42	TPR_S_W=1/(1+EXP(-AW1/2)); SE_TPR_W=AW2/(8*(COSH(AW1/4))**2);
43	IF _NAME_^='INTERCEP' THEN DO; A_L=.; A_U=.;
44	OR_SROC=.; TPR_S_W=.; SE_TPR_W=.; END;
45	DATA XXX1; SET XX1; IF _NAME_^='INTERCEP' THEN DELETE; PROC PRINT;
46	PROC REG DATA=SROC OUTEST=NW OUTSEB SIMPLE;
47	MODEL D=S;
48	TITLE '******unweighted general linear regression model *********';
49	DATA NW1; SET NW;
50	PROC TRANSPOSE DATA=NW PREFIX=A OUT=WW;

SAS code 1. (*Continued*).

```
Number                                  SAS Code
  51   DATA XX2 ;           SET WW;   OR_SROC=EXP(A1);
  52   A_L=A1-A2*TINV(1-0.05/2,&N-1);A_U=A1+A2*TINV(1-0.05/2,&N-1);
  53    TPR_STAR=1/(1+EXP(-A1/2)); SE_TPR=A2/(8*(COSH(A1/4))**2);
  54   IF _NAME_^='INTERCEP' THEN DO; A_L=.; A_U=.;
  55    OR_SROC=.; TPR_STAR=.; SE_TPR=.; END;
  56   DATA XXX2; SET XX2;IF _NAME_^='INTERCEP' THEN DELETE; PROC PRINT;
  57     DATA XXX; MERGE XXX1 XXX2;
  58     KEEP TPR_S_W TPR_STAR SE_TPR_W SE_TPR Z_SROC P_SROC;
  59        Z_SROC=(TPR_S_W-TPR_STAR)/(SE_TPR_W**2+SE_TPR**2)**0.5;
  60   P_SROC=2*(1-PROBNORM(Z_SROC)); PROC PRINT;
  61   TITLE 'compare the TPR_STAR between unweighted and weighted regression
  62    model ';
  63   DATA SROCS; SET SROC;PROC SORT; BY S;
  64   DATA BS1;    KEEP II S D; SET SROCS;
  65   II+1;      IF II>N_RL THEN DELETE ;
  66   PROC UNIVARIATE DATA=BS1 NOPRINT; VAR S; OUTPUT OUT=A1 MEDIAN=S1;
  67   PROC UNIVARIATE DATA=BS1 NOPRINT; VAR D; OUTPUT OUT=A2 MEDIAN=D1;
  68   DATA BS2;    KEEP II S D; SET SROCS;
  69   II+1;      IF II<=&N-N_RL THEN DELETE ;
  70   PROC UNIVARIATE DATA=BS2 NOPRINT; VAR S; OUTPUT OUT=A3 MEDIAN=S2;
  71   PROC UNIVARIATE DATA=BS2 NOPRINT; VAR D; OUTPUT OUT=A4 MEDIAN=D2;
  72     DATA AA; MERGE A1 A2 A3 A4;
  73   A_ROBUST=(D1*S2- D2*S1)/(S2-S1); B_ROBUST=(D2-D1)/(S2-S1);
  74    OR_ROB=EXP(&A_ROB); TPR_ROB=1/(1+EXP(-&A_ROB/2)); PROC PRINT;
  75   TITLE '*****ROBUST REGRESSION METHOD******************';
  76     DATA AAA;    KEEP B_ROBUST A_ROBUST;    SET AA;
  77   DO J=1 TO &N;    B_ROBUST=B_ROBUST; A_ROBUST=A_ROBUST; OUTPUT;END;
  78   DATA AAAA;    KEEP A_ROBUST B_ROBUST COUNT0-COUNT2;
  79   MERGE SROC AAA;
  80   Y_HAT=&A_ROB+B_ROBUST*S;
  81   SIGN=D-Y_HAT;
  82   IF SIGN>0 THEN COUNT1+1;    IF SIGN=0 THEN COUNT0+1;
        IF SIGN<0 THEN COUNT2+1;
  83   TITLE 'COUNT1 and COUNT2 are the number of scatter lie above or below
       respectively, COUNT0 is the number of scatter on the line';
            PROC PRINT; RUN;
```

SAS code 1 can solve the regression parameter A and B of SROC curve in S, D plane using the above weighted method, unweighted method, and robust method. The common odds ratio and TPR^* which reflect the diagnostic accuracy of the test are also calculated using the parameters obtained above.

For similar data in Table 14, you need to change the number of primary studies n in 02nd row of SAS code 1, and the number of true positive (TP),

false negative (FN), false positive (FP), true negative (TN) from 11th to 30th row.

To obtain exactly the intercept of robust regression, according to the request of precision of intercept of robust regression (e.g. decimal digits), the value of 3rd row after "A_ROB=" must be changed again and again after running the SAS code. After each running, the number of scatter points lies above line (COUNT1) and lies below line (COUNT2) must be observed. If COUNT1 = COUNT2, running SAS is end, the intercept A is the intercept of robust regression.

The file of SROC is obtained using the row from 04th to 33rd which include the value of variables of TP, FN, FP, TN, sensitivity, 1-specificity, weight, D and S. The parameter and related value of weighted regression model are obtained through the rows from 34th to 44th. The parameter and related value of unweighted regression model are obtained through the rows from 45th to 55th. The result comparing the diagnostic accuracy TPR^* s between weighted and unweighted methods is gained through the rows from 56th to 60th. The parameter and related value are obtained through the rows from 61st to 83rd.

4.1.6. *Other practical issues of SROC analysis*

TPR^* in ROC analysis is often an important summary statistics for meta-analysis. While it is useful, it may not always relevant clinically. For example, if all previous studies had false positive rates less than 20%, while the false positive rate of TPR^* is in 30%, TPR^* becomes irrelevant because it is out of the clinical range of practical uses. In such a case, a backward translation of mean D and mean S into ROC curve can provide a more informative summary statistics. This summary point is simply expressed as

$$\overline{TPR} = \frac{\exp\{(\bar{S} + \bar{D})/2\}}{1 + \exp\{(\bar{S} + \bar{D})/2\}}$$

and

$$\overline{FPR} = \frac{\exp\{(\bar{S} - \bar{D})/2\}}{1 + \exp\{(\bar{S} - \bar{D})/2\}},$$

which is always on the SROC curve. Another relevant alternative summary statistics for SROC is the area under the curve (AUC). Like we use AUC of a ROC curve to compare diagnostic tests, the AUC of SROC does not depend on the selected threshold that TPR^* used. It is particularly useful when two SROC curves cross to each other. More useful is the conditional AUC

when the upper limit of false positive rates is given. For example, we are only interested in the performances of diagnostic tests when its specificity is above 85%. This corresponds to the AUC of SROC in the section of FPR being less than 15%.

One advantage of SROC is to relate non-linear relationship between TPR and FPR to linear regression S and D. While formula (66) compares two TPR^* points as an approach to compare two diagnostic tests, it did not take advantage of linear relationship between S and D fully. Alternatives include the use of analysis of covariance in the regression step of S and D. By adding an additional covariate X to indicate different diagnostic modalities, linear modal theory can be used to test statistical significance of different modalities. In addition, we can add other covariates, such as the year of publication and the design of the studies, into the linear model to assess the effects of other uncontrollable factors on the diagnostic utilities. When meta-analysis includes multi-modality studies, i.e. among studies that one patient being evaluated by several diagnostic techniques, a random-effects model of individual study can be built into the linear model to control for correlated results reported in these papers. Several examples of using these generalized linear models can be found in literature.

In meta-analysis, if each individual accepts several diagnostic tests, in order to dispel the correlation among several diagnostic tests, the random-effect model can be established. Some researchers suggest using generalized linear model to control the correlation.[32–34]

4.2. *Other methods of estimating log odds ratio of diagnostic test*

Both Mantel-Haenszel method and exact-based logit method[35] can be used to calculate the log odds ratio of diagnostic test.

4.2.1. *Mantel-Haenszel method*

Assume there are h primary studies of diagnostic test and the symbol a_i, b_i, c_i, d_i, and n_i represent true positive, false positive, false negative, true negative and the total number from ith study $(i = 1, 2, \ldots, h)$ respectively. Adjusted odds ratios of Mantel-Haenszel method OR_{MH} is expressed as:

$$OR_{MH} = \sum_{i=1}^{h} \left(\frac{a_i d_i}{n_i} \right) \Bigg/ \sum_{i=1}^{h} \left(\frac{b_i c_i}{n_i} \right). \tag{68}$$

Using the formula

$$X^2_{MH} = \sum_{i=1}^{h} \left(\frac{a_i d_i - b_i c_i}{n_i} \right)^2 \Bigg/ \sum_{i=1}^{h} \left(\frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{(n_i - 1) n_i^2} \right) \tag{69}$$

performs the test of statistical significance. $100(1-\alpha)\%$ confidence interval of the adjusted odds ratios of Mantel-Haenszel OR_{MH} is

$$\left(OR_{MH}^{1-U_{1-\alpha/2}/\sqrt{X^2_{MH}}},\ OR_{MH}^{1+U_{1-\alpha/2}/\sqrt{X^2_{MH}}} \right). \tag{70}$$

$U_{1-\alpha/2}$ is a quantile from the standard normal distribution under test level α, $U_{1-\alpha/2}$ of 95% confidence interval is $U_{1-\alpha/2} = 1.96$.

4.2.2. *Exact-based logit confidence interval*

The method was proposed by Woof in 1955, so it was named Woof method. The odds ratio OR_L can be expressed as:

$$OR_L = \exp \left[\left(\sum_{i=1}^{h} (w_i \ln OR_i) \right) \Bigg/ \sum_{i=1}^{h} w_i \right]. \tag{71}$$

The $100(1-\alpha)\%$ confidence interval is

$$\left(OR_L \exp \left[-U_{1-\alpha/2} \Bigg/ \sqrt{\sum_{i=1}^{h} w_i} \right],\ OR_L \exp \left[U_{1-\alpha/2} \Bigg/ \sqrt{\sum_{i=1}^{h} w_i} \right] \right), \tag{72}$$

where OR_i is the odds ratio of ith study

$$w_i = \mathrm{var}(\ln OR_L))^{-1} = (1/a_i + 1/b_i + 1/c_i + 1/d_i)^{-1}.$$

If there are 0 in any cell of a study, each cell of the study is added a small value, e.g. 0.5.

To test if the odds ratio of the primary studies is homogeneity, the Breslow-Day test of homogeneity can be used. The Breslow-Day statistic is expressed as:

$$Q_{BD} = \sum_{i=1}^{h} [a_i - E(a_i | OR_{MH})]^2 / \mathrm{var}(a_i | OR_{MH}), \tag{73}$$

where E and var represent expected value and variance respectively. Statistic Q_{BD} is an approximate chi-squared statistic with freedom degree $df = h - 1$.

SAS code 2. Calculated odds ratio using Mantel-Haenszel method and logit method.

```
Number                          SAS Code
  1     DATA C;              SET SROC;
  2     A=1 ; B=1; F=TP;OUTPUT;      A=1 ; B=2; F=FN;OUTPUT;
  3     A=2 ; B=1; F=FP;OUTPUT;      A=2 ; B=2; F=TN;OUTPUT;
  4     TITLE 'CRUDE ODDS RATIO';
  5     PROC FREQ DATA=C ;           WEIGHT F;
  6     TABLES A*B/ALL RISKDIFF RELRISK NOPRINT;
  7     TITLE 'MANTEL-HAENSZEL ODDS RATIO and LOGIT ODDS
  8     RATIO';
  9     PROC FREQ DATA=C ;           WEIGHT F;
  9     TABLES I*A*B/ALL RISKDIFF RELRISK NOPRINT ;RUN;
```

4.2.3. *An example*

Use the file of SROC (data see Example 5) of the SAS code 1 from 01st to 31st and the SAS code 2, the odds ratio of the diagnostic test is estimated by Mantel-Haenszel method and Exact-based logit method. In the SAS code 2, the code of row from 1 to 3 is used to transform the SROC file into the required data format. The code of row from 5 to 6 are used to calculate the crude odds ratio. The code of row from 8 to 9 is used to calculate the adjusted odds ratios of Mantel-Haenszel method and odds ratio of exact-based logit method.

The FREQ procedure in the rows from 4 to 6 calculates the summary Mantel-Haenszel statistics of 59 studies. The results are $\chi^2_{MH} = 2829.032$, $df = 1$, $P \leq 0.001$. The crude odds ratio of Mantel-Haenszel method is 5.542, and 95% confidence interval is (5.203, 5.903). The crude odds ratio of logit method is 5.542, and 95% confidence interval is (5.193, 5.915).

The FREQ procedure in the rows from 7 to 9 calculates the summary Mantel-Haenszel statistics of 59 studies. The results are $\chi^2_{MH} = 2231.929$, $df = 1$, $P \leq 0.001$. The adjusted odds ratio of Mantel-Haenszel method is $OR_{MH} = 5.573$, and 95% confidence interval is (5.189, 5.984). The adjusted odds ratio of logit method is $OR_L = 5.557$, 95% confidence interval is (5.137, 6.010).

These results are similar to the odds ratio of weighted regression model. Breslow-Day test of homogeneity is performed for the data. We have $Q_{BD} = 394.286$, $df = 58$, $P \leq 0.001$. These suggest the difference among 59 primary studies have statistical significance.

The above results of analysis suggest that the diagnostic accuracy of Pap test is similar with those obtained by several methods. TPR^* is about

0.7, The odds ratio of positive diagnostic result is about 5. These suggest that the test plays an important role in cervical precancer, but these results suggest that the diagnostic accuracy of the test is not high.

The methods above assumed that the golden standard is perfect. If the golden standard is imperfect, the diagnostic accuracy of the test must be adjusted. Walter *et al.* proposed the method estimating the SROC curves of test with imperfect reference standards in 1999.[29]

Although someone proposed the meta-analysis method of diagnostic test using the area under curve (AUC), how to use both AUC and the data of sensitivity and specificity need to be studied further.

5. Meta-analysis for Linkage Studies

Recently, linkage studies are rapidly becoming numerous. At the same time, conflicting claims of linkage also sprout in genome wide scans. Serious discussion has begun regarding how to control false positives or spurious linkages. Meta-analysis can quantitatively synthesize results from multiple independent studies into a pooled measure of the overall effect of genetic linkage. But because there may exist too many differences between linkage studies, such as different ascertainment of pedigrees, different disease definition, different genetic markers or different statistical techniques, a common effect size is difficult to be found and extracted. And so the general meta-analysis methods are difficult to be applied directly. We here introduce some meta-analysis methods that are appropriate for linkage studies.

5.1. *Meta-Analysis of P Values*

5.1.1. *Statistical method*

Assume that there are m independent studies assessing linkage of a disease or trait to a maker. Let P_i denote the P value associated with the ith study $(i = 1, 2, \ldots, n)$, then n independent P values can be combined into a single test of significance.

$$X^2 = -2\sum_{i=1}^{n} \ln(P_i)\,. \tag{74}$$

If the null hypothesis is true, i.e. if there are no genes underlying the trait near the marker locus, this quantity has a χ^2 distribution with two degrees of freedom as proposed by Fisher in 1954. Alternatively, a weight may be

assigned to each individual study indicating its importance. Assigning a weight v_i to the ith study, and form the product

$$P_w = P_1^{v_1} P_2^{v_2} \cdots P_n^{v_n} . \tag{75}$$

The validity of the omnibus null hypothesis is tested using the cumulative distribution of P_w, $\text{Prob}(P_w \leq q) = \sum_{k=1}^{n} (q^{1/v_k})/a_k$, where $a_k = \Pi_{i=1, i\neq k}^{n} (v_k - v_i)/v_k$. A simple choice for the weight is $v_k = \frac{1}{n_k} / \sum_i \frac{1}{n_i}$, where n_i is the number of sib-pairs used in the ith study (in sib-pair tests).[36] One may assign a different level of importance to each individual study based on the presumption that some designs are more powerful than others. For example, if 1000 random sib-pairs are needed for a power of 80%, and the same power could be achieved by using 40 ED sib-pairs or 200 affected sib-pairs, then all three studies would have equal weights for importance, although their sample size are considerably different.

If all studies we want to summarize have identical genotyped markers and same linkage analysis method being used, then one can apply Fisher's method to combine P value directly. Since in gene mapping studies, genetic markers are used only as references to infer the location of the putative disease gene at the chromosome or infer whether a disease gene is located at a specific region of the genome, different studies may use different genetic markers, although their objectives are same. Moreover, they may use different linkage analysis method. If we want to synthesize this kind of studies, we must firstly extract a single P value for the region from each study. We will take the summarization of 4 practical studies concerning linkage of BMI with markers in the human OB gene region as an example illustrate some techniques in the following paragraph.

5.1.2. *The extraction of P value*

(i) For study with a single marker, no correction needs to be applied. For example, Borecki *et al.*[37] used only one marker in the area of the human OB gene. This one marker was KELL, located at 7q33. Four hundred pairs of sibling pairs were included and the Haseman-Elston procedure was used to yield a p value of 4.8×10^{-6}. It could be used directly.

(ii) If a separate P value for each of several markers is reported in a chromosome region, we could convert each P value to a corresponding (standard normal) Z-score by means of the inverse standard normal distribution function $Z = \Phi^{-1}(1 - P)$. The correlation between any two of them is equal to the correlation of corresponding IBD status between them. For

example, the correlation between Z_i and Z_j is $r_{ij} = (1 - 2\theta_{ij})^2$, where θ denotes recombination fraction. θ_{ij} can be determined according to the distance between ith and jth markers. One centimorgan (cM) or 1 million base pairs (bp) is approximately equal to θ of 0.01. We could use statistic $S_k = \sum_{i=1}^{k} Z_i$ to summarize the information of all markers, the variance of the sum is the sum of variances plus twice the sum of the covariances for all component, that is $\mathrm{var}(S_k) = k + 2\sum_{i<j} r_{ij}$. So statistic $T = \frac{S_k}{\sqrt{\mathrm{var}(S_k)}}$ distributed as standard normal, and it can be used to derive a single P value for the study.

Example 6. Clement *et al.*[38] evaluated linkage to BMI dichotomized as "greater than 35" or "less than or equal to 35" with 8 markers ranging from D7S651 to D7S509 using sib-pair method. A part of results are displayed in Table 16.

Table 16. Proportion of alleles IBD in OB markers for concordant (obese-obese) sib-pairs.

Marker	n	$\bar{\pi}$	t	P	P^*	$Z_i = \Phi^{-1}(1 - P_i^*)$
D7S651	66	0.57	1.98	0.03	0.025970	1.943627
D7S692	59	0.52	0.68	NS	0.249605	0.675734
D7S677	46	0.49	−0.29	NS	0.386574	0.288260
D7S680	57	0.59	2.47	0.008	0.008292	2.395791
D7S514	53	0.59	2.44	0.009	0.009066	2.362904
D7S530	65	0.59	2.96	0.002	0.002155	2.854504
D7S640	57	0.55	0.99	NS	0.163216	0.981324
D7S509	56	0.54	1.01	NS	0.158459	1.00081
Total						12.50295

P is the P value reported in the original literature.
P^* is the P value recovered according to t value and degree of freedom $(n - 1)$.

The distances (cM) between every two adjacent markers in Table 16 are 13, 3, 7, 0, 2, 5, 5 respectively. We can get $\sum_{i<j} r_{ij} = 16.3756$, $\mathrm{var}(S_k) = 8 + 2 \times 16.3756 = 40.7512$. The calculation of Z_i is showed in Table 16. Statistic T can be calculated as

$$T = \frac{S_k}{\sqrt{\mathrm{var}(S_k)}} = \frac{12.50295}{\sqrt{40.7515}} = 1.958585\,, \quad P = 0.0251\,.$$

(iii) If a single P value was provided from a multipoint procedure, then Lander-Kruglyak correction could be applied to get a corrected P value,

$$P^* = 1 - \exp(-\mu(T))\,, \tag{76}$$

where $\mu(T) = [C + 2\rho GT^2]\alpha(T)$; $T = \Phi^{-1}(1 - P)$ is a standard normal Z-score corresponding to cumulative probability $1 - P$; C is the number of chromosome; G is the genome length measured in Morgans; $\alpha(T) = P$ is the pointwise significance; ρ is the crossing over rate between the genotypes being compared.

Example 7. Duggirala *et al.*[39] examined the linkage of BMI to markers spanning a 211 cM (D7S531 to D7S483) using a multipoint procedure, and resulted in a combined P value of 0.003.

In this example, $C = 1$(one chromosome used for the study), $\rho = 2$ (for sib-pair tests), $\alpha(T) = 0.003$, $T = \Phi^{-1}(1 - 0.003) = 2.747765$,

$$\mu(T) = (1 + 2 \times 2 \times 2.11 \times 2.747765^2) \times 0.003 = 0.194171\,,$$

$$P^* = 0.1765\,.$$

(iv) Sometimes researchers may use multiple cutoff points or multiple criteria to define the affected or unaffected in one study. If the analysis methods they have used are one-side sib-pair tests, the process of extracting a single P value is similar to that of (ii). Notice that here the multiple criteria of classification are concerned but not the multiple markers. The estimation of correlation is different, for example, the correlation between Z_i and Z_j is calculated as $r_{ij} = \sqrt{\frac{\min(n_i,n_j)}{\max(n_i,n_j)}}$, where n_i, n_j are the number of sib-pairs having been used in ith and jth classification respectively.

Example 8. Reed *et al.*[40] examined linkage of BMI to 8 markers contained in and surrounding the interval D7S1873 through D7S1875 using two methods (sib-pair analysis and TDT). Three cutoff points were used to define obese and linkage analysis has been performed respectively. The main results are displayed in Table 17 and 18.

If a study used a two-side TDT (Table 18), we could convert the chi-squares to Z-scores by taking their square root, just like the column 6 in Table 18. The correlation among the Z's can again be estimated as the square root of the proportion of subjects in a subset divided by the number of subjects in the larger set, For example, the estimated correlation between the Z-score in subjects with a BMI ≥ 40 and the Z-score for subjects with a BMI ≥ 30 is $\sqrt{70/121} = 0.761$. If there are m Z-scores, then statistic $Q = ZR^{-1}Z'$ has a chi-square distribution with the degree of freedom equal

Table 17. Mean proportion of the OB gene haplotypes (D7S1873-D7S1875) identical by descent for obese-obese sib-pairs.

Obese Cutoff	Pairs(n)	Proportion of IBD	t	P	$Z_i = \Phi^{-1}(1 - P_i^*)$
≥ 30	213	0.51 ± 0.33	0.24	0.4038	0.243524
≥ 35	135	0.50 ± 0.35	0.03	0.4333	0.167979
≥ 40	59	0.60 ± 0.33	2.28	0.0132	2.220277
Total					2.63178

Proportion of IBD is expressed as mean±SD.

Table 18. Transmission disequilibrium of a haplotype (D7S504-D7S1875) flanking the OB locus.

BMI of Sibling	1–5 Transmitted/ Not Transmitted	%Transmitted	χ_1^2	P	$Z_i = \sqrt{\chi_1^2}$
≥ 30	71/50	58.7	3.64	0.056	1.907878
≥ 35	60/39	60.6	4.45	0.035	2.109502
≥ 40	46/24	65.7	6.91	0.009	2.628688

to $m-1$. Where $Z = (Z_1, Z_2, \ldots, Zm)$ and R is the correlation matrix. With the data in Table 18, we get $Z = (1.907878\ 2.109502\ 2.628688)$

$$Q = ZR^{-1}Z' = Z \begin{pmatrix} 1 & 0.904530 & 0.760600 \\ 0.904530 & 1 & 0.84875 \\ 0.760600 & 0.840875 & 1 \end{pmatrix} Z' = 6.944753$$

$P = 0.0310$.

In Example 8, Reed combined the marker information into haplotypes and conducted their analysis by looking at sharing of haplotypes rather than alleles. This aspect of their analysis simplifies the extraction of a single P value since significance is assessed only for IBD sharing at the single haplotype rather than at each individual locus, so the P values need not be corrected with Lander-Kruglyak method. With the data in Table 2, we get $\sum_{i<j} r_{ij} = 1.983509$, $\mathrm{var}(S_k) = 3 + 2 \times 1.983509 = 6.967017$. The calculation of Z_i are displayed in column 6 of Table 17. Then the statistic

$$T = \frac{S_k}{\sqrt{\mathrm{var}(S_k)}} = \frac{2.63178}{\sqrt{6.967017}} = 0.997071\,, \quad P = 0.1594\,.$$

(v) If a study has performed more than one test with the same data, just like Example 3, we still have two p values after combination, one from sib-pair

test and one from TDT. If the correlation between these two tests could be determined, then one could combine these into a single P value. However, it is not immediately apparent about how to estimate this correlation. Allison *et al.* (1998) propose several alternatives[41]: First, one could, on some *a priori* grounds of preference, choose one test over another. For example, one might argue that because all of the other studies are using a sib-pair approach rather than TDT it would be more appropriate to combine sib-pair data rather than the TDT data and be consistent with the others. Second, one could multiply the lowest P value by two (the number of test) as a form of Bonferroni correction. However, this is overly conservative because it does not take the correlation between the two tests into account. Third, one could estimate the correlation via simulation. Fourth, one could conduct the overall meta-analysis with the results of each test.

The results of meta-analysis for the above four studies are displayed in the last row of 4th and 5th columns in Table 4. When using Reed *et al.*[40] sib-pair test result, the overall $P = 4.9047 \times 10^{-6}$ (d.f. = 8); when using TDT result, the overall $P = 1.1999 \times 10^{-6}$ (d.f = 8). Besides these, we have conducted sensitivity analysis also in this example, the sensitivity analysis means that each study result was removed from the analysis, and the chi-square statistic with 6 d.f. (from the remaining study results) was computed. The corresponding P values are given in first 5 rows of 4th and 5th columns in Table 19. This table shows that Borecki *et al.*[37] study has a great influence to the overall P value. But even excluding this study, the remaining results still provide a significant value ($P < 0.05$). So this study suggests that there is evidence for linkage of BMI to somewhere in the OB region. Note that this meta-analysis is only an example, we have not collected all of possible literatures.

Table 19. The results of overall meta-analysis and sensitivity analysis.

Reference		P Value	χ^2 (P Value)[a]	χ^2 (P Value)[b]
Borecki *et al.* (1994)[37]		4.8×10^{-6}	14.51(0.0244)	$17.79(6.7893 \times 10^{-3})$
Clement *et al.* (1996)[38]		0.0251	$31.64(1.9165 \times 10^{-5})$	$34.91(4.4856 \times 10^{-6})$
Dugirala *et al.* (1996)[39]		0.1765	$35.54(3.3919 \times 10^{-6})$	$38.81(7.7944 \times 10^{-7})$
Reed *et al.* (1996)[40]	Sibpair	0.1594	$35.33(3.7152 \times 10^{-6})$	
	TDT	0.0310		$35.33(3.7152 \times 10^{-6})$
Overall			$39.01(4.90 \times 10^{-6})$	$42.28(1.20 \times 10^{-6})$

a: Using Reed *et al.*[40] sib-pair test result.
b: Using Reed *et al.*[90] TDT result.

5.2. *The Meta-analysis for Genome Search*

The Genome Search Meta-analysis method (GSMA) uses a non-parametric ranking procedure to identify genetic regions that show consistently increased sharing statistics or lod scores among several genome screens.[42,43] This method splits the whole chromosomes into bins of approximately equal length and ranks these bins according to the lod scores, Z-statistics or P values with the most significant result having the highest rank within each genome screen. Then the ranks for each bin are summed across screens. For any bin, the null hypothesis is that no susceptibility loci exist within the bin, and the ranks are assigned randomly. For m studies and n bins, the probability that the sum of ranks (X_i) is equal to a value R is given by

$$P\left(\sum_{i=1}^{m} X_i = R\right) = \begin{cases} 0 & R < m \\ \dfrac{1}{n^m}\displaystyle\sum_{k=0}^{d}(-1)^k \binom{R-kn-1}{m-1} \times \binom{m}{k} & m \le R \le mn \\ 0 & R > mn\,, \end{cases} \tag{77}$$

where d is the integer part of $(R-m)/n$. From this distribution, we can calculate the probability that a summed rank of R or greater within a bin under the null hypothesis.

The choice of bin width has several constraints: The bin width must be appropriate for all chromosomes, with at least two bins on the smallest chromosome, and at least one marker should be genotyped within each bin. To ensure the independence of lod score or P value for adjacent markers, Wise *et al.* proposed to use 30 cM as the width of each bin.

Since some literatures may report only the most significant results, the information for some bins is lost. This will not bias the results of the GSMA, provided a strict lod score or P value cut-off has been used and all chromosomes have been genotyped. If ranks can be assigned to the top bins, the remaining bins could be given equal ranks of $(120 - x + 1)/2$. If different genome search contributes differently to the meta-analysis, a weight may be assigned to each screen, such as $\log(N)$, where N is the number of pedigrees or sib pairs in each study. Although the above probability distribution for the summed ranks under null hypothesis will no longer hold, the P-value can be generated through simulation of the weighted ranks.

5.3. *Conclusion*

The major forte of Fisher's combining P value method are its simplicity in calculation and its flexibility in pooling results from studies which may examine slightly different hypotheses or use different outcome measures. However, it also has many drawbacks, sometimes its result is difficult to explain because only one highly significant P value from a single study may determine the significance of the Fisher test statistic; it cannot be used to make inferences about the average effect size or the consistency of results across studies. But in practice, published results from heterogeneous studies are likely to report P values only. When nothing else is available, combining P values can provide an overall assessment of linkage.

GSMA allows systematic integration of data from several genome screens. The major strength of the GSMA is its application to a diversity of study designs, it is not restricted by different phenotype definitions, family structures, markers, or analysis methods across studies. Wise *et al.*[42,43] have applied this method to four genome screens in multiple sclerosis and across 11 screens from autoimmune disorders, which showed that the GSMA is a valuable data exploration tool to obtain an overview of the genome search results within and across disease phenotypes.

To ensure the quality of meta-analysis, the pre-analysis process is very important, we must set strict literature inclusion standard according to professional knowledge, and collect literatures through multiple ways to reduce as much publication bias as we can.

6. Bias in Meta-Analysis

6.1. *Source of bias*

Meta-analysis should be viewed as an observational study of the evidence. In epidemiology, bias may be defined as any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions that are systematically different from the truth. Bias often cause conflicting results of meta-analysis and threaten its internal validity and reliability. In each step of meta-analysis, like locating and selecting studies for inclusion in meta-analysis, or extracting accurate study data, bias may be introduced. As noted by Felson, there are at least three types of bias involved in meta-analysis: Sampling bias, selection bias and within study bias.[44]

6.1.1. *Sampling bias*

The validity of a meta-analysis depends on complete sampling of all the studies performed on a particular topic. Any incomplete sampling is potential to bias. Sampling bias arise when retrieving the relevant studies, which consists of:

(1) Studies with significant results are more likely to get published than studies without significant results, leading to *publication bias*.
(2) In the process of retrieving published studies using computerized database, *indexing bias* and *search bias* may occurr. The former is defined as biased indexing of published studies, which means indexing error or indexing variability. Indexing bias is not under the meta-analysts control. Search bias is another type of sampling bias due to inadequate or incomplete search. Index bias or search bias can lead to failure to capture all indexed studies in a database.
(3) Relying heavily on references published in other articles or in review of literature may cause *reference bias* or *citation bias* into a meta-analysis.
(4) *Multiple publications bias* occurs when studies whose results are published in a series of articles are more likely to be sampled than those published only once. *Multiple publications bias* can induce meta-analyst confusion when the publications do not have the same first author or when one publication does not refer to the prior one. *Multiply used subjects bias* can occur when the same subjects are reported in two separate studies when they actually a part of only one study.
(5) The included studies in meta-analysis based exclusively on reports in English may leads to *English language bias*.

In practice, to reduce or avoid sampling bias require that the meta-analyst embarking on a database search chooses appropriate index terms and conducts the search with a systematic strategy.

6.1.2. *Selection bias*

Selection bias occurs when eligible studies are chosen in a meta-analysis, according to the criteria of inclusion and exclusion. In this process, two types of bias may be introduced, one is *inclusion criteria bias*, and the other is *selector bias*. If the inclusion criteria is developed by an investigator familiar with the area under study, the criteria can be influenced by knowledge of the results of the set of potential studies, and this would cause bias. *Inclusion criteria bias* is difficult to avoid since a good knowledge of

a topic is a prerequisite to develop an inclusion criteria. In *selector bias*, inclusion criteria have been set, although they may not be so specific as to dictate which studies are included or excluded from the meta-analysis. This leaves the meta-analyst selector free to choose studies, a choice which is susceptible to bias.

Selection bias of studies is probably the central reason for discrepant results in meta-analyses. For example, in 1992, two meta-analyses published in BMJ (*British Medical Journal*) and *Lancet*, respectively. Both compared low molecular weight heparins and standard heparin in the prevention of thrombosis after surgery, but the conclusions were widely divergent.[45,46] One concluded that "low molecular weight heparins seem to have a higher benefit to risk ratio than unfractionated heparin in preventing perioperative thrombosis", whereas the other considered that "there is at present no convincing evidence that in general surgery patients low molecular weight heparins, compared with standard heparin, general a clinically important improvement in the benefit to risk ratio". Egger pointed out that the conflicting results of two meta-analyses were mainly related to the selection of studies.[47] Nurmohamed *et al.*[46] based their analysis on a subgroup of trials that they considered possess the highest methodological strength, while Leizorovicz *et al.*[45] included all trials in their analysis. Many other elements, for example, language restrictions or use of unpublished material — could contribute to conflicting conclusions.

Criteria for including studies in a meta-analysis may be influenced by knowledge of the results of the set of potential studies and lead to inclusion bias.

One important way to avoid selection bias is to create extremely specific and clear study inclusion criteria, so that the selector has little chance to inject bias into the selection decision. Blind method is also suggested to limit selector bias. The most common is to blind the methods and results of studies to make it hard for the meta-analyst selector to determine the inclusion of a study through results. In this method, there are often two selectors who work independently. Any disagreement in study selection is solved by a joint meeting or by a third selector. This process certainly decreases the chance of selector bias, but it does not eliminate it.

Another way of handling the selection bias is to include all studies that meet basic entry criteria then perform sensitivity analyses with regard to the different possible entry criteria. Any conclusions from a meta-analysis that are highly sensitive to altering the entry criteria should be treated with caution.

6.1.3. *Within study biases*

After studies are selected for a meta-analysis, data should be accurately extracted from the study. There are several opportunities for bias, the most likely bias is *extractor bias*, which can create systematically biased results. There may be considerable inter- and intra-observer variability in extracting data from studies. To minimize *extractor bias*, an extraction sheet should lay out specific rules for data extraction with clarity.

Meta-analyst bias may affect the scoring of studies for quality. If study results are weighted for quality in the analysis, *a bias in scoring study quality* may have a real impact in meta-analysis results. Giving rigid rules on how to measure the quality of trials may help lessen observer variability and mitigate bias.

The primary study paper included in the meta-analysis itself may not accurately report the study's result. For example, the study has several outcomes which were measured, but the only results reported are those which reach statistical significance, and this can introduce a *reporting bias*. Unfortunately, the prevalence of *reporting bias* is unknown, but it is a widespread problem which could serve to substantially bias meta-analysis results.

6.2. *Publication bias*

Publication bias is usually used to refer to the greater likelihood of research with statistically significant results to be submitted and published compared with non-significant and null results. More generally, publication bias is the systematic error in a statistical inference by conditioning on the achievement of publication status. Publication bias occurs because published studies are not representative of all studies that have ever been done.

6.2.1. *The causes and consequence of publication bias*

Publication bias has long been recognized and much discussed. Publication bias can originate from three sources: The authors, the sponsors of the study, and the editor or reviewers of the journal to which the paper is submitted. First, authors may be less likely submit papers if the results are not significant. Second, the editors of the journal may favor publication of positive results. Finally, the sponsor may play an important role in generating publication bias, especially if it is a pharmaceutical company funded study. The implication is that the pharmaceutical industry discourages the

publication of studies which have negative findings. In addition, multicenter studies are more likely to be published than studies from a single center.

Existence of a bias in favor of publication of statistically significant results is well documented. Easterbrook *et al.*[48] carried out a retrospective study of 285 research projects that had been approved by the Central Oxford Research Ethics Committee between 1984 and 1987. They found 154 studies had statistically significant results and 131 did not. Of the 154 studies with statistically significant results, 60.4% had been published, whereas only 34.4% of the studies that did not have statistically significant results had been published. Using logistic regression and adjusting for relevant covariates, they found that studies with statistically significant results were more likely to have been published and/or presented than those with non-significant results (OR = 3.56, 95%CI 1.82–6.99).[48]

Publication bias may seriously distort the findings of a meta-analysis, and certainly threaten the validity and reliability of results. For example, in a meta-analysis about the effect of an alkylating agent alone comparing with combination chemotherapy on survival in patients with advanced ovarian cancer, Simer found that the conclusion based on the published studies is different from that based on studies registered in the International Cancer Research Data Bank. The pooled results from published trials showed significant efficacy, while data from prospectively registered trials (both published and unpublished) showed no significant advantage of combination chemotherapy over single agent treatment.[49]

6.2.2. *Methods of detecting and correcting for publication bias*

Although searching for relevant unpublished studies is important and may sometimes alleviate publication bias, identifying such studies may be difficult. Hence we need methods to assess the magnitude of publication bias in a meta-analysis, based on the data in the available studies. In fact, various methods have been devised to attempt to detect and correct publication bias, but none of the available methods is entirely satisfactory for dealing with this problem. Here, commonly used methods are described as following.

6.2.2.1. Funnel Plot

Funnel Plot, or, funnel graph, is the frequently used method for detecting the publication bias. The basic idea is that if the point estimates from individual studies are plotted against the inverse of the variances, or another

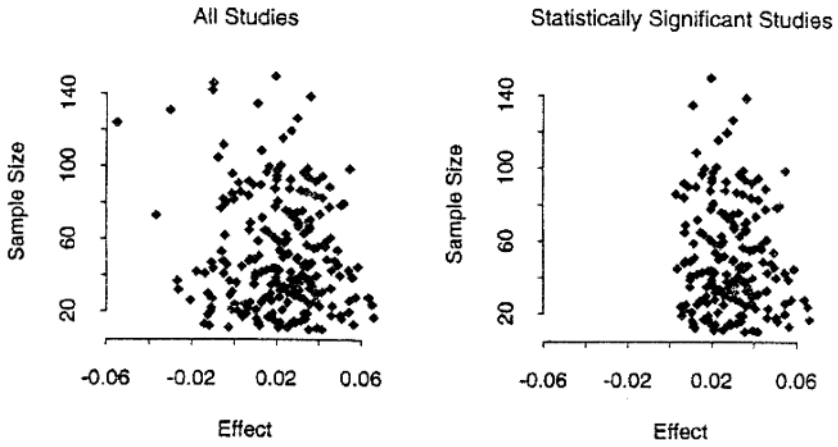


Fig. 9. Two funnel plots based on simulated data. The left plot displays absence of publication, and the right displays the presence of publication bias.
Source: Normand[10] *Stat. Med.* **18**: 339.

surrogate for sample size, the points visualized together should produce a funnel shape, so they are scattered around the true value of the point estimate with the scattering narrowing as the standard errors decrease.[50] That is, in such a plot, the effect size of studies is plotted against study sample size. If there is no publication bias, the plot would resemble an inverted funnel with a wide dispersion of results among studies of small size and a narrower range of study results for large studies. If the plot shows an asymmetrical and skewed shape, publication bias may present. This usually takes the form of a gap in the wide part of the funnel, which indicates the absence of small studies showing no benefit or harm. Figure 9 demonstrates two funnel plots based on simulated data. The left plot displays the simulated summaries for all the studies, which means absence of publication bias. The right plot displays the simulated summaries for studies that are statistically significant at the 0.05 level, which suggests the presence of publication bias.

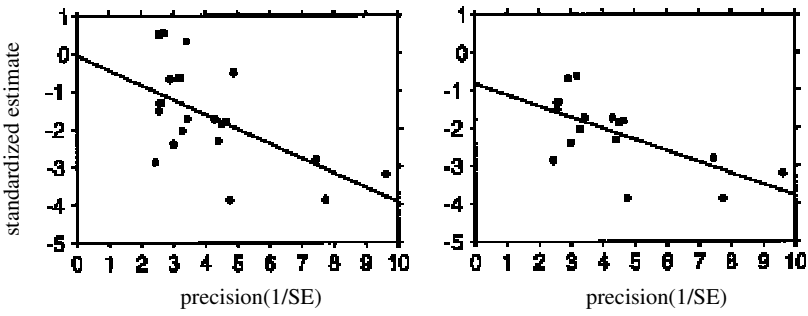
In fact, the funnel plot is a graphical test for any type of bias that is associated with sample size. The publication bias and sampling bias are more likely to affect smaller studies than large trials and may thus lead to funnel plot asymmetry. Another source of asymmetry arises from differences in the methodological quality. Smaller studies are, on average, conducted and analyzed with less methodological rigor than larger studies, and trials of lower quality tend to show larger effects. Other factor, such as hetero-

geneity in treatment effect between low and high risk groups can also lead to asymmetry in the funnel plot.

The major advantage of funnel plot is that it is easy to be performed which only requires published data. But the method is practically limited to meta-analysis with large enough numbers of studies to allow one to visualize (as opposed to fantasize) a funnel shape to the data. The symmetry of funnel plot is defined informally. So, if the number of studies included in a meta-analysis is small, it is difficult to detect the symmetry of funnel plot through visual examination.

6.2.2.2. Egger's linear regression method[51]

Egger proposed a linear regression model to measure funnel plot asymmetry. It is a formal test for asymmetry in funnel plot. The standard deviate y_i, ($y_i = t_i/s_i$, t_i is the effect size, s_i is standard error for study i) is regressed on precision $x_i (x_i = 1/s_i)$, then the significance of intercept differing from zero (at $\alpha < 0.1$) is tested. That is, $y_i = a + bx_i$. The points from a homogeneous set of trials, not distorted by publication bias (or other bias), will thus scatter about a line that runs through the origin at standard normal deviate zero ($a = 0$), with the slope b indicating the size and direction of effect. This situation corresponds to a symmetry funnel plot [Fig. 10(a)]. If it is asymmetric, with smaller studies showing effects that differ systematically from larger studies, the regression line will not pass through the origin [Fig. 10(b)]. The intercept a provides a measure of asymmetry — the larger it deviate from zero the more the asymmetric.



(a) no publication bias (b) publication bias

Fig. 10. Example of the Egger's regression method using a simulated meta-analysis. *Source*: Macaskill (2001). *Stat. Med.* **20**: 644.

Egger examined whether the regression method predicts discordance of results when meta-analyses were compared to large trials.[51] They found in the eight pairs of meta-analysis and large trial, there were four concordant and four discordant pairs. In all case discordant cases, meta-analyses show larger effects. Funnel plot asymmetry was present in three out of four discordant pair but none of concordant pairs. They also found, in 14 (38%) journal meta-analysis (from four famous medical journals) and 5 (13%) Cochrane reviews, funnel plot asymmetry indicating that there was bias.

But the statistical properties of Egger's linear regression method are not described, and the test may itself be biased. This method violates the usual assumptions of simple linear regression. There is measurement error in the independent variable because the standard errors were estimated from the observed data, and is therefore subject to sampling error. This results in a biased estimate of the regression slope.

6.2.2.3. Begg's rank correlation test[52]

Begg's method uses Kendall's tau to test for correlation between the standardized treatment effect t_i^*, and the variance of the treatment effect (v_i), where

$$t_i^* = (t_i - \bar{t})/\sqrt{v_i^*},$$

$$\bar{t} = \sum(t_j/v_j)\Big/\sum(1/v_j) \quad \text{and} \quad v_i^* = v_i - 1\Big/\sum(1/v_j)\,.$$

Alternatively, the test can be based on the correlation between t_i^* and the sample size for each study (n_i). Treatment effects are standardized to obtain a set of estimates that can be assumed to be independent and identically distributed under the null hypothesis of no publication bias.

The rank correlation test has been described as a direct statistical analogue of the funnel plot. But the power of the test varies along with the unknown characteristics in meta-analysis. Even though the result is not significant, publication bias cannot be ruled out in small meta-analyses.

6.2.2.4. Fail-safe number[53]

Rosenthal's "fail-safe number" (N_{FS}), is the number of unpublished null studies needed to remove the significance from the finding of a meta-analysis. The method involves computing the standardized normal deviate Z, associated with each published study and then calculating a combined deviate Z_s. The values of N_{FS} required to bring the new overall P-value to

any desired level can then be calculated, an implausibly high value being regarded as evidence against the file-drawer hypothesis (publication bias). It has been suggested that N_{FS} should be presented for all meta-analyses, as an aid in the assessment of the degree of confidence that can be placed in the results.

However, plausibility of existence of certain number of unpublished studies is judged subjectively. Furthermore, this method assumes published and unpublished studies are of similar sizes. Even in similar sized studies, this method will be misleading if the average effect of unpublished studies is in opposite direction to published studies.[54]

Besides the methods describes above, another kind of methods pursues truncated sampling model to deal with publication bias, where it is assumed that statistically non-significant results do not get published. Hedges developed a model of the selection process involving a step function relating the P-value to the probability of selection in the context of a random-effects model. The model permits the estimation of a weight function representing selection along with the means and variances of effects. Dear and Begg's semi-parametric method is quite similar to that of Hedgess model, in which the selection publication is modeled also using a weight function on two-sided P-value scale.[55] The difference is that Hedge's pre-specifying the region of the P-value scale within which the weight function is assumed to be constant. Gleser proposed two general models that revisit Rosenthal's attempts to explore the number of unpublished studies and introduce several frequentist methods for interval estimates.[56] These methods take advantage of the fact that under the null hypothesis of interest, P-values from experiments testing this H_0 have a common known distribution which is independent of each experiment's design, sample size, and concomitant variables. But these methods are not widely accepted and are not recommended.

Recently, source augmentation method has been developed for detecting and correcting the publication bias. Givens used a Bayesian model to augment observed data by simulating the outcomes for missing studies, thereby creating a "complete" data for meta-analysis.[57] The author described how the random-effects model may be extended to account for publication bias, assuming that in addition to the n observed studies there are further m studies that are not observed. The number m and relative risks found from these studies are unknown and must be estimated, and uncertainties about these estimates are reflected in the final meta-analysis inference by treating them as parameters in a Bayesian analysis.

In fact, none of the available methods is entirely satisfactory for dealing with the publication bias so far. Thus, we should consider other ways to avoid publication bias. First, results of large studies most closely approximate the average results of all studies, whether published or unpublished. Furthermore, large studies, even with null results, are almost always published. Therefore, the meta-analyst can test the pooled results of studies to see if they approach the overall pooled result. Second, a meta-analyst can also attempt to obtain data from unpublished studies, an endeavor recommended. Nonetheless, finding those studies can be very difficult. Finally, one important solution to publication bias may be the establishment a clinical trial registries, a movement to register all initiated studies has begun among those in clinical trials field but not yet among those conducting observational studies.

References

1. Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Education Research* **5**: 3–8.
2. Huque, M. F. (1988). Experiences with meta-analysis in NDA submissions. *Proceedings of the Biopharmaceutical Section of the American Statistical Association* **2**: 28–33.
3. Sackett, D. L., Richardson, W. S., Rosenberg, W. M. *et al.* (2000). *Evidence-based Medicine*: *How to practice and teach EBM*, 2nd edn., Churchill Livingstone, London.
4. Egger, M. and Smith, G. D. (1997). Meta-analysis: potentials and promise. *British Medical Journal* **315**: 1371–1374.
5. Beecher, H. K. (1955). The powerful placebo. *Journal of the American Medical Association* **159**: 1602–1606.
6. Jones, D. R. (1995). Meta-analysis: Weighing the evidence. *Statistical in Medicine* **14**: 137–139.
7. Stampfer, M. J., Goldhaber, S. Z. and Yusuf, S. (1982). Effects of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials. *New England Journal of Medicine* **307**: 1180–1182.
8. Egger, M. and Smith, G. D. (1997). Meta-analysis: Principles and procedures. *British Medical Journal* **315**: 1533–1537.
9. Boyd, N. D., Martin, L. J. and Noffel, M. (1993). A meta-analysis of studies of dietary fat and breast cancer. *British Journal of Cancer* **68**: 627–636.
10. Normand, S. L. (1999). Tutorial in biostatistics. Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistical in Medicine* **18**: 321–359.
11. Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, Orlando.
12. Petitti, D. B. (2000). *Meta-Analysis*, *Decision Analysis*, *and Cost-Effectiveness Analysis*: *Methods for Quantitative Synthesis in Medicine*, 2nd edn., Oxford University Press, New York.

13. Hardy, R. J. and Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistical in Medinice* **17**: 841–856.
14. Robinn, J., Greenland, S. and Breslow, N. E. (1986). A general estimator for the variance of the Mantel-Haenszel odds ratio. *The American Journal of Epidemiology* **124**: 719–723.
15. Yusuf, S., Peto, R., Lewis, J. *et al.* (1985). β-blockade during and after myocardial infartion: An overview of the randomized trials. *Progress in Cardiovascular Diseases* **27**: 335–371.
16. Feiss, J. L. and Gross, A. J. (1991). Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology* **44**: 127–139.
17. DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**: 177–188.
18. Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychology Bulletin* **92**: 490–492.
19. DuMouchel, W. (1990). Bayesian meta analysis, in *Statistical Methodology in the Pharmaceutical Sciences*, ed. D. Berry Marcel Dekker, New York, 509–529.
20. Carlin, J. B. (1992). Meta-analysis for 2×2 tables: A Bayesian approach. *Statistical in Medinice* **11**: 141–158.
21. Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* **B55**: 3–23.
22. Gelfand, A. E., Hills, S. E., Racine-Poom, A. *et al.* (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**: 972–985.
23. Spiegelhalter, D., Thomas, A., Best, N. *et al.* (1996). *WinBUGS User Manual* (*Version 1.1.1*). MRC Biostatistics Unit, Institute of Public Health, Cambridge.
24. Irwig, L., Tosteson, A. N. A., Gatsonis, C. *et al* (1994). Guideline for meta-analysis evaluating diagnostic tests. *Annals of International Medicine* **120**(8): 667–676.
25. Moses, L. E., Shapiro, D. and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some addition considerations. *Statistical in Medicine* **12**: 1293–1316.
26. Fahey, M. T., Irwig, L. and Macaskill, P. (1995). Meta-analysis of Pap test accuracy. *The American Journal of Epidemiology* **141**: 680–689.
27. Kinkel, K., Hricak, H., Lu, Y. *et al.* (2000). US charaterization of ovarian masses, a meta-analysis. *Radiology* **217**: 803–811.
28. Kinkel, K., Kaji, Y., Yu, K. K. *et al.* (1999). Radiologic staging in patients with endometrial cancer: A meta analysis. *Radiology* **212**(3): 711–718.
29. Walter, S. D., Irwig, L. and Glasziou, P. P. (1999). Meta analysis of diagnostic tests with imperfect reference standards. *Journal of Clinical Epidemiology* **52**(10): 943–951.

30. Frei, K. A., Kinkel, K., Bonel, H. M. *et al.* (2000). Endometrial cancer: Frequency of myometrial invasion per grade and incremental value of pre-operative MRI in specialist referral: Meta- and Bayesian-analysis. *Radiology* **216**(2): 444–449.
31. Yu, C. H. (2000). The methods of ROC analysis and applications in medical research. Doctoral Dissertation of Fourth Military Medical University, 111–122.
32. Eleftherios, C. V. (1998). Meta-analysis of studies of the diagnostic accuracy of laboratory tests. *Pathology and Laboratory Medicine* **122**: 675–685.
33. Kester, A. D. and Buntinx, F. (2000). Meta-analysis of ROC curves. *Medical Decision Making* **20**(4): 430–439.
34. Vamvakas, E. C. (1998). Meta-analysis of studies of the diagnostic accuracy of laboratory tests: A review of the concepts and methods [see comments]. *Archives of Pathology and Laboratory Medicine* **122**(8): 675–686.
35. SAS Institute Inc (1996). *SAS/STAT Software*: *Changes and Enhancements through Release 6.11*, SAS Institute Inc., Cary, NC, 221–230.
36. Gu, c., Province, M., Todorov, A. *et al.* (1998). Meta-analysis methodology for combining non-parametric sibpair linkage results: Genetic homogeneity and identical markers. *Genetic Epidemiology* **15**: 609–626.
37. Borecki, I. B., Perusse, L. *et al.* (1994). An exploratory investigation of genetic linkage with body composition and fatness phenotypes: The Quebec family study. *Obesity Research* **2**: 213–219.
38. Clement, K., Garner, C., Hager, J. *et al.* (1996). Indication for linkage of the human OB gene region with extreme obesity. *Diabetes* **45**: 687–690.
39. Duggirala, R., Michael, P. and Mitchell, B. D. (1996). Quantitative variation in obesity-related traits and insulin precursors linked to the OB gene region on human chromosome 7. *American Journal of Human Genetics* **59**: 694–703.
40. Reed, D. R., Ding, Y., Xu, W. *et al.* (1996). Extreme obesity may be linked to markers flanking the human OB gene. *Diabetes* **45**: 691–694.
41. Allison, D. B. and Heo, M. (1998). Meta-analysis of linkage data under worst-case conditions: A demonstration using the human OB region. *Genetics* **148**: 859–865.
42. Wise, L. H., Lanchbury, J. S. and Lewis, C, M. (1999). Meta-analysis of genome searches. *Annals of Human Genetics* **63**(3): 263–272.
43. Wise, L. H. and Lewis, C. M. (1999). A method for meta-analysis of genome searches: Application to simulated data. *Genetic Epidemiology* **17**(suppl 1): 767–771.
44. Felson, D. (1992). Bias in meta-analytic research. *Journal of Clinical Epidemiology* **45**: 885–892.
45. Leizorovicz, A., Haugh, M. C., Chapuis, F. R. *et al.* (1992). Low molecular weight heparin in prevention of perioperative thrombosis. *British Medical Journal* **305**: 913–920.
46. Nurmohamed, M. T., Rosendaal, F. R. and Bueller, H. R. (1992). Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: A meta-analysis. *Lancet* **340**: 162–156.

47. Egger, M. and Smith, D. G. (1998). Meta-analysis: Bias in location and selection of studies. *British Medical Journal* **316**: 61–66.
48. Easterbrook, P. J., Berlin, J. A., Gopalan, R. *et al.* (1991). Publication bias in clinical reaserch. *Lancet* **337**: 867–872.
49. Simes, J. R. (1986). Publication bias: The case for an international registry of trials. *Journal Clinical Oncology*, 1529–1541.
50. Light, R. J. and Pillemer, D. B. (1984). Quantitative procedures. *Summing Up*: *The Science of Reviewing Research*. Harvard University Press, Cambridge, MA.
51. Egger, M., Smith, G. D., Schneidet, M. *et al.* (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**: 629–634.
52. Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**: 1088–1099.
53. Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychology Bulletin* **86**: 638–641.
54. Thornton, A. and Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology* **53**: 207–216.
55. Dear, K. B. G. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Sciences* **7**: 237–245.
56. Gleser, L. J. and Olkin, I. (1996). Model for estimating the number of unpublished studies. *Statistical in Medicine* **15**: 2493–2507.
57. Given, G. F., Smith, D. D. and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Sciences* **12**: 221–250.

About the Author

Xuyu ZHOU obtain her bachelor degree in Medicine (1990) from Tongji Medical University, Master degree in Biostatistics (1999) from Sun Yat-Sen University of Medical Sciences (SUMS). She is the director of the Information Retrieval Department in Medical Information Institute in SUMS. Her research interests are meta-analysis, evidence-based medicine and clinical epidemiology.

This page intentionally left blank

CHAPTER 8

DESCRIBING DATA, VARIABILITY AND OVER-DISPERSION IN MEDICAL RESEARCH

MING TAN

University of Maryland, Greenebaum Cancer Ctr,
Division of Biostatistics, 22 S. Greene St.,
Baltimore MD 21201-1544, USA
Tel: (410) 328-7516; mtan@umm.edu

1. Introduction

In an era of rapid advances in molecular biology and genetics, medical research at all levels (from basic science, to translational research and to clinical research) produces a wealth of data at an amazing speed. These data themselves are useless unless they are converted into information and knowledge. What distinguishes Statistics as a scientific discipline is that it aims to make inference about the unknown population from analyzing the sample data. A key concept that is constant in statistical theory and practice is that of variability. It is inherent in our daily lives, our data and in statistical estimates derived from the data. Because every person is different, a wonderful drug or therapy may only work for some but not all patients. Our blood pressures vary all the time. It can be influenced by when and how they are taken, whether you are worried or anxious about it, or you are in good health or not, and some other unknown factors (*the random variation*). Random variation is the unexplained variation, the noise part. In fact, controlling variability due to different possible factors is the subject of statistical experimental design. As we cannot possibly control all factors, the random variation is always there. Statistical methods provide justifications to how many subjects (how large a sample) will be needed to separate noise from trend, and statistical estimate of variability quantifies the uncertainty in biomedical findings. This knowledge can be further used to tailor treatment strategy for patients.

Clearly, describing the data and understanding the variability ties closely to the experimental design and biomedical process the data arises from and also the study design (the deliberate process of generating data for scientific investigation). Understanding the biological and medical process is essential to understanding and making sense of the data. Thus, methods and tools to describe and model the variability in a succinct way allow us to easily convey information in the data. The key is to understand the variability underlining the data. Common data analytical techniques are now well summarized.[1,7] The focus of this chapter is to introduce some more advanced methods for best describing and understanding variability.

2. Methods for Describing Data

The first step towards understanding data and making any inference is to understand what type of data we are dealing with since different types of data require different statistical methods for analysis. This is a fact sometimes easily overlooked by non-statisticians. We shall first review some common types of data in biomedical research with special emphasis on those not often discussed about in textbooks but appears increasingly often in medical research. We shall also point out the methods that ought to be used to analyze them.

2.1. *Types of data*

Although the data in biomedical research often is complex, they do fall into several common categories. Understanding them will guide us to choose the right methods for summarizing and analyzing the data. The type of data determines what methods will be used for analyzing the data and making inference. In addition to reviewing the basic types of data,[1] we shall describe other types of data that occur increasingly common in modern biomedical research.

2.1.1. *Categorical data*

When a patient or her conditions are classified into different categories, those observations would give rise to categorical data (or sometimes called dichotomous or attribute data). The simplest examples are the two-category (yes/no observations) such as if a patient has responded to cancer therapy or whether the patient is smoker or nonsmoker or whether the patient has colon cancer or not. This type of data is sometimes under the name

of binary data or 0–1 data. Data of three or more categories include blood types (A, B, AB, O), combined categories such as female and male leukemia and non-leukemia patients. Since there is no apparent order among these categories (blood types, gender/disease classifications), this type of data is also called nominal data. They can be analyzed with methods for contingency tables or a generalized linear model. Another type of categorical data include the classification of smokers (total none, occasional, heavy), the stages of breast cancer (I, II, II, IV), the degree of improvement after therapy (none, moderate, great, full), and the degree of pain (minimal, moderate, severe, unbearable) as subjectively assessed. Here, there is an apparent order among all the categories, these data are called ordinal data. However, just like in nominal data, arithmetic does not make sense in ordinal data although some of them may appear to be numeric, e.g. it is hard to say unbearable pain is twice as bad as severe pain.

One case such distinction may become obscure is the score data where scores are assigned to certain outcomes that does indicate an equal increment from one point to one point higher.

2.1.2. *Continuous data*

Continuous data arise when some form of measurements is taken, e.g. body weight and temperature, blood pressures and most of blood chemistry test (bilirubin, hemoglobin, cholesterol etc.). Oftentimes, these observations or its transformation (e.g. its logrithm) are considered normally distributed. Statistical methods and models for analyzing continuous data are most comprehensively developed. However, the accuracy of these measurements, knowledge about the reliability of these measurements is important to make valid inference.[4] Especially, it should be noted that when these observations are used as independent variables in the analysis, an errors-in-variable (or measurement error) model may be necessary.[2]

2.1.3. *Ratios*

Ratio data arise when we take ratio of two variables. For example, ejection fraction, an important cardiac function index, is the ratio of the difference between end systolic and diastolic volumes to end systolic volume, cardiac output, the percent change in renal function (e.g. ,the glomerular filtration rate) from certain baseline. More recently, the microarray gene expression ratio has become a focus of many cutting-edge medical research. The microarray technology has allowed fast large scale (up to thousands

of genes) analysis of gene expression. In these experiments, the ratios of gene expression from one color (red) signal to that another color (green) signal are expressed as spot for each gene. Then, the analysis of these gene expression ratios must take into account how the ratio is derived and an appropriate corresponding (in fact, Gamma) distribution should be used for analysis.[3,10] The influence and importance of measurement error are usually not well addressed in elementary textbook. Recent methodological research has further extended the measurement error models in generalized linear models and survival models.[2]

2.1.4. *Continuous proportional data*

This is really a subtype of ratio data when the ratio is a percentage between 0 and 1. It includes data such as the percentage of decrease in renal functions at different follow-up times from the baseline, and percentage of change from pre-treatment to post-treatment in terms of certain physiological variables or some molecutar or genetic targets. Statistical methods to directly model the means of the proportional responses have just emerged[12,13] using the simplex distribution of Barndorff-Nielsen and Jorgensen.[6] The simplex distribution takes into account the fact that such responses are percentages restricted between 0 and 1 and may as well have large dispersion. It has been discovered recently that there may well be large dispersion in this kind of data.

2.1.5. *Repeated measures*

In medical studies, subjects are often followed overtime either in natural history study of certain disease or therapeutic studies, or measurements or observations are obtained within certain experimental units or clusters (e.g. eyes or limbs of an individual). These observations are called repeated measures data, or if they are obtained over different times from the same individual, they are sometimes call longitudinal data. This kind of design is often necessary in order to assess how patients do overtime. For example, we may be interested how certain physiological variables (glomerular filtration rate) or genetic variables (for instance, telomere length) change over time, or whether certain events (e.g. ear infection) occur overtime.

The key issue here is that the within patient or with cluster correlation needs to be accounted for one in the experimental design and data analysis. For example, children who have ear infection in one of their ears may be more likely to have infection in their other ears. Thus, 10 patients with each

patient having 10 repeated measures do not the same power as 100 patients alone.

Depending on the type of response variable of interest, we may have repeated continuous data or categorical data or ordinal data. Different statistical models need to be used to analyze diffrernent kind of repeated measures data although the method is now unified with generalized linear models (GLIM). For repeated ordinal data, you may have to use models outside of GLIM, for example, the proportional odds model.[11,14]

2.1.6. *Censored and truncated data*

When we are not able to measure a variable precisely and only know that an observation is beyond some threshold, we call the observation censored. The most common censored data in biomedical research is the survival data, broadly defined, data of time to the occurrence of certain event, e.g. Epstein-Barr infection, or the death of a patient. This is perhaps one of the most common types of data in medical research, since we often want to know if a new drug regiment or a surgical or a medical procedure can save more lives than does a conventional treatment. Special techniques are needed in the analysis of survival data for several reasons. First survival data is generally not symmetrically distributed so not normally distributed, it is more satisfactory to use an alternative distribution in the model. Secondly, at the time of analysis, the survival endpoint (either it be death or remission of cancer) of some patients have not been observed yet, and the survival status may never be known since some patients may be lost to follow up.

2.2. *Variability*

Variability is one of the fundamentally important concepts that underlie all statistics theory and methods. As the world is full of uncertainty, it is fortunate to have statistics to study uncertainty scientifically and statisticians are also fortunate for uncertainty. Often a biologically active agent only has 5% chance to make to the clinic due to the variability experiment and mostly in human. Variability makes statistics and statisticians indispensable in medical research. So related another essential concept is to the probability distribution that is used to describe and analyze data. Often we assume that the observations are from certain distribution that is known except for some unknown parameters. The most prominent distribution is the normal distribution, which is fundamentally important in statistics because the central limit theorem suggests that most common

statistics be asymptotically normally distributed. Methods that do not assume parametric form is called nonparametric methods. The advantage of a parametric model is its simplicity and efficiency. Sometimes an intermediate (semi-parametric) approach is taken in that characteristics of main interest are assumed of a parametric model. So far parametric and semi-parametric methods are the most commonly used methods in medical research.

2.3. *Basic techniques*

The most common techniques for data description are mean and standard deviation, which is often associated with parametric description of the data. Normal distributions are completely specified by its mean and standard deviation. The mean is a measure of the central location and standard deviation is a measure of variability. Because of the importance of normal distribution based theory in statistical inference, these two numbers have special meaning. However, if the distribution of the variable under study is not normal, then they do not necessarily give good inferential values. Sometimes the variability may beyond what the assumed distribution can describe (the so-called *over-dispersion*).

Another commonly used statistics to describe data is the five number summary statistics, which are the minimum, maximum and 75%, 50% (the median) and 25% percentiles. Together with the mean and standard deviation, the five-number summary statistics give a good summary about the distribution of the data. For example, if the distribution is symmetric, then the mean and median should be equal. If the mean is greater than the median, the distribution is skewed to the right; and if the mean is less than the median, the distribution is skewed to the left.

2.3.1. *Example* 1 *(Phase I Clinical Trials and Pharmacokinetics Studies of Topotecan in Solid Tumors)*

Topotecan is a new molecular target based anti-cancer agent. It is a semi-synthetic water-soluble derivative of camptothecin whose anti-tumor effect is mediated by inhibiting topoisomerase activity by binding to the DNA topoisomerase I complex. This drug has shown promising anti-tumor activity in preclinical and clinical studies of adult and pediatric solid tumors.[5,15] The goal of the study is to determine if variability in topotecan lactone systemic exposure can be reduced by a dose adjustment strategy in a phase I clinical trial using pharmacokinetics (PK) guided dose escalation. Intravenous topotecan were given to 15 children with relapsed solid over

30 minutes 5 days a week for 2 consecutive weeks. Doses were individualized based on the patient's topotecan systemic clearance to maintain a single day plasma topotecan lactone area under the plasma concentration-time curve (AUC) of 150 ± 30 ng/ml*hr (**Cohort #1 for the first 8 patients**) or 100±20 ng/ml*hr (**Cohort #2 for 9 patients**) where two patients who had been in Cohort 1 were moved to Cohort 2 due to excessive toxicities. In fact, the AUC target was lowered to 100 ± 20 ng/ml*hr in general for toxicity concerns. Plasma samples were collected before at 0.25, 0.5, 1, 3, and 6 hours after completion of the topotecan infusion, which give one PK study using a two-compartment model. For each cycle of treatment at each dosage, PK studies were planned to be done on day 1, 3, 6, 8, 10.

2.4. *Graphic methods*

Indeed sometimes a picture is worth thousand words. Graphic methods are commonly used in statistics and medical research to depict the data and illustrate the methods. For example, the five numbers are commonly

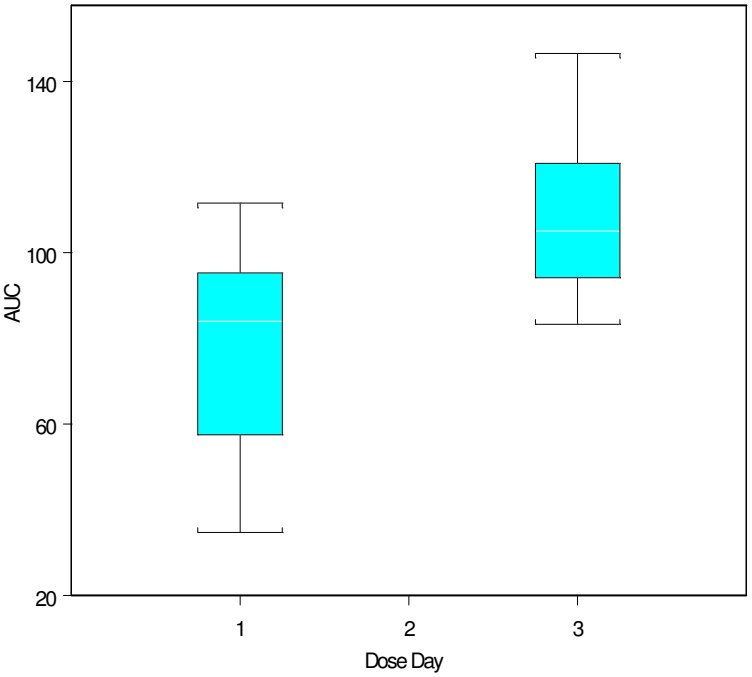


Fig. 1. Comparison of cycle 1 AUC: Days 1 and 3.

plotted as the box-and-whisker plot,[1] where the central line represents the median, the box represents the 25% and 75% percentiles (or the lower to higher quartiles), and the whiskers are the minimum and the maximum. To see the distribution and variability of AUC, which measures patient's systemic exposure to the drug, Fig. 1 gives box-and-whisker plots for AUC at days 1 and 3 for the 15 patients. As shown, the AUC is not symmetric and after dose given on day 1, AUC is skewed to the right, but at day 3, AUC becomes more symmetric, which partly represents the effect due to drug dose targeting based on pharmacokinetics.

To describe the distribution and variability of the data, histogram and some version of smoothing technique is often used. A spline smoothing estimator, a nonparametric estimate of the density, provides a better description of the probability density of the distribution. With modern statistical software, it is very easy to generate such estimate and overlay on the histogram. Figure 2 gives the histograms for the AUCs from the eight patients in Cohort 1 in Example 1. In the fixed group, the 36 AUCs were calculated alternatively using a fixed dose of 4 mg/m^2 divided by the patient's topotecan lactone clearance, and in the targeted group, the 8 PK studies from the first dose of the first cycle and one PK study from the second dose of the

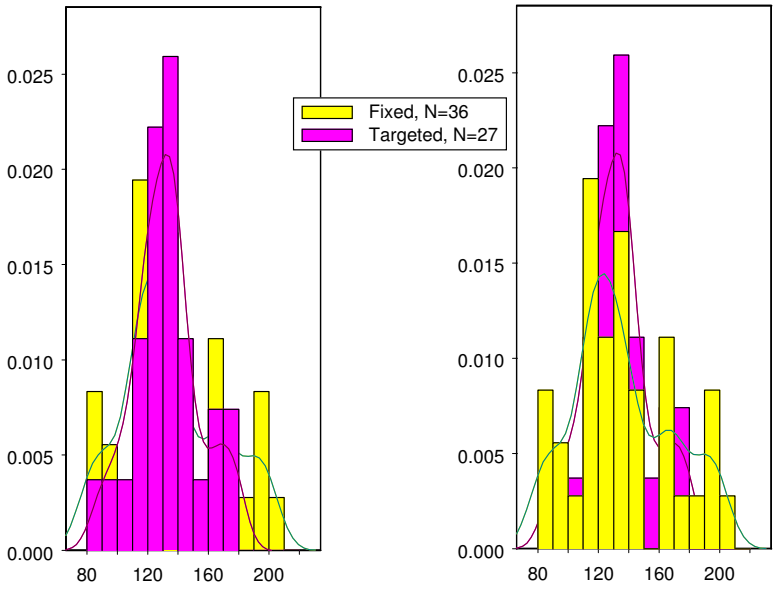


Fig. 2. Fixed and targeted AUC 150.

first cycle of one patient were excluded, so there were 27 PK studies from the eight patients whose AUCs were adjusted to the target AUC range by varying doses.

3. Describing Data via Adjusting for Factors with a Model

As alluded to early, effective description of the data depends on the design. Sometimes summarizing the data may not be so straightforward because of the complicated design under which the data are produced, e.g. dependence of the observations and missing values. A straightforward mean and standard deviation may be misleading. In Example 1, because patients have multiple cycles of chemotherapy and PK studies were performed at multiple doses and some patients missed some PKs, this gives rise to an unbalanced repeated measures data structure. We used a mixed-effects model to estimate the PK parameters and compare those whose AUCs fall into the targeted ranges and those whose AUCs were not. The summary statistics (such as the means and standard deviations) will need to account for within patient correlation. Table 1 gives both the estimates that accounted for such correlation and those that did not for comparison purpose. As shown in this table, the summary statistics using all the data based on the model may be different from a straightforward calculation and the ones that accounted for the within patient correlation should be considered for making inference.

More elaborate estimates based on models are often needed in order to avoid bias. Several other examples can be found in Meyers *et al.*[8] and Nelson *et al.*[9] where a mixed effects spline model was used to estimate the

Table 1. Mean and standard deviation estimates according to dose adjustment requirements derived from a mixed effect model.

	Estimated mean (SD)		
PK Parameter	Adjusted	Unadjusted	*p*-value
Vc	31.67 (2.51)	30.90 (2.10)	0.76
Kel	1.28 (0.12)	1.21 (0.11)	0.30
Kcp	0.90 (0.13)	0.83 (0.13)	0.59
Kpc	0.68 (0.06)	0.68 (0.04)	0.54
Alpha	2.51 (0.27)	2.36 (0.25)	0.51
T1/2_alpha	0.34 (0.04)	0.37(0.03)	0.44
Beta	0.34 (0.02)	0.33(0.02)	0.44
T1/2_beta	2.28 (0.19)	2.34 (0.19)	0.58
CL	33.88 (1.62)	32.14 (1.30)	0.20
Vdss	66.13 (3.78)	64.55 (3.11)	0.55

mean glomerular filtration rate for diabetic patients at different stages of their lives and the associated standard errors.

4. Over-Dispersion Issues

The term over-dispersion refers to the phenomenon that the observed variability (the variance) is more than the nominal variability (variance) under a presumed model. Statistically speaking, over-dispersion depicts that the mean-variance relationship of the assumed distribution is not correct. Although it is generally recognized that over-dispersion occurs in discrete data models under the binary and Poisson distribution assumptions. It also occurs in continuous proportional data shown recently in Song and Tan.[12] The existence of over-dispersion is noticed a long time ago in statistics. Fisher noticed a lot of data in practice is over-dispersed in 1951. Several natural questions arise such as what is the consequence of ignoring over-dispersion in the analysis and what are the appropriate techniques to detect and model the dispersion. In this section, we shall discuss thee questions in several distributions including the more familiar binary and Poisson data and the recent developments on proportional data.

4.1. *Binomial data*

Binary outcome, e.g. success/failure of therapy, response to a cancer drug, etc. is one of the most common outcomes in medical research. Generically, let the success probability be p and the binary (0–1) outcome of each of the n binary sequences (e.g. n cells, n mice, and perhaps n patients). Then the binomial outcome is $Y = \sum_{i=1}^{n} Y_i$. Over-dispersion arises when the empirical variance is greater than the binary variance $np(1-p)$, which is a function of the mean p. In this case, the distribution is completed determined by the mean parameter p. Then the variance of the binomial is

$$\sum_{i=1}^{n} \operatorname{var}(Y_i) + 2\sum_{i<j}^{n} \operatorname{cov}(Y_i, Y_j) = np(1-p) + 2\sum_{i<j}^{n} \operatorname{cov}(Y_i, Y_j)\,.$$

Therefore, when the binary sequences are not independent of each other, namely, $\operatorname{cov}(Y_i, Y_j)$ is not zero, over-dispersion would occur. The consequence would depend on how much the over-dispersion is. Generally, over-dispersion can not be ignored.

Testing if over-dispersion presents can be obtained through generalized linear models. With recent development in generalized linear mixed effects

model and Bayesian hierarchical model, over-dispersion can be accounted for directly in the modeling process.

4.2. *Poisson data*

Similar to binomial data, Poisson distribution is determined by its mean parameter. Since Poisson model belongs to the generalized linear model, similar test statistics and modeling methods can be used in testing and modeling over-dispersion.

4.3. *Continuous proportional data*

The continuous proportional data have not been talked about much is the continuous proportional data and the directional data. The continuous proportional data arise when the response of interest is a percentage between 0 and 1, for instance, the percentage of decrease in renal functions at different follow-up times from the baseline, or the percentage of decrease in blood pressures from the baseline. The usual practice has been just to treat them as normal distribution. However, as shown in Song and Tan[12] the variability in the response percentage is far beyond what the normal distribution can describe. In fact, although when the dispersion parameter is small, the dispersion models are approximately normal,[6] real world data are often with large dispersion as studied by Fisher in 1953. Here the normal model is usually not appropriate since if two variables are normally distributed, an assumption which is often considered plausible, the ratio of the two is generally not.

4.3.1. *Example* 2 *(A prospective ophthalmology study on the use of intraocular gas in retinal repair surgeries*[8]*)*

The outcome variable of the study was the percentage of gas left in the eye. The gas was injected into the eye before surgery for a total of 31 patients. The patients were then followed three to eight (average of 5) times over a three-month period. The volume of the gas in the eye at the follow-up times was recorded as a percentage of the initial gas volume in that eye. An important issue was to estimate the kinetics of the disappearance of the gas (e.g. decay rate of the gas). Clearly the response variable here is confined between 0 and 1. Although, for instance, a logit transformation results in a transformed response in, linear regression models with nonlinear transformed responses are often difficult to interpret. Particularly

the serial correlation structure of the nonlinear transformed responses can not be easily converted to that of the original responses. Our goal was to be able to model the dependence of mean gas decay on certain covariates directly. A common practice has been to assume that the response variable is normally distributed and ignore the fact that the responses are percentages confined between 0 and 1. However, as shown later, the variability in the response percentage is far beyond what the normal distribution can describe. In fact, although when the dispersion parameter is small, the dispersion models are approximately normal,[6] real world data are often with large dispersion.

A moment estimator of the dispersion parameter σ^2 may be obtained by using he fact that the expected value of $d(Y;\mu) = \sigma^2$ Therefore,

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{m} n_i - p} \sum_{i=1}^{n} \sum_{j=1}^{n_i} d(y_{ij}, \hat{\mu}_{ij}) \,,$$

which is a consistent estimator of σ^2 as m tends to infinity provided that $\hat{\mu}_{ij}$'s are consistent.

In Example 2, the estimate of dispersion parameter $\sigma^2 = 14.2$. The p-value based on a χ^2 distribution with 2 degree of freedom is 0.0008, suggesting that the dispersion parameter is significantly greater than 0, that is, significantly greater than the dispersion of a normal distribution. Thus, the gas volume is not normally distributed at all. In fact, graphically, the simplex density function with this large dispersion parameter indicates the density has a dominant mass between 0.8 and 1, which is consistent with the feature of the data, that is, over 40% of observations are in this range. Therefore, indeed, the dispersion is needed to analyze this kind of data.

Acknowledgments

The author thanks his collaborators Drs. Victor Santana and Clinton Stewart for making their data available for inclusion as an illustrative example, Dr. Peter Song for joint work on continuous proportional data and Kevin Liu and Catherine Billups for assistance with the data analysis. The author also acknowledges partial support by U.S. National Cancer Institute Comprehensive Cancer Center Support Grant CA21765 and by American, Lebanese, Syrian Associated Charities (ALSAC).

Appendix

The density of a simplex distribution,[6] with mean (location parameter) $\mu \in (0,1)$ and dispersion parameter $\sigma^2 > 0$,is given by

$$p(y;\mu,\sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{1/2} \exp\{-d(y;\mu/(2\sigma^2)\}\,, \quad y \in (0,1)\,,$$

where

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\,.$$

The advantage of using this distribution is that the simplex distribution is a dispersion model,[6] where the response has density function of the form

$$a(y;\sigma^2)\exp\{-d(y;\mu)/(2\sigma^2)\}\,, \quad y \in (0,1)\,.$$

The density for this dispersion model seems analytically similar to that of a normal distribution (see Jorgensen,[6] for details) and it also includes a large class of distributions confined in (0, 1), ranging from highly skewed to very flat distributions (see, e.g. Fig. 1.7 of Jorgensen[6]).

The dispersion model is more general than the familiar generalized linear model based on exponential family of distributions.

References

1. Altman, D. G. (1991). *Practical Statistics for Medical Research*, Chapman and Hall, London.
2. Carroll, R., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall, London.
3. Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarrays. *Nature Genetics Supplement* **21**: 33–37.
4. Gleser, L. J. (1994). The importance of assessing measurement reliability in multivariate regression, *Journal of the American Statistical Association* **87**: 696–707.
5. Houghton, P. J, Chesire, P. J, Myers, L. *et al.* (1992). Evaluation of 9-dimethylaminomethyl-10-hydroxycamptothecin against xenografts derived from adult and childhood solid tumors. *Cancer Chemother. Pharmacol.* **31**: 229-239.
6. Jorgenes, B. (1997). *Dispersion Model*, Chapman and Hall/CRC, London.
7. Moore D. S. and McCabe, G. P. (1989). *Introduction to the Practice of Statistics,* New York.
8. Myers, B. D., Nelson, R. G., Tan, M., Beck, G. J., Bennett, P. H., Knowler, W. C., J., Blouch, K. and Mitch, W. E. for the Diabetic Renal Disease Study. (1989). Progression of overt nephropathy in non-insulin-dependent diabetes. *Kidney Int.* **47**: 1781–1789.
9. Nelson, R. G., Bennett, P. H, Beck, G. J., Tan, M., Knowler, W. C., Mitch, W. E., Hirschman G. H. and Myers, B. D. for the Diabetic Renal Disease Study. (1996) Development and progression of renal disease in Pima Indians with non-insulin-dependent diabetes mellitus. *NEJM* **335**: 1636–1642.
10. Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2000). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **000**: 000–000.

11. Qu, Y. and Tan, M. (1998). Analysis of clustered ordinal data with subclusters via a Bayesian hierarchical model. *Comm. Statist. A*: *Theory and Method* **27**: 1461–1475.
12. Song, P. and Tan, M. (2000). Marginal model for continuous proportional data. *Biometrics* **56**: 496–502.
13. Tan, M. (2001). Using dispersion models in molecular pharmacology and genetics. Invited Presentation at *Joint Statistical Meetings*, Atlanta, GA, August 7.
14. Tan, M., Qu, Y., Mascha, E. and Schubert, A. (1999). A Bayesian hierarchical model for multi-level repeated ordinal Data: Analysis of oral practice examinations in a large anesthesiology training program. *Statist. Med.* **18**: 1983–1992.
15. Zamboni, W. C., Bowman, L. C., Tan, M. *et al.* (1999). Interpatient variability in bioavailability of the intravenous formulation of topotecan given orally to children with recurrent solid tumors. *Cancer Chemother. Pharmacol.* **43**: 454–460.

About the Author

Ming Tan is Professor of Biostatistics and Head of Biostatistics Division of the Greenebaum Cancer Center of the University of Maryland School of Medicine. He was Associate Member of Biostatistics at the Department of Biostatistic, St. Jude Children's Research Hospital and biostatastics director for the hospital's Developmental Therapeutics for Solid Malignancies Program. He was Assistant (1990–1996) and Associate Staff/Professor (1996–1997) of Biostatistics at the Department of Biostatistics and Epidemiology at The Cleveland Clinic Foundation, where he collaborated in multiple disease research areas ranging from cardiology to neurology and as biostatistician of coordinating centers for multi-center longitudinal studies funded by NIH and industry. He earned his PhD in Statistics from Purdue University in 1990. His current research interests include statistical methods in cancer drug development, group sequential designs for clinical trials, latent variable, random-effects and hierarchical Bayesian models for longitudinal data and accuracy of diagnostic tests. He serves as associate editor for *Biometrics and Communcations in Statistics*: *Theory and Methods* and as member for US NCI (National Cancer Institute) Clinical Oncology Study Section and member of FDA (Food and Drug Administration) advisory committee.

CHAPTER 9

TIME SERIES ANALYSIS AND ITS APPLICATIONS IN MEDICAL SCIENCES

JINXIN ZHANG and YINGDONG ZHENG

Department of Medical Statistics, School of Public Health, The Northern Campus of Sun Yat-sen University, 2nd Zhongshan Road, Guangzhou 510080, PR China
Tel: 86-020-87330673; jasonty@263.net; zjx@gzsums.edu.cn

DEJIAN LAI

School of Public Health, Texas State University, USA

1. Introduction to Time Series Analysis

When we try to observe dynamic variables $x_1, x_2, x_3, \ldots, x_i, \ldots$ in a medical research, these variables can be regarded as a stochastic process, since there are a considerable number of adventitious factors that may have effects on the data themselves with uncertainty. For example, the vital readings obtained from a monitor and from prevalence or mortality rates of some diseases in a particular region across time. The series of these observed values is called a time series. In Fig. 1, a time series of the number of outpatient visits in the Second Affiliated Hospital of Shanxi Medical University from January 1980 to December 1999 is shown. Generally speaking, the observed results of a series may not be expressed by a deterministic function; they can be treated as a realization of a stochastic process due to the influence of random factors. Let $\{x_t\}$ denote the stochastic process with $x_1, x_2, x_3, \ldots, x_i, \ldots$. Here t does not necessarily represent time; it may be the index of a space, temperature or vector.

Any observed result at a particular time is determined by many influential factors. Because of the interaction of these factors, the analysis of time series becomes quite complicated. The frequently encountered factors are the mode of trend, seasonality, periodicity or irregularity. In order to

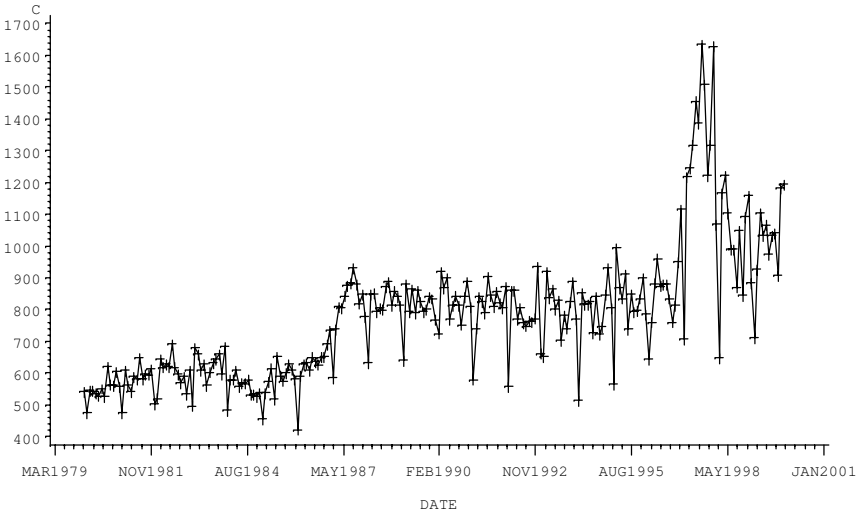


Fig. 1. The number of outpatient visits in the Second Affiliated Hospital of Shanxi Medical University in 1980–1999.

effectively apply the time series models, all these factors above must be taken into considerations.

Statistical predictions are usually based on previous and present information. The prediction derived from statistical models can guide our future decisions, avoid unnecessary mistakes and minimize the loss. Many phenomena, not only in nature, but also in life sciences are of statistical relationship. It is feasible in theory to treat those observed dynamic results as a time series with special properties.[1]

Statistical predictions can be divided into qualitative and quantitative areas. The necessary assumption is that the dynamics in trend, speed, etc., will be of relatively persistent in a long period of time. As they are changing almost all the time to some extent, the assumption becomes really fragile. The accumulation of changes in quantity may lead to a leap of quality; as a result, the relationship before and after the leap may be quite different. The prediction in quality is possible anyway. It is often true that more abundant collection of data can provide more reliable predictions. Thus, it is necessary to collect data as much and precise as possible.

If there are causal relationships among the variables, to establishment of these relationships using statistical models will make the prediction possible. However, neither finding the causal relationship nor collecting sufficient data to construct the model is simple.

Hannan[2] gave a detailed discussion on time series theory and the spectral analytical methodology. There are many strategies to calculate p — the order of autocorrelation and q — the order of moving average and relatively satisfied results may be found only when $q = 0$. Box and Jenkins summarized and presented some experiential principles from their experience to determine p and q in 1970.[2] Some more strategies have been reported after that, but they are generally based on exploratory trials.

In a series of reports given by S. M. Pandit and Wu, all stationary processes can be expressed as ARMA$(n, n-1)$, a simple mathematical form that we will explain in details later. Even if a stationary process is not in this simple format, it can still be approximated by an ARMA$(n, n-1)$ with acceptable accuracy.[3] Furthermore, many practically observed series can be represented by ARMA(2, 1). The fitness of ARMA$(n, n-1)$ to time series can be performed easily and effectively, even when the series comes from a ARIMA(p, d, q) or a ARMA(n, m) (here, $m \neq n$), two more complicated versions that will be discussed later. This shows it is unnecessary to worry that a time series may have a very complicated form in modeling and makes us confident that ARMA models can satisfactorily fit with common time series.

The linear trend is sometimes caused by unduly small intervals of sampling and this is called pseudo-trend. The solution is to use a smaller interval (when the observation cannot be done in a longer period) or to extend observation for longer period (when the intervals cannot be shortened). If neither the intervals nor the period can be changed, we need be aware of those tendencies, especially when the special explanation is difficult to be drawn.

The lag-free difference and seasonal difference are both helpful to change a nonstationary series to become a stationary one. Modeling is based on the attributes of autocorrelation function and periodograms or even the attributes of the original data itself. When data show a trend or seasonality, the autocorrelation will not attenuate rapidly and the corresponding periodograms tend to be distorted. In this situation, the lag-free and seasonal difference may make the identification of ARMA model less difficult. One useful strategy is that when there is any modulus of roots equal to 1 in the equation of model, differential operation is introduced into the equation in order to make full use of the provided information in time series.

One of the essential characteristics of time series is the correlation between observations, which is a basis for further analysis.[2] The procedure of analysis is generally divided into the followings: (1) model selection and

parameter estimation; (2) adaptability of the model; (3) prediction. The commonly used models are ARIMA models, exponential average, linear or multiple regression, growth curve, Markov chain and gray model.

Our past research experience focuses less in frequency domain but more in time domain. The effective prediction is also needed when missing observations exist. The observations in practice can be regarded as a realization of a stochastic process. As it is a sample from the whole process, the periodicity needs to be examined via a hypothesis test. Nonlinear analysis of time series has been popular recently in time series and nonlinear theory is needed to identify the model.[4]

1.1. *Models for time series analysis*

Time series analysis has been applied in economy, meteorology, geology, hydrology, military and other different fields of science successfully. Medical statisticians are also trying to utilize it in medical research.

The much-concerned research is not only on the essential conditions for applications of particular forms of models, but also on ideal fitness and prediction of those models. There are two types of seasonality — definite and indefinite. The definite seasonality means that the fitted model includes a term, which is the summation of periodic function and stationary noise. The indefinite seasonality means that the correlation between observations is significant with periodic intervals. For the definite seasonality, the difference will make the fitting and predicting difficult. For the indefinite seasonality, the difference is a necessary procedure for stationary. The research given by Bell and Hillmer[5] shows that the business per month may change because of the difference in numbers of Sundays in different months. Easter in western countries and Spring Festival in China may be located in different month according to the Gregorian calendar, and this also leads to the variation of business in that month. The effectiveness of the model may be improved when the above-mentioned situation is considered for monthly-based observations. The emendatory form of ARIMA model is $Z_t = \sum_{i=1}^{7} \beta_i T_{it} + \alpha H(\tau, t) + \frac{\theta(B)}{\phi(B)\delta(B)}\alpha_t$. The first and the second terms in the right hand of the equation are correspondent with Sundays and Easters.

Multivariate time series analysis is another point of much interest.[6] It can be summarized into two aspects[7]: (1) to determine the mode of correlation, such as circumstance, causation, or feedback; (2) to improve precision of predictions. When the predicted variable contains some

information from other external variables, the prediction will become more effective if these variables are included in the model. Tiao and Box[2] have pointed out the significance of spectral analysis: (1) to detect the correlation (lags may exist) in time series; (2) to be helpful to the explanation of the model. Chan and Wallis[8] put forward reformed vector autoregression model to prove the interactions between variables and to simplify the model (the explanation of interactions is coincident with professional knowledge although the variance of residuals increases). Ahn[9] discussed the low order components of scalar quantity after the first difference so as to improve the estimation of parameters in the model.

The parameters in time series models may be treated as time-varying sometimes and this has been verified by some practical experience. Stock[10] studied the elements λ in $V_t = \tau v_t = (\lambda/T)v_t$, which is the change of parameter. He has deduced the asymptotic unbiased estimator for the median of λ. Conditional heterogeneity appears as the change of variation situation along with time. Engle summarized this seminar paper in 1982[11] and presented ARCH model. The research on this topic followed with much interest from then on, especially in economic areas.[12,13] Many researchers have attempted to use semi-parametric or nonlinear nonparametric methods to fit the time series and the goodness of fit has been discussed a lot.[14–16]

It is known that ARIMA models have short-term effects. For long-term effects, autoregressive fractionally integrated moving average (ARFIMA) models are needed.[17,18] The models can be expressed as $\varphi(B)(1-B)^{\delta}Z_t = \theta(B)\varepsilon_t$, in which we have $(1-B)^{\delta} = \sum_{j=0}^{\infty} C_j(\delta)B^j$, where $\delta \in (-1, 0.5)$. As a matter of fact, with the term δ the observations in the infinite past may also have effects on the present value. ARFIMA is an example of long-memory time series model.

ARIMA model can be regarded as a transformation from original data into white noise. The residuals after modeling are the estimation of error and they are asymptotic to the error when the original series are long enough. The statistic $Q = n\sum_{k=1}^{m} \hat{r}_k^2$ is constructed and it is asymptotically of a χ^2 with degree of freedom $\nu = m - p - q$ when the series belongs to ARMA(p, q).[19] McLeod and Li[20] found that the variance of Q tends to be smaller and the precision of evaluation to the goodness of fit is improved coinstantaneously when the sample size becomes larger. Ljung and Box[21] presented that Q still can be used as the measurement of goodness of fit even though the errors ε_t may not be normal distribution. The statistics $S_1 = T^{-1}\lambda^{*\prime}V^{*-1}\lambda^*$ and $S_2 = T^{-1}\lambda^{*\prime}(V^* + G_1G_1')^{-1}\lambda^*$ have been constructed by Poskitt and Tremayne[22] in doing diagnostic test to the model. The studies

on noise have involved the nonlinearity and chaos theory. This makes the dimensions of the dynamic system to be fractions.[23]

1.1.1. *The memory in time series*

In statistical point of view, the dynamic characteristic appears as the correlation between the present events and the historical events. The correlation function is used to portray the characteristic in time series.

In the view of systems theory, memory means dynamic characteristic, with which the subsequent outputs are influenced by the present input. The system has dynamic characteristic of first order when any particular observation only effects the next observation following it. Similarly, the system has dynamic characteristic of n order when any particular observation can effect the next n observations after it. For example, a patient takes analgesic drug at the time of T; it can be regarded as an input to the system at T.

When the drug reacts only at the next observing time, as it may be illustrated as Fig. 2, we say the metabolism system is of the first order. It shows that after taking drug at T the situation of the observation next to it becomes very well, but becomes worse after that point. When the drug is effective during the next four observing points although becomes less effective gradually in this period (Fig. 3), this is called the fourth order system.

When the input does not only influence the present output (by the intensity of φ_0) but also the next output (by the intensity of φ_1), the model

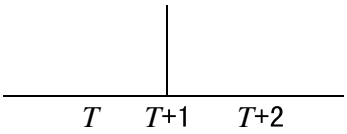


Fig. 2. The effectiveness of an analgesic drug (memory of 1st order).

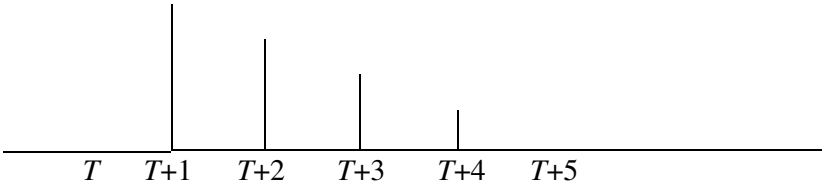


Fig. 3. The effectiveness of an analgesic drug (memory of 4th order).

can be denoted as $X_t = \varphi_0 W_t + \varphi_1 W_{t-1}$. The generalized form is

$$X_t = \varphi_0 W_t + \varphi_1 W_{t-1} + \varphi_2 W_{t-2} + \cdots ,$$

Where $\varphi_j (j = 0, 1, 2 \cdots)$ are the influential intensity of W_{t-j} to X_t and φ is called memory function. As a matter of fact, the memorial characteristic is the basis for us to establish models for the system and predict the future situation.

1.1.2. *The collection of data*

Observing and recording the output from the system with given intervals is called sampling. The sampling intervals are often denoted by Δ. The sampled observation after kth intervals is denoted as X_k. It is the discretized results $X(t_0 + k\Delta)$ from $X(t)$. Sometimes a time series may be the values of accumulated results. For example, the numbers of births in a month in a region or the daily urine output from a patient can form a time series.

As time series can be viewed as the output from a dynamic system, the systems theory can be used to analyze the dynamic structure and evolutive relationship. However, the discretized results may lose the information between $t_0 + (i-1)\Delta$ and $t_0 + i\Delta$. The shorter the interval Δ is, the less information is lost. At the same time, we get more observations and may waste our resource without much additional useful information. To choose proper interval lengths to sample the underlying dynamic system is also a critical procedure for the researchers.

1.1.3. *The pre-treatment of time series*

Just like any other statistical tools, time series analysis deserves careful treatment. It is not recommended to model the time series blindly without careful check and pre-treatment.

Take the time series of the natural growth rates of Chinese population for a simple example. The growth rates are the ratios of the changed numbers and the average population size in a given period. In the consideration of population management, the length of the period is usually the calendar year from January 1 to December 31.

The definition of the variable should keep consistent. Although Hong Kong has returned back to China in 1997, considering of the consistency, "Chinese population" still means the population in the Mainland China after that year. Another consideration is to the calculation method for the

annual average population. For example, we can use half of the sum of the population at the beginning and the population at the end of the year, i.e.

$$\text{the mid year population} = \frac{\text{population at the beginning} + \text{population at the end}}{2}.$$

1.1.4. *Missing values and the interpolation*

During sampling of observations, missing values may occur due to malfunction of instruments, mal-operation to the instrument or the unexpected observing conditions. They may also occur when trying to subdivide the sampling intervals.

When such missing values exist, the time series is corrupt. The fragmentary series is hard to be analyzed with commonly used time series models. However, it may not be possible to repeat the history values. An easy remedy is to interpolate the observations according to the tendency of the series. On the other hand, the models that are effective to the series of unequal intervals are beneficial in such situation.

1.1.5. *Stationary process*

The stationary process is a process[2] that has steady statistic characteristics. When the following equation holds to any continuous $t_1, t_2, \ldots, t_n$ and any given ε,

$$F_n(x_1, x_2, \ldots, x_n;\ t_1, t_2, \ldots, t_n) = F_n(x_1, x_2, \ldots, x_n;\ t_1 + \varepsilon, \ldots, t_n + \varepsilon)$$

where $F_n(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n)$ is the distribution function of x_1, $x_2, \ldots, x_n$ at time $t_1, t_2, \ldots, t_n$ for any n. Then $\{X_t\}$ is called a strict stationary process.

As the distribution function describes the statistical characteristics perfectly, the above equation means that all statistical characteristics will not change along with time. It is so called strongly stationary process. It seems that these characteristics may be used to establish a principle to judge whether a process is stationary. Unfortunately, this will be difficult for practical use. Stationarity of the process indicates the environment and main influential factors retain relatively stable along the period of time. For example, when manufacturing drugs, the output can be regarded as a stationary process as the raw material, the functions of product line, proficiency of workers are the same.

A better workable stationarity of the time series is the weak stationary defined as followings[24]:

$$\begin{cases} E[X_t] = a & \forall\, t \in T \\ E[X_{t+\tau} - a][X_t - a] = R(\tau) & \forall\, t, t+\tau \in T\,. \end{cases}$$

Here, $R(\tau)$ is the covariance function of X_t, which is independent of t.

1.1.6. *Test for stationarity*[25]

As we have mentioned in Sec. 1.1.5, we now discuss the stationarity in two forms — weak and strong stationarity. In practice, we may consider not only the statistics but also dynamic system characteristics. One useful way is to check the absolute value of the latent roots λ s. If there is a $|\lambda| > 1$, that indicates that stationarity is not tenable.[2]

A trend may be random or deterministic. The deterministic trend has a consistent influence and makes the series non-stationary.[2] However, the system can still be treated as stationary when the trend drifts randomly.

(1) **Plot** We can examine the periodic trend by the plot of X_t changing along with t to check the stationarity. The series can be treated as stationary if there is no evidence of periodicity. This strategy is easy to understand and perform. However, the performer needs plenty of experience and the results may be different from each other.

(2) **Autocorrelation and partial autocorrelation** The autocorrelation and partial autocorrelation of a standardized time series ($Ex(t) = 0$) are either tailed or cut off. If the two functions belong to neither of the above situations, the series may be nonstationary. For example, autocorrelation decreasing gradually (periodically or not) indicates that a particular trend or periodicity may exist.

(3) **Eigenvalue** Fitting the series with a model and then calculate the eigenvalues of the eigenfunction corresponded with the model. If all the eigenvalues satisfy $|\lambda| < 1$, the series is stationary. Otherwise, it is nonstationary.

(4) **Parameters** Autocorrelations can be used to define the stationarity. We can check the model of time series and calculate the autocorrelations. The following array can be obtained, where $\varphi_0 = -1$. The parameters in the first row are autocorrelations, in the second row are autocorrelations

row	parameters						
1	φ_0	φ_1	φ_2	$\cdots$			φ_n
2	φ_n	φ_{n-1}	φ_{n-2}	$\cdots$			φ_0
3	a_0	a_1	a_2	$\cdots$		a_{n-1}	
4	a_{n-1}	a_{n-2}	a_{n-3}	$\cdots$		a_0	
5	b_0	b_1	b_2	$\cdots$	b_{n-2}		
6	b_{n-2}	b_{n-3}	b_{n-4}	$\cdots$	b_0		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$		$\vdots$	
$2n-3$	l_0	l_1	l_2				

ordered inversely, in the third row are

$$a_i = \begin{vmatrix} \varphi_0 & \varphi_{n-i} \\ \varphi_n & \varphi_i \end{vmatrix} = \varphi_0\varphi_i - \varphi_n\varphi_{n-i}\,, \quad i = 0, 1, 2, \ldots, n-1\,.$$

In the determinant a_i, the first column is the elements located in the first two rows and the first column of the table, the second column is the elements located in the first two rows and the ith column. In the fourth row are the same elements as in the third row but ordered inverse-wise. In the fifth row are,

$$b_i = \begin{vmatrix} a_0 & a_{n-1-i} \\ a_{n-1} & a_i \end{vmatrix} = a_0a_i - a_{n-1}a_{n-1-i}\,, \quad i = 0, 1, 2, \ldots, n-2\,.$$

The elements in the sixth row are same as those in the fifth row but ordered inverse-wise. The other rows are calculated in the similar way. Only three elements are left in the $(2n-3)$th row. The series is stationary when the following three conditions are satisfied.

$$\begin{cases} \varphi_1 + \varphi_2 + \varphi_3 + \cdots + \varphi_n < 1 \\ -\varphi_1 + \varphi_2 - \varphi_3 + \cdots + (-1)^n\varphi_n < 1 \\ \quad |\varphi_n| < |\varphi_0|\,, \quad |a_{n-1}| < |a_0| \\ |b_{n-2}| < |b_0| \\ \qquad \cdots \\ \quad |l_2| < |l_0|\,. \end{cases}$$

(5) **Inverse order test**[25] Inverse order test is a method to detect special tends of mean or variance. The procedure is as follows.

- Cut the series into M parts and calculate their means or variances and the results are analyzed.

- Count the numbers of inversed orders. An inversed order is defined as that, there is a value in the series greater than value that situated formerly, $y_j > y_i (j > i)$. The number of inversed orders of y_i is denoted as A_i. The total number of inverse orders is $A = \sum_{i=1}^{M-1} A_i$.
- Construct a statistic for hypothesis test. The expectation and variance of the test statistics under the null hypothesis of no trend are as the followings:

$$E(A) = \frac{1}{4} M(M-1) \tag{1}$$

$$D(A) = \frac{M(2M^2 + 3M - 5)}{72}. \tag{2}$$

Here M is the length of the series y_i. We then establish a statistic Z,

$$Z = \frac{[A + \frac{1}{2} - E(A)]}{\sqrt{D(A)}} \tag{3}$$

which distributes asymptotically as $N(0, 1)$. The original series x_i is stationary when $|Z| < 1.96$ under the significant level $\alpha = 0.05$. Otherwise, x_i is nonstationary.

The series x_i contains an increasing trend when A is large and contains a decreasing trend when A is small.

The hypothesis test mentioned here is effective to those monotonic trends. As to those complicated trends, other strategies are needed.

(6) **Hypothesis test based on number of runs**[25] Assume that the mean of $\{X_t\}$ is $\bar{X}$ and we transform the original series into a series of signs. Those values equal or greater than $\bar{X}$ are changed to be "+" and the rest to be "−". A piece of the new series composed of continuous and same signs is called a run. For the series X_t,

$$5 \quad 6 \quad 6 \quad 9 \quad 5 \quad 6 \quad 4 \quad 8 \quad 3 \quad 8$$

the mean is $\bar{X} = 6$ and the new series is,

$$- \quad + \quad + \quad + \quad - \quad + \quad - \quad + \quad - \quad +$$

there are 8 runs in it.

The basic logic of run test is that the observations take values randomly around the mean if the time series is stationary. If there are too few numbers of run, the observations continuously get values higher or lower than the mean, which indicates the existence of some monotonic trends or periodic fluctuations. If there are numerous numbers of run, some nonrandom factors

may also exist. For example, if there are $n-1$ runs in a time series with sample size n, it indicates that the observations are correlated negatively with the first order.

Let N_1 be the number of "+"s and N_2 be the number of "−"s in the transformed series and the total number of runs be r. To a random series, we have

$$E(r) = \frac{2N_1N_2}{N} + 1\,,$$

$$D(r) = \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)}\,.$$

When both N_1 and N_2 are bigger than 15, the statistic $Z = \frac{r-E(r)}{\sqrt{D(r)}}$ is distributed asymptotically as $N(0,1)$.

Under the significance level α, if $r_L < r < r_U (r_L = E(r) - 1.96\sqrt{D(r)}$ and $r_U = E(r) + 1.96\sqrt{D(r)})$ or when $|Z| < 1.96$ holds, the series is stationary. Otherwise, the series is nonstationary.

1.2. *Inverse autocorrelation and its application in the identification of ARMA models*

Two important considerations to fit ARMA models are the goodness of fit and abstention of parameters. Let's assume that $\{x_t\}$ is a stationary time series in which mean and auto-covariance both satisfy the weak stationary conditions. The following is well known to us: If autocorrelation $r(k)$ or auto-covariance $\gamma(k)$ is quite close to zero (i.e. cut off) after $k = q$ and partial autocorrelation $p(k)$ decreased gradually (i.e. tailed down), $\{x_t\}$ is often an MA(q) series. On the other hand, if partial autocorrelation $p(k)$ cut off at $k = p$ and autocorrelation $r(k)$ tailed down, $\{x_t\}$ is often a AR(p) series. If both autocorrelation $r(k)$ and partial autocorrelation $p(k)$ tails down, then $\{x_t\}$ is often an ARMA(p, q) series. When we try to establish a proper model for time series, we wish a parsimonious model without too much lack of fitness and nor over fitness may happen. Inverse autocorrelation is helpful for us to find a relatively optimal model as it may show much structural information of the series.

1.2.1. *Definition*

We denote the spectral density of $\{x_t\}$ as $S(f)$, auto-covariance of as $\{x_t\}\gamma(k)$ autocorrelation of $\{x_t\}$ as $r(k)$ and k the number of lag

$(k = 0, 1, \ldots)$. We have,

$$\gamma(k) = \int_0^1 e^{2\pi\, ikf} S(f) df\,,$$

$$r(k) = \gamma(k)/\gamma(0)\,.$$

Let $Si(f) = 1/S(f)$ and $1/S(f)$ is integrable, then inverse autocorrelation is defined as

$$\gamma_i(k) = \int_0^1 e^{2\pi\, ikf} Si(f) df \tag{4}$$

$$ri(k) = \gamma i(k)/\gamma i(0)\,. \tag{5}$$

Here, $ri(k)$ can be comprehended as the autocorrelation of a time series corresponded to a spectral density $Si(f)$.

1.2.2. *Some characteristics of* $ri(k)$

(1) When $\{x_t\}$ fits an AR(p) model,

$$x_t + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \mu_t = \varepsilon_t$$

we have

$$ri(k) \neq 0 \quad k \leq p$$

$$ri(k) = 0 \quad k > p$$

If $\{x_t\}$ fits an MA(q) model, the $ri(k)$ tends to tail down gradually.

(2) The inverse autocorrelation $ri(k)$ can be used together with $r(k)$ to estimate the parameters in an ARMA(p, q) model. Assume that $\{x_t\}$ is a series satisfying ARMA(p, q) model $x_t + \sum_{j=i}^{p} \alpha_j x_{t-j} + \mu = \varepsilon_t + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}$ and we have the following difference equations:

$$r(k) + \alpha_1 r(k-1) + \cdots + \alpha_p r(k-p) = 0 \quad k > q$$

$$ri(k) + \beta_1 ri(k-1) + \cdots + \beta_q ri(k-q) = 0 \quad k > p$$

When $r(k)$'s and $ri(k)$'s are substituted with $\hat{r}(k)$'s and $\hat{r}i(k)$'s respectively. The estimations of α_j's and β_j's are obtained by solving the above simultaneous linear equations. The least squares solution may be used when k is large enough.

(3) If any parameters in an ARMA model equal to zero and the number of necessary estimators are fewer than $(p + q + 2)$, the model changes into

an ARMA with sparse coefficients.

$$x_t + \alpha_{i1}x_{t-i1} + \alpha_{i2}x_{t-i2} + \cdots + \alpha_{ip}x_{t-ip} + \mu$$
$$= \varepsilon_t + \beta_{j1}\varepsilon_{t-j1} + \beta_{j2}\varepsilon_{t-j2} + \cdots + \beta_{jq}\varepsilon_{t-jq}\,, \quad \text{where } ip \geq p,\ jq \geq q\,.$$

These kinds of models commonly happen in seasonal time series. It is difficult to judge the formation of the suitable model with only the information from autocorrelation and partial autocorrelation. Fortunately, if the α_k's are zero, the corresponding $\hat{r}i(k)$'s are approximately zero. This will be helpful for us to find a model with abstentious parameters and simplify the calculation of parameter estimation.

1.2.3. *Estimation of* $\hat{r}i(k)$

Assume that $\{x_t\}$ belongs to a model with pth order autocorrelation. Any invertible ARMA series can be changed into an AR series with a higher order (possibly infinite).[26] $\hat{\sigma}^2$ and $\hat{\alpha}_j (j = 1, 2, \ldots, p)$ are estimates to σ^2 and α_j according to the information in the original series. $ri(k)$'s can be estimated by,

$$\hat{r}i(k,p) = \frac{\sum_{j=0}^{p-k} \hat{\alpha}_j \hat{\alpha}_{j+k}}{\sum_{j=0}^{p} \hat{\alpha}_j^2}\,, \quad \text{where } \begin{matrix} \hat{\alpha}_0 = 1\,, \\ k = 1, 2, \ldots, p\,. \end{matrix}$$

The $\hat{\alpha}_j$'s can be solved by Yule-Walker's Eq. (7) or Durbin's recurrence formulas.[2]

In practice, the order p is unknown at the beginning. When $\{x_t\}$ is a pure AR process, p can be estimated by partial autocorrelation; When $\{x_t\}$ is not a pure AR process, then inspect the $\hat{r}i(k,p)$'s with different p. At the place where p stops to fluctuate, the corresponding p is a suitable order. $\hat{r}i(k)$ can also be estimated by periodogram.[2]

1.2.4. *The use of* $\hat{r}i(k)$

The procedure to find suitable order of ARMA(p, q) is usually based on try and error.[2] In practice, p is the value where the residual variance begins to be stable. If p is too small, the estimation of $\hat{r}i(k;p)$ is more likely to be biased. If p is too big, the standard error is large. A good strategy is to calculate a series of $\hat{r}i(k;p)$ with different p. After the proper p, $\hat{r}i(k;p)$ tends stable.

With larger p and q, the ARMA(p, q) model may fit the data better. However, larger order may lead to the increase of the estimation errors of

the parameters. Take AR(p) model for example. When $\alpha_{p+1} = \alpha_{p+2} = \cdots = \alpha_{\hat{p}} = 0$, the real model AR($p$) becomes a specific form of AR($\hat{p}$) when $\hat{p} > p$. The precision of parameter estimation declines because of the extra estimation for $\alpha_{p+1}, \alpha_{p+2}, \ldots, \alpha_{\hat{p}}$.

With the information provided by $\hat{r}i(k)$, the fitted model has a more solid foundation. Especially to sparse coefficient models and seasonal autocorrelation models, the effectiveness of the estimation will be improved considerably. Some other principles such as FPE, AIC and BIC are also commonly applied in model selection.[27] No matter which principle is used, the diagnostic tests on the residuals are necessary. If the residuals can pass the tests of randomness, the model is an acceptable one. Otherwise, more investigations are needed to find an effective model.

2. Predictions in Time Series

A condition for time series prediction is that the series can be summarized with a set of parameters and they are consistent after the observing time. The research by Box and Tiao[28] showed that the predictive errors will increase when the model fails to describe the series. Assume that the prediction residuals are $a_1, a_2, \ldots, a_m$, the variance of noise is $\hat{\sigma}^2$ and then the statistic $\hat{Q} = \hat{\sigma}^{-2} \sum_{l=1}^{m} a_1^2$ belongs to a F distribution with degree of freedom m and $(n-p)$, where p is the number of parameters in the model. When $\hat{Q}$ is larger than the critical value, we conclude that the model is lack of fit.

When the original assumptions of the time series are changed and the model fails to describe the time series, under the new conditions, we say a structural break happens in the dynamic system. It is reported that[29] we can decrease or offset the changes in the conditions with innovation of intercept or difference of series. Structural breaks are ubiquitous caused by known or unknown reasons. Granger and his colleagues have provided some suggestions[30] for model construction: If the structural breaks is expected, the different models to the separate periods should be assigned correspondingly. If it is unexpected, the preparation for dealing with the breaks should also be considered before hand.

How to perform prediction after the breaks? Clements and Hendry[31] have done some research on structural breaks. They conclude that an ideal goodness of fit may not necessarily lead to a satisfactory predictions; a definite change (e.g. the unification of the Western and Eastern Germany) may not produce an item in the commonly used forecasting model. A

precipitate innovation to the intercept may improve the predictive precision. When structural breaks happen, the first thing for the researcher to do is to find it as soon as possible. Then it is necessary to update the model effectively. They both deny that nonlinear models have obvious significance to economic data although the nonlinear theory has been developed extensively.

The subjective prediction based on experience and the objective prediction based on statistical modeling are both important in practice.[32] With the development of computing power, objective prediction has moved forward greatly. However, the performance of the prediction is based on assumptions and the application is also constrained. For the objective method, the assumptions are relatively weak and easy to apply. Many special topics are available in the literature like how to detect and analyze the trend in time series, the skills for dealing with seasonality, the flexibility in time series, the way to treat noises in time series data, the effectiveness corresponded with the length of historical values, how to choose a proper lead time in prediction, the feedback effects in time series and how to present results with an essential form.

Bewley[33] has discussed how to combine these two aspects with the examples of diffusion models and vector autoregressive models. Ten items of principle is summarized for performing statistical predictions.

Seasonality is an important ingredient in time series analysis. Seasonal difference and X-11 method for ARIMA series are used to detect particular periodicity or other deterministic elements.[27] The following model is needed to describe the situation where some periodic elements exist.

$$x_t = \sum_{s=1}^{S} v_s D_{s,1} + \sum_{s=1}^{S} \phi_{s,1} + \sum_{s=1}^{S} \phi_{s,1} D_{s,t} x_{t-1} + \cdots + \sum_{s=1}^{S} \phi_{s,p} D_{s,t} x_{t-p} + \mu_t \,,$$

where $D_{s,t} (s = 1, 2, \ldots, S)$ are dummy variables. When t is in the sth season, $D_{s,t} = 1$. Otherwise, $D_{s,t} = 0$. This model above is called univariate periodic time series model, which makes the prediction more effective.[34] However, any inadequate seasonal adjustments will distort the characteristics of the series in trend, periodicity and non-linearity.[35] As to the unit root test, it tends to accept the null hypothesis and lead to the abuse of difference.[36,37]

Wallis and Whitley[38] reviewed the predictive errors occurred in economic prediction in England from 1984 to 1988. They found that the theoretical characteristics and practical efficiency are quite different. Innovation is needed to supply necessary information to the model for prediction. If the

conditions of the original series have changed considerably, the prediction errors will certainly increase. It is also reported that[39] trend may appear as the autocorrelation with lower orders, introducing into the model with an autocorrelation item instead of a trend item will lead to the increase of residuals. Welch suggested[40] that the correlation between closely-located values not be taken seriously into account but the shift of mean be paid more attention. In one word, there is no such model that can substitute the others. The predictive efficiency is correlated with special conditions and the ideal model is only locally optimal.[41]

The time series analysis has used achievements in other disciplines. The prediction with ARMA model has borrowed the principle from system theory to process a signal with a filter, which has a particular form of transfer function.[42] The state transfer function and the measurement function are $x_r = Fx_{t-1} + G\varepsilon_{t-1}$ and $y = H'x_t + \varepsilon_t$, where $G = (\phi_1 - \theta_1, \phi_2 - \theta_2, \ldots, \phi_r - \theta_r)'$, $H = (1, 0, \ldots, 0)'$. We have,

$$F = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 & 0 \\ \phi_2 & 0 & 1 & & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ \phi_{r-2} & 0 & 0 & \cdots & 1 & 0 \\ \phi_{r-1} & 0 & 0 & \cdots & 0 & 1 \\ \phi_r & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad F_\theta = \begin{bmatrix} \theta_1 & 1 & 0 & \cdots & 0 & 0 \\ \theta_2 & 0 & 1 & & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ \theta_{r-2} & 0 & 0 & \cdots & 1 & 0 \\ \theta_{r-1} & 0 & 0 & \cdots & 0 & 1 \\ \theta_r & 0 & 0 & \cdot & 0 & 0 \end{bmatrix},$$

and we can get the state transfer function as,

$$P_{t+1,t} = F_\theta \left\{ P_{t,t-1} - P_{t,t-1} H \sum_{t}^{-1} H' P_{t,t-1} \right\} F_\theta' .$$

Swanson[43] tried to fit economic time series with several other models. His results showed that flexible specification models and less flexible fixed specification linear models both tend to capture the shifting trend easily, especially when the lead time is longer than 1. When the strategy for model selection has changed, the model may become less optimal and special cost functions are needed for the evaluation of the models.

The prediction precision is one important consideration for model selection. The selection principles can be separated into the aggregate selection rule and the individual selection rule. The former one is to select a uniform model for all variables and the later one is to select different models for all variables respectively. Shah[44] managed to apply individual selection

rule combined with discriminate analysis. The result is that the individual selection based on the scores of discriminate analysis is better than any aggregate selections.

The disputes on the efficiency of established models are concentrated on whether the mathematical models can summarize the causal or contextual relations hidden in the time series. Lim and O'Connor[45] have investigated this issue. Their conclusion is that the effectiveness of prediction is not improved if the information is not reliable. Otherwise, it will be better than so called optimal model by the proceeding selection. However, some researchers[46] still engage themselves to obtain an optimal model that may summarize the causal and contextual relation. The innovation to predictive values is neglected.

2.1. *ARIMA model and its application to the prediction of medical supplies in a hospital*

The sufficient supply of medical consumed material in polyclinics should be provided to serve for the diagnostic and treatment activity. The prediction is needed in order to avoid conflict between supply and demand.[26]

2.1.1. *The method for prediction*

The medical material demanded in a hospital is influenced by many factors, which are difficult to be modeled with. However, the observed time series of the consumed material can be treated as one realization of stochastic process.[27]

The theoretical and practical researches on quantitative prediction have been attracted more and more attention. The strategies such as moving average, trend fitting, exponential move, seasonal trend model, Markov chain, gray models and ARIMA models are all widely used.

2.1.2. *ARIMA model*

The ARMA(p, q) model was put forward synthetically by Box and Jenkins in 1970[2] and is also called the Box-Jenkins model. The model can be expressed as the following,

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{6}$$

where $\varphi_1, \varphi_2, \ldots, \varphi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ are the coefficients of autoregression and moving average. It can be simplified as $\varphi(B)y_t = \theta(B)a_t$, where

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$
$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

and

$$By_t = y_{t-1}\,.$$

B is called the back shift operator. If the dth difference transform can make a nonstationary time series change into ARMA(p, q),[2] the model is an ARIMA(p, d, q) model.

2.1.3. *Identification of the model*

Autocorrelation, inverse autocorrelation and partial autocorrelation are three main resources for us to select models for a stationary time series.[27] We have mentioned some principles about this at the beginning part in Sec. 1.2. For those nonstationary series, using ARIMA(p, d, q) models may be applicable. ARIMA$(p, d, q)(P, D, Q)$s models are useful to the series that contains seasonality.[27] The effectiveness of fitting is evaluated by analysis of residuals. When the residuals are accepted as white noise, the model fits the time series well.

2.1.4. *Estimation of parameters and diagnostic test*

The sample autocorrelation r_k is the correlation of the time series with the same series with a lag of k defined as

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}\,.$$

By the relations of sample autocorrelations with the coefficients φ_k and θ_k, the estimation of the coefficients in the model can be realized using Yule-Walker equations.[26]

To the pth ordered autoregression process AR$(p)x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t$, we have the following simultaneous equations which

are called Yule-Walker equations.

$$\begin{cases} \rho_1 = \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1}\,, \\ \rho_2 = \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-1}\,, \\ \quad \vdots \\ \rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p\,, \end{cases} \tag{7}$$

where $\rho_k = \mathrm{cov}(x_t, x_{t-k})$. The estimated results of autoregression coefficients from (7) are called Yule-Walker estimators. Let

$$\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix},\ \rho_p = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix},\ P_p = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix}. \tag{8}$$

Using the sample autocorrelations we have $\hat{\rho}_p = \hat{P}_p\hat{\phi}$. The Eqs. (7) can be denoted as $\phi = P_p^{-1}\rho_p$, where

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{pmatrix}, \quad \hat{P}_p = [r]\,, \quad \hat{p} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix}.$$

When $p = 2$, we can estimate ϕ_1 and ϕ_2 with the following formulas:

$$\hat{\phi}_1 = \frac{r_1(1 - r_2)}{1 - r_1^2}\,,$$

$$\hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}$$

The relationship between partial autocorrelations ϕ_{ki} and autocorrelations ρ is[2]

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{k(k-1)}\rho_{j-k+1} + \phi_{kk}\rho_{j-k}\,, \quad j = 1, 2, \ldots, k\,,$$

where not all the ϕ_{kj}'s are zero.

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix}, \tag{9}$$

i.e.

$$P_k \phi_k = \rho_k \, .$$

The solutions of the equations can be deduced when $k = 1, 2, 3, \ldots$

$$\phi_{11} = \rho_1 \, , \quad \phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \, ,$$

$$\phi_{33} = \begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix} \div \begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix} \cdots \, .$$

This is how we get partial autocorrelation ϕ_{kk}.

The hypothesis test for the validity of the model includes at least the followings:

- **Stationarity and invertibility.** This is to make sure that all the roots in $\varphi(B) = 0$ and $\theta(B) = 0$ are within the unit circle.
- **The hypothesis test on residuals.** When all the absolute values of the sample autocorrelations of the residual series are smaller than $1.96/\sqrt{n}$, the series are regarded as being fitted well enough. Another method is to use the statistic $Q = n \sum_{k=1}^{m} r_k^2$, where Q is asymptotically $\chi^2_{(k-p-q)}$ distribution. Here p and q are the orders of autoregression and moving average, while $n = N - d$, N is the length of the series and d is the order of difference.
- **The hypothesis test to overfitting.** All the redundant parameters are supposed to be excluded from the model although we need to increase the orders so as to reduce the sum of squared residuals.[2]

2.1.5. *Prediction*

With the model $\varphi(B) y_t = \theta(B) a_t$, we can get

$$y_t = \varphi^{-1}(B)\theta(B) a_t = \sum_{k=0}^{\infty} \psi_k a_{t-k}$$

and

$$y_{t+l} = \psi_0 a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + \sum_{j=0}^{\infty} \psi_{l+j} a_{t-j} \, .$$

The predicted value with the lead time l is

$$\hat{y}_t(l) = \sum_{j=0}^{\infty} \psi^*_{l+j} a_{t-j} \, .$$

The weight coefficients ψ^*_{l+j}'s can be calculated by the principle of least squared predicted values. The estimator of the predicted value is $\hat{y}_y(l) = \sum_{j=0}^{\infty} \psi_{l+j} a_{t-j}$. The variance is $E[a_t(l)]^2 = (\psi_0^2 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_a^2$.[2] Here $\hat{\sigma}_a^2 = \frac{\sum_{t=1}^{n} a_t^2}{n-p-q}$ is the variance of the white noise. The 95% confidence interval of the prediction is

$$\hat{y}_t(l) \pm 1.96 \left(1 + \sum_{j=1}^{l-1} \hat{\psi}_j^2\right)^{1/2} \hat{\sigma}_a \, .$$

2.1.6. *An example*

In the management of medical material (take X-ray film for example), demand exceeds supply and supply exceeds demand are both the situations that hospital wants to avoid. It is beneficial for hospital and patients to forecast the demand of X-ray film.

In this section, we present an example on predicting the seasonal demand of X-ray film in the Second Affiliated Hospital of Shanxi Medical University using the information of seasonal demand from 1987 to 1997. After transformation of Box-Cox with $\lambda = 0.192$, ARIMA model is used and the resulting model is ARIMA$(0, 1, 1)(0, 1, 1)s$.[27] The predicted values in the first and the second season are 212.01 and 274.61. The relative errors are -5.77% and -10.55%, respectively.

2.2. *The efficiency of prediction*

One of the purposes of modeling a time series is for predicting the future. The prediction variance increases along with the increase of the lead time l. The prediction precision declines gradually as l increases. If the prediction needs be performed with a high value l, it is suggested combine the ARIMA model with other techniques.[47]

2.2.1. *Prediction errors and the confidence interval of prediction*

The prediction error with lead time one can be defined as $e_{t-1}(1) = x_t - \hat{x}_{t-1}(1) = a_t$. At the time $t-1$, x_t is a random variable, so is the error.

The error is $e_{t-1}(1) \sim NID(0, \sigma_a^2)$. The conditional distribution of the observations for an AR(1) model is[2]

$$(x_t|x_{t-1}) \sim NID(\phi_1 x_{t-1}, \sigma_a^2)\,.$$

The confidence interval of the prediction with confident level 95% is $\hat{x}_{t-1}(1) \pm 1.96\sigma_a$, or $\phi_1 x_{t-1} \pm 1.96\sigma_a$.

2.2.2. *Correlation between prediction errors*

The prediction errors from a fitted time series model are correlated. Take an AR(1) model for example, the errors $e_t(2) = a_{t+2} + \phi_1 a_{t+1}$ and $e_{t+1}(2) = a_{t+3} + \phi_1 a_{t+2}$ are not independent. In fact, the covariance between them is,

$$\mathrm{Cov}[e_t(2), e_{t+1}(2)] = E(a_{t+2} + \phi_1 a_{t+1})(a_{t+3} + \phi_1 a_{t+2}) = \phi_1 \sigma_a^2\,,$$

and the correlation can be expressed as

$$\mathrm{Cov}[e_t(2), e_{t+1}(2)] = \frac{\mathrm{Cov}[e_t(2), e_{t+1}(2)]}{\{\mathrm{Var}[e_t(2)]\mathrm{Var}[e_{t+1}(2)]\}^{1/2}} = \frac{\phi_1 \sigma_a^2}{(1+\phi_1^2)\sigma_a^2} = \frac{\phi_1}{1+\phi_1^2}\,.$$

The prediction error to a ARMA(p, q) with lead time l is

$$e_t(l) = a_{t+l} + G_1 a_{t+l-1} + \cdots + G_{l-1} a_{t+1}\,,$$

where G is called Green function. For the prediction errors $e_t(l)$ and $e_{t+j}(l)$, we have covariance and correlation function as the following:

$$\mathrm{Cov}[e_t(l), e_{t+j}(l)] = E[(a_{t+l} + G_1 a_{t+l-1} + \cdots + G_{l-1} a_{t+1})\,.$$

$$\begin{aligned}
&(a_{t+l-j} + G_1 a_{t+l-j-1} + \cdots + G_{l-1} a_{t+j+1})] \\
&\quad = \sigma_a^2 (G_j + G_1 G_{j+1} + G_2 G_{j+2} + \cdots + G_{l-j-1} G_{l-1}) && (j < l) \\
&\quad = 0 && (j \geq l)\,.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}[e_t(l), e_{t+j}(l)] &= \frac{G_j + G_1 G_{j+1} + G_2 G_{j+2} + \cdots + G_{l-j-1} G_{l-1}}{1 + G_1^2 + G_2^2 + \cdots + G_{l-1}^2} && (j < 1) \\
&= 0 && (j \geq l)\,.
\end{aligned}$$

2.2.3. *Improve the prediction precision by indicator series*

Indicator series is helpful for the improvement of prediction although an ARMA model may fit a stationary time series with any needed precision theoretically.[2] However, take one or more correlated series into consideration for the model, the prediction efficiency may not be necessarily ameliorated.

For a given time series, the important foundations to establish the model are dynamic characteristic, memory or correlation. One may not be able to take into account all of the characteristics and that will lead to disappointed results. In this situation, the indicator series may provide help for modeling in some aspects. Firstly, indictor may supply some necessary supplementary information to the series especially when the prediction series is short. Another situation is that, the probability structure is changing along the time and the indicator may be useful to model the change.[48] Scientific knowledge in the subject area is needed to decide whether an indicator series should be included into the model or not.

2.3. *Combined predictions*

Combined predictions are the methods with which to perform prediction with the combination of several kinds of models so as to improve the prediction efficiency. As long as the combination is properly organized, the aim could be reached effectively.[49] All the combined models have some rational components in it and this is the foundation to expect the improvement of efficiency by combination of models.

2.3.1. *Unequal weights*

As the efficiency of a model may change at different segments of the time series, unequal weights are more reasonable comparing with the constant ones. If a model is poor at all segments, the model should be excluded. If a model is perfect all the time, it should be maintained as the only desired model. Combination becomes meaningless under these two situations. The unequal weights are coincided with the characters of the models.[49]

Comparing the combined predictions with unequal weights $\sum_{i=1}^{n} w_i(t)$ $\hat{y}_i(t)$ to those with constant weights $\sum_{i=1}^{n} w_i \hat{y}_i(t)$, we can construct objective functions for the purpose of minimizing the sum of squared errors. It sounds reasonable for the combined predictions with unequal weights to

reflect the dynamic correlation between observations and this makes the predictions with unequal weights adapt to more types of time series.

2.3.2. *Optimization of the weights*

If a vector of weights K_n can minimize the sum of squared errors, then K_n is called the optimized weights vector. Many Chinese statisticians have done a lot concerning optimized combination. Tang *et al.*[50] has paid much attention on how to optimize the weights.

Although we have optimized values for the unequal weights, it is difficult to calculate them. According to the definition of optimized combined prediction, the vector of optimal weights is an n-dimension column vector, in which only one element is 1 and other elements equal to 0. The location of 1 is uncertain. Most of the time the vector cannot be solved out since it is difficult for any single strategy of combination to summarize the properties of all kinds of methods. Under different circumstances, we may have to apply different combination to conform prediction precision.

3. Spectral Analysis

3.1. *Considerations on seasonality (periodicity, circadian rhythm) in time series*

Seasonality (periodicity) is a commonly observed phenomenon in time series and it is an important basis for us to establish models. In the frequency domain, seasonality can be identified by "the peaks in a periodogram located at certain particular frequencies". In the time domain, it shows as the regular cycles caused by seasonal factors (for example the climate, religious festivals, etc.). The features include external variables that cannot be controlled by artificial means but may be predicable to some extent. The consideration on seasonality is more significant in long-term predictions than in short-term predictions in some cases.[51] Fisher and Wallis indicated some main factors, such as external variables, residual innovation, some dynamic characteristics of the series and the annual projects, are the direct causations to seasonality.[52] The adjustments on these causations may counteract the influence of seasonality. In one medical research on the relations of mood with sunshine and temperature, many patients are detected to be influenced by these two factors after the seasonality has been taken into account.

Albertson and Avlen[53] compared the effects of different models when using seasonality. Their results showed that goodness of fit and prediction can be improved considerably if seasonality was modeled. The periodic autoregressive model is only good in short-term predictions. However, ARIMA models with dummy variables work fine in many situations. It is suggested introduce terms with bigger lags so as to capture the property with longer intervals of periodicity.[2]

US Census Bureau developed the X-11 model to process time series including periodicity of month or season.[27] The original series is denoted as a summation model or multiplication model. Take multiplication model ($O_t = S_t C_t D_t I_t$) for example, C_t is the term of trend, which also includes other long-term. S_t is the change that happens within a year and the value of S_t is constant or shift slowly in every year. The item D_t corresponds to trade date, which may locate at different positions in the calendar. The irregularity is denoted by I_t that is the residuals left and cannot be explained by S_t, C_t and D_t. C_t and O_t are at a similar quantitative scale. The three other terms may have values near 1.0 (or in the percentage form, 100.0). Eliminating the seasonal component will be helpful to reveal the difference between two months or two seasons. The adjusted series may show more clearly the importance of trend. In general, the application of this strategy has improved the effectiveness of seasonal adjustment.[54–56]

Hillmer and Tiao have discussed the seasonality in ARIMA models.[57] Let $Z_t = S_t + T_t + N_t$ be the summation-formed model, where S_t, T_t and N_t are seasonality, trend and random noise respectively. Assume that these three components belong to an ARIMA model themselves, $\phi_s(B)S_t = \eta_s(B)b_t, \phi_T(B)T_t = \eta_T(B)C_t$ and $\phi_N(B)N_t = \eta_N(B)d_t$. As to Z_t, we have $\varphi(B)Z_t = \theta(B)a_t$. Here the highest order in $\phi_s(B), \phi_T(B)$ and $\phi_N(B)$ is the same as the order of $\varphi(B), \theta(B)$ and σ_a^2 can be obtained by the following equation:

$$\frac{\theta(B)\theta(F)\sigma_a^2}{\varphi(B)\varphi(F)} = \frac{\eta_s(B)\eta_s(F)\sigma_b^2}{\phi_s(B)\phi_s(F)} + \frac{\eta_T(B)\eta_T(F)\sigma_c^2}{\phi_T(B)\phi_T(F)} + \frac{\eta_N(B)\eta_N(F)\sigma_d^2}{\phi_N(B)\phi_N(F)}.$$

The research of Burridge and Wallis[58] has showed that seasonality adjustment and Kalman filter may reserve and prolong the information about the difference of variances in all the seasons.[59] They have deduced out the calculation of variances for season-adjusted data. For predictions to seasonal series, Chen[60] studied the robustness of different models with the Monte Carlo method. Under the parsimonious principle, he found that Holt-Winters method (a model with consideration of trend and seasonality) and

ARIMA models are good enough in terms of robustness to most seasonal influences. If parsimony is not concerned, ARIMA models, classical regression models and structural component models are not robust. One of the hottest research areas in time series is to decompose trend and seasonality from the series. Smoothness Priors-State Space Model is put forward by Kitagawa and Gersch[61] to portray information on these two aspects. They used the statistic Q to measure the goodness of fit and the focus of attention was on residuals. The selection principles (e.g. AIC) focus on the abilities of the models.

3.2. *The basic concepts on spectral analysis*

The spectral analysis on stationary models is to infer the distribution functions according to the observed series, such as the estimation or hypothesis test to spectral density or characteristic peaks in the periodogram. The spectrum of stationary series is a description of its statistical characteristics. As to multiple time series, principal component analysis and canonical correlation can be used for detecting frequency components in the series. Window functions are needed to improve the characteristics of the estimation. The spectral density function $\hat{f}(\omega)$ is called the estimation with a spectral window.[24]

The squared amplitude I_i in the periodogram corresponds with the variance that the ith component contributes to the total variance. $G(r) = I(r)/\sum_{i=1}^{M} I_i$ is constructed to determine whether the rth biggest component is of statistical significance.[49]

The autocorrelation function in time domain and the spectrum in frequency domain are equivalent mathematically. They both are important foundation in time series modeling.

3.3. *The application to time series*

It is well known that a beam of sunlight can be decomposed into red, orange, yellow, green, blue and violet colors. A particular color is corresponding to a wave with particular frequency. The similar situations exist in time series. The vibrations in time series can be decomposed into sine (or cosine) waves with different frequencies and amplitudes. Studying the spectral characteristics in medical time series is helpful for revealing the nonrandom information that conceals in the series and benefits the effectiveness of fitting.[49]

3.3.1. *Time domain and frequency domain*

Time series $\{x_t\}$ can be regarded as the observed result of a dependent variable while the corresponding independent variable is time. Some functions such as autocorrelation, partial autocorrelation can be constructed in time domain in order to describe the series.[24] The analysis that only uses the functions whose corresponding dependent is time is called the analysis in time domain. With Fourier transformation, the dependent variable turns into frequency and the related analysis is performed in frequency domain.

For the stationary time series $\{x_t\}$,

$$\varphi(B)x_t = \theta(B)a_t$$

where

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p\,,$$
$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q\,,$$

we have the spectral density function as the following:

$$S_{\mathrm{ARMA}}(f) = \sigma_a^2 \left| \frac{1 - \sum_{K=1}^{q} \theta_K e^{-i2\pi Kf}}{1 - \sum_{K=1}^{p} \varphi_K e^{-i2\pi Kf}} \right|^2 \quad \left(-\frac{1}{2} \le f \le \frac{1}{2}\right), \tag{10}$$

3.3.2. *White noise*

For white noise, the spectral density is $s_a(f) = \sum_{K=-\infty}^{\infty} r_K e^{-i2\pi Kf} = \sigma_a^2$. The density function becomes a constant just like white light contains all kinds of light with equal amplitudes.

If a time series fits a model very well, the residuals will be white noise series. The goodness of fit becomes a hypothesis test on residuals.

The hypothesis test on white noise is to judge whether all the autocorrelations $\rho_K(a) = 0 (K \neq 0)$ in time domain. While in frequency domain, it is the test to judge whether $s_a(f)$ is a constant.[24]

$H_0 : \{a_t\}$, $t = 1, 2, \ldots, N$ is white noise.

When N is large enough, the M components (M is an integer, $M \le \frac{N}{4}$) are

$$\sqrt{N}\hat{\rho}_1(a), \sqrt{N}\hat{\rho}_2(a), \ldots, \sqrt{N}\hat{\rho}M(a)$$

distribute as $N(0,1)$ approximately under the null hypothesis of H_0. The test on the independency of $\{a_t\}$ becomes the test on whether the M estimates are distributed as $N(0,1)$ asymptotically. The statistic $Q = \sum(\sqrt{N}\hat{\rho}_j(a))^2 = N\sum_{j=1}^{M} \hat{\rho}_j^2(a)$ is a X^2 distribution with freedom of $(M - p - q)$. When we have $Q \le \chi^2_{\alpha(M-p-q)}$, we say the model is well fitted.[2]

3.4. *Spectral analysis for outpatient flow*

Spectral analysis has become an important skill in data process and systematic analysis in engineering and other related fields, especially after Cooley and Tukey introduced Fast Fourier Transform (FFT) in 1965 and the availability of the computing power. Here an example in hospital management is given to illustrate the application.

3.4.1. *Transformation from time domain to frequency domain*

With Fourier transformation, a time series can be transformed from time domain to frequency domain.[24]

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e_{-i\omega t} dt \,. \tag{11}$$

The inverse transformation can also be fulfilled with the following:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{i\omega t} d\omega \,, \tag{12}$$

which is called inverse Fourier transformation.

When the observations are obtained at discrete time, the definition of Fourier transformation is,

$$X_k = \frac{1}{N} \sum_{t=0}^{N-1} x_t e^{-i2\pi k/N} \,, \quad k = 0, 1, 2, \ldots \tag{13}$$

X_k is called the finite discrete Fourier transformation (DFT) and can be abbreviated as $X_k = \mathrm{DFT}[x_t]$. The inverse discrete Fourier transformation is,

$$x_t = \sum_{k=0}^{N-1} X_k e^{i2\pi k/N} \,. \tag{14}$$

Assume that the discrete series $\{x_s\}$, $s = 0, 1, 2, \ldots, N-1$ is sample from the continuous procedure $x(t)$ in $[-\frac{T}{2}, \frac{T}{2}]$ with equal intervals. Here N is the sample size, T is the length of sampling time and $\Delta = T/N$ is the length of interval for sampling. Fourier transformed form of x_s and $x(t)$ have the relation as the following:

$$X(\omega) = \lim_{\Delta \to 0} \lim_{T \to \infty} T K_k \,.$$

3.4.2. *Spectral density function and its estimation*

For the time series $\{x_t\} = 0, 1, 2, \ldots, N-1$ with sample size N, the power spectral function can be defined as[24]

$$P_N = \frac{1}{N} \sum_{t=0}^{N-1} x_t^2 \,.$$

When X_k is the Fourier transformation of x_t, the following can be proved:

$$P_N = \sum_{k=0}^{N-1} |X_k|^2 \,.$$

The spectral function

$$S_x^*(k) = T|X_k|^2 \,, \tag{15}$$

$$S_x^*(k) = T \cdot \mathrm{DET}\left[\frac{1}{N} \sum_{t=0}^{N-1} x_t \cdot x_{t+\tau}\right] = T \cdot \mathrm{DET}[C_\tau] \,. \tag{16}$$

The spectral function is an important statistical description of a stationary process. The estimation of spectral density function is based on the information from the observed values. Spectral windows are used to construct consistent estimate of the special density function of the time series.[62,63]

3.4.3. *An example*

The outpatient attendances with respiratory, gastrointestinal and cardiovascular diseases in the second affiliated hospital of Shanxi Medical University from May 1989 to December 1998 are studied. Bartlett Window is used to smooth the periodogram.

The spectral analysis was done with SAS6.12 software. The statistic of Kolmogorov-Smirnov is larger than $a\sqrt{1/(m-1)} = 1.36\sqrt{1/(116-1)} = 0.1268$ (m is the sample size). The hypothesis, that the original series is a white noise, is rejected.

The characteristics of Figs. 4 and 5 are similar. There are two peaks at 4 months and 12 months, respectively, although the heights are different. These two peaks correspond with two periodic components in the series and we have one more obvious peak at 24 months in respiratory patient flow. In these series, the peaks at 12 and 24 months are related to the annual periodicity corresponding to physiological regularities of human being, pathogenic microorganism and other pathogenic factors. In spring, pollen in the air is

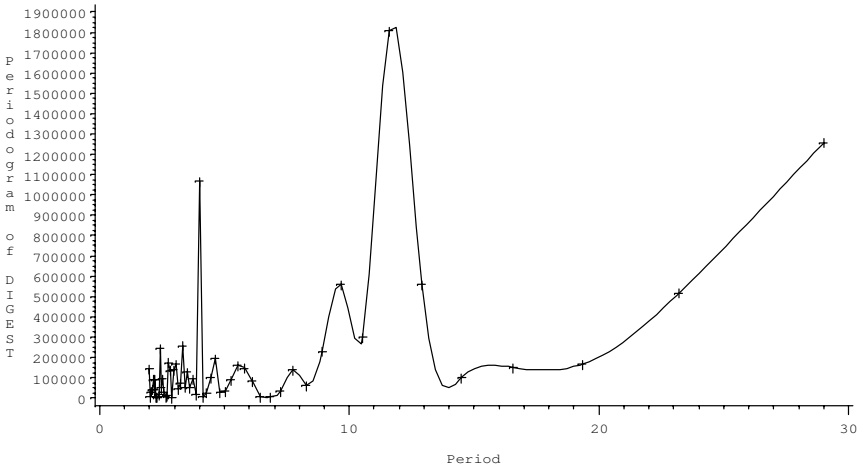


Fig. 4. Periodogram of the outpatient attendance of gastrointestinal disease in the second affiliated hospital of Shanxi Medical University.

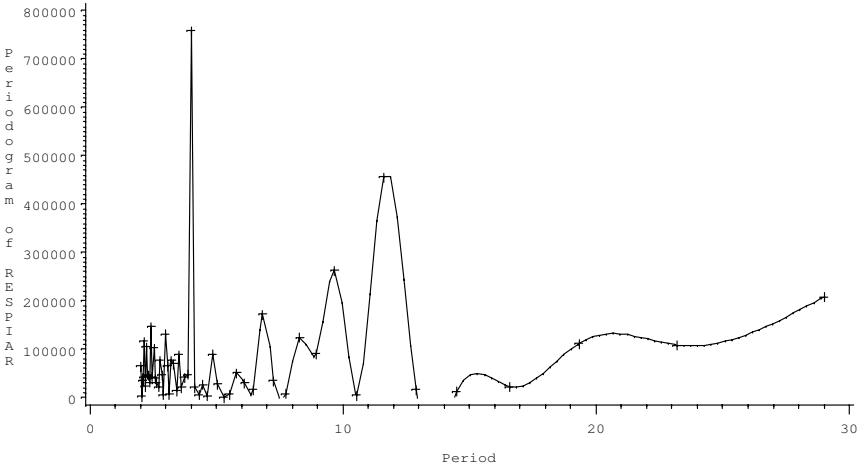


Fig. 5. Periodogram of the outpatient attendance of respiratory disease in the second affiliated hospital of Shanxi Medical University.

high; in autumn, artemisia plants are prosperous; in winter, the temperature is very low. These three important pathogenic factors cause outbreak in the interval of 4 months and they leads to the peak at four months in the periodogram. As to the periodogram for gastrointestinal patient flow there is a peak at four months. At the end of summer and beginning of autumn

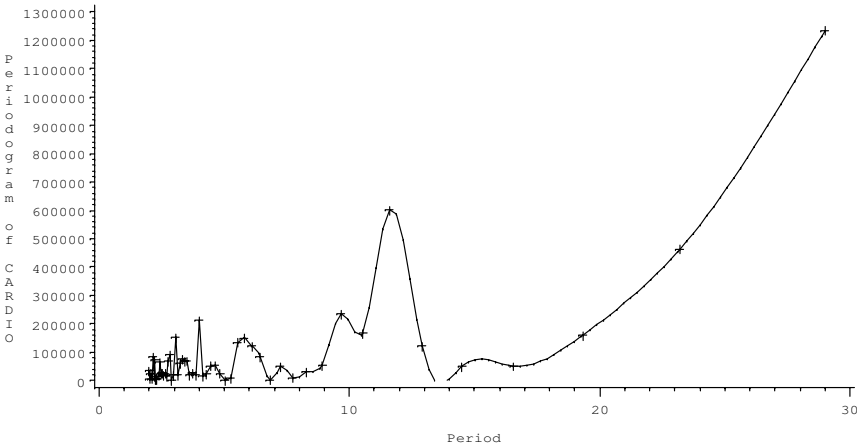


Fig. 6. Periodogram of the outpatient attendance of cardiovascular disease in the second affiliated hospital of Shanxi Medical University.

we have a high prevalence rate on enteritis because of much more raw food and cold drink. During the shifting time from winter to spring and from fall to winter, the prevalence rates of gastritis and gastric ulcer diseases turn to be high (further pathology explanation is needed). The three events above have nearly the interval of four months.

In the periodogram of cardiovascular patient attendance, there is a 12-months peak corresponding to seasonality. This may be caused by physiological regularities of human being and other pathogenic factors. The characteristics of periodograms corresponding to different factors are quite different.

3.5. *The identification of potential periodicity in time series*

It is important to identify potential periodicity in time series analysis in frequency domain.

3.5.1. *The mathematic model to describe the periodic components*

A time series $\{x_t\}$ that consists periodicity can be described as the following

$$x_t = \sum_{i=1}^{k} (a_i \cos 2\pi f_i t + b_i \sin 2\pi f_i t) + \varepsilon_t \,. \tag{17}$$

The estimations of $\hat{a}_i, \hat{b}_i, i = 1, 2, \ldots, k$ may be obtained by the least squared method. Set some particular frequencies $f_1, f_2, \ldots$ and construct the model as (17). Take frequency f_1 for example, the estimations of $\hat{a}_1$ and $\hat{b}_1$ will make $\hat{a}_1^2 + \hat{b}_1^2$ significantly larger than zero if f_1 locates near to a potential frequency. Otherwise, we say that they are not significantly different from zero and the corresponding frequency may not exist.

3.5.2. *Hypothesis test for the peaks in the periodogram*

Not all the peaks located in a periodogram are significantly different from zero. It is true that we may find some peaks in the periodogram even though the corresponding time domain series is a white noise due to the variability of the estimate. Hypothesis test is needed to statistically test the peaks that are caused by nonrandom variation.

The hypothesis test intends to test the amplitude $c_i = \sqrt{a_i^2 + b_i^2}$.

$$H_0 : c_i = 0\,, \quad i = 1, 2, \ldots, k\,.$$

Let $s = [\frac{N}{2}]$, I_j is the periodogram ordinate. The statistic for hypothesis test is constructed as

$$g = \frac{\max_{1 \le j \le s}(I_j)}{\sum_{j=1}^{s} I_j}\,, \tag{18}$$

which is called a Fisher statistic. Under the assumption of H_0 Fisher proved that the distribution of g is,

$$P[g > Z] = s(1 - Z)^{s-1} - \frac{s(s-1)}{2}(1 - 2Z)^{s-1} + \cdots + (-1)^s \frac{s!}{a!(s-a)!}(1 - aZ)^{s-1}\,. \tag{19}$$

Here, a is the maximum integer less than $1/Z$.

Fisher test deals with only the highest peak in the periodogram. Whittle[24] popularized it to the second highest peak. Let I_{j1} is the first highest peak, I_{j2} is the second highest peak and the corresponding statistic is

$$g_2 = \frac{I_{j2}}{(\sum_{j=1}^{s} I_j) - I_{j1}}\,.$$

The distribution of g_2 above is similar with that of (19). If the hypothesis test to I_{j2} is significant, we continue to test the next highest peak until all the significant peaks are detected.

3.5.3. *The cleavage of a peak*

A single peak may split into several nearly located peaks when the Fourier frequencies $2\pi f_i$ do not contain the non-Fourier frequency of the cyclic mechanism of the time series. It can be proved that periodogram is an unbiased estimation of power spectrum. However, it is not a consistent estimator of the spectral density function.

4. Nonlinear Model

4.1. *Threshold autoregressive model*

This model is brought forward by Tong[64] in 1978. The general formation is,

$$X_t = \varphi_0^{(j)} + \sum_{i=1}^{P_j} \varphi_i^{(j)} X_{t-i} + \varepsilon_t^{(j)} ,$$
$$\text{while } r_{j-1} \le X_{t-d} \le r_j \quad j = 1, 2, \ldots, k . \tag{20}$$

The set of r_i's $(-\infty = r_0 < r_1 < \cdots < r_k = \infty, r_j, j = 1, 2, \ldots, k-1)$ are called threshold values, d is called the parameter of delay. $\{\varepsilon_t^{(j)}\}$ is a white noise series that has a variance of σ_j^2. $\{\varepsilon_t^{(j)}\}$ and $\{\varepsilon_t^{(j')}\}$ are independent to each other when $j \neq j'$.

The Eq. (20) shows that the threshold values divide the axis $(-\infty, \infty)$ into k intervals. In every intervals, X_t is expressed by an autoregressive model with the order of p_j. Actually, the model is composed of k autoregressive models with different orders. It can be denoted as SETAR $(d, k, p_1, \ldots, p_k)$.

The most commonly used threshold autoregressive model is SETAR $(d, 2, p_1, \ldots, p_k)$. This model can be expressed as,

$$X_t = \begin{cases} \varphi_0^{(1)} + \varphi_1^{(1)} + \cdots + \varphi_{p_1}^{(1)} X_{t-p_1} + \varepsilon_t^{(1)} & X_{t-d} \le r_1 , \\ \varphi_0^{(2)} + \varphi_1^{(2)} + \cdots + \varphi_{p_2}^{(2)} X_{t-p_2} + \varepsilon_t^{(2)} & X_{t-d} > r_1 . \end{cases} \tag{21}$$

In medical research, the thresholds can be determined by professional knowledge. For example, in a relatively fixed population, the prevalence rate of tuberculosis is correlated with the average antibody level r_1. For the cases of the prevalence higher or lower than the critical value, the dynamics of prevalence are quite different. In this situation, a threshold autoregression model is applicable and the average antibody level r_1 is used as the threshold.

4.2. *Bilinear model*

This model is raised in economic field.[65] For example, the output in the tth year is x_t and it is used as the input of the next year. The rate of recovery y_t is an MA (1) model

$$y_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \varepsilon_t + \theta\varepsilon_{t-1}\,,$$

The output of this year is,

$$x_t = x_{t-1} + \varepsilon_t x_{t-1} + \theta\varepsilon_{t-1}x_{t-1}\,,$$

which is called a bilinear model.

A generalized form of it is

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + \sum_{i=0}^{q} \theta_i \varepsilon_{t-i} + \sum_{k=0}^{Q}\sum_{l=1}^{P} \beta_{kl}\varepsilon_{t-k}x_{t-l}\,.$$

Here, $\{\varepsilon_t\}$ is a white noise series. ε_t and ε_s are dependent random variables, with means are zero and variance is σ_s^2. (p, q, P, Q) are the orders of the model.

This model is a linear function of x_t when ε_t is given and is a linear function of ε_t when x_t is given. That is why it is called a bilinear model.

4.3. *Exponential autoregressive model*

It was put forward by Ozaki.[65] The generalized form of this model is,

$$x_t = \sum_{i=1}^{p} (\varphi_i + \psi_i e^{-rx_{t-1}^2})x_{t-i} + \varepsilon_t\,.$$

Here, $\varphi_i, \psi_i, r > 0$ are all constants and $\{\varepsilon_t\}$ is a white noise series.

This model is used to describe some medical time series where the amplitudes are closely correlated.

4.4. *State dependent model*

This model was raised by Priestley[65] in 1980. The generalized form is,

$$x_t = \mu(x_{t-1}) + \sum_{j=1}^{p} \varphi_j(x_{t-1})x_{t-j} + \sum_{i=1}^{q} \theta_i(x_{t-1})\varepsilon_{t-i} + \varepsilon_t\,,$$

where $x_{t-1} = (\varepsilon_{t-q}, \ldots, \varepsilon_{t-1}, x_{t-p}, \ldots, x_{t-l})^\tau$ and the model can be denoted as SDM(p, q).

SDM(p, q) is a widely used model.

(1) It becomes an ARMA(p, q) when $\mu(x_{t-1}), \varphi_j(x_{t-1})$ and $\theta_i(x_{t-1})$ are dependent on x_{t-1}.

(2) When $\theta_i(x_{t-1}) \equiv 0 (i = 1, 2, \ldots, q)$, $x_{t-d} \in R^{(i)}$, we have $\varphi_j(x_{t-1}) = \varphi_j^{(i)}, \mu(x_{t-1}) = \mu^{(i)}$, $j = 1, 2, \ldots, p$. $R^{(i)} = (r_{i-1}, r_i]$, $i = 1, 2, \ldots, l, -\infty = r_0 < r_1 < \cdots < r_{l-1} < r_l = \infty$. In this situation, the SDM(p, q) becomes a threshold autoregressive model.

$$x_t = \mu^{(i)} + \sum_{j=1}^{p} \varphi_j^{(j)} x_{t-j} + \varepsilon_t^{(j)}, \quad \text{while } x_{t-d} \in R^{(i)} \quad i = 1, 2, \ldots, l\,.$$

(3) When $\theta_i(x_{t-1}) = 0 (i = 1, 2, \ldots, q)$, $\mu(x_{t-1}) = 0$, $\varphi_j(x_{t-1}) = \varphi_j + \psi_j e^{-r x_{i-1}^2} (j = 1, 2, \ldots, p)$, the SDM$(p, q)$ becomes an exponential autoregressive model.

(4) When $\mu(x_{t-1}), \varphi_j(x_{-1}) (j = 1, 2, \ldots, p)$ are constants and $\theta_i(x_{t-1}) = \psi_j + \sum_{k=1}^{p} \beta_{jk} x_{t-k}$, $j = 1, 2, \ldots, \max(q, Q)$. Here p and Q are both positive integers. When $q < Q$, $\theta_j = 0$; when $q > Q$, $\beta_{jk} = 0$, $j = 1, 2, \ldots, \max(q, Q)$. SDM$(p, q)$ becomes to be a bilinear model.

5. Multivariable ARMA Model

To those complicated medical phenomena, multivariable time series analysis is useful.[65]

5.1. *The concept*

Let $X_t = (X_{1t}, X_{2t}, \ldots, X_{kt})^\tau$ is a k-dimension stationary time series with mean zero (to every $X_{it}, i = 1, 2, \ldots, k$). X_t satisfies,

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-p} \tag{22}$$

Here, $\varphi_i (i = 1, 2, \ldots, p)$, $\theta_j (j = 1, 2, \ldots, q)$ are all $k \times k$ matrix. $\varphi_p \neq 0$, $\theta_q \neq 0$, $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{kt})^\tau$ is k-dimension white noise series. That is,

$$\mathrm{E}\varepsilon_t = 0, \ \mathrm{E}\varepsilon_t \varepsilon_s^\tau = \begin{cases} S\,, & \text{when } t = s \\ 0\,, & \text{when } t \neq s\,. \end{cases}$$

Here S is a $k \times k$ positive definite matrix. $E\varepsilon_t X_{t-j}^\tau = 0$, $j = 1, 2, \ldots$.

The polynomial of operator B is also a $k \times k$ matrix.

$$\varphi(B) = I - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

$$\theta(B) = I - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \,.$$

Here, I is a $k \times k$ unit matrix. When all the roots of $\det(\varphi) = 0$ are located in the unit circle, the series is stationary. When all the roots of $\det(\theta) = 0$ are located in the unit circle, the series is invertible. The original series is called an ARMA(p, q) series. It is well known that any ARMA(p, q) series can be approximated by an AR(p) model.

5.2. *The Yule-Walker estimation of the parameters*

Assume that stationary $\{X_t\}$ is an AR(p) series with mean 0,

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = \varepsilon_t \,, \quad t = 1, 2, \ldots, n \,. \tag{23}$$

X_{t-h}^{τ} is multiplied at the both sides of the equation. After the expectation is performed, we have

$$\gamma_0 = \sum_{i,j=1}^{P} \varphi_j \gamma_{i-j} \varphi_i^{\tau} + S \,,$$

$$\gamma_h = \sum_{j=1}^{P} \varphi_j \gamma_{h-j} \,, \quad h = 1, 2, \ldots \,. \tag{24}$$

Here $\gamma_h, h = 1, 2, \ldots$ are all positive definite matrixes. Let $h = 1, 2, \ldots, p$. Because of $\gamma_{-h} = \gamma_h^{\tau}$, we have the linear equations as the following:

$$\begin{bmatrix} \gamma_1^{\tau} \\ \gamma_2^{\tau} \\ \cdots \\ \gamma_p^{\tau} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1^{\tau} & \gamma_0 & \cdots & \gamma_{p-2} \\ \cdots & \cdots & & \\ \gamma_{p-1}^{\tau} & \gamma_{p-2}^{\tau} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \varphi_1^{\tau} \\ \varphi_2^{\tau} \\ \cdots \\ \varphi_p^{\tau} \end{bmatrix} . \tag{25}$$

The estimation can be obtained when we substitute γ_h above with $\hat{\gamma}_h$. As to the $\hat{\gamma}_h$'s they can be derived by the following formula

$$\hat{\gamma}_h = \sum_{t=h+1}^{n} x_t x_{t-h}^{\tau} / n \,. \tag{26}$$

Let

$$\hat{\Gamma}_p = \begin{bmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \cdots & \hat{\gamma}_{p-1} \\ \hat{\gamma}_1^\tau & \hat{\gamma}_0 & \cdots & \hat{\gamma}_{p-2} \\ \cdots & \cdots & & \\ \hat{\gamma}_{p-1}^\tau & \hat{\gamma}_{p-2}^\tau & \cdots & \hat{\gamma}_0 \end{bmatrix}, \ \hat{\xi}_p = \begin{bmatrix} \hat{\gamma}_1^\tau \\ \hat{\gamma}_2^\tau \\ \cdots \\ \hat{\gamma}_p^\tau \end{bmatrix}, \ \Phi_p = \begin{bmatrix} \hat{\varphi}_1^\tau \\ \hat{\varphi}_2^\tau \\ \cdots \\ \hat{\varphi}_p^\tau \end{bmatrix} \tag{27}$$

be the estimation of (25) can be fulfilled by

$$\hat{\Phi}_p = \hat{\Gamma}_p^{-1} \xi_p \,, \tag{28}$$

which is called Yule-Walker estimation (or moment estimation).

In practice, we can use least square estimation, recursive algorithm to estimate the parameters.[65]

5.3. *Predictions and errors*

Assume that the AR(p) model is denoted as

$$X_t - \varphi_{p1} X_{t-1} - \cdots - \varphi_{pp} X_{t-p} = \varepsilon_t \,.$$

The prediction with the lead of one is $\hat{X}_{t-1}(1)$, that is

$$\hat{X}_{t-1}(1) = \varphi_{p1} X_{t-1} + \varphi_{p2} X_{t-2} + \cdots + \varphi_{pp} X_{t-p} \,,$$

and the prediction error is $\tilde{X} = X_t - \hat{X}_{t-1}(1) = e_t$. The variance matrix of the prediction is $\mathrm{E}\tilde{X}_t \tilde{X}_t^\tau = \mathrm{E} e_t e_t^\tau = S_p$ If the parameters are estimated by $\hat{\varphi}_{p1}, \hat{\varphi}_{p2}, \ldots, \hat{\varphi}_{pp}$, the prediction with lead of one is $\hat{x}_{t-1}(1) = \hat{\varphi}_{p1} X_{t-1} + \hat{\varphi}_{p2} X_{t-2} + \cdots + \hat{\varphi}_{pp} X_{t-p}$. The variance of the prediction error $D_p = \mathrm{E}\hat{X}_t \hat{X}_t^\tau$ can be obtained from

$$\hat{D}_p = \left(1 + \frac{kp}{n}\right) \left(1 - \frac{kp}{n}\right)^{-1} \left(\gamma_0 - \sum_{j=1}^{p} \hat{\varphi}_{pj} \hat{\gamma}_j^\tau\right). \tag{29}$$

6. Some Supplementary Topics

The research on time series has been driven by applications. There are many new development: (1) from linearity to nonlinearity; (2) apply the time series theory to the unbalanced models and establish dynamic unbalanced models to perform prediction; (3) combine Bayes theory with time series analysis to detect the changing point in dynamic data.

6.1. *Nonlinear time series model*

6.1.1. *The smooth state transform model*[65]

$$X_{t+1} = \alpha_1^\tau W_t + \varphi(r^\tau W_t)(\alpha_2^\tau W_t) + \varepsilon_t \,.$$

$\varphi(\cdot)$ is a monotonous limited function on (0, 1), $r^\tau W_t$ is the transforming variable, which is applied to fulfill the transform between two states.

The model above can be expressed as the following,

$$\begin{cases} \text{state 1}: & X_{t+1} = \alpha_1^\tau W_t + \varepsilon_t \,, & r^\tau W_t = 0 \quad \text{and} \quad \varphi(0) = 0 \\ \text{state 2}: & X_{t+1} = (\alpha_1^\tau + \alpha_2^\tau) W_t + \varepsilon_t \,, & r^\tau W_t = 1 \quad \text{and} \quad \varphi(0) = 1 \,. \end{cases}$$

6.1.2. *The model with time variable parameters*

$$X_{t+1} = \alpha_t W_t + \varepsilon_t \,. \tag{30}$$

Here, α_t is a parameter that change across time and it is not a function of $W_{t-j} (j \neq 0)$. For example $\alpha_t = \varphi\alpha_{t-1} + a_t$, while a_t is white noise and unit roots exist.

α_t can be regarded as a marginal cost parameters. Time variable parameters can be estimated by Kalman algorithm.[65]

6.1.3. *Projective pursuing model*

$$X_{t+1} = \alpha^\tau W_t + \sum_{j=1}^{q} r_j \varphi_j(\beta_j^\tau W_j + \theta_j) + \varepsilon_{t+1} \,. \tag{31}$$

Here, $\varphi_j(\cdot)$ is a smooth function and the estimation is performed first to those given values.

6.1.4. *The system of neural network*

$$X_{t+1} = \alpha^\tau W_t + \sum_{j=1}^{q} r_j \varphi(B_j^\tau W_t + \theta_j) + \varepsilon_{t+1} \,. \tag{32}$$

Here, $\varphi(\cdot)$ is a monotonic limited function e.g. $\varphi(z) = (1 + \exp(-z))^{-1}$.

The models (31) and (32) are both a kind of weighted projecting processes, which can be used in complicated medical time series. Even to outliers from the regular scatters, these models work fine.

6.1.5. *Product model*[65]

$$X_{t+1} = \alpha^{\tau} W_t + X_{t-j}\varepsilon_{t-k} + Y_{i,t-j}\varepsilon_{t-k}\,.$$

Here W_t contains the lagged X and lagged Y. The product item will lead to nonlinearity.

6.1.6. *Flexible Fourier model*[65]

$$X_{t+1} = \alpha^{\tau} W_t + \sum_{j=1}^{q} r_j \cos(\beta_j^{\tau} W_t + \theta_j) + \varepsilon_{t+1}\,.$$

The estimation to the parameters is relatively complicated. Some particular transformations are needed to produce a linear or nonlinear model. Then, the estimation will be meliorated.

6.2. *Vector autoregression model (VAR)*

Sims[65] put forward this model, as he doesn't deem the assumption is helpful that some variables be treated as external variables, nor the assumption that some parameters equal to zero reasonable.

Sims suggested that the same numbers of variables are needed in the structural equations in order to find all the possible interaction. The main difference of the classical viewpoints and Sims's idea is that whether a variable can be defined as a internal or external one.

Generally speaking, a VAR model has the following shortcomings: The model will depend on the transformation of data and the series is assumed to be stationary; there is a gap between the theory and practice; the simplified multinomial should be orthogonalized when it is used to prediction. As no feedback relation is included in the model, the influence with a delay cannot be described.

6.3. *Bayesian theory in medical time series prediction*

To the ordinary linear model $Y = XB + U$, all frequentist's methods are based on an assumption that all the estimated parameters are fixed. Unfortunately, it seems that this assumption does not hold all the time. In the view of Bayesian theory, the parameters B's are random. The posterior distribution is determined by prior knowledge. The change of model structure is detected and the theoretical hypothesis is tested. The most special point in Bayesian methods is that prior knowledge is used in prediction.

In recent years, a new idea is to apply time series theories into Bayesian models to predict the change points in the dynamic data with prior information.[65]

6.4. *Unbalanced time series model*

This model was used to predict the economy situation in Poland by Bowditch[65] in 1987. Because of the confinement from the unbalanced theory and the modeling skills, the application of this method keeps at a logjam. However, it is always a possible topic in economic prediction.

Some researchers begin to combine time series theory with unbalanced model and a new direction is formed.[65] This direction is paid much attention as it can explain the complex medical time series very well under some particular situation.

References

1. Jiang, Q. L. (1978). (translated by Fang, J. Q. into Chinese) *The Principle of Stochastic Process and the Related Models in Life Sciences*, Shanghai Publishing Company for Translated Books, Shanghai
2. Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1997) (translated into Chinese by Gu, L. *et al.*) *Time Series Analysis*: *Prediction and Control*, Chinese Publishing Hall of Statistics, Beijing.
3. Pandit, S. M. and Wu, X. M. (1998). *The Analysis and Some Applications of Time Series and the Related System*(in Chinese), The Publishing Hall for Mechanical Engineering, Beijing.
4. An, H. Z and Chen, M. (1998). *Nonlinear Time Series Analysis* (in Chinese), Shanghai Publishing Hall of Science and Technology, Shanghai.
5. Bell, W. R. and Hillmer, S. C. (1983). Modeling time series with calendar variation. *Journal American Statistical Association* **78**: 526–534.
6. Quenouille, M. H. (1957). *The Analysis of Multiple Time-Series*, Charles Griffin and Company Limited.
7. Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *Journal American Statistical Association* **76**: 802–816.
8. Chan, W.-Y. T. and Wallis, K. F. (1978). Multiple time series modeling: Another look at the mink-muskrat interaction. *Applied Statistics* **27**(2): 168–175.
9. Sung, K. A. (1997). Inference of vector autoregressive models with cointegration and scalar components. *Journal American Statistical Association* **92**: 350–356.
10. Stock, J. H. and Watson, M. W. (1998). Median unbiased estimation of coefficient variance in a time-varying parameter model. *Journal American Statistical Association* **93**: 349–358.

11. Ling, Shiqing and Li, W. K. (1997). On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *Journal American Statistical Association* **92**: 1184–1194.
12. Krämer, W. and Michels, S. (1997). Autocorrelation- and heteroskedasticity-consistent t-values with trending data. *Journal of Econometrics* **76**: 141–147.
13. West, K. D. (1997). Another heteroskedasticity- and autocorrelation-consistent covariance matrix estimator. *Journal of Econometrics* **76**: 171–191.
14. Gallant, A. R., Hsieh, D. and Tauchen, G. (1997). Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* **81**: 159–192.
15. Drost, F. C. and Klaassen, C. A. J. (1997). Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* **81**: 193–221.
16. Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics* **81**: 223–242.
17. Koop, G. and Ley, E. (1997). Bayesian analysis of long memory and persistence using ARFIMA models. *Journal of Econometrics* **76**: 149–169.
18. Breidt, F. J., Crato, N. and Lima, P. D. (1988). The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics* **83**: 325–348.
19. Box, G. E. P. and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal American Statistical Association* **65**: 1509–1526.
20. McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARIMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**(4): 269–273.
21. Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**(2): 297–303.
22. Poskitt, D. S. and Tremayne, A. R. (1980). Testing the specification of a fitted autoregressive-moving average model. *Biometrika* **67**(2): 359–363.
23. Szpiro, G. G. (1997). Noise in unspecified non-linear time series. *Journal of Econometrics* **78**: 229–255.
24. Chen, Z. G. (1988). *Time Series and its Spectral Analysis* (in Chinese), Chinese Publishing Hall of Science, Beijing.
25. Wang, Z. S. (2000). *Time Series Analysis* (in Chinese), Chinese Publishing Hall of Statistics, Beijing.
26. Feng, W. Q. (1994). *The Techniques for Economic Predictions and Decision-Making* (in Chinese), The Publishing Hall of Wuhan University, Wuhan.
27. Gao, H. X. (1998). *The User's Manual of SAS/ETS Software* (in Chinese), Chinese Publishing Hall of Statistics, Beijing.
28. Box, G. E. P. and Tiao, G. C. (1976). Comparison of forecast and actuality. *Applied Statistics* **25**(3): 195–200.
29. Clemens, M. P. and Hendry, D. F. (1998). Forecasting economic process. *International Journal of Forecasting* **14**: 111–131.
30. Granger, C. W. J., Drmerod, P. and Smith, R. (1998). Comments on forecasting economic process. *International Journal of Forecasting* **14**: 133–137.

31. Clements, M. P. and Hendry, D. F.(1998). Forecasting economic process — A reply. *International Journal of Forecasting* **14**: 139–143.
32. Webby, R. and O. Connor, M. (1996). Judgmental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting* **12**: 91–118.
33. Bewley, R. (1997). The forecast process and academic research. *International Journal of Forecasting* **13**: 433–437.
34. Herwartz, H. (1997). Performance of periodic error correction models in forecasting consumption data. *International Journal of Forecasting* **13**: 421–431.
35. Gooijer, J. G. D. and Franses, P. H. (1997). Forecasting and seasonality. *International Journal of Forecasting* **13**: 303–305.
36. Kulendran, N. and King, M. L. (1997). Forecasting international quarterly tourist flows using error-correction and time-series models. *International Journal of Forecasting* **13**: 319–327.
37. Shen, C.-H. (1996). Forecasting macroeconomic variables using data different periodicities. *International Journal of Forecasting* **12**: 269–282.
38. Wallis, K. F. and Whitley, J. D. (1991). Sources of error in forecasting and expectations. UK economic models 1984–1988. *Journal of Forecasting* **10**: 231–253.
39. Saligari, G. R. and Snyder, R. D. (1997). Trends Lead times and forecasting. *International Journal of Forecasting* **13**: 477–488.
40. Welch, E., Bretschneider, S. and Rohrbaugh, J. (1998). Accuracy of judgmental extrapolation of time series data characteristics: Causes and remediation strategies for forecasting. *International Journal of Forecasting* **14**: 95–110.
41. Veloce, W. (1996). An evaluation of the leading indicators for the Canadian economy using time series analysis. *International Journal of Forecasting* **12**: 403–416.
42. Burridge, P. and Wallis, K. F. (1998). Prediction theory for autoregressive-moving average processes. *Econometric Reviews* **7**(1): 65–95.
43. Swanson, N. R. and White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting* **13**: 439–461.
44. Shah, C. (1997). Model selection in univariate time series forecasting using discriminant analysis. *International Journal of Forecasting* **13**: 489–500.
45. Lim, J. S. and O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting* **12**: 139–153.
46. Harvey, N. and Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgmental forecasting. *International Journal of Forecasting* **12**: 119–137.
47. Pindyck, R. S. and Rubinfeld, D. L. (1998). *Econometric Models and Economic Forecasts*, The McGraw-Hill Companies, Inc.
48. Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis*, Springer-Verlag New York, Inc.
49. Gu, L. (1994). *Time Series Analysis and its Applications in Economic Sciences* (in Chinese), Chinese Publishing Hall of Statistics, Beijing.

50. Tang, X. W, Cao, C. X. and Jin, D. Y. (1994). The further research on the vector of the optimal weights. *Prediction* (in Chinese) **13**(2): 48–49.
51. Wallis, K. F. (1974). Seasonal adjustment and relations between variables. *Journal American Statistical Association* **69**: 18–31.
52. Albert, P. S. (1993). A model for seasonal changes in time series regression relationships with an application in psychiatry. *Statistics in Medicine* **12**: 1555–1568.
53. Albertson, K. and Aylen, J. (1996). Modeling the Great Lakes freeze: Forecasting and seasonality in the market for ferrous scrap. *International Journal of Forecasting* **12**: 345–359.
54. Wallis, K. F. (1995). Models for X-11 and X-11 forecast procedures for preliminary and revised seasonal adjustment. *Time Series Analysis Macroeconometric Modeling*.
55. Wallis, K. F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11 method. *Journal of Royal Statistical Society A* **145**(1): 74–85.
56. Burridge, P. and Wallis, K. F. (1984). Unobserved-components models for seasonal adjustment filters. *Journal of Bussiness and Economic Statistics* **2**(4): 350–359.
57. Hillmer, S. C. and Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal American Statistical Association* **77**: 63–70.
58. Burridge, P. and Wallis, K. F. (1990). Seasonal adjustment and Kalman ltering: Extension to periodic variances. *Journal of Forecasting* **9**: 109–118.
59. Burridge, P. and Wallis, K. F. (1985). Calculating the variance of seasonally adjusted series. *Journal American Statistical Association* **80**: 541–552.
60. Mary, E. W.-T., Atay, D. Ö. and Pollock, A. C. (1997). Currency forecasting: An investigation of extrapolative judgement. *International Journal of Forecasting* **13**: 509–526.
61. Kitagawa, G. and Gersch, W. (1984). A smoothness priors-state space modeling of time series with trend and seasonality. *Journal American Statistical Association* **79**: 378–389.
62. Gottlieb, D. and Orszag, S. A. (1977). *Numerical Analysis of Spectral Methods: Theory and Applications*, Society for Industrial and Applied Mathematics.
63. Stoica, P. and Moses, R. L. (1997). *Introduction to Spectral Analysis*, Prentice Hall, Inc. Simon and Schuster/A. Viacom Company.
64. Tong H. (1990). *Non-linear Time Series*, Oxford University Press.
65. Ma, S. C. and Gao, B. Q. (1997). *Economic Time Series Analysis* (in Chinese), The Publishing Hall of Liaoning University, Shenyang.

About the Author

Jinxin Zhang was born in Yuci Shanxi, the People's Republic of China in 1966. He obtained his BS in Biomedical Engineering (1989) from Tianjin

University, MS in Biostatistics (1997) from Shanxi Medical University, and PhD in Biostatistics (2000) from the Fourth Military Medical University. He did postdoctoral research at Sun Yat-sen University of Sciences from January 2001 to January 2003. More than 20 theses relating to his research on time series and multivariate analysis have been published.

This page intentionally left blank

CHAPTER 10

APPLICATIONS OF STATISTICAL METHODS IN MEDICAL IMAGING

JESSE S. JIN

School of Information Technologies,
The University of Sydney, NSW 2006, Australia
Tel: +61-2-9351-3766; jesse@it.usyd.edu.au

Medical images are two-dimensional stochastic signals. There are many common issues of stochastic signals such as noise removal, signal restoration, signal sampling, etc. There are also many special issues which are relevant to high dimensional signals only, such as segmentation, clustering, etc. This chapter discusses issues of medical imaging. In particular, we will discuss the application of statistical methods in this area.

1. Introduction

Medical imaging is a fast growing area with the richest source of information and variety of modalities such as Magnetic Resonance Imaging (MRI), X-ray Transmission Imaging (X-ray), Computerised Tomography (CT), ultrasound images (both 2D and 3D), Positron Emission Tomography (PET), Single-Photon Computed Tomography (SPECT), Magnetic Source Imaging (MSI), Electrical Source Imaging (ESI), X-ray Mammography (MG), Orthopantomograms (OPG), and many others.

MRI is one of the most powerful non-invasive techniques in diagnostic clinical medicine and biomedical research. The technique is an application of nuclear magnetic resonance (NMR), a well-known analytical method of chemistry, physics and molecular structural biology. MRI is primarily used as a technique for producing anatomical images, but MRI also gives information on the physical-chemical state of tissues, flow diffusion and motion information. Magnetic Resonance Spectroscopy (MRS) gives chemical/composition information. MRI has revolutionised imaging of the brain, spine and the musculoskeletal system. Superb soft tissue contrast and

spatial resolution have made MRI the investigation of choice in many neurologic and orthopaedic diseases.

X-rays are generated by the interaction of accelerated electrons with a target material (usually tungsten). X-rays are deflected and absorbed to different degrees by the various tissues and bones in the patient's body. The amount of absorption depends on the tissue composition. For example, dense bone matter will absorb many more X-rays than soft tissues, such as muscle, fat and blood. The amount of deflection depends on the density of electrons in the tissues. Tissues with high electron densities cause more X-ray scattering than those of lower density. Thus, since less photons reach the X-ray film after encountering bone or metal rather than tissue, the X-ray will look brighter for bone or metal.

CT became generally available in the mid 1970s and is considered one of the major technological advances of medical science. X-ray CT gives anatomical information on the positions of air, soft tissues, and bone. Three-dimensional imaging is achieved by rotating an X-ray emitter around the patient, and measuring the intensity of transmitted rays from different angles.

Ultrasound, as currently practiced in medicine, is a real-time tomographic imaging modality. Not only does it produce real-time tomograms of the position of reflecting surfaces (internal organs and structures), but it can be used to produce real-time images of tissue and blood motion.

The history of PET can be traced to the early 1950s, when workers in Boston first realized the medical imaging possibilities of a particular class of radioactive isotopes. Whereas most radioactive isotopes decay by release of a gamma ray and electrons, some decay by the release of a positron. A positron can be thought of as a positive electron. Widespread interest and an acceleration in PET technology was stimulated by development of reconstruction algorithms associated with X-ray CT and improvements in nuclear detector technologies. By the mid-1980s, PET had become a tool for medical diagnosis, for dynamic studies of human metabolism and for studies of brain activation.

PET has a million fold sensitivity advantage over other techniques used to study regional metabolism and neuroreceptor activity in the brain and other body tissues. In contrast, magnetic resonance has exquisite resolution for anatomic studies and for flow or angiographic studies. In addition, magnetic resonance spectroscopy has the unique attribute of evaluating chemical composition of tissue but in the millimolar range rather than the nanomolar range. Since the nanomolar range is the concentration range of

most receptor proteins in the body, positron emission tomography is ideal for this type of imaging. The major clinical applications of PET have been in cancer detection of the brain, breast, heart, lung and colorectal tumors. Another application is the evaluation of coronary artery disease by imaging the metabolism of heart muscle.

SPECT, like PET, acquires information on the concentration of radionuclides introduced to the patients body. SPECT dates from the early 1960s, when the idea of emission traverse section tomography was introduced by D. E. Kuhl and R. Q. Edwards prior to either PET, X-ray CT, or MRI.

Iron currents arising in the neurons of the heart and the brain produce magnetic fields outside the body. These fields can be measured by arrays of SQUID (Superconducting QUantum Interference Device), detectors that are placed on or near the head or chest. The recording of magnetic fields of the head is known as MagnetoEncephaloGraphy (MEG) while that of the heart is called MagnetoCardioGraphy (MCG). Magnetic Source Imaging (MSI) is the general term for the reconstruction of current sources in the heart or brain from the measurements of external magnetic fields.

Electrical source imaging (ESI) is an emerging technique for reconstructing electrical activity in the brain or heart from electric potentials measured on the scalp or torso. Standard ElectroEncephaloGraphic (EEG), ElectroCardioGraphic (ECG) and VectorCardioGraphic (VCG) techniques are limited in their ability to provide information on regional electrical activity or localize bioelectrical events within the brain and heart. Noninvasive ESI of the brain requires simultaneous electric potential recordings from 20 or more electrodes for the brain and 100 to 250 torso electrode sites to map the body surface potential from the heart.

X-ray mammography (MG) is an effective method to diagnose the breast cancer. A low dose X-ray screening mammograms are performed on a woman's breasts with no symptoms to detect breast cancer at an early stage. The practice can perform diagnostic mammography. Breast needle localisation prior to surgery can be performed to provide location information and fine tissue information.

Orthopantomograms (OPG) and lateral cephalograms are the latest techniques for dental or orthodontic assessment.

Medical images are 2D stochastic signals. There are many common issues of stochastic signals such as noise removal, signal restoration, signal sampling, etc. There are also many special issues which are relevant to high dimensional signals only, such as segmentation, clustering, etc. We will discuss issues of medical imaging. In particular, we will discuss the

image sampling and compression in Sec .2, filtering in Sec .3, segmentation in Sec .4 and registration in Sec .5. Finally, there is a conclusion.

2. Sampling and Compression Using Statistical Features of Images

Computer-based advanced medical imaging techniques such as Positron Emission Tomography (PET) have been playing a crucial and expanding role in modern medical research and diagnosis. However, these powerful techniques have being accompanied by the growing size of image data sets as well. For example, a routine dynamic PET study using the CTI 951 scanner usually acquires 31 cross-sectional image planes of 128×128 pixels each, at 20 to 30 time points. It results a 4D data set containing up to 11 million data points with approximately 22 Mbytes storage space. As the resolution of current PET imaging improves, the large volume of related data will further increase. It has therefore, prompted significant recent interest in developing efficient image compression techniques which can contribute to the current expansion in medical digitalization, image database management and telemedicine.

Taking advantage of domain specific physiological kinetic knowledge related to dynamic PET images and physiological tracer kinetic modeling, this paper presents a novel knowledge-based near-lossless data compression algorithm for dynamic PET images. The proposed compression algorithm consists of three stages: (a) compression in the temporal domain using optimal image sampling schedule design; (b) compression in the spatial domain through cluster analysis; and (c) index image compression using standard still image compression techniques. In this section, clinical human brain PET studies using the $[^{18}F]$ 2-*fluoro-deoxy-glucose* (FDG) tracer are presented to illustrate the proposed compression algorithm. The technique can be easily applied to other PET studies with different tracers. The conventional[22] and proposed techniques are implemented on clinical dynamic PET images. Empirical results are given to illustrate the compression performance and the image quality.

2.1. *Tracer kinetic modeling and functional imaging*

Tracer kinetic techniques with PET are widely applied to extract valuable information from dynamic processes in the body. This information is usually defined in terms of a mathematical model $u(\boldsymbol{t}|\boldsymbol{p})$, where $\boldsymbol{t} = 1, 2, \ldots, T$ and $\boldsymbol{p}$ are the model parameters. The parameters describe the delivery,

transport and biochemical transformation of the tracer. The driving function for the model is the plasma blood input function, which is often obtained from blood sampling.[22] Measurements acquired by PET define the tissue time activity curve (TAC), or output function, denoted $z_i(\boldsymbol{t})$, where $\boldsymbol{t} = 1, 2, \ldots, T$ are discrete sampling times of the measurements, and $i = 1, 2, \ldots, I$ corresponds to the ith pixel in the imaging region. The purpose of dynamic PET image analysis is to obtain tracer TACs and parameter estimates for each pixel in the imaging region. These parameters can then be used to define physiological parameters, such as the local cerebral metabolic rate of glucose (LCMRGlc).

The conventional method uses the complete set of acquired PET projection data. Through the parameter estimation on a pixel-by-pixel basis using certain rapid estimation algorithms,[16,22,36] functional images can be generated. In this section, the Patlak method[35,36] was used to generate the LCMRGlc functional images for the purpose of comparing the estimation accuracy of the original and compressed data.

2.2. *Sampling and compression in temporal and spatial domains*

The sampling and compression scheme using statistical features of tracer kinetics consists of three stages.[6]

2.2.1. *Stage* 1 : *Compression in the temporal domain using optimal image sampling schedule*

In dynamic PET studies, the reliability of temporal frames is directly influenced by the sampling schedules and duration used to acquire the data. The longer the duration and greater the radio-activity counts, the more reliable the temporal frames. However, in order to obtain quantitative information from the dynamic processes, a certain number of temporal frames are required. Recently, it has been shown that the minimum number of temporal frames required is equal to the number of model parameters to be estimated.[26] Based on this, an algorithm that automatically determines optimal image sampling schedule (OISS) and maximizes the information content of the acquired PET data was developed.[26] The algorithm utilizes the accumulated/integral PET measurements.

In the design of OISS, a new objective function based on the *Fisher Information Matrix*,[10] was proposed to limit the loss of dynamic information. This objective function was used to discriminate between different

experimental protocols and sampling schedules. OISS can be directly applied to acquisition of PET projection data. This reduces the number of temporal frames obtained and therefore, reduces data storage. Furthermore, as fewer temporal frames are reconstructed the computational burden posed by image reconstruction is reduced. Details of this algorithm can be found in Li *et al.*[26]

2.2.2. *Stage* 2 : *Compression in the spatial domain through cluster analysis*

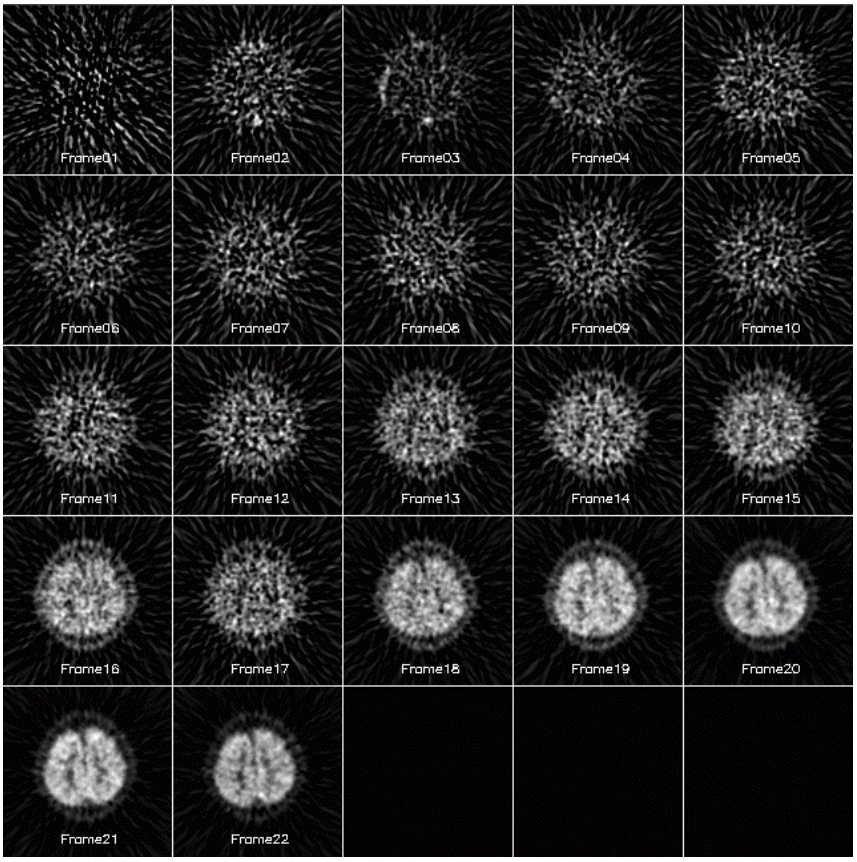
The prior knowledge has the form of tracer kinetic model to a time series of PET tracer uptake measurements. From the model, using cluster analysis, the image-wide TACs can be extracted and further classified into a certain numbers of TAC groups which corresponding to different tissue regions, according to the similarity of their kinetics.

Cluster analysis aims at grouping and classifying image-wide TACs, $z_i(t)$ (where $i = 1, 2, \ldots, I$), into C_j cluster groups (where $j = 1, 2, \ldots, J$ and $J \ll I$) by measuring the magnitude of natural association (similarity characteristics). It is expected that TACs with high degrees of natural association will belong to different groups.[8] It should be noted that each TAC must be assigned uniquely to a cluster group. In this paper, a hierarchical-agglomerative clustering algorithm based on the Euclidean distance measurement was used to classify the clinical dynamic PET image data.

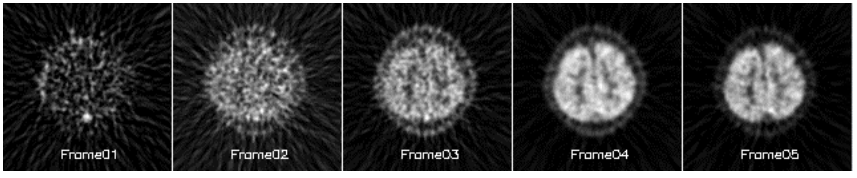
Using the results of cluster analysis, an index table containing the mean TAC within each cluster and an indexed image can be formed. The indexed image represents a mapping from the cluster to its respective pixel TAC locations. This image together with the index table forms the basis of the compressed temporal/spatial data. With PET, the number of distinguishable clustering groups may generally not exceed 64. This means that an 8-bit indexed image is sufficient to represent the cluster mapping.

2.2.3. *Stage* 3 : *Index image compression*

A lossless compression scheme is considered in this paper for further reduction of the indexed image. The PNG (Portable Network Graphics)[11] format was used to compress and store the indexed image obtained from cluster analysis. The coding technique presently defined and implemented for PNG is based on deflate/inflate compression with a 32-Kb sliding window. The PNG format was chosen over other lossless image compression file formats



(a)



(b)

Fig. 1. Compression results. (a) A set of 22 temporal-frame images (scaled) for the 15th plane from one patient study. (b) Results of the proposed compression method in temporal domain: 5 temporal-frame images (scaled), obtained from 1(a).

due to its portability, flexibility and being legally unencumbered. Details on the PNG format can be found in Crocker.[11]

Human dynamic FDG-PET brain studies were performed using an eight-ring, fifteen-slice PET scanner (GE/Scanditronix PC4096-15WB). This scanner contains 4096 detectors and achieves axial and trans-axial resolutions of 6.5-mm full width at half maximum (FWHM) at the center of the field of view. Between 200 and 400 mBq (approximately 0.5 mg) of FDG was injected intravenously and arterial blood sampling commenced immediately thereafter. The blood samples (each 2–3 ml) were taken at 8×0.25 minute intervals for the first 2 minutes, then at 2.5, 3, 3.5, 7, 10, 15, 20, 30, 60, 90 and 120 minutes. These samples were immediately placed on ice and the plasma was subsequently separated for the determination of plasma FDG and "cold" glucose concentration. Figure 1(a) shows a set of temporal frames for the 15th plane from one patient study. Due to the lower tracer concentration in the first few frames, these images were scaled to be visible.

3. Noise Reduction Using Statistical Anisotropic Diffusion

Diffusion processes have been widely used in quantum physics, material science, fluid dynamics, nuclear science, medicine and chemical physics. Perona and Malik[38,39] introduced it to image processing and proposed a multi-scale smoothing and edge detection scheme. It has the good property of eliminating noise while preserving high frequency components, namely edges.[2]

Diffusion is an iterative process. The degree of diffusion depends on the threshold of diffusion, i.e. the contrast cut-off. A contrast above the threshold will be enhanced during the diffusion process and that below the threshold will be smoothed out. The selection of the threshold is vital to the filtering process. However, the threshold varies from image to image. The problem compounds with the contrast variation from region to region and with intensity distortion of the same region in an image. It is thus desirable to have an adaptive criterion for selecting a threshold.

The threshold in a diffusion process is closely correlated with the contrast of the edges in an image. Selecting the threshold is a process of analysing local contrast. In low contrast images, especially when noise is present and the signal-noise ratio (SNR) is low, the contrast between regions is not significant and will be very difficult to pick up. The difficulty lies in the noise presence, unknown distribution of a stochastic signal,

and unknown combination of multiple interferences. In most of these cases, the histogram of the region shows a single peak. Many automatic threshold selection mechanisms require a bi-peak histogram such as Tsai and Chen[46] and Bhandari *et al.*[5] A bi-peak or multi-peak histogram may not exist in many cases. Luijendijk[29] proposed an automatic threshold selection using two histograms based on the count of 4-connected regions. Tseng and Huang[47] proposed to select the threshold using edge information, i.e. the intensity along edge intervals. Nagawa and Rosenfeld[33] fitted the histogram with two Gaussian functions, and Cho *et al.*[7] applied bias correction factors. Glaseby[18] combined them with an amendment using iteration. The assumption of Gaussian distribution is weak and correction does not make up this vital defect. Furthermore, iteration makes the computation very expensive.

Another difficulty is due to intensity distortion. The applicability of histogram analysis is based on the assumption that all image pixels which have a similar grey level correspond to one object or region of interest in the image. However, this assumption is not always true for most images. Rodriguez and Mitchell[41] used an adaptive thresholding method that extracts the background in two phases. The first step uses a global threshold to extract the structure of the regions and the second step refines the segmentation. Parker[34] used a local threshold to grow a region after finding a seed pixel in an object. Spann and Horne[44] grow regions from low resolution to high resolution in a quadtree structure. The adaptive scheme is a proper way to combat the distortion of intensity. However, the above mentioned methods have a try-and-error nature and do not have a solid theoretical foundation.

This section describes an adaptive diffusion scheme by applying the Central Limit Theorem. Regression is used to separate the distribution of the major object in a local window from other objects in a single-peak histogram. The separation will help to automatically determine the threshold. We have applied the algorithm to X-ray angiogram (XRA) images to extract brain arteries. The algorithm works well for single-peak distributions where there are no valleys in the histograms. It has also been used for filtering microscope images of kidneys where there are multiple visual objects and the contrast between objects is very low. The scheme shows that a fully automatic filtering process can be achieved. It works well with images which have texture patterns and are contaminated with noise while the distribution of noise is unknown. These kinds of images have posed a significant problem for traditional filtering schemes such as wavelet based de-noising.[13]

3.1. *Non-linear anisotropic diffusion*

Low-pass filters have been used to remove noise. Most filters are isotropic. Isotropic filtering tends to smear the corners and loses the accuracy of edges. To examine the problem carefully, we notice that the gradient along an edge is not isotropic. It has the highest value perpendicular to the edge and is dilated along the edge. It is therefore proper to increase the smoothing function parallel to the edge and stop the smoothing perpendicular to the edge. Non-linear anisotropic diffusion provides such a function. It takes the form

$$\frac{\partial}{\partial t} I(x, y, t) = \operatorname{div}(g(\nabla I) \nabla I)\,, \tag{1}$$

where $I(x, y, t)$ is the signal and $g(\nabla I)$ is a dilation function of gradients. There are two frequently used dilation functions:

$$g_1(x, y, t) = \frac{1}{1 + \frac{\nabla I(x,y,t)}{k}}\,, \tag{2}$$

$$g_2(x, y, t) = \exp\left\{ -\left(\frac{\nabla I(x, y, t)}{k} \right)^2 \right\}. \tag{3}$$

Calculation of diffusive filtering can be performed by a difference operation

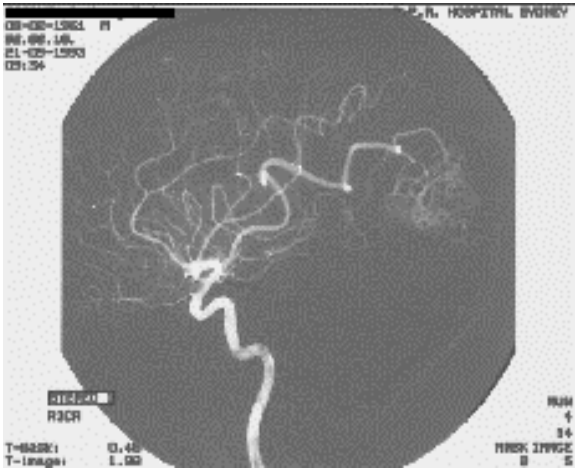
$$\begin{aligned}
\frac{\partial}{\partial t} I(x, y, t) &= \operatorname{div}[g(x, y, t) * \nabla I(x, y, t)] \\
&= \frac{\partial}{\partial t} \left[g(x, y, t) * \frac{\partial}{\partial x} I(x, y, t) \right] + \frac{\partial}{\partial y} \left[g(x, y, t) * \frac{\partial}{\partial y} I(x, y, t) \right] \\
&= g(x + 1, y, t)[I(x + 1, y, t) - I(x, y, t)] \\
&\quad + g(x, y, t)[I(x - 1, y, t) - I(x, y, t)] \\
&\quad + g(x, y + 1, t)[I(x, y, +1, t) - I(x, y, t)] \\
&\quad + g(x, y, t)[I(x, y, -1, t) - I(x, y, t)] \\
&= \Phi'_e + \Phi'_w + \Phi'_s + \Phi'_n\,.
\end{aligned} \tag{4}$$

Diffusion encourages intra-region smoothing in preference to smoothing across boundaries. The basis of this method is to suppress smoothing at boundaries by selecting locally adaptive diffusion strengths. The parameter κ plays an important role in diffusion. If the κ value is set to too high the filter will act as a smoothing filter, diffusing across the edge boundary; while if κ is too low, small dilation will result in many iterations. At some κ

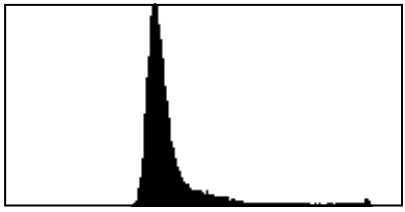
values, an extra edge will be introduced between the region of high intensity and region of low intensity. Therefore, the vital question in our design is the selection of κ.

3.2. *Selection of the cut-off contrast*

Images requiring processing often have very low contrast with many intensity layers. Determining an appropriate threshold for such images is difficult. Figure 2 shows an XRA image of the brain artery (a) and its histogram (b) which is a single peak histogram. The selection of a threshold value from such a histogram is ambiguous and not viable by trial and error. We have developed a region-based method to dynamically select a threshold using regression.



(a)



(b)

Fig. 2. Histogram analysis on background. (a) An XRA image; (b) Its histogram.

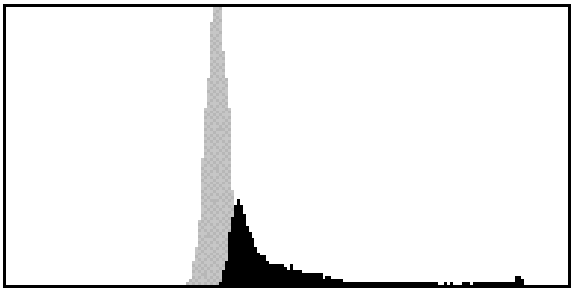


Fig. 3. Segmenting histogram using Gaussian regression.

3.2.1. *Selecting the threshold by regression and likelihood classification*

Our scheme is based on the Central Limit Theorem. It is difficult to segment brain arteries from the background because of the low contrast and an overwhelming proportion of the background. We do not know the histogram distribution of the background. However, from the Central Limit Theorem we know that if $x_1, x_2, \ldots, x_n$ are independent, identically distributed random variables with expectation μ and finite variance σ^2, then $y = \frac{1}{n}\sum_{i=1}^{n} x_i$ is asymptotically normal (μ, σ^2) when n is large enough.[42] Regression using a Gaussian distribution can separate the background histogram from the foreground histogram, as shown in Fig. 3, where shaded area shows the background histogram and the darker area is the foreground histogram. After separating the histogram, it is easy to select a threshold for image segmentation and to analyse foreground objects.

The sampling data for regression is obtained from partial histogram. We calculate the mean value of the histogram and take the half with less variance. Then we find the modal of that half histogram. The sampling data, $h_i, i \in S$, is on the same side with the modal against the mean value. The regression is obtained by

$$\begin{cases} \mu = \max\limits_{i \in s}(h_i) \\ \sigma = \sqrt{2\sum\limits_{i \in s}(h_i - \mu)^2}\,. \end{cases}$$

However, when the number of background pixels is not large enough, it is improper to use the Gaussian distribution in regression. Figure 4 shows another XRA image (a) and its histogram (b). Figure 4(c) is the histogram

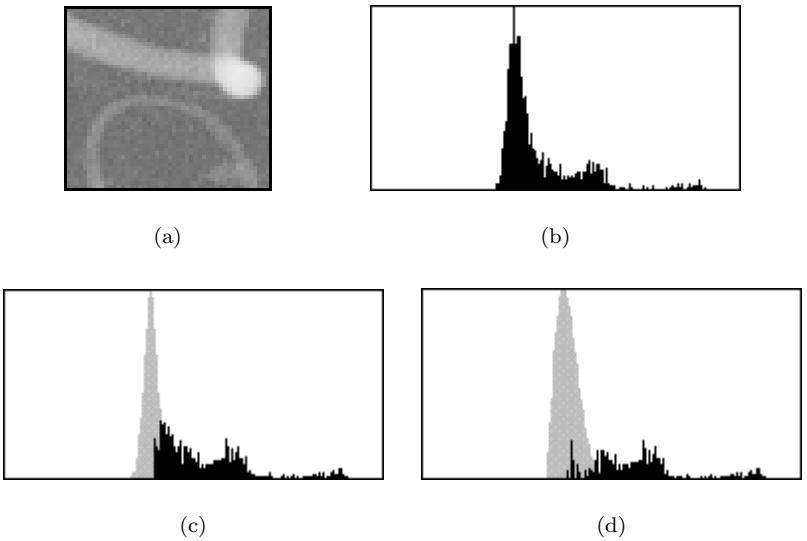


Fig. 4. Segmentation using regressions. (a) The original image; (b) Its histogram; (c) Gaussian regression does not show clear separation; (d) Rayleigh regression shows a clear separation.

after regression using Gaussian distribution over the background. It does not show a valley between two peaks as we expect, which means there is no clear separation. In this situation, we apply the Rayleigh distribution in regression. Probability theory states that when n is not large enough, $x = \bar{x}_n$ satisfies Rayleigh distribution:

$$f(x) = \begin{cases} \dfrac{x}{\mu^2} e^{\frac{x^2}{2\mu^2}} & x \geq 0 \\ 0 & x < 0\,. \end{cases}$$

The Rayleigh regression is obtained by

$$\begin{cases} \mu = \sqrt{\dfrac{\pi}{2}} \max\limits_{i \in s}(h_i) \\ \sigma = \sqrt{\dfrac{4-\pi}{2}\mu^2}\,. \end{cases}$$

3.2.2. *Extracting a cut-off contrast*

The diffusion process is critically depended on the κ value in functions (2) and (3). The parameter κ can be associated with the contrast. The following

discussion on extracting κ will be based on diffusion function (2) but it can be easily converted to function (3) by dividing κ by $\sqrt{2}$. Although we can obtain a proper estimation of the distribution of one visual object in the image, e.g. background in XRA images, we do not have information on other objects, e.g. vessels in XRA images. It is very difficult to estimate the average contrast between two objects. We use likelihood classification[45] to separate pixels from two objects after we separate the background histogram from the foreground histogram. These two histograms are used as the probability distributions of two clusters in likelihood classification. We calculate κ value from the following

$$\kappa = \left\{ l \left| \max_{l \in \{0..255\}} \left(\sum_{P \in p}^{N_l} \nabla_p I(x,y)/N_l \right) \right. \right\},$$

where N_l is the pixel number with gray level l, and P is a set of neighboring pixel pairs whose two pixels belong to different clusters. This calculation can be restricted to a local region. If two neighbor pixels belong to two clusters, we accumulate their difference into a difference histogram. The contrast can be extracted from the modal of differences within a local region.

4. Medical Imaging Segmentation

Segmentation is the process in which an image is divided into constituent objects or parts. It is often the first and most vital step in an image analysis task. Effective segmentation can usually dictate eventual success of the analysis. For this reason, many segmentation techniques have been developed by researchers worldwide.[19] Segmentation of intensity images usually involves four main approaches, namely thresholding, boundary detection, region-based and hybrid methods.

Thresholding techniques[43] are based on the postulate that all pixel whose value lie within a certain range belongs to one class. Such methods neglect all of the spatial information of the image and do not cope well with noise or blurring at boundaries.

Boundary-based methods are sometimes called edge-detection,[12] because they assume that pixel values change rapidly at the boundary between two regions. The basic method is to apply a gradient filter to the image. High values of this filter provide candidates for region boundaries, which must then be modified to produce closed curves representing the boundaries between regions.

Region-based segmentation algorithms postulate that neighbouring pixels within the same region have similar intensity values, of which the

split-and-merge[21] technique based on homogeneity criterion is probably the most well know. It includes seeded region growing[32] and unseeded region growing.

Hybrid methods combine one or more of the above-mentioned criteria. This class includes the morphological watershed[32] segmentation, variable-order surface fitting[4] and active contour[24] methods.

This section presents two methods among which statistical features are used in segmentation.

4.1. *Probilistical segmentation using expectation-maximization*

Intensity-based classification of MR images has proven problematic, even when advanced techniques are used. Intra-scan and inter-scan intensity inhomogeneities are a common source of difficulty. While reported methods have had some success in correcting intra-scan inhomogeneities, such methods require supervision for the individual scan. This section describes a new method called adaptive segmentation that uses knowledge of tissue intensity properties and intensity inhomogeneities to correct and segment MR images. Use of the EM algorithm leads to a method that allows for more accurate segmentation of tissue types as well as better visualization of MRI data, that has proven to be effective in a study that includes more than 1000 brain scans. Implementation and results are described for segmenting the brain in the following types of images: axial (dual-echo spin-echo), coronal (3DFT gradient-echo T1-weighted) all using a conventional head coil; and a sagittal section acquired using a surface coil. The accuracy of adaptive segmentation was found to be comparable with manual segmentation, and closer to manual segmentation than supervised multi-variate classification while segmenting gray and white matter.

Advanced applications that use the morphologic contents of MRI frequently require segmentation of the imaged volume into tissue types. Such tissue segmentation is often achieved by applying statistical classification methods to the signal intensities[25,49] in conjunction with morphological image processing operations.[9,17]

Conventional intensity-based classification of MR images has proven problematic, however, even when advanced techniques such as non-parametric, multi-channel methods are used. Intra-scan intensity inhomogeneities due to RF coils or acquisition sequences (e.g. susceptibility artifacts in gradient echo images) are a common source of difficulty. Although MRI images may appear visually uniform, such intra-scan

inhomogeneities often disturb intensity-based segmentation methods. In the ideal case, differentiation between white and gray matter in the brain should be easy since these tissue types exhibit distinct signal intensities. In practice, spatial intensity inhomogeneities are often of sufficient magnitude to cause the distributions of signal intensities associated with these tissue classes to overlap significantly. In addition, the operating conditions and status of the MR equipment frequently affect the observed intensities, causing significant inter-scan intensity inhomogeneities that often necessitate manual training on a per-scan basis.

Intra- and inter-scan MRI intensity inhomogeneities is modeled with a spatially-varying factor called the *gain field* that multiplies the intensity data. The application of a logarithmic transformation to the intensities allows the artifact to be modeled as an additive *bias* field. If the gain field is known, then it is relatively easy to estimate tissue class by applying a conventional intensity-based segmenter to the corrected data. Similarly, if the tissue classes are known, then it is straightforward to estimate the gain field by comparing predicted intensities and observed intensities. It may be problematic, however, to determine either the gain or the tissue type without knowledge of the other. It will be shown that it is possible to estimate both using an iterative algorithm (that converges in five to ten iterations, typically).

A Bayesian approach is used to estimating the bias field that represents the gain artifact in log-transformed MR intensity data. First, a logarithmic transformation of the intensity data is computed as follows:

$$Y_i = g(X_i) = (\ln([X_i]_1), \ln([X_i]_2), \ldots, \ln([X_i]_m))^T , \tag{5}$$

where X_i is the observed MRI signal intensity at the ith voxel, and m is the dimension of the MRI signal.

Similar to other statistical approaches to intensity-based segmentation of MRI,[9,17] the distribution for observed values is modeled as a normal distribution (with the incorporation of an explicit bias field):

$$p(Y_i|\Gamma_i, \beta_i) = G_{\psi\Gamma_i}(Y_i - \mu(\Gamma_i) - (\beta_i) , \tag{6}$$

where

$$G_{\psi\Gamma_i}(x) = (2\pi)^{-\frac{m}{2}} |\psi_{\Gamma_i}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^T \psi_{\Gamma_i}^{-1} x\right)$$

is the m-dimensional Gaussian distribution with variance ψ_{Γ_i} and where

Y_i is the observed log-transformed intensities at the ith voxel;

Γ_i is the tissue class at the ith voxel;
$\mu(x)$ is the mean intensity for tissue class x;
ψ_x is the covariance matrix for tissue class x;
β_i is bias field at the ith voxel.

Here, Y_i, $\mu(x)$, and β_i are represented by m-dimensional column vectors, while ψ_x is represented by an $m \times m$ matrix. Note that the bias field has a separate value for each component of the log-intensity signal at each voxel. In words, (6) states that the probability of observing a particular image intensity, given knowledge of the tissue class and the bias field is given by a Gaussian distribution centered at the biased mean intensity for the class.

A stationary prior (before the image data is seen) probability distribution on tissue class is used, it is denoted as $p(\Gamma_i)$.

If this probability is uniform over tissue classes, our method devolves to a maximum-likelihood approach to the tissue classification component. A spatially-varying prior probability density on brain tissue class has been studies.[23] Such a model might profitably be used within this framework.

The entire bias field is denoted by $\beta = (\beta_0, \beta_1, \ldots, \beta_{n-1})^T$, where n is the number of voxels of data. The bias field is modeled by a n-dimensional zero mean Gaussian prior probability density. This model allows us to capture the smoothness that is apparent in these inhomogeneities:

$$p(\beta) = G_{\psi_\beta}(\beta)\,, \tag{7}$$

where

$$G_{\psi_\beta}(\beta) = (2\pi)^{-\frac{n}{2}} |\psi_{\beta_i}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^T \psi_{\beta_i}^{-1} x\right)$$

is the n-dimensional Gaussian distribution. The $n \times n$ covariance matrix for the entire bias field is denoted ψ_β. Although ψ_β will be too large to manipulate directly in practice, tractable estimators can result when ψ_β is chosen so that it is banded.

It is assumed that the bias field and the tissue classes are statistically independent, this follows if the intensity inhomogeneities originate in the equipment. Using the definition of conditional probability the joint probability on intensity and tissue class can be obtained as follows:

$$p(Y_i, \Gamma_I | \beta_i) = p(Y_i | \Gamma_i, \beta_i) p(\Gamma_i)\,, \tag{8}$$

and we may obtain the conditional probability of intensity alone by computing a marginal over tissue class:

$$p(Y_i|\beta_i) = \sum_{\Gamma_i} p(Y_i, \Gamma_i|\beta_i) = \sum_{\Gamma_i} p(Y_i|\Gamma_i, \beta_i) p(\Gamma_i)\,. \tag{9}$$

This expression may be written more compactly as

$$\left[\sum_j W_{ij}[\psi_j^{-1}(Y_i - \mu_j - \beta_i)]_k + \frac{\frac{\partial}{\partial[\beta_i]_k}p(\beta)}{p(\beta)}\right]_{\beta=\hat{\beta}} = 0 \quad \forall\, i, \kappa \tag{10}$$

with the following definition of W_{ij}, (which are called the weights),

$$W_{ij} \equiv \frac{\lfloor p(\Gamma_i)G_{\psi_{\Gamma_i}}(Y_i - \mu(\Gamma_i) - \beta_i\rfloor_{\Gamma_i=\text{tissuee-class-}j}}{\sum_{\Gamma_i} p(\Gamma_i)G_{\psi_{\Gamma_i}}(Y_i - \mu(\Gamma_i) - \beta_i)}\,. \tag{11}$$

where subscripts i and j refer to voxel index and tissue class respectively, and defining

$$\mu_j \equiv \mu(\text{tissue-class-}j)$$

as the mean intensity of tissue class j. The mean residual is defined as

$$\bar{R}_i \equiv \sum_j W_{ij}\psi_j^{-1}(Y_i - \mu_j)\,, \tag{12}$$

and the mean inverse covariance is

$$\overline{\psi^{-1}}_{ik} \equiv \begin{cases} \sum_j W_{ij}\psi_j^{-1} & \text{if } j = \kappa \\ 0 & \text{otherwise}\,. \end{cases} \tag{13}$$

The result of the statistical modeling in this section has been to formulate the problem of estimating the bias field as a non-linear optimization problem embodied in

$$\bar{R} - \overline{\psi^{-1}}\hat{\beta} - \psi_\beta^{-1}\hat{\beta} = 0$$

or

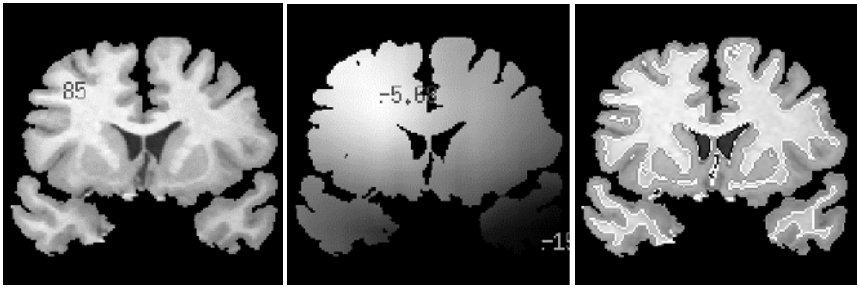
$$\hat{\beta} \equiv (\overline{\psi^{-1}} + \psi_\beta^{-1})^{-1}\bar{R}\,. \tag{14}$$

This optimization depends on the mean residual of observed intensities and the mean intensity of each tissue class, and on the mean covariance of the tissue class intensities and the covariance of the bias field.

The expectation-maximization (EM) algorithm is used to obtain bias field estimates from the non-linear estimator of (10). The EM algorithm iteratively alternates evaluations of the expressions appearing in models (11) and (14),

$$W_{ij} \leftarrow \frac{\lfloor p(\Gamma_i)G_{\psi_{\Gamma_i}}(Y_i - \mu(\Gamma_i) - \beta_i)\rfloor_{\Gamma_i=\text{tissue-class-}j}}{\sum_{\Gamma_i} P(\Gamma_i)G_{\psi_{\Gamma_i}}(Y_i - \mu(\Gamma_i) - \beta_i)}\,, \tag{15}$$

$$\hat{\beta} \leftarrow (\overline{\psi^{-1}} + \psi_\beta^{-1})^{-1}\bar{R}\,. \tag{16}$$



(a) (b) (c)

Fig. 5. Segmentation using expectation-maximization. (a) Original MRI brain slide; (b) Bias fied estimation; (c) Segmentation result.

In other words, model (15) is used to estimate the weights given an estimated bias field, then model (16) is used to estimate the bias, given estimates of the weights.

The adaptive segmentation can be applied to spin-echo and gradient-echo images. Examples are shown for the coronal (3DFT gradient-echo T1-weighted) images. All of the MR images shown in this section were obtained using a General Electric Signa 1.5 Tesla clinical MR imager [General Electric Medical Systems, Milwaukee, WI]. An anisotropic diffusion filter described in Sec. 3 was used as a pre-processing step to reduce noise.

Figure 5(a) shows the input image, a slice from a coronal 3DFT gradient-echo T1-weighted acquisition. The brain tissue ROI was generated manually. Figure 5(b) shows the final bias field estimate. The largest value of the input data was 85, while the difference between the largest and smallest values of the bias correction was about 10. Figure 5(c) shows the segmentation resulting from adaptive segmentation.

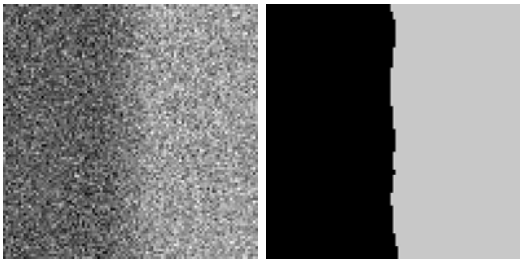
Note the significant improvement in the right temporal area. In the initial segmentation the white matter is completely absent in the binarization.

4.2. *Unseeded region growing*

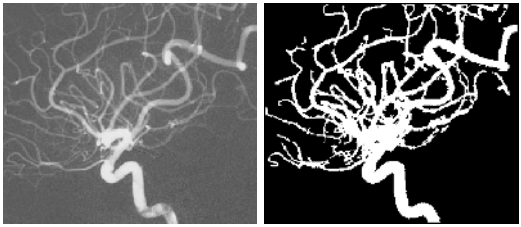
Unseeded region growing is similar to seeded region growing except that no explicit seed selection is necessary: the seeds can be generated by the segmentation procedure automatically. Therefore, this method can achieve fully automatic segmentation with the added benefit of robustness from being a region-based segmentation.

Formally, the segmentation process initializes with region A_1 containing a single image pixel, and the running state of the segmentation process consist of a set of identified regions, $A_1, A_2, \ldots, A_n$. Let T be the set of all unallocated pixels which borders at least one of these regions

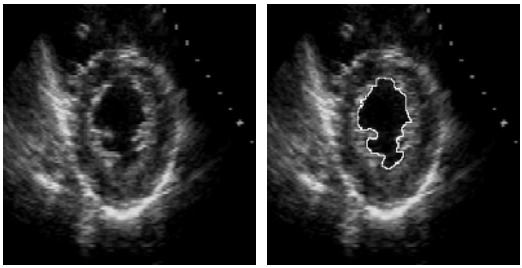
$$T = \left\{ x \notin \bigcup_{i=1}^{n} A_i \wedge \exists k : N(x \cap A_k) \neq \emptyset \right\},$$



(a)



(b)



(c)

Fig. 6. Segmenation using unseeded region growing. (a) Noisy image ($\sigma = 10.0$); (b) X-ray angiogram; (c) Ultrasound heart image.

where $N(x)$ are immediate neighboring pixels of point x. Further, we define a difference measure

$$\delta(x, A_i) = |g(x)\text{-mean}_{y \in A_i}[g(y)]|\,,$$

where $g(x)$ denotes the image value at point x, and i is an index of the region such that $N(x)$ intersect A_i.

The growing process involves selecting a point $z \in T$ and region A_j where $j \in [1, n]$ such that

$$\delta(x, A_i) = \min_{x \in T, \kappa \in [1,n]} \{\delta(x, A_i)\}\,.$$

If $\delta(z, A_j)$ is less than the predefined threshold t, then the pixel is added to A_j. Otherwise, we must choose the most substantially similar region $\boldsymbol{A}$ such that

$$\boldsymbol{A} = \arg\min_{A_k} \{\delta(x, A_k)\}\,.$$

If $\delta(z, \boldsymbol{A}) < t$, we can assign the pixel to $\boldsymbol{A}$. If neither of these two conditions above apply, then it is apparent that the pixel is significantly different from all the regions found so far, so a new region, A_{n+1} would be identified and initialized with pint z. In all three cases, the statistic of the assigned region must be updated once the pixel has been added to the region.

The URG segmentation procedure is inherently iterative, and the above process is repeated until all pixels have been allocated to a region. To ensure correct behavior with respect to the homogeneity criterion, the region growing operation requires the determination of the "best" pixel each time a region statistic is changed. The details of implementation can be found in Lin *et al.*[28] The segmentation results can be seen in Fig .6.

5. Improving Confidence Intervals of Image Registration Using 3-D Monte Carlo Simulations

Clinical diagnosis and treatment usually require registration of images with multiple modalities. Most of the medical image registration methods[30,31,48] minimize or maximize values of certain cost functions to achieve the global optimized match. These functions are usually the sum of squares of the distances between certain homogenous features in the two image sets to be registered. The sum of distances between homogenous point pairs of the two image sets,[15] distances between skin surfaces of CT, MR and PET images of the head in the "head-hat" method,[37] the absolute difference between pixel values of PET image and pixel values of image simulated by MR

image,[28] and the ratio between pixel values and their means in the same tissue class[3,51] are examples of these cost functions. However, most of these cost functions do not directly reflect the distance between the actual and estimated positions of targets, i.e. the target registration error (TRE). Most medical applications demand accuracy and precision assessment methods to justify their results. Internal consistency measures were used by Woods *et al.*[51] to place limits on registration accuracy for MRI data. Almost all other registration accuracy assessment methods fall into two broad categories: qualitative evaluations by visual inspection and quantitative evaluation by reference to results from a gold standard registration method. The former methods require special expertise and extensive experience, while the latter methods require an extremely accurate gold standard that cannot be easily achieved. Different methods may not always be comparable to each other under identical criteria.

Using the terminology of nonlinear regression analysis,[14] we can refer the problem of image registration as a nonlinear least sum of squares estimation of the transformation parameters that result in the optimal fitting of one set of image (function) to the other set of image (data). For least square estimation methods, the cost function could be assumed to be linear around the neighborhood of the current parameter values. So that we can calculate the confidence intervals or regions using the following equation[14]:

$$(\theta - \theta_0) \sum (f') \leq (\sigma^2 (n-1)) F(p, n-p, 1-\alpha)\,, \tag{17}$$

where F is a chosen F-test value of the corresponding confidence level, σ^2 is the residual sum of squares (registration cost function) value at the location of the estimated parameters, and $\sum(f')$ represents the sum of the derivatives of the reference model image to the transformation parameters. θ and θ_0 are the parameters corresponding to the confidence level and the optimal parameters found by the registration procedure, respectively.

Since all the data points involved in the calculation of (17) should be statistical independent to each other, and the data points in the images are correlated, the number of points in the image could not be used directly as n and the effective number of independent data points needs to be estimated.

To determine the effective number of independent data points involved in the estimation of confidence intervals, we first used one Monte Carlo simulation study based on normal conditions. The same number n selected according to this simulation results was found to be consistent for both the 95% and 90% confidence levels. We have further investigated the validity of the selected number n in various simulated conditions in other parts of the study.

Monte Carlo studies to simulate 2D PET images and subsequent registrations of the simulated images were conducted. The resulted distributions of the estimated transformation parameters were used to assess the consistency of 90%, 95% and 99% confidence intervals with the distributions in the parameter space. 2D grey matter and white matter sinograms of the segmented 2D Hoffman brain phantom[20] were combined with the grey-to-white ratios of 2:1, 3:1 and 4:1 before reconstruction to see whether the discrepancies of the ratios in two images can affect the confidence intervals. Then, filtered back-projection reconstruction programs with various filters (i.e. Hanning, Ramp, Butter-worth, Ham, Parzen and Shepp-Logan filters) were employed to reconstruct images of size 128×128. Various amounts of spatial displacements (i.e. rotations of 0.3, 0.8, 1.2 and 3.3 degrees, and translations of 0.16, 0.8, 1.6 and 2.4 mm) were introduced. Various levels of Poisson noise (i.e. total counts of 5×10^5, 1×10^6 and 2×10^6) were simulated. A Gaussian smoothing filter with a FWHM of 5 mm is applied to both sets of images before registration. The Powell's algorithm[40] was selected as the optimization procedure.

In the cases of extreme noise conditions and large contrast discrepancies, the residual sum of squares (RSS) consists of two parts: the systematic error and the error due to statistical noise:

$$RSS = RSS_{\text{system}} + RSS_{\text{noise}} \,. \tag{18}$$

The systematic error is contributed by the innate difference between the two images, inappropriate registration method, precision error of the program, etc. Such errors are independent of the initial displacements and noise. The second part of the residual sum of squares is due to statistical noise. If the systematic error is relatively large compared to the noise term, i.e. for cases with very low noise levels and high grey-to-white ratio discrepancies, the estimated residual sum of squares needs to be adjusted for systematic error.

Since the systematic component in RSS is much less sensitive to spatial smoothing than the other component in Eq. (18), it can be estimated by applying smoothing filters to both sets of images with relatively large FWHMs when the parameters are found. By removing the systematic component, the result RSS provides an estimation of the noise component in Eq. (18).

The calculated confidence intervals based on statistical regression are consistent with the simulation results for sample distributions of the transformation parameters of image co-registration. Varying the amount of displacement, reconstruction processes, noise levels, or tracer distributions have little impacts on the validity of the calculated confidence intervals.

After adjusted for systematic errors in the estimated residual sum of squares, confidence intervals can be calculated accurately even for very noisy conditions and with large distribution discrepancies between the two sets of images. Since multi-modality registration can be viewed as mono-modality registration of one image set with another simulated from the other image modality, this method is also expected to be applicable to multi-modality registration. Hence, visual inspection and validations by experts are not necessary for assessing the precision of the registration results. The results indicate the use of statistical confidence intervals has a potential to provide an automatic and objective assessment of individual image registration.

6. Conclusion

We have attempted a brief summary of the applications of statistical methods in image processing in general and medical imaging in particular. The issues cover image sampling, compression, filtering, segmentation and registration. Methods have been discussed in theory and illustrated in empirical results. Statistical methods are powerful tools in many signal processing applications. We hope this summary will provide an insight for the further use of statistical methods in image processing.

References

1. Adam, R. and Bischof, L. (1994). Seeded region growing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(6): 641–647.
2. Alvarze, L. and Mazorra, L. (1994). Signal and image restoration using shock filters and anisotropic diffusion. *SIAM Journal on Numerical Analysis* **31**(2): 590–594.
3. Ardekani, B. A., Braun, M. *et al.* (1995). A fully automatic multimodality image registration algorithm, *Journal Computer Assistant Tomography* **19**(4): 615–623.
4. Besl, P. J., and Jain, R. C. (1988). Segmentation through variable-order surface fitting, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **10**(2): 167–192.
5. Bhandari, D., Pal, N. R. and Majumder, D. D. (1992). Fuzzy divergence, probability measure of fuzzy events and image thresholding. *Pattern Recognition Letter* **13**: 857–867.
6. Cai, W., Feng, D. and Fulton, R. (1998). Clinical investigation of a knowledge-based data compression algorithm for dynamic neurologic FDG-PET images, *Proceedings of the 20th Annual International Conference of the IEEE*

Engineering in Medicine and Biology Society (EMBS'98), Vol. 20, part 3, 1270–1273, Hong Kong, October 29–November 1.

7. Cho, S., Haralick, M. R. and Yi, S. (1989). Improvement of Kittle and Illingworth's minimum error thresholding. *Pattern Recognition* **22**: 609–617.
8. Ciaccio, E. J., Dunn, S. M. and Akay, M. (1994). Biosignal pattern recognition and interpretation systems: Methods of classification, *IEEE Engineering in Medicine and Biology* **13**: 129–135.
9. Cline, H. E., Lorensen, W. E., Kikinis, R. and Jolesz, F. (1990). Three-dimensional segmentation of MR images of the head using probability and connectivity. *JCAT* **14**(6): 1037–1045.
10. Cobelli, C., Ruggeri, A., DiStefano, III, J. J. and Landaw, E. M. (1985). Optimal design of multioutput sampling schedules — software and applications to endocrine–metabolic and pharmacokinetic models, *IEEE Transactions on Biomedical Engineering* **32**(4): 249–256.
11. Crocker, L. D. (1995). PGN: The portable network graphic format. *Dr. Dobb's Journal*, 36-49, July.
12. Davis, S. L. (1975). A survey of edge detection techniques. *Computer Graphics Image Processing* **4**: 248–270.
13. Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transaction Information Theory* **IT-41**: 613–627.
14. Draper, N. R. (1981). *Applied regression Analysis*, 2nd edn., Wiley, New York.
15. Evans, A. C., Marrett, S., Collins, L. and Peters, T. M. (1989). Anatomical-functional correlative analysis of the human brain using three dimensional imaging systems. In *Medical Imaging*: *Image Processing*, eds. R.H. Schneider, S.J. Dwyer III AND R.G. Jost, SPIE Press, Bellingham, WA., **1092**: 264–274 vol. 1092, pp. 264-274.
16. Feng, D., Ho, D., Chen, K., Wu, L., Wang, J., Liu, R. and Yeh, S. (1995). An evaluation of the algorithms for constructing local cerebral metabolic rates of glucose tomographical maps using positron emission tomography dynamic date. *IEEE Transaction on Medical Imaging* **14**(4): 697–710.
17. Gerig, G., Kuoni, W., Kikinis, R. and Kubler, O. (1993). Medical imaging and computer vision: An integrated approach for diagnosis and planning. *Proceedings of the 11th DAGM Symposium*, 425–443.
18. Glaseby, C. A. (1985). An analysis of histogram based thresholding algorithms. *CVGIP*: *Graphical Models and Image Processing* **55**: 532–533.
19. Haralick, R. M. and Shapiro, L. G. (1985). Image segmentation techniques. *Computer Graphics Image Processing* **29**: 100–132.
20. Hoffman, E. J., Cutler, P. D., Guerrero, T. M., Digdy, W. M. and Mazziotta, J. C. (1991). Assessment of accuracy of PET utilizing a 3-D phantom to simulate the activity distribution of [18F] fluorodeoxyglucose uptake in the human brain. *Journal of Cerebral Blood Flow and Metabolism* **11**: 17–25.
21. Horowitz, S. L. and Pavlidis, T. (1974). Picture segmentation by a directed split-and-merge procedure. *Proceedings 2nd International Joint Conference On Pattern Recognition*, 424–433.

22. Huang, S. C., Phelps, M. E., Hoffman, E. J., Sideris, K., Selin, C. and Kuhl, D. E. (1980). Non-invasive determination of local cerebral metabolic rate of glucose in man. *American Journal of Physiology* **238**: E69–E82.
23. Kamber, M., Collins, D., Shinghal, R., Francis, G. and Evans, A. (1992). Model-based 3D segmentation of multiple sclerosis lesions in dual-echo MRI data. *SPIE Vol. 1808*, *Visualization in Biomedical Computing*.
24. Kass, M., Witkin, A. and Terzonpoulos, D. (1987). Snakes: Active contour models. *Proceedings International Conference On Computer Vision*, London.
25. Kohn, M., Tanna, N., Herman, G. *et al.* (1991). Analysis of brain and cerebrospinal fluid volumes with MR imaging. *Radiology* **178**: 115–122.
26. Li, X., Feng, D. and Chen, K. (1996). Optimal image sampling schedule: A new effective way to reduce dynamic image storage space and functional image processing time. *IEEE Transactions* **15**: 710–718.
27. Lin, J. Z., Jin, J. S. and Hugo, T. (2001). Unseeded region growing. *Proceedings Workshop on Visual Information Processing*, 2000, Sydney.
28. Lin, K. P., Huang, S. C. *et al.* (1994). A general technique for interstudy registration of multifunction and multimodality images. *IEEE Tran Nuclear Science* **41**(6): 2850–2855.
29. Luijendijk, H. (1991). Automatic threshold selection using histograms based on the count of 4-connected regions. *Pattern Recognition Letter* **12**: 219–228.
30. Maintz, J. B. A. and Viergever, M. A. (1998). A survey of medical image registration. *Medical Image Analysis* **2**(1): 1–36.
31. Maurer, C. R. and Fitzpatrick, J. M. (1993). A review of medical image registration, In *Interactive Imageguided Neurosurgery*, *American Association of Neurological Surgeons*, ed. R.J. Maciunas, Parkridge, IL, 17–44.
32. Meyer, F. and Beucher, S. (1979). Morphological segmentation. *Journal of Visual Communication And Image Representation* **1**: 21–46.
33. Nagawa, Y. and Rosenfeld, A. (1979). Some experiments on variable thresholding. *Pattern Recognition* **11**: 191–204.
34. Parker, J. R. (1991). Gray level thresholding in badly illuminated images. *IEEE Transaction PAMI* **13**: 813–819.
35. Patlak, C. S. and Blasberg, R. G. (1985). Graphical evaluation of blood to brain transfer constaints from multiple-time uptake data generalizations. *Journal of Cerebral Blood Flow and Metabolism* **5**: 584–590.
36. Patlak, C. S., Blasberg, R. G. and Fenstermacher, J. (1983). Graphical evaluation of blood to brain transfer constants from multiple-time uptake data. *Journal of Cerebral Blood Flow and Metabolism* **3**: 1–7.
37. Pelizzari, C. A., Chen, G. T. Y. *et al.* (1989). Accurate three-dimensional registration of CT, PET and/or MR images of the brain. *Journal Computer Assistant Tomography* **13**: 20–26.
38. Perona, P. and Malik, J. (1987). Scale-space and edge detection using anisotropic diffusion. *Proceedings IEEE Workshop Computer Vision*, Miami, FL, 16–22.
39. Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transaction PAMI* **12**: 629–639.

40. Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Juornal* **7**: 155–163.
41. Rodriguez, A. A. and Mitchell, O. R. (1991). Image segmentation by successive background extraction. *Pattern Recognition* **24**: 409–420.
42. Ross, S. M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*, John Wiley and Sons, NY.
43. Sahoo, P. K., Soltani, S. and Wong, A. K. C. (1988). A survey of threhsolding techniques. *Computer Graphics Image Processing* **41**: 230–260.
44. Spann, M. and Horne, C. (1989). Image segmentation using a dynamic thresholding pyramid. *Pattern Recognition* **22**: 719–732.
45. Tou, J. T. and Gonzalez, R. C. (1972). *Pattern Recognition Principle*, Addison-Wesley.
46. Tsai, D. M. and Chen, Y. (1992). A fast histogram-clustering approach for multi-level thresholding. *Pattern Recognition Letter* **13**: 245–252.
47. Tseng, D. C. and Huang, M. Y. (1993). Automatic thresholding based on human visual perception. *Image and Vision Computing* **11**: 539–548.
48. Van den Elsen, P. A., Pol, E. J. D. and Viergever, M. A. (1993). Medical image matching — A review with classification. *IEEE Engineering in Medicine and Biology* **12**: 26–39
49. Vannier, M., Butterfield, R., Jordan, D., Murphy, W. *et al.* (1985). Multispectral analysis of magnetic resonance images. *Radiology* **154**: 221–224.
50. Wells III., W. M., Grimson, W. E. L., Kikinis, R. and Jolesz, F. A. (1996). Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging* **15**(4): 429–443
51. Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R. and Mazziotta, J. C. (1998). Automated image registration: I. General methods and intrasubject, intramodality validation. *Journal Computer Assistant Tomography* **22**(1): 139–152.

About the Author

Jesse Jin graduated with a BEng from Shanghai Jiao Tong University and a PhD University of Otago, New Zealand. He is an Associate Professor and postgraduate coordinator in the Department of Computer Science, University of Sydney, and an Adjunct Associate Professor and Director of the Visual Information Processing Laboratiry in the School of Computer Science and Engineering, University of South Wales, Australia. Dr. Jin is an international renowed expert on multimedia technology and visual information retrival and processing. He has published 135 articles and 6 books. He also has one patent and is in the process of filing 3 more patents. He established a spin-off company and the company won the 1999 ATP Vice-Chancellor New Business Creation Award. He is a consultant of

many companies such as Motorola, Silicon Graphics, Computer Associates, ScanWorld, Proteome Syestems, HyperSoff, CyberView, etc. He was a visiting professor in MIT, UCLA, HKPU and Tsinghua University. He is also a Vice-President of Ausinan Science & Technology Society. His research interests include image processing and multimedia.

Section 2

Statistical Methods in Pharmaceutical Research

This page intentionally left blank

CHAPTER 11

STATISTICS IN PHARMACOLOGY AND PRE-CLINICAL STUDIES

TZE LEUNG LAI

Department of Statistics, Stanford University, Room 127, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA
Tel: 650-7232622; lait@stat.stanford.edu

MEI-CHIUNG SHIH

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02215, USA

GUANGRUI ZHU

Aventis Pharmaceuticals, P O Box 6800, Bridgewater, NJ 08807-0800, USA

1. Introduction

Pharmacology[1] as the science dealing with interactions between living systems and molecules, especially chemicals introduced from outside the system. This broad definition includes clinical pharmacology, whose objective is to prevent, diagnose and treat diseases with drugs, and the pathogenesis of diseases due to chemicals in the environment. A drug is defined in[1] as a small molecule that, when introduced into the body, alters the body's function. The component of a cell or organism that interacts with a drug and initiates the chain of biochemical events leading to the drug's therapeutic and toxic effects is called a *receptor*. The receptor concept has become the central focus of investigation of *pharmacodynamics* — the study of drug effects and their mechanisms of action. The relation between the dose of a drug and its clinically observed effects can be quite complex. In carefully controlled *in vitro* systems, however, the relation between the concentration of a drug at the site(s) of action and its effects can often be described by relatively simple mathematical models.

How a drug dose produces its effects involves not only pharmacodynamics but also *pharmacokinetics*. The latter is concerned with the concentration-time curve that is associated with the following "history" of a single adminstration of a drug:

(i) absorption phase of the drug into the body — transfer of the drug from its site of administration (via oral, or inhalational, or intravenous, or other route) into the bloodstream,
(ii) distribution phase — distribution of the drug to different compartments of the body, including receptor binding sites in the target tissue, and resulting in rapid decline in plasma concentration,
(iii) elimination phase — excretion of chemically unchanged drug or elimination via metabolism that converts the drug into one or more metabolites (e.g. at the liver).

Section 2 presents an overview of the basic principles, models and statistical methods in pharmacokinetics and pharmacodynamics. An active area of research in the field is pharmacometrics and Sec. 2 also gives some recent trends in this area. Particular attention will be directed to population pharmacokinetics and its interactions with several branches of modern statistics, including nonlinear mixed effects models, hierarchical and empirical Bayes methods, and generalized linear mixed effects models.

Section 2 also discusses the role of pharmacokinetic and pharmacodynamic studies in drug development. Specifically they are used to determine the *dosage regimen* of the drug (i.e. how much and how often it should be taken). These studies are initially performed *in vitro* and then on animals to come up with rough guesses of a region of dosage regimens in which clinical studies on human subjects are to be performed. The *in vitro* and animal studies are called *pre-clinical* and precede the clinical studies that are classified as Phase I studies (on healthy volunteers) and Phases II and III clinical trials (on patients).

Other statistical applications in pharmacology and pre-clinical studies include bioequivalence and bioavailability (treated in Sec. 3), assay development and validation (summarized in Sec. 4), drug discovery (reviewed in Sec. 5) and toxicology (treated in Chapter 13).

2. Pharmacokinetics and Pharmacodynamics

Drug administration can be divided into two phases, a pharmacokinetic (PK) phase in which the kinetics of drug absorption, distribution, and

elimination translate into drug concentration-time relationships in the body, and a pharmacodynamic (PD) phase in which the drug concentration at the site(s) of action leads to the response/effects produced. Knowledge of both phases is important for the design of a dosage regimen to achieve the therapeutic objective. Since both the desired response and toxicity of the drug are functions of the drug concentration at the site(s) of action, the therapeutic objective can be achieved only when the drug concentration lies within a "therapeutic window," outside which the therapy is either ineffective or has unacceptable toxicity. Drug concentrations, however, can rarely be measured directly at the sites of action and are typically measured at the plasma, which is a more accessible site. An optimal dosage regimen can therefore be defined as one that maintains the plasma concentration of a drug within the therapeutic window. This can be achieved for many drugs by giving an initial dose to yield a plasma concentration within the therapeutic window and then maintaining the concentration within this window by periodic doses to replace the drug lost over time.

2.1. *PK/PD models*

Many PK and PD models have been developed in clinical pharmacology. The monographs[1–5] give a comprehensive introduction to these models and their applications. The PK models can be roughly classified as "mechanistic" or "empirical," while mechanistic models can be classified as "physiologic" or "compartmental." In physiologic models, the body is viewed in physiologic terms, making use of *a priori* knowledge of physiology, anatomy and biochemistry. Although the tissues or organs differ from one another, they share many qualitative features. As an illustrative example, consider how anatomy affects elimination. First "clearance" CL is defined as the rate of elimination divided by the concentration of the drug. If the organs of elimination are in parallel, then CL is the sum of the CL_i over the elimination organs i. On the other hand, if the organs of elimination are in series (working sequentially one after another), then CL is proportional to $1 - \Pi(1 - E_i)$, where E_i is the extraction ratio of the drug at organ i. In particular, since the gut-liver system is in series for portal circulation whereas the portal and arterial systems into the liver are in parallel, it follows that

$$\mathrm{CL} = Q_H\{f_{HP}[1 - (1 - E_{\mathrm{gut}})(1 - E_{\mathrm{liver}})] + (1 - f_{HP})E_{\mathrm{gut}}\}, \qquad (1)$$

where f_{HP} is the fraction of total hepatic blood flow Q_H that enters the liver via the hepatic portal vein.

In compartmental models, the body is viewed in terms of kinetic compartments between which the drug distributes and from which elimination occurs. The kinetics is often described by a linear system of ordinary differential equations, which have explicit solutions involving exponential functions. On the other hand, the rate constants of a compartmental model may be functions of the concentration of the drug itself or another metabolite/interacting drug, leading to a system of nonlinear differential equations that have to be solved numerically. Empirical PK models are typically poly-exponential models of the form $\Sigma\alpha_i e^{-\lambda_i t}$. It is well known that different compartmental models may imply the same poly-exponential models, leading to identifiability difficulties with compartmental models in empirical work.[7,8]

A basic goal of PD models is to describe and quantify the steady-state relationship of drug concentration (C) at an effector site to the drug effect (E). The simplest PD model for one drug is the so-called "Emax model" defined by

$$E = e_{\max} C/(C + c_{50})\,, \tag{2}$$

where $e_{\max}$ is the maximum effect that the drug can produce and c_{50} is the concentration that yields 50% of $e_{\max}$. Note that this equation is the same as the Langmuir model in thermodynamics or the Michaelis–Mantern model in enzyme kinetics, in which the equilibrium state of ligand binding reactions is given by

$$B = \nu F/(\alpha + F)\,, \tag{3}$$

where B and F are the concentrations of the bound and free ligand, respectively, ν is the capacity of the binding site and $1/\alpha$ is the affinity constant. In fact, assuming that E is proportional to B, Eq. (2) follows from Eq. (3). A variant of Eq. (2) to incorporate the baseline effect e_0 is

$$E = e_0 + e_{\max} C/(C + c_{50})\,. \tag{4}$$

When the effect decreases response, $e_0 = e_{\max}$ and Eq. (4) has the form

$$E = e_0 - e_0 C/(C + c_{50}) = e_0 c_{50}/(C + c_{50})\,.$$

A convenient surrogate for the drug concentration at an effector site, which is difficult to measure directly, is dose (D). In empirical work, the Emax model is often reformulated as

$$E = e_0 + e_{\max} D/(D + \mathrm{ED}_{50})\,. \tag{5}$$

A more general form of Eq. (2) is

$$E = bX/(X + a)\,. \tag{6}$$

For $b = e_{\max}$, $a = 1$ and $X = (C/c_{50})^{\gamma}$ with $\gamma > 0$, Eq. (6) is called the "Sigmoid-Emax model." While the special case $\gamma = 1$ of such models reduces to Eq. (2), the inclusion of γ gives an additional adjustable parameter in fitting the model from data.

A general Emax model for two drugs, with concentrations C and C^*, incorporating both competitive and noncompetitive interactions is of the form

$$E = e_{\max}\left\{\frac{(C/c_{50}) + \alpha(C^*/c^*_{50}) + \beta(C/c_{50})(C^*/c^*_{50})}{1 + (C/c_{50}) + (C^*/c^*_{50}) + \delta(C/c_{50})(C^*/c^*_{50})}\right\}, \tag{7}$$

where $0 \le \alpha \le 1$, $0 \le \delta \le 1$, and $\beta \ge 0$ with $\beta = 0$ if $\delta = 0$. In particular, for $\delta = 1$ and $\beta = 1 + \alpha$, the right hand side of Eq. (7) can be written as a sum of $(C/c_{50})/\{1 + (C/c_{50})\}$ and $\alpha(C^*/c^*_{50})/\{1 + (C^*/c^*_{50})\}$, yielding additive effects of the two drugs. The case $\beta = \delta = 0$ gives a "competitive interaction model," which can be written as a linear combination of two terms of the form in Eq. (6) with $b = e_{\max}$ and $(X, a) = (C/c_{50}, 1 + C^*/c^*_{50})$ or $(C^*/c^*_{50}, 1 + C/c_{50})$. The case $\beta > \delta > 0$ shows synergism between the two drugs, while $\delta > \max(\beta, 0)$ shows antagonism. In particular, the case $\beta = 0$ and $\delta = 1$ gives a "non-competitive antagonism model," which can be written as a linear combination of two terms of the form in Eq. (6) with $b = e_{\max}$ and

$$(X, a) = (C/c_{50}, 1 + C^*/c^*_{50} + CC^*/c_{50}c^*_{50}) \quad \text{or}$$

$$(C^*/c^*_{50}, 1 + C/c_{50} + CC^*/c_{50}c^*_{50})\,.$$

Non-competitive antagonism can be explained by using receptor theory as follows. A drug interacts with two sites, one of which activates a receptor which may still interact with a second drug to form another non-activated receptor.

2.2. *PK parameters and their nonparametric estimates*

Several physiologic (e.g. maturation of organs in infants) and pathologic (e.g. kidney failure, heart failure) processes require dosage adjustments in individual patients to modify specific PK parameters. Two basic parameters in this connection are **clearance** (a measure of the ability of the body to eliminate the drug) and **volume of distribution** (a measure of the apparent space in the body available to contain the drug).

Drug clearance principles are similar to clearance concepts in renal physiology, in which creatinine or urea clearance is defined as the rate of elimination of the compound in the urine relative to the plasma concentration. Thus clearance CL of a drug is the rate of elimination by all routes relative to the concentration C of the drug in a biologic fluid:

$$\text{CL} = \text{Rate of elimination}/C\,. \tag{8}$$

The commonly used biologic fluid in Eq. (8) is plasma, for which CL is, strictly speaking, "plasma clearance." When C is C_b (blood concentration) or C_u (unbound or free drug concentration), then Eq. (8) gives "blood clearance" or "clearance based on unbound drug concentration," respectively. In healthy subjects, the clearance of amikacin is 91 ml/min, with 98% of the drug excreted in the urine unchanged. This means that the kidney is able to remove this drug from approximately 89 ml of plasma per minute. Propranolol is cleared at the rate of 840 ml/min, almost exclusively by the liver. This means that the liver is able to remove this drug from 840 ml of plasma per minute. For most drugs, clearance is constant over the plasma or blood concentration range in clinical settings, so the rate of elimination of the drug is proportional to its concentration C, in view of Eq. (8).

Clearance is perhaps the most important PK parameter to be considered in defining a rational drug dosage regimen. In most cases, the clinician would like to maintain steady-state drug concentrations C_{ss} within a known therapeutic window. Steady state will be achieved when the dosing rate (rate of active drug entering the systemic circulation) equals the rate of drug elimination. Therefore,

$$\text{Dosing rate} = \text{CL} \times C_{ss}\,. \tag{9}$$

The two major sites of drug elimination are the kidneys and the liver. Clearance of unchanged drug in the urine represents renal clearance. Within the liver, drug elimination occurs via biotransformation of the drug to one or more metabolites, or excretion of unchanged drug into the bile, or both. When no other organs are involved in elimination of the drug, $\text{CL} = \text{CL}_{\text{renal}} + \text{CL}_{\text{liver}}$ since the liver and kidneys work in parallel. The rate of elimination of a drug by a single organ can be defined in terms of the blood flow entering and exiting from the organ and the concentration of drug in the blood. The rate of presentation of the drug to the organ is the product of blood flow (Q) and entering drug concentration (C_i), while the rate of exit of drug from the organ is the product of blood flow and exiting

drug concentration (C_o). The difference between these rates at steady state is the rate of drug elimination:

$$\text{Rate of elimination} = Q \times C_i - Q \times C_o\,. \tag{10}$$

Dividing Eq. (10) by the concentration C_i of the drug entering the organ yields

$$\text{CL}_{\text{organ}} = \frac{Q \times C_i - Q \times C_o}{C_i} = Q \times \frac{C_i - C_o}{C_i}\,. \tag{11}$$

The expression $(C_i - C_o)/C_i$ is called the extraction ratio (ER) of the drug.

Bioavailability is the fraction of unchanged drug reaching the systemic circulation after its administration by any route. For an intravenous dose of the drug, bioavailability is 1. For a drug administered orally, bioavailability may be less than 1 since the drug may be incompletely absorbed, or metabolized in the gut, the portal blood or the liver prior to entry into the systemic circulation. If a drug is metabolized in the liver or excreted in bile, some of the active drug absorbed from the gastrointestinal tract will be inactivated by hepatic processes before the drug can reach the general circulation and be distributed to its sites of action. If the metabolizing or biliary excreting capacity of the liver is great, the so-called "first-pass effect" on the extent of availability will be substantial. The systemic bioavailability (F) of a drug that is completely absorbed and eliminated only by metabolism in the liver is given by

$$F = 1 - \text{ER}\,, \tag{12}$$

where $\text{ER} = \text{CL}_{\text{liver}}/Q_{\text{liver}}$ is the hypatic extraction ratio.

The **AUC** (area under the plasma or blood concentration-time curve) is a commonly used measure of the extent of absorption or availability of the drug absorbed in the body. It is usually calculated using the trapezoidal rule based on the blood or plasma concentrations obtained at various blood sampling times. Yeh and Kwan[8] considered spline and Lagrange interpolation schemes in lieu of the linear interpolation implied by the trapezoidal rule and compared these methods. Let $C_0, C_1, \ldots, C_k$ be the plasma or blood concentrations obtained at times $0, t_1, \ldots, t_k$, respectively. The AUC from time 0 to t_k, denoted by AUC_{0,t_k}, can be obtained via the trapezoidal rule as

$$\text{AUC}_{0,t_k} = \sum_{i=1}^{k} (t_i - t_{i-1})(C_i + C_{i-1})/2\,. \tag{13}$$

Typically t_k should be chosen so that C_k does not fall below the so-called "limit of quantitation" (LOQ) that will be defined in Sec. 4. In principle, the AUC should be calculated from 0 to ∞ (not just to the time of the last blood sample), and the portion of the remaining area from t_k to ∞ can be large. An estimate of AUC ($= \mathrm{AUC}_{0,\infty}$) is

$$\mathrm{AUC} = \mathrm{AUC}_{0,t_k} + C_k e^{-\lambda t_k}/\lambda\,, \tag{14}$$

where λ, called the *elimination rate constant*, is estimated from the elimination phase of the graph of log-concentration versus time by linear regression, assuming that it is linear so that λ corresponds to the slope of the fitted regression line; (see Ref. 2, Chapter 3 and Appendix A). The United States Food and Drug Administration (FDA) regulations require that sampling be continued through at least 3 half-lives of the active drug ingredient, measured in blood or urine, so that the remaining area beyond time t_k is only a small proportion of AUC_{0,t_k}.

The AUC also provides a simple relationship between the volume of distribution and dose. The volume of distribution (V) is defined as

$$V = \text{Amount of drug in body}/C\,, \tag{15}$$

where C is the concentration of the drug in blood or plasma, depending on the fluid measured. It reflects the apparent space available in both the general circulation and the tissue of distribution. It does not represent a real volume but should be regarded as the size of the pool of blood fluids that would be required if the drug were distributed equally throughout all parts of the body. From mass balance and steady state considerations, V is related to clearance via $\mathrm{CL} = \lambda V$, where λ is the elimination rate constant in Eq. (14). Moreover, $F \times \mathrm{Dose} = \mathrm{CL} \times \mathrm{AUC}$ (= total amount eliminated), where F is the systematic bioavailability in Eq. (12).[2] Hence,

$$V = \mathrm{CL}/\lambda = (F \times \mathrm{Dose})/(\lambda \times \mathrm{AUC})\,. \tag{16}$$

Besides CL, V, and AUC (measuring bioavailability), another PK variable, called the **elimination half-life** and denoted by $\mathrm{t}_{1/2}$, has to be considered when designing drug dosage regimens. It is given by

$$\mathrm{t}_{1/2} = (\ell n 2)/\lambda = 0.693\ V/\mathrm{CL} \tag{17}$$

and corresponds to the time taken for the concentration to drop to half of its initial level, assuming a one-compartment model for the drug's elimination phase in the body, as is usually done in designing drug dosage regimens.

In view of Eq. (17), $\mathrm{t}_{1/2}$ can be estimated by $(\ell n 2)/\hat{\lambda}$, where $\hat{\lambda}$ is an estimate of the elimination rate constant described after Eq. (14). When

$F = 1$, we can estimate CL by $\widehat{\text{CL}} = \text{Dose}/\text{AUC}$. Without assuming F to be 1, we have to replace Dose above by $\hat{F} \times \text{Dose}$, where $\hat{F}$ is an estimate of F. Once CL has been estimated, we can estimate V by $\widehat{\text{CL}}/\hat{\lambda}$. To estimate F, we need additional data following an intravenous dose D^*, yielding AUC^* and whose F^* can be assumed to be 1. Then F can be estimated from the original (extravascular) dose D and AUC by

$$\hat{F} = \min\left\{\frac{\text{AUC}/D}{\text{AUC}^*/D^*}, 1\right\}.$$

The above PK parameters are considered in a single dose trial. In practice, drugs are most commonly prescribed to be taken at fixed and equal time intervals, each of width τ. The maximum, minimum, and average concentration of the drug in steady state, denoted by $C_{ss,\max}$, $C_{ss,\min}$ and $C_{ss,av}$, respectively, are considered in conjunction with the steady-state volume of distribution and AUC during a dosing interval in steady state. See Chapter 7 of Rowland and Tozer,[2] which also shows how to develop a dosage regimen from knowledge of these PK parameters and the therapeutic window of a drug. Data obtained on multiple dosing can be used to estimate the PK parameters of a drug as follows. The most useful information derived from a multiple dosing study is the ratio of clearance to availability. It is obtained from

$$\frac{\text{CL}}{F} = \frac{(\text{Dose}/\tau)}{C_{ss,av}}, \tag{18}$$

where $C_{ss,av}$ is determined from the area under the plasma concentration-time curve within a dosing interval at steady state divided by τ. Occasionally, the drug is given as a multiple intravenous regimen, in which case the ratio $(\text{Dose}/\tau)/C_{ss,av}$ is simply clearance, since $F = 1$. The accuracy of the clearance estimate depends on the number of plasma concentrations measured in the dosing interval and on the ratio of $\tau/\text{t}_{1/2}$. The estimate can be improved by using several dosing intervals in steady state. Equation (18) is also useful for determining the relative availability of a drug administered extravascularly, between two treatments (e.g. dosage forms) A and B. Assuming that clearance remains unchanged, we have

$$\text{Relative availability} = \frac{(C_{ss,av})_B}{(C_{ss,av})_A} \cdot \frac{(\text{Dose}/\tau)_A}{(\text{Dose}/\tau)_B}. \tag{19}$$

2.3. *Parametric and population PK/PD models*

The nonparametric estimates of PK parameters described above assume that the blood (or urine) samples are collected frequently through at least 3 half-lives of the active ingredient, so that the curve between successive times t_k and t_{k+1} is well approximated by the line joining its values at these two points. When the experiment does not meet such conditions, the nonparametric estimates of AUC, CL and $t_{1/2}$ become unreliable and there are no satisfactory ways to evaluate the bias and standard error of such estimate. In this case it is preferable to use a parametric approach, based on the commonly used one-compartment model

$$y_j = \frac{Dk_a}{V(k_a - k_e)}(e^{-k_e t_j} - e^{-k_a t_j}) + \epsilon_j\,, \quad 1 \le j \le n\,, \tag{20}$$

in which y_j is the concentration at time t_j after the administration of a single oral dose D. Here V, k_a, k_e are the volume of distribution, absorption rate constant and elimination rate constant, respectively. Note that model (20) has the form of a bi-exponential model $\alpha_1 e^{-\lambda_1 t} + \alpha_2 e^{-\lambda_2 t}$ with $\alpha_1 = \alpha_2$.

Lai[7] gives a review of the literature on fitting the poly-exponential regression model $y_j = \beta + \sum_{k=1}^{k} \alpha_k e^{-\lambda_k t_j} + \epsilon_j$, in which the errors ϵ_j are assumed to be independent with zero means and

(i) $\mathrm{var}(\epsilon_j) = \sigma^2$ (constant variance error models), or
(ii) $\mathrm{var}(\epsilon_j) = f_\theta^2(t_j)\sigma^2$ (constant coefficient of variation error models), or
(iii) $\mathrm{var}(\epsilon_j) = f_\theta(t_j)\sigma^2$ (Poisson-type error models),

where $\theta = (\lambda_1, \ldots, \lambda_k; \alpha_1, \ldots, \alpha_k, \beta)$ and $f_\theta(t) = \beta + \sum \alpha_k e^{-\lambda_k t}$. We can estimate θ by weighted least squares, i.e. by minimizing

$$S(\theta) = \sum_{j=1}^{n} w_j [y_j - f_\theta(t_j)]^2\,. \tag{21}$$

For fixed $\lambda_1, \ldots, \lambda_k$, $f_\theta(t)$ is linear in the parameters $\beta, \alpha_1, \ldots, \alpha_k$ and standard formulas in multiple linear regression can be used to find least squares estimates of the linear parameters $\beta, \alpha_1, \ldots, \alpha_k$. This reduces the problem of minimizing $S(\theta)$ to that of minimizing

$$S^*(\lambda_1, \ldots, \lambda_k) = \min_{\beta, \alpha_1, \ldots, \alpha_k} S(\theta)\,.$$

In the case of the Poisson-type or constant coefficient of variation error model, the weights w_j also involve the unknown parameter θ and can be determined at each iteration from the previous iterate. It is shown in Lai[7] that S^* not only provides a relatively stable numerical algorithm for finding

the least squares estimates but also sheds light on the range of models that are compatible with the data. Depending on the experimental design, S^* can be very flat over a broad region containing the minimum or can decrease steeply to the minimum. It is also shown in Lai[7] that although the parameter vector θ may be poorly estimated because S^* is relatively flat, the function $f_\theta(\cdot)$ is typically well estimated by weighted least squares. Therefore derived parameters like AUC can still be well estimated from the estimated $f_{\hat{\theta}}$ even though $\hat{\theta}$ does not estimate θ well because of the experimental design.

Parametric modeling also facilitates the evaluation of standard errors and construction of confidence intervals. For the Emax model (2), which can be rewritten as $E/C = aE + b$ with $a = -1/c_{50}$ and $b = e_{\max}/c_{50}$, Scatchard[9] proposed to estimate a and b by linear regression of the observed E/C on C. This simple method is usually adequate for point estimation because of the large signal-to-noise ratio in the measurements. It is, however, unsatisfactory for constructing confidence intervals of the unknown parameters, as has been noted in the ligand-binding literature related to the mathematically equivalent model (3). Lai and Zhang[10] give a review of the literature and propose a new approach using nonlinear least squares and bootstrap methods to construct confidence regions for the parameters. The numerical studies reported in Lai and Zhang[10] show that these confidence regions are markedly different from the elliptical confidence regions based on asymptotic normal approximations.

So far we have considered estimation of the PK/PD parameters of a subject from the data in a study on the subject. In many PK/PD studies, however, data are collected from a number of subjects, some of whom may have intensive blood sampling while others only have sparse data. A primary objective of these studies is to study the PK/PD characteristics of the entire population, such as how they vary with certain covariates. This requires embedding the individual parametric PK/PD models in a population model. For example, the y_j in model (20) are now replaced by y_{ij}, where i denotes the subject number. Since the dose, volume of distribution, absorption and elimination rate constants may vary from subject to subject, we also have to replace D, V, k_a, k_e, n by D_i, V_i, k_{ai}, k_{ei} and n_i in model (20). Let θ_i be the vector consisting of the logarithms of the PK parameters V_i, k_{ai}, k_{ei}. The unknown θ_i may vary with certain covariates, such as the subject's age and body weight. How can the individual subjects' data be used to analyze such relationships for the target population, of which the subjects can be regarded as a sample? We shall show that

nonlinear mixed effects modeling provides a valuable tool to address this problem.

Returning to the PD model (2), the variable C refers to concentration at an effector (tissue) site. It is usually impossible to measure C directly, so some surrogate for C has to be used, as in model (5). On the other hand, if one has a kinetic model for C, then it can be used to impute the value of C from the blood/urine measurements. Chapter 9 of Davidian and Giltinan[11] illustrates how population PK/PD models can be synthesized for such tasks.

2.4. *Nonlinear mixed effects models*

The preceding population PK/PD models are special cases of nonlinear mixed effects models (NONMEM) of the form

$$y_{ij} = f_i(t_{ij}, \theta_i) + \varepsilon_{ij}\,, \quad \theta_i = g(x_i, \beta) + b_i \ (1 \le j \le n_i, 1 \le i \le K)\,, \qquad (22)$$

in which θ_i is a $1 \times r$ vector of the ith subject's parameters whose regression function on the subject's observed covariate x_i is given by $g(x_i, \beta)$ with $1 \times s$ parameter vector β, which is the "fixed effect" to be estimated. The "random effects" b_i in model (22) are assumed to be independent and identically distributed, having common distribution G with mean 0. The ith subject's response y_{ij} at t_{ij} has mean $f_i(t_{ij}, \theta_i)$, in which f_i is a known function. Given θ_i, the random errors ε_{ij} are assumed to be normal with mean 0 and standard deviation $\sigma w(\theta_i)$, in which w is a given function and σ is an unknown parameter. The regression function g relates θ_i to the ith subject's physiologic characteristics that constitute the covariate vector x_i in model (22). The first equation of (22) is often called the *individual measurement model* and the second equation the *population structure model*. The population distribution G is usually assumed to be normal with mean 0 and covariance matrix Σ so that β, σ, Σ can be estimated by maximum likelihood. However, unlike linear mixed effects models in which the normal assumption on G yields closed-form expressions of the likelihood, the normality of G in nonlinear mixed effects models leads to computationally intensive likelihoods that involve K integrals. A commonly used approach, as adopted in the software package NONMEM[12] or the `nlme` procedure in S-Plus, is to develop iterative schemes based on first-order approximations of $f_i(t_{ij}, g(x_i, \beta) + b_i)$ in model (22) so that the normal assumption on G can be used to reduce the problem to that of a linear Gaussian mixed effects model at each iterative step.

Unless otherwise stated, we shall assume throughout the sequel that the random errors ε_{ij} in model (22) have common variance σ^2 (so $w(\theta) \equiv 1$). The likelihood function $L(\beta, \sigma, \Sigma)$ is proportional to

$$|\Sigma|^{-K/2} \prod_{i=1}^{K} \int_{\mathbb{R}^r} \sigma^{-n_i} \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, g(x_i, \beta) + b_i)]^2 - \frac{1}{2} b_i \Sigma^{-1} b_i^T \right\} db_i \,, \tag{23}$$

where $|\Sigma|$ denotes the determinant of Σ. For the case of more general $w(\theta_i)$, simply replace σ in model (23) by $\sigma w(g(x_i, \beta) + b_i)$. Computing the maximum likelihood estimate of (β, σ, Σ) via numerical integration and nonlinear optimization becomes difficult for large K. Letting $\eta = (\sigma, \Sigma)$, Lindstrom and Bates[13] proposed the following iterative procedure that involves successive linear approximations to $f_i(t_{ij}, g(x_i, \beta) + b)$. At the mth iteration, the Lindstrom–Bates procedure consists of a pseudo-data step and a linear mixed effects (LME) step.

(a) *The pseudo-data step*: Given the current estimate $\hat{\eta}^{(m)}$ of η, compute $\hat{\beta}^{(m)} = \hat{\beta}(\hat{\eta}^{(m)})$ and $\hat{b}_i^{(m)} = \hat{b}_i(\hat{\eta}^{(m)})$, $1 \le i \le K$, that jointly minimize

$$\sum_{i=1}^{K} \{(\hat{\sigma}^{(m)})^{-2} \, S_i(b, \beta) + b_i (\hat{\Sigma}^{(m)})^{-1} b_i^T / 2\} \,, \quad \text{where}$$

$$S_i(\beta, b) = \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, g(x_i, \beta) + b)]^2 \Big/ 2 \,. \tag{24}$$

This can be carried out by modifying a standard nonlinear least squares routine; see Sec. 6.1 of Lindstrom and Bates.[13] Define the $s \times n_i$, $r \times n_i$ and $1 \times n_i$ matrices

$$X_i^{(m)} = \left(\frac{\partial f_i}{\partial \beta}(t_{ij}, g(x_i, \beta) + \hat{b}_i^{(m)})|_{\beta = \hat{\beta}^{(m)}} \right)_{1 \le j \le n_i} ,$$

$$Z_i^{(m)} = \left(\frac{\partial f_i}{\partial b_i}(t_{ij}, g(x_i, \hat{\beta}^{(m)}) + b_i)|_{b_i = \hat{b}_i^{(m)}} \right)_{1 \le j \le n_i} ,$$

$$Y_i^{(m)} = (y_{ij} - f_i(t_{ij}, g(x_i, \hat{\beta}^{(m)}) + b_i^{(m)}))_{1 \le j \le n_i} + \hat{\beta}^{(m)} X_i^{(m)} + \hat{b}_i^{(m)} Z_i^{(m)} \,.$$

(b) *The LME step*: Linear approximation to $f_i(t_{ij}, g(x_i, \beta) + b_i)$ around $(\hat{\beta}^{(m)}, \hat{b}_i^{(m)}$ leads to the linear mixed effects model

$$Y_i^{(m)} = \beta X_i^{(m)} + b_i Z_i^{(m)} + (\varepsilon_{i1}, \ldots, \varepsilon_{in_i}) \,. \tag{25}$$

The integrals in expression (23) for the likelihood function of the linear mixed effects model (25), instead of model (22), have closed-form expressions, yielding maximum likelihood estimates of the form

$$\hat{\beta} = \left(\sum_{i=1}^{K} Y_i^{(m)} V_{i,m}^{-1} X_i^{(m)T} \right) \left(\sum_{i=1}^{K} X_i^{(m)} V_{i,m}^{-1} X_i^{(m)T} \right)^{-1} , \tag{26}$$

where $V_{i,m} = Z_i^{(m)T} \hat{\Sigma} Z_i^{(m)} + \hat{\sigma}^2 I_{n_i}$ and $\hat{\eta} = (\hat{\sigma}, \hat{\Sigma})$ is computed via the Newton–Raphson algorithm to maximize the likelihood; see Sec. 6.2 of Lindstrom and Bates[13] where a restricted maximum likelihood (REML) variant of the procedure is also given.

Several alternatives to the linearization approach have been proposed in the literature. One is Monte Carlo integration, whose accuracy and computational complexity depend critically on how and how many samples are drawn. Importance sampling and periodic updating of the importance weights during iterative maximization of the likelihood have been proposed.[14–16] Another alternative, proposed by Pinheiro and Bates,[17] is to use an adaptive version of Gaussian quadrature based on ideas similar to importance sampling in Monte Carlo integration. A third approach is to use MCEM (Monte Carlo EM) in which the E-step of the usual EM algorithm is replaced by an empirical estimate based on a random sample generated from the conditional distribution.[15]

Instead of applying Monte Carlo methods to compute the integrals in the likelihood function to be maximized in the maximum likelihood approach, it seems more direct to apply Markov Chain Monte Carlo (MCMC) to evaluate the posterior distribution of (β, σ, Σ) when *a prior* distribution on these parameters is assumed. MCMC enables one to generate a sequence of random samples whose limiting distribution is the target distribution (in this case the posterior distribution of (β, σ, Σ)) and thereby avoids the calculation of normalizing constants and the numerical integration associated with any probability statements of interest. The most popular MCMC method used in the mixed effects model framework is the Gibbs sampler. This is because the (hierarchical) Bayes model allows a natural grouping of the vector of all unknown or unobserved parameters into

subvectors β, σ, Σ and $(\theta_i,\ i = 1,\ldots,n)$, where drawing samples for each component is much easier than drawing samples for the whole vector. Successful usage of Gibbs sampler for NONMEM in population PK studies has been reported in Refs. 11, 18–22. The relative efficiencies of different MCMC procedures have been investigated by Bennett *et al.*[23] and Shih.[24] In addition to considerations in choosing transition functions, there are other practical issues one has to deal with when implementing MCMC, such as the number of chains to run, the length of burn-in sequences, and how to monitor convergence. These are no general answers to these questions and they often need to be addressed empirically by numerical experiments; see Chapter 26.

The normality assumption on the population distribution G has been weakened by Davidian and Gallant,[25] who assume that G has a density function of the form of a product of a multivariate normal $N(0,\Sigma)$ density function and the square of a polynomial of degree p, which was introduced in another context and called the "smooth nonparametric" (SNP) model by Gallant and Nychka.[26] The coefficients of the polynomial and the components of the matrix Σ can be estimated by maximum likelihood, while the degree p of the polynomial can be chosen via standard model selection criteria like BIC, AIC or the Hannan-Quinn criterion. Magder and Zeger[27] proposed an alternative method that uses mixtures of normals, while Fattinger *et al.*[28] modeled each component of b_i as a data-dependent monotone spline transformation of the corresponding component of a multivariate normal vector. All these methods require considerably more intensive computation to maximize the likelihood function than the case of normal G assumed before.

Since the normality assumption on G only provides numerically tractable maximum likelihood estimates after various approximations and since attempts to relax that assumption have led to even more computationally intensive procedures, a natural alternative is to try estimating G nonparametrically (by a distribution with finite support, with the number of support points depending on the sample size). However, even for the simple case $n_i \equiv n$ and $f_i(t_{ij},\theta_i) = \theta_i$ with known β and σ, it is difficult to estimate G well since the optimal rate of convergence of the estimate to G is very slow when G has a smooth density function, as pointed out by Carroll and Hall[29] and Fan.[30] When G has fixed support, Chen[31] showed that the optimal convergence rate is $K^{-1/2}$ if the number of support points is known but decreases to $K^{-1/4}$ otherwise as $K \to \infty$. Lindsay[32] showed

that the nonparametric maximum likelihood estimate $\hat{G}$ of G is unique and discrete, with no more than K support points, and Mallet[33] made use of this and other properties of $\hat{G}$ to develop an algorithm to compute $\hat{G}$. The situation becomes considerably worse when σ and β are unknown and $f_i(t_{ij}, \theta_i)$ is nonlinear in θ_i, for which little is known about the performance of nonparametric estimates and it is also difficult to compute $\hat{G}$.

One way to ensure that β and σ can be well estimated is to require the dataset to contain a subset from subjects whose θ_i can be well estimated. This idea was introduced in the work of Ibragimov and Has'minskii[34] who consider estimation of (α, G) from independent random variables $y_1, \ldots, y_J$ such that the conditional density function of y_i given θ_i has the parametric form $f_\alpha(\cdot|\theta_i)$, in the presence of another "direct" sample $\theta_1, \ldots, \theta_I$ from G. Let $K = I + J$. They show that under certain regularity conditions, a variant of the nonparametric maximum likelihood estimate that is initialized at a $\sqrt{n}$-consistent estimate of (α, G) is asymptotically efficient. Their model of the data $\{\theta_1, \ldots, \theta_I; y_1, \ldots, y_J\}$ is commonly called the Ibragimov-Has'minskii (IH) model. We shall relax the model assumptions and extend them to our setting, providing what will be called an "Ibragimov-Has'minskii (IH) environment."

In an IH environment, there are I ($\leq K$) subjects whose θ_i can be well estimated by the nonlinear least squares estimate $\tilde{\theta}_i$ based on (y_{ij}, t_{ij}), $1 \leq j \leq n_i$. Without loss of generality we can assume that these are the first I subjects. We can determine from the data the standard error of each component of $\tilde{\theta}_i$ using the asymptotic formulas in nonlinear regression.[35] The ith study is deemed "good" if all components of $\tilde{\theta}_i$ have reasonably small standard errors relative to their absolute values. A consistent estimate of σ^2 is given by

$$\tilde{\sigma}^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (w(\tilde{\theta}_i))^{-2} (y_{ij} - f_i(t_{ij}, \tilde{\theta}_i))^2 \Bigg/ \sum_{i=1}^{I} (n_i - r)\,. \tag{27}$$

Such IH environments arise in most population PK studies, which use combined data from several Phases I, II and III trials. The subjects in Phase I trials are usually healthy volunteers or patients with the intent-to-treat disease, from whom intensive blood sampling is conducted, and thus provide natural candidates for good studies.

Lai and Shih[36] developed the following iterative scheme to compute the MLE of (β, σ, G) in an IH environment. First note that in the case $w(\theta) \equiv 1$ the likelihood function is proportional to

$$L(\beta, \sigma, G) = \prod_{i=1}^{K} \sigma^{-n_i} \sum_{m=1}^{M} \alpha_m \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, g(x_i, \beta) + \zeta_m)]^2 \right\} \quad (28)$$

when G has a finite support $\{\zeta_1, \ldots, \zeta_M\}$ and puts mass α_m at ζ_m. For the case of more general $w(\theta_i)$, simply replace σ in model (28) by $\sigma w(g(w_i, \beta) + \zeta_m)$. The initial estimate $(\hat{\beta}^{(0)}, \hat{\sigma}^{(0)}, \hat{G}^{(0)})$ is obtained as follows: Let $\hat{\sigma}^{(0)} = \tilde{\sigma}$ and $\hat{\beta}^{(0)}$ be the least squares estimate $\hat{\beta}^{(0)}$ which minimizes $\sum_{i=1}^{I} (\widetilde{\theta}_i - g(x_i, \beta))^T (\widetilde{\theta}_i - g(x_i, \beta))$. Let $\widetilde{b}_i = \widetilde{\theta}_i - g(x_i, \hat{\beta}^{(0)})$, $1 \le i \le I$, denote the residuals, and let $\widehat{b}_i = \widetilde{b}_i - (\sum_{j=1}^{I} \widetilde{b}_j)/I$ be the centered residuals. Let $\hat{G}^{(0)}$ be the distribution putting weight $1/I$ at each centered residual. $(\hat{\beta}, \hat{\sigma}, \hat{G})$ is computed via an iterative procedure in which the following two steps are used to compute $(\hat{\beta}^{(k)}, \hat{\sigma}^{(k)}, \hat{G}^{(k)})$ from $(\hat{\beta}^{(k-1)}, \hat{\sigma}^{(k-1)}, \hat{G}^{(k-1)})$; see Ref. 36 where a termination criterion and numerical examples are given.

Step 1. Suppose $\hat{G}^{(k-1)}$ puts mass α_j at ζ_j $(j = 1, \ldots, M_{k-1})$. Find the maximizer $(\hat{\beta}^{(k)}, \hat{\sigma}^{(k)})$ of $L(\beta, \sigma, \hat{G}^{(k-1)})$.

Step 2. Use Mallet's algorithm[33] to maximize $L(\hat{\beta}^{(k)}, \hat{\sigma}^{(k)}, G)$ over the set of distributions G with no more than K support points.

2.5. *Empirical Bayes methods for individualization and diagnostics*

We now consider the prediction problem of estimating a function $h(\theta)$ of the unobservable parameter θ for a new subject with covariate x and from whom some data have been collected. For example, in population PK studies, it is believed that efficacy and toxicity of a drug are directly related to the drug concentrations at the target site, which are generally not available but for which blood concentrations are often good surrogates; therefore the criteria for designing the dosing regimen for a specific subject often involve functions of individual concentrations, or equivalently, functions of the individual parameter θ. The subject's data are often too sparse to provide an adequate estimate $\hat{\theta}$ of θ so that $h(\hat{\theta})$ can be used to estimate $h(\theta)$. If β, σ and G are known, then a natural estimate of $h(\theta)$ in the mixed effects model is the posterior mean $E_{\beta,\sigma^2,G}[h(\theta) \,|\, \text{subject's data}]$. Without assuming β, σ^2 and G to be known, the empirical Bayes approach in Ref. 36

replaces them by their estimates $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{G}$ from the K studies so that $h(\theta)$ is estimated by

$$\widehat{h(\theta)} = E_{\hat{\beta},\hat{\sigma}^2,\hat{G}}[h(\theta) \,|\, \text{subject's data}]\,. \tag{29}$$

This idea of borrowing information from other subjects is in fact one of the main motivations for building population structure models. In particular, because of ethical and practical reasons, intensive blood sampling is often not feasible for clinical patients, for whom this individualization of dosing regimen can be obtained by combining the patient's sparse data and characteristics (as measured by x) with the large database for the population model. See also Berzuini[38] for an example of medical monitoring.

Empirical Bayes ideas can also be used to derive diagnostics for the regression model (22). If the individual parameters θ_i were observed, the residuals $r_i = \theta_i - g(x_i, \hat{\beta})$ would provide approximations for the unobservable i.i.d. random variables b_i. Therefore substantial deviation of these residuals from i.i.d. patterns would suggest inadequacies and possible improvements of the assumed regression model. Since the θ_i are not observed, we propose to replace them by the empirical Bayes estimate $E_{\hat{\beta},\hat{\sigma}^2,\hat{G}}(\theta_i \,|\, y_{i1}, \ldots, y_{in_i}, t_i, x_i)$, leading to the following generalized residuals in the sense of Cox and Snell[39]:

$$\hat{r}_i = E_{(\hat{\beta},\hat{\sigma}^2,\hat{G})}(\theta_i \,|\, y_{i1}, \ldots, y_{in_i}, t_i, x_i) - g(x_i, \hat{\beta})\,, \quad i = 1, \ldots, K\,. \tag{30}$$

The $\hat{r}_i$ can be interpreted as estimates of the independent zero-mean random variables $r_i = E_{(\beta,\sigma^2,G)}(\theta_i \,|\, y_{i1}, \ldots, y_{in_i}, t_i, x_i) - g(x_i, \beta)$.

Instead of using the posterior mean in Eq. (30), it is popular in population PK studies to use the posterior mode

$$\hat{s}_i = \arg\max_{b_i} p_{(\hat{\beta},\hat{\sigma}^2,\hat{G})}(b_i \,|\, y_{i1}, \ldots, y_{in_i}, t_i, w_i) \tag{31}$$

to form the residuals $\hat{s}_i - g(x_i, \hat{\beta})$, where $p_{(\beta,\sigma^2,G)}$ denotes the posterior density in the Bayesian model with given β, σ^2 and G. This was first suggested by Maitre *et al.*[40] in connection with linearization methods under the assumption of normality for the population distribution, but is also used as a general strategy in the semiparametric models of Davidian and Gallant[25] and the hierarchical Bayesian models of Wakefield and Racine-Poon.[44] For linear Gaussian mixed effects models, the mean and the mode of the conditional distribution of θ_i given $y_{i1}, \ldots, y_{in_i}$ coincide since the conditional distribution is Gaussian, so the theoretical justification for

the r_i via an empirical Bayes point of view applies also to the s_i. In the case of nonlinear mixed effects models, the posterior mean and mode no longer coincide, and $\hat{r}_i$ is usually easier to compute and more robust. In the above empirical Bayes approach, we have replaced (β, σ^2, G) in the posterior mean $E_{\beta,\sigma^2,G}[h(\theta) \,|\, \text{subject's data}]$ by an estimate $(\hat{\beta}, \hat{\sigma}^2, \hat{G})$. This estimate $(\hat{\beta}, \hat{\sigma}^2, \hat{G})$ can be either parametric, as in Lindstrom and Bates,[13] or nonparametric, as given above.

We next list some examples of using empirical Bayes/hierarchical Bayes/posterior mode estimates in NONMEM to quantify covariate effects on PK parameters in the literature:

(a) Population PK analysis of felbamate in epileptic patients[42]: Apparent clearance of felbamate was found to decrease with age for children (age ≤ 12) and to stay relatively constant beyond 13 years of age. There were 1–17 blood samples per subject. This study, undertaken by Zhu and his collaborators at Schering-Plough Research Institute and Wallace Laboratories, led to the FDA approval of the labeling of felbamate for its prescription to children.
(b) Population PK analysis of quindine in hospitalized patients treated for atrial fibrillation over ventricular arrhythmias[19,25,43]: The effects of dichotomized creatinine clearance, body weight and α_1-acid glucoprotein concentration on clearance were analyzed from a study consisting of 1–11 blood samples per subject.
(c) Population PK analysis of phenobarbital in neonates[11,25,44]: The effects of birth weight and 5-minute Apgar score on clearance and volume were analyzed from a study with sparse PK data in each subject (having only 1–6 concentration measurements).

Model validation methodology for population PK analysis is still in its infancy. One commonly used approach is to use m-fold cross-validation or bootstrap to estimate the prediction errors based on a fitted model. Here the prediction error may be associated with prediction of concentrations or prediction of PK parameters (that can be estimated nonparametrically only from subjects with intensive measurements). Given the computational complexity associated with fitting nonlinear mixed effects models, m-fold cross-validation (with $m \leq 20$) appears to be more feasible than the bootstrap (for which the FDA recommends using at least 200 bootstrap samples).

2.6. *The Lindstrom–Bates algorithm and related statistical methods*

Vonesh[45] proposed an alternative to the Lindstrom–Bates algorithm (consisting of the pseudo-data and LME steps described in Sec. 2.4) by applying, for fixed β and η, Laplace's asymptotic formula

$$\int_{\mathbb{R}^r} e^{l(b)}\, db \sim (2\pi)^{r/2} |-\ddot{l}(\hat{b})|^{-1/2} e^{l(\hat{b})} \tag{32}$$

to each integral in expression (23), where $\ddot{l}$ denotes the Hessian matrix of second partial derivatives of l and $\hat{b}$ maximizes $l(b)$. Earlier, Wolfinger[46] derived the pseudo-data step of the Lindstrom–Bates algorithm by applying for fixed η Laplace's asymptotic formula to the multiple integral

$$\int \cdots \int \exp\left\{\sum_{i=1}^{K} l_i(b_i;\beta)\right\} d\beta db_1 db_2 \cdots db_K\,, \tag{33}$$

and then used a Gauss–Newton approximation of $-\ddot{l}$ to derive the REML version of the LME step. Laplace's asymptotic formula has also been used by Breslow and Clayton[47] and Lee and Nelder[48] to derive their estimators in *generalized linear mixed models* (GLMM) and *hierarchical generalized linear models* (HGLM), respectively. The HGLM involves independent random vectors (y_i, x_i^T, z_i^T) such that the conditional density function of y_i given a $1 \times K$ vector of random effects b has the GLM (generalized linear model) form

$$f(y|b, z_i, x_i) = c(y, \phi) \exp\{(\theta_i y - \psi(\theta_i))/a(\phi)\}\,, \tag{34}$$

in which ϕ is a dispersion parameter, θ_i is the canonical parameter such that $E(y|b, z_i, x_i) = g(\beta x_i + b z_i)$ and g is the inverse of a monotone link function. Letting f_α be the density function of b with unknown parameter α, Lee and Nelder[48] define the hierarchical likelihood (h-likelihood) by

$$h(b, \beta, \phi, \alpha) = \log f_\alpha(b) + \sum_{i=1}^{n} \log f(y_i|b, z_i, x_i)\,. \tag{35}$$

They propose to estimate β, ϕ, α by an iterative procedure whose mth iteration consists of the following two steps:

(i) Given the current estimate $(\hat{\phi}^{(m)}, \hat{\alpha}^{(m)})$ of (ϕ, α), compute the maximizer $(\hat{b}^{(m)}, \hat{\beta}^{(m)})$ of $h(b, \beta, \hat{\phi}^{(m)}, \hat{\alpha}^{(m)})$ by solving the score equations $\partial h/\partial\beta = 0$ and $\partial h/\partial b = 0$.

(ii) Given the current estimate $(\hat{b}^{(m)}, \hat{\beta}^{(m)})$ of (b, β), maximize the adjusted profile h-likelihood $h_A(\phi, \alpha) = h(\hat{b}^{(m)}, \hat{\beta}^{(m)}, \phi, \alpha) + (\log|2\pi\phi H^{-1}|)/2$ with

$$H = \left(\begin{array}{cc} \partial^2 h/\partial\beta^2 & \partial^2 h/\partial\beta\partial b \\ \partial^2 h/\partial b\partial\beta & \partial^2 h/\partial b^2 \end{array} \right) \Bigg|_{(b,\beta)=(\hat{b}^{(m)},\hat{\beta}^{(m)})},$$

by solving the score equations $\partial h_A/\partial\phi = 0$ and $\partial h_A/\partial\alpha = 0$.

For the special case of normal f_α with mean 0 and covariance matrix $\Sigma(\alpha)$, the HGLM reduces to the GLMM considered by Breslow and Clayton[47] who make use of the normality assumption to come up with an explicit expression for Laplace's approximation to the likelihood function $\int e^{h(b,\beta,\phi,\alpha)} db$, yielding an algorithm similar to that of Lindstrom and Bates for NONMEM. The Lee-Nelder procedure above is somewhat different and is motivated by generalizing Henderson's[49] joint likelihood for linear models with normal random effects. It can be derived by applying Lapalce's approximation to $\iint e^{h(b,\beta,\phi,\alpha)} db d\beta$, analogous to integral (33).

Let β_0 and σ_0 denote the true values of β and σ. A sufficient condition for the validity of Laplace's asymptotic formula (32) is that $l(b) = N\lambda(b)$, where $N \to \infty$ and λ is a fixed smooth function with a unique maximum. The integral for the ith subject in model(23) has the form

$$\int_{\mathbb{R}^r} \exp\{l_i(b|\beta, \sigma, \Sigma)\} db\,, \quad \text{where}$$

$$l_i(b|\beta, \sigma, \Sigma) = -S_i(b, \beta)/\sigma^2 - b\Sigma^{-1}b^T/2 - n_i \log\sigma\,, \tag{36}$$

in which S_i is computed via Eq. (24) from n_i observations (y_{ij}, t_{ij}), $1 \le j \le n_i$. If these observations are sufficiently informative about the ith subject's parameter vector $\theta_i = g(x_i, \beta_0) + b_i$, then for (β, σ) near (β_0, σ_0), $S_i(b, \beta)$ becomes peaked around b_i and can be approximated by a quadratic function in a neighborhood of the maximizer $\hat{b}_i = \hat{b}_i(\beta, \sigma, \Sigma)$ of $l_i(b|\beta, \sigma, \Sigma)$. Laplace's asymptotic formula basically replaces l_i in integral (36) by the approximating quadratic function of b as $\lambda_{\min}(-\ddot{l}_i(\hat{b}_i|\beta, \sigma, \Sigma)) \to \infty$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a symmetric matrix.

When the ith subject has sparse data (y_{ij}, t_{ij}), $S_i(b, \beta)$ is no longer peaked around $\hat{b}_i$ and Laplace's asymptotic formula may be a poor approximation to integral (36). A better way to compute integral (36) in this case is to use Monte Carlo, expressing integral (36) as the expectation

$$E_\Sigma\{\exp(-S_i(b, \beta)/\sigma^2)\}\,, \tag{37}$$

where E_Σ denotes expectation under the probability measure for which b is a normal random vector with mean 0 and covariance matrix Σ. Lai and Shih[37] proposed the following hybrid method for evaluating model (36). Take $c > 10$ and let $V_i = -\ddot{l}(\hat{b}_i|\beta, \sigma, \Sigma)$.

(i) If $\lambda_{\min}(V_i) < c$, evaluate integral (36) by Monte Carlo approximation to expression (37):

$$B^{-1} \sum_{j=1}^{B} \exp\{-S_i(b_{ij}, \beta)/\sigma^2\},$$

where b_{ij}, $j = 1, \ldots, B$, are independent samples from the $N(0, \Sigma)$ distribution.

(ii) If $\lambda_{\min}(V_i) \geq c$, evaluate integral (36) by its Laplace approximation

$$(2\pi)^{r/2} |V_i|^{-1/2} \exp\{l_i(\hat{b}_i|\beta, \sigma, \Sigma)\}.$$

By performing simple diagnostics on the appropriateness of using Laplace's asymptotic formula to evaluate the integral in expression (23) for the ith subject, the hybrid approach preserves the computational simplicity of Laplace's method when it can be used and switches to the Monte Carlo method when Laplace's method fails. In practice, the actual population distribution G of the random effects b_i may differ substantially from the assumed normal distribution with unknown covariance matrix, which at best can only be regarded as an approximation to G. If the ith subject has only sparse data so that $S_i(b, \beta)$ is relatively flat in b, then applying the Monte Carlo approach to the subject is tantamount to choosing a certain random distribution G_i, which is the empirical distribution of a sample of size B from a normal distribution, to approximate G. Since the assumed normal distribution is itself also an approximation to G, there is no need for a "high resolution" in the random distribution used to approximate the normal distribution, so using $50 \leq B \leq 200$ samples in the Monte Carlo method should be able to provide enough statistical detail so that the resultant estimator of (β, σ, Σ) still has a low computational cost comparable to that of the Lindstrom–Bates estimator. On the other hand, if the ith subject has enough data so that $S_i(b, \beta)$ is peaked around $\hat{b}_i$ for β near β_0, the Monte Carlo approach becomes unreliable unless B is sufficiently large and importance sampling is needed to generate the B samples from a distribution that is peaked around $\hat{b}_i$, so Laplace's method gives a much better approximation to (36) in this case. Thus the Monte Carlo and Laplace's methods complement each other in the hybrid approach, which uses either

$N(0, \Sigma)$ or the empirical distribution of a sample of size B from $N(0, \Sigma)$ as the approximation $G_i(\cdot|\Sigma)$ to the unknown (and possibly non-normal) mixing distribution G. Using this hybrid approach to compute expression (23) approximately, Lai and Shih[37] make use of numerical differentiation and iterative optimization schemes such as conjugate gradient and quasi-Newton methods[50] to maximize this approximation to expression (23), providing the estimator $(\hat{\beta}, \hat{\sigma}, \hat{\Sigma})$ of (β, σ, Σ). Good starting values in this iterative scheme to compute $(\hat{\beta}, \hat{\sigma}, \hat{\Sigma})$ can be obtained by running several steps of the Lindstrom–Bates `nlme` procedure.

Lai and Shih[37] also develop an asymptotic theory of the hybrid estimator $(\hat{\beta}, \hat{\sigma})$ as the number K of subjects becomes infinite. This theory does not require all subjects to have sufficient data to estimate their θ_i consistently, nor does it require the actual G to be normal. Under the assumption that a sufficiently large subset of the subjects have good studies in the sense that their $\lambda_{\min}(V_i)$ exceeds the threshold c for applicability of Laplace's approximation to evaluate integral (36) and some additional regularity conditions, $(\hat{\beta}, \hat{\sigma})$ is shown to converge with probability 1 to (β_0, σ_0) as $K \to \infty$. Let $n = n_1 + \cdots + n_K$. It is also shown in Lai and Shih[37] that $\sqrt{n}(\hat{\beta}_n - \beta_0, \hat{\sigma}_n - \sigma_0)$ has a limiting normal distribution as $K \to \infty$ under these and some other conditions. Moreover, this hybrid estimator and its asymptotic theory have been extended in Lai and Shih[37] to the HGLM of Lee and Nelder[48] and the GLMM of Breslow and Clayton.[47]

3. Bioavailability and Bioequivalence

Generic drug products (manufactured by other companies that are not the innovator) have become increasingly popular since the 1960s. For the approval of a generic drug product, the FDA usually does not require a regular new drug application (NDA) submission to demonstrate the efficacy and safety of the product. Instead, it requires the generic drug company to submit bioavailability (BA) information on the generic drug and to provide evidence of its bioequivalence (BE) to the standard (or reference) drug in an "abbreviated new drug application" (ANDA), following certain regulations that became effective in 1977 and are codified in 21 CFR 320, in which BA is defined as "the rate and extent to which the active ingredient or active moiety is absorbed from a drug product and becomes available at the site of action." In Sec. 2.2 we have discussed how the PK data of a drug can be used to measure its BA. This section focuses on BE and the statistical methods in BE studies.

Two drug products are said to be *bioequivalent* if they contain either identical amounts of the same active ingredient (i.e. are "pharmaceutical equivalents") or an identical therapeutic moiety and if their rates and extents of absorption are not significantly different when administered at the same dose under similar experimental conditions. BE studies are conducted not only for ANDAs of generic drugs but also for formulation change of an approved drug. For example, clinical trials for the NDA of a drug usually use the drug produced in a laboratory setting. After approval, commercial batches produced from manufacturing plants have to be demonstrated to be bioequivalent to the clinical trial batches. Moreover, there may also be changes from tablet to capsule formulations so that BE studies are needed.

BE studies typically use healthy normal subjects and do not involve Phases II and III trials. A pilot study using a small number (e.g. 6) of subjects can be carried out in advance to assess inter-subject and intra-subject variabilities, sample size, time intervals to collect blood or urine samples and to provide other information. Instead of the commonly used randomized designs in Phases II and III studies, in which each subject is randomly assigned to one and only one formulation of a drug (parallel designs), BE studies typically use the crossover design, which is a modified randomized block design in which each block (consisting of a subject or a group of subjects) receives more than one formulation of a drug at different time periods. Crossover designs have the following advantages in BE studies:

(a) Each subject serves as his/her own control, allowing a within-subject comparison between formulations.
(b) Inter-subject variability is removed from the comparison between formulations.
(c) With proper randomization of subjects to the sequence of formulation administrations, a crossover design can provide the best unbiased estimates of the differences (or ratios) between formulations. On the other hand, care must be taken to address the "carry-over" effects in crossover designs. In BE studies, the "washout" period, which is defined as the rest period between two treatment periods for the effect of the preceding treatment period to taper off, must be long enough so that the carry-over effect from one treatment period to the next is negligible. There is an extensive literature on crossover designs for clinical trials,[51–57] and for BE studies.[58]

Although parallel designs are infrequently used in BE studies since crossover designs usually provide much better ways of identifying and removing the inter-subject variability from the comparison between formulations based on a sample of typically 18–24 subjects, there are situations in which a parallel design is preferable to a crossover design, e.g. when (i) the inter-subject variability is relatively small compared to the intra-subject variability, or (ii) the drug has long elimination half-life so that the long washout period in a crossover design prolongs the study and increases the chance of drop-out of the subjects, or (iii) the cost of increasing the number of subjects is smaller than that of adding an additional treatment period, or (iv) extensive blood collection is not feasible from the subjects.

Suppose there are two formulations, one of which is a test formulation (T) and the other a reference (or standard) formulation (R) of a drug. For a standard 2×2 crossover design, each subject is randomly assigned to either the first sequence RT or the second sequence TR at two dosing periods. A subject assigned RT receives R at the first dosing period and T at the second period. The dosing periods are separated by a washout period of sufficient length to rule out carry-over effects. More generally, an $m \times n$ crossover design involves m sequences of formulations that are administered at n time periods. Examples are the 2×4 crossover design consisting of the two sequences TRTR and RTRT, and Balaam's 4×2 crossover design[51] consisting of the four sequences TT, RR, RT and TR.

A widely used statistical model to perform inference in these designs is the linear mixed effects model

$$y_{ijk} = \mu + a_j + \eta_{ik} + b_{jk} + c_{j-1,k} + \epsilon_{ijk}\,, \tag{38}$$

where i refers to the subject number, j the period number and k the sequence number. Here μ is the overall mean, a_j is the fixed effect of the jth period (with $\Sigma a_j = 0$), η_{ik} is the random effect (assumed to be normal with mean 0) of the ith subject in the kth sequence, b_{jk} is the fixed effect of the formulation in the jth period of the kth sequence, and ϵ_{ijk} is the within-subject random error which is assumed to be normal with mean 0. In particular, for a standard 2×2 crossover design, b_{jk} is the fixed effect of R (resp. T) if $j = k$ (resp. $j \neq k$). Note that model (38) assumes first-order (i.e. one-period) carry-over effects: $c_{j-1,k}$ represents the (fixed) residual effect carried over from period $j-1$ to period j in the kth sequence. For two-period designs, carry-over effects can only occur in the second period. It is also assumed that the η_{ik} and ϵ_{ijk} are independent with $\mathrm{var}(\eta_{ik}) = \sigma_\eta^2$ and $\mathrm{var}(\epsilon_{ijk}) = \sigma_\epsilon^2$. Standard ANOVA techniques can be used to construct

unbiased estimates and confidence intervals of linear contrasts of the fixed effects, while the variance parameters σ_η^2 and σ_ϵ^2 can be estimated by the method of moments or restricted maximum likelihood.[38,39] The y_{ijk} in model (38) is typically some transformation of the observed response (e.g. logarithm of the AUC) to make it approximately normal. Note that the logarithmic transformation converts multiplicative effects into additive effects, as assumed in model (38).

Although model (38) leads to standard F-tests of equality between the formulations T and R, it has been recognized since the 1970s that testing the usual hypothesis of equality is inappropriate for BE, whose purpose is to verify that the two formulations have no "biologically significant" differences.[60,61] One way to address this difficulty is to change the null hypothesis of equality (versus the alternative hypothesis of inequality) into a null hypothesis of the form $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$, with an interval alternative hypothesis $H_1 : \theta_1 < \theta < \theta_2$, where θ is the parameter of interest and the interval (θ_1, θ_2) is a biological indifference zone. Schuirmann[62,63] and Anderson and Hauck[64] have developed test procedures for what is now called *average bioequivalence*. Instead of relying on hypothesis testing, Westlake[61] proposed the following confidence interval procedure to assess average bioequivalence. Let $\mu_T(\mu_R)$ denote the mean response of a subject receiving treatment T(R) in model (38). If a $(1-2\alpha)\times 100\%$ confidence interval for $\mu_T - \mu_R$ is within the acceptance limits as recommended by the regulatory agency, then accept the test formulation T as bioequivalent to the reference formulation R.

Average bioequivalence only compares the means of the marginal distributions of the PK parameters of interest, such as AUC or $C_{\max}$, associated with the two formulations. Under normality assumptions, the equivalence between distributions is characterized by the equivalence of their means and variances. *Population bioequivalence* therefore also compares the variances of the two formulations. The intra-subject variability, particularly associated with switching from one formulation to another, leads to another criterion in assessing BE, called *individual bioequivalence*. To explain the underlying motivation, suppose a patient switches from R to T that has a much higher intra-subject variability than R. This may push the AUC of the patient outside the established therapeutic window of R. Consequently, population BE does not guarantee that the two formulations are exchangeable and therapeutically equivalent and individual BE is needed.

To see what is involved in assessing these three criteria of BE, assume for simplicity that there are no carry-over effects and no period-sequence interactions so that (38) can be reduced to a form that involves T and R more directly as

$$y_{i\delta\nu} = \mu_\delta + \alpha_{i\delta} + \epsilon_{i\delta\nu}\,, \tag{39}$$

where i is the subject number, δ = T or R, ν denotes the number of times that δ appears in a sequence and therefore $1 \leq \nu \leq n_\delta$ (= largest number of times that δ appears in the available sequences). The $\epsilon_{i\delta\nu}$ are independent normal with mean 0, variance σ_ϵ^2 and are independent of the random effects $(\alpha_{iT}, \alpha_{iR})$ that are independent and have a bivariate normal distribution with $\mathrm{var}(\alpha_{iT}) = v_T$, $\mathrm{var}(\alpha_{iR}) = v_R$ and $\mathrm{var}(\alpha_{iT} - \alpha_{iR}) = \sigma_D^2$. According to the 1999 FDA Guidance on Bioequivalence, average BE is established if the 90% confidence limits for e^{μ_T}/e^{μ_R} are 4/5 and 5/4, or equivalently, if $\pm\ell n(1.25)$ are the 90% confidence limits for $\mu_T - \mu_R$. Note that the total variance of the T formulation is $\sigma_T^2 = \sigma_\epsilon^2 + v_T$, while that of the R formulation is $\sigma_R^2 = \sigma_\epsilon^2 + v_R$. Population BE is established if the 95% upper confidence bound for $\{(\mu_T - \mu_R)^2 + (\sigma_T^2 - \sigma_R^2)\}/\sigma_R^2$ falls below the FDA specified limit of $\{(\ell n 1.25)^2 + 0.02\}/(0.2)^2 = 1.745$. Individual BE is established if the 95% upper confidence bound for $\{(\mu_T - \mu_R)^2 + \sigma_D^2\}/\sigma_\epsilon^2$ falls below another FDA specified limit. These upper confidence bounds can be obtained by appealing to the central limit theorem and using the delta method to compute the asymptotic standard errors. Alternatively, bootstrap methods can be used to compute the confidence bounds and confidence intervals; see in particular Chapter 25 of Efron and Tibshirani[65] and Sec. 4.5.3 of Chow and Liu.[58] The inclusion of population BE and individual BE besides average BE by the FDA in its guidelines for the pharmaceutical industry reflects its concerns about prescribability and switchability of generic drug products. Prescribability means that when a physician prescribes a generic drug product to a patient for the first time, they should both be assured that the drug product yields safety and efficacy results comparable to that of the reference product in the patient population. Switchability means that when a physician switches a reference product to a generic product for a patient, they should both be assured that the generic product will yield comparable safety and efficacy results for the same individual.

Nonparametric and Bayesian approaches to BE have also been developed in the literature.[58] There are intriguing theoretical problems concerning BE in statistical decision theory.[66] Crossover designs and average BE for more than two formulations have also been studied.[58]

4. Assay Development and Validation

The availability of reliable assays is central to determining the drug concentrations in blood, urine, etc., in PK studies. When a pharmaceutical compound is discovered, it is necessary to develop an assay method to measure the substance levels in plasma, serum, etc. The substance that is being measured is called an *analyte*, and the objective is to determine the analyte's *potency*, which refers to its content or activity (e.g. number of particles, gravitometric mass, percent of impurity). There are three types of assays that are commonly used in the pharmaceutical industry: (i) chemical assays such as HPLC (high performance liquid chromatographs), (ii) immunoassays (e.g. radioimmunoassays, enzyme-linked immunosorbent assays), (iii) biological assays (measuring the analyte's potency relative to some standard drug in terms of the magnitudes of their effects on responses from living subjects).

For the development of an assay method of a pharmaceutical compound, the FDA requires that the assay method meet the established specifications, for which instrument calibration is essential. A common approach to calibration is to have a number of known standard concentration preparations put through the instrument to obtain the corresponding responses. Fitting an appropriate statistical model to the data yields an estimated calibration curve, called the *standard curve*. Simple linear regression of the response on the standard is perhaps the most widely used statistical model. The standard curve is used to determine the unknown potency.[67]

Validation of an assay method is the process by which it is established, in laboratory studies, that the performance characteristics of the method indeed meets the specified criteria. As specified in Chow and Liu,[67] these criteria include (i) accuracy (no systematic error in the assay method), (ii) precision (measurement error of the method), (iii) limit of detection/quantitation (LOD/LOQ, which is the lowest concentration of analyte in a sample that can be detected/determined with acceptable precision under the specified experimental conditions), (iv) range (reliable range of the method), (v) linearity (whether the assay generates results that are directly proportional to the concentration of analyte within a given range), (vi) specificity (whether the assay measures the analyte and no other substance in the specimen), (vii) ruggedness (degree of reproducibility of assay results under a variety of normal test conditions, such as different laboratories, assay temperatures, days). Commonly used statistical methods for assay validation include:

(a) regression analysis (particularly with respect to accuracy, linearity and LOD/LOQ),
(b) analysis of variance (particularly with respect to ruggedness); see Chapter 3 of Chow and Liu.[67]

Lin[68] introduces a concordance correlation coefficient to evaluate reproducibility and ruggedness, while Chapter 10 of Davidian and Giltinan[11] applies nonlinear mixed effects models to the analysis of assay data.

5. Drug Discovery

As pointed out in the preceding section, assay development is an important facet in the drug discovery process. Another important facet is of a biological nature and involves the identification of a biological target or pathway. In recent years, advances in bioinformatics and genomics have provided new tools and opportunities in this direction. Besides applications to assay development and bioinformatics, statistical methods are also useful in screening compounds for clinically active drugs, and in searching for novel, active compounds.

A pharmaceutical company typically has a large inventory of compounds, of which an unknown small proportion is truly active. Dunnett[69] developed a model that takes into account the costs and benefits of any screening procedure to derive an optimal procedure; see also the subsequent work of Bergman and Gittins[70] in this direction. Colton[71] and King[72] considered multistage screening procedures, while Redman and King[73] proposed group screening that uses balanced and partially balanced incomplete block designs to increase the rate of compound screening without reducing necessary replication.

Numerical topology is the assignment of numerical values to topologically invariant features of molecules. There is an isomorphism between two-dimensional molecular diagrams and connected graphs; the edges and vertices of the graphs correspond to bonds and atoms of molecules, yielding numerical representation of compounds or parts of compounds. With this representation, search for active compounds involves a very large set of graphs. Moreover, there may also be a large number of potential chemical modifications at different sites that one may want to experiment with. Experimental design techniques are particularly useful for such problems.[74,75]

References

1. Katzung, B. G. (1992). *Basic and Clinical Pharmacology.* Prentice Hall, Englewood Cliffs, NJ.

2. Rowland, M. and Tozer, T. N. (1989). *Clinical Pharmacokinetics*, 2nd edn., Lea and Febiger, Philadelphia.
3. Evans, W. E., Schentag, J. J. and Jusko, W. J. (eds). (1986). *Applied Pharmacokinetics*, 2nd edn., Applied Therapeutics Inc., SF.
4. Welling, P. G. (1986). *Pharmacokinetics*: *Processes and Mathematics*, American Chemical Society, Washington, DC.
5. Goldstein, A., Aronow, L. and Kalman, S. M. (1974). *Principles of Drug Action*, 2nd edn., Wiley, NY.
6. Wagner, J. G. (1976). Linear pharmacokinetic equations allowing direct calculation of many needed pharmacokinetic parameters from the coefficients and exponents of polyexponential equations which have been fitted to data. *Journal of Pharmacokinetics and Biopharmaceutics* **4**: 443.
7. Lai, T. L. (1985). Regression analysis of compartmental models. *Journal of Research of the National Bureau Standard* **90**: 525.
8. Yeh, K. C. and Kwan, K. C. (1978). A comparison of numerical integration algorithms by trapezoidal, Lagrange and spline approximations. *Journal of Pharmacokinetics and Biopharmaceutics* **6**: 79.
9. Scatchard, G. (1949). The attractions of protein for small molecules and ions. *Annals of New York Academy of Science* **51**: 660.
10. Lai, T. L. and Zhang, L. (1994). Statistical analysis of ligand-binding experiments. *Biometrics* **50**: 782.
11. Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, NY.
12. Beal, S. M. and Sheiner, L. B. (1992). *NONMEM User's Guide*. NONMEM Project Group, UCSF.
13. Lindstrom, M. J. and Bates, B. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**: 673.
14. Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series* **B54**: 657.
15. Quintana, F. A., Liu, J. S. and del Pino, G. E. (1999). Monte Carlo EM with importance reweighting and its applications in random effects models. *Computational Statistics and Data Analysis* **29**: 429.
16. Gelman, A. (1995). Method of moments using Monte Carlo simulation. *Journal of Computational and Graphical Statistics* **4**: 36.
17. Pinhero, J. C. and Bates, D. M. (1995). Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics* **4**: 12.
18. Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994). Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. *Applied Statistics* **43**: 201.
19. Wakefield, J. (1996). The Bayesian analysis of population pharmacokinetic models. *Journal of the American Statistics Association* **91**: 62.
20. Wakefield, J. and Benett, J. (1996). The Bayesian modeling of covariates for population pharmacokinetic models. *Journal of the American Statistical Association* **91**: 917.

21. Gelman, A., Bois, F. and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **91**: 1400.
22. Müller, P. and Rosner, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* **92**: 1279.
23. Bennett, J. E., Racine-Poon, A. and Wakefield, J. C. (1996). MCMC for nonlinear hierarchical models. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gills, S. Richardson and D. J. Spiegelhalter, Chapman and Hall, NY, 339–357.
24. Shih, M. (1999). Estimation in nonlinear mixed effects models: Parametric and nonparametric approaches. PhD. dissertation, Dept. Statistics, Stanford University.
25. Davidian, M. and Gallant, A. R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with applications to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics* **20**: 529.
26. Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**: 363.
27. Magder, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixture of gaussians. *Journal of the American Statistical Association* **91**: 1141.
28. Fattinger, K. E., Sheiner, L. B. and Verotta, D. (1995). A new method to explore the distribution of interindividual random effects in nonlinear mixed effects models. *Biometrics* **51**: 1236.
29. Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association* **83**: 1184.
30. Fan, J. Q. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals Statistics* **19**: 1257.
31. Chen, J. (1994). Optimal rate of convergence for finite mixture models. *The Canadian Journal of Statistics* **22**: 387.
32. Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Annals of Statistics* **11**: 86.
33. Mallet, A. (1986). A maximum likelihood estimation method for random coefficient regression models. *Biometrika* **73**: 645.
34. Ibragimov, I. A. and Has'minskii, R. Z. (1983). On asymptotic efficiency in the presence of an inifinite-dimensional nuisance parameter. In *Lecture Notes in Mathematics 1021*, Springer-Verlag, NY, 195–220.
35. Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, NY.
36. Lai, T. L. and Shih, M.-C. (2003). Nonparametric estimation in nonlinear mixed effects models. *Biometrika* **90**, in press.
37. Lai, T. L. and Shih, M.-C. (2003). A hybrid estimator in nonlinear and generalised linear mixed effects models. *Biometrika* **90**, in press.
38. Berzuini, C. (1996). Medical monitoring. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks and S. Richardson, Spiegelhalter, Chapman and Hall, 1996, 321–337.

39. Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society Series* **B30**: 248.
40. Maitre, P. O., Buhrer, M., Thomson, D. and Stanski, D. R. (1991). A three-step approach combining Bayes regression and NONMEM population analysis: Application to midazolam. *Journal of Pharmacokinetics and Biopharmaceutics* **19**: 377.
41. Wakefield, J. and Racine-Poon, A. (1995). An application of Bayesian pharmacokinetic/pharmacodynamic models to dose recommendation. *Statistics in Medicine* **14**: 971.
42. Banfield, C. R., Zhu, G. R., Jen, J. F., Jensen, P. K., Schumaker, R. C., Perhach, J. L., Affrima, M. B. and Glue, P. (1996). The effect of age on the apparent clearance of felbamate: A retrospective analysis using nonlinear mixed effects modeling. *Therapeutic Drug Monitoring* **18**: 19.
43. Verme, C. N., Ludden, T. M., Clementi, W. A. and Harris, S. C. (1992). Pharmacokinetics of quinidine in male patients: A population analysis. *Clinical Pharmacokinetics* **22**: 468.
44. Boeckmann, A. J., Sheiner, L. B. and Beal, S. L. (1992). Part V (Introductory Guide) of *NONMEM User's Guide*. NONMEM Project Group, UCSF.
45. Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed effects models. *Biometrika* **83**: 447.
46. Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed effects models. *Biometrics* **80**: 791.
47. Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9.
48. Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series* **B58**: 619.
49. Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**: 423.
50. Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C*: *The Art of Scientific Computing*, Cambridge Univ. Press.
51. Balaam, L. N. (1968). A two-period design with t^2 experimental units. *Biometrics* **24**: 61.
52. Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics* **36**: 69.
53. Cheng, C. S. and Wu, C. F. (1980). Balanced repeated measurements designs. *The Annals of Statistics* **8**: 1272.
54. Laska, E. M., Meisner, M. and Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics* **39**: 1089.
55. Laska, E. M. and Meisner, M. (1985). A variational approach to optimal two-treatment crossover designs: Applications to carryover effect methods. *Journal of the American Statistical Association* **80**: 704.
56. Jones, B. and Kenward, M. G. (1989). *Design and Analysis of Crossover Trials*, Chapman and Hall, NY.

57. Fleiss, J. L. (1989). A critique of recent research on the two-treatment crossover design. *Controlled Clinical Trials* **10**: 237.
58. Chow, S. C. and Liu, J. P. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, NY.
59. Chinchilli, V. M. and Esinhart, J. D. (1996). Design and analysis of intra-subject variability in cross-over experiments. *Statistics in Medicine* **15**: 1619.
60. Metzler, C. M. (1974). Bioavailability: A problem in bioequivalence. *Biometrics* **30**: 309.
61. Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences* **61**: 1340.
62. Schuirmann, D. J. (1981). On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval. *Biometrics* **37**: 617.
63. Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**: 657.
64. Anderson, S. and Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics Series* **A12**: 2663.
65. Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, NY.
66. Brown, L. D., Hwang, J. T. and Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics* **25**: 2345.
67. Chow, S. C. and Liu, J. P. (1995). *Statistical Design and Analysis in Pharmaceutical Science*. Marcel Dekker, NY.
68. Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**: 255.
69. Dunnett, C. W. (1961). Statistical theory of drug screening. In *Quantitative Pharmacology*, ed. H. DeJonge, North Holland, Amsterdam.
70. Bergman, S. W. and Gittins, J. C. (1985). *Statistical Methods for Planning Pharmaceutical Research*. Marcel Dekker, NY.
71. Colton, T. (1963). Optimal drug screening plans. *Biometrika* **50**: 31.
72. King, E. P. (1964). Optimal replication in sequential drug screening. *Biometrika* **51**: 110.
73. Redman, C. E. and King, E. P. (1965). Group screening utilizing balanced and partially balanced incomplete block designs. *Biometrics* **21**: 865.
74. Thornber, C. W. (1979). Isosterism and molecular modification in drug design. *Chemical Society Reviews* **8**: 563.
75. Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K. and Moose, W. H. (1995). Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *Journal of Medicinal Chemistry* **38**: 1431.

About the Author

Tze Leung Lai born in Hong Kong in 1945, received his BA degree in Mathematics (1967) from the University of Hong Kong and PhD in

Statistics (1971) from Columbia University. He then remained on the faculty of Columbia University as Assistant Professor (1971–1974), Associate Professor (1974–1977), Professor (1977–1986) and Higgins Professor (1986–1987) of Mathematical Statistics. Since 1987, he has been Professor of Statistics at Stanford University, where he is currently Chairman of the Department of Statistics. His research interests include adaptive control and learning systems, sequential experimentation, fault detection and quality control, stochastic optimization, biostatistics, probability theory and mathematical finance. He received the Committee of Presidents of Statistical Societies (COPSS) Award and the Guggenheim Fellowship in 1983. He is an elected member of Academia Sinica (Taipei) and International Statistical Institute.

CHAPTER 12

STATISTICS IN BIOPHARMACEUTICAL RESEARCH

SHEIN-CHUNG CHOW

StatPlus, Inc. Heston Hall, Suite 206, 1790 Yardley-Langhorne Road, Yardley, PA 19067, USA
Tel: 215-321-1082; scchow@aol.com

ANNPEY PONG

Biostatistics, Forest Laboratories, Inc. 18th Floor, Haborside Financial Center, Plaza V, Jersey City, NJ 07311, USA
Tel: 201-4278318; Annpey.Pong@frx.com

1. Pharmaceutical Research and Development

In the process of research and development of a pharmaceutical entity, statistics are necessarily applied at various critical stages of the process to meet regulatory requirements for the effectiveness, safety, identity, strength, quality, purity, stability, and reproducibility of the pharmaceutical entity under investigation. A pharmaceutical entity could be a drug product, a biological product, a medical device, or a combination of a drug product, a biological product and a medical device. The critical stages of the process of pharmaceutical research and development include pre-IND (Investigational New Drug Application), IND, NDA (New Drug Application) and post-NDA. The role of statistics at these critical stages is briefly described below.

At the very early stage of pre-IND, pharmaceutical scientists may have to screen thousands of potential compounds in order to identify a few promising compounds. An appropriate use of statistics with efficient screening and/or optimal designs will assist pharmaceutical scientists to cost effectively identify the promising compounds within a relatively short period of time. As indicated by the United States Food and Drug Administration (FDA), an IND should contain information regarding chemistry,

manufacturing, and controls (CMC) of the drug substance and drug product to ensure the identity, strength, quality, and purity of the investigational drug. In addition, the sponsors are required to provide adequate information regarding pharmacological studies for absorption, distribution, metabolism, and excretion (ADME) and acute, subacute, and chronic toxicological studies and reproductive tests in various animal species to support that the investigational drug is reasonably safe to be evaluated in clinical trials in humans. At this stage, statistics are usually applied to (i) validate a developed analytical method, (ii) establish drug expiration dating period through stability studies, and (iii) assess toxicity through animal studies. Statistics are required to meet standards of accuracy and reliability.

Before the drug can be approved, the FDA requires that substantial evidence of the effectiveness and safety of the drug be provided in the Technical Section of Statistics of an NDA submission. Since the validity of statistical inference regarding the effectiveness and safety of the drug is always a concern, it is suggested that a careful review be performed to ensure an accurate and reliable assessment of the drug product. In addition, in order to have a fair assessment of the efficacy and safety of the investigational drug, the FDA also establishes advisory committees, each consisting of clinical experts, pharmacological experts, statistical experts, and one advocate (not employed by the FDA) in designated drug classes and specialties, to provide a second but independent review of the submission. The responsibility of the statistical expert is not only to ensure that a valid design is used but also to evaluate whether statistical methods used are appropriate for addressing the scientific and medical questions regarding the effectiveness and safety of the drug.

After the drug is approved, the FDA also requires that the drug product be tested for its identity, strength, quality, purity, and stability before it can be released for use. For this purpose, the current Good Manufacturing Practice (cGMP) is necessarily implemented to (i) validate the manufacturing process, (ii) monitor the performance of the manufacturing process, and (iii) provide quality assurance of the final product. At each stage of the manufacturing process, the FDA requires that sampling plans, acceptance criteria, and valid statistical analyses be performed for the intended tests such as potency, content uniformity, and dissolution.[69] For each test, sampling plan, acceptance criteria, and valid statistical analysis are crucial for determining whether the drug product pass the test based on the results from a representative sample.

In this chapter, we will not only introduce some key statistical concepts commonly encountered in pharmaceutical research and development, but also provide a comprehensive review of some important topics such as assay validation, stability design and analysis, individual bioequivalence, statistical principles for good clinical practice, and statistics in diagnostic imaging. Detailed information regarding the application of statistics at various critical stages during the process of pharmaceutical research and development can be found in Chow.[9]

2. Key Statistical Concepts

Key statistical concepts in the design and analysis of studies that are commonly conducted at various stages of pharmaceutical research and development are described below.

2.1. *Bias and variability*

For approval of a drug product, regulatory agencies usually require that the results of the studies conducted at various stages of drug research and development must be accurate and reliable to provide a valid and fair assessment of the treatment effect. The accuracy and reliability are usually referred to as the closeness and the degree of the closeness of the results to the true value (i.e. true treatment effect). Any deviation from the true value is considered a bias, which may be due to selection, observation, and statistical procedures. Pharmaceutical scientists should make any attempts to avoid bias whenever possible to ensure that the collected data are accurate. The reliability of a study is an assessment of the precision of the study, which measures the degree of the closeness of the results to the true value. The reliability reflects the ability to repeat or reproduce similar outcomes in the targeted population. The higher precision a study is, the more likely the results would be reproducible. The precision of a study can be characterized by the variability incurred during the conduct of the study.

In practice, since studies are usually planned, designed, executed, analyzed, and reported by a team consisting of pharmaceutical scientists from different disciplines, bias and variability inevitably occurs. It is then suggested that possible sources of bias and variability be identified at the planning stage of the study not only to reduce the bias but also to minimize the variability.

2.2. *Type I error, significance level, and power*

In statistical analysis, two different kinds of mistakes are commonly encountered when performing hypotheses testing. As an example, consider the example of pharmaceutical application. Suppose that a pharmaceutical company is interested in demonstrating that a newly developed drug is efficacious. The null hypothesis is often chosen as that the drug is inefficacious vs. the alternative hypothesis of that the drug is efficacious. The objective is to reject the null hypothesis and conclude the alternative hypothesis that the drug is efficacious. Under the null hypothesis, a type I error is made if we conclude that the drug is efficacious when in fact it is not. This error is also known as consumer's risk. The acceptable level of probability of committing type I error is known as the significance level. If the probability of observing type I error based on the data is less than the significance level, we conclude that a statistically significant result is observed. The probability of observing type I error is usually referred to as p-value of the test. Similarly, a type II error is committed if we conclude that the drug is inefficacious when in fact it is. This error is referred to as the producer's risk. The power is defined as the probability of correctly concluding that the drug is efficacious when in fact it is. For assessment of drug effectiveness and safety, a sufficient sample size is often selected to have a desired power with a pre-specified significance level. The purpose is to control both type I error (significance level) and type II error (power).

2.3. *Confounding and interaction*

In pharmaceutical research and development, there are many sources of variation, which have impact on the evaluation of the treatment. If these variations are not identified and properly controlled, then they may be mixed up with the treatment effect for which the studies are intended to demonstrate. In this case, the treatment effect is confounded with the effects due to these variations. Statistical interaction is to investigate whether the joint contribution of two or more factors is the same as the sum of the contributions from each factor when considered alone. If an interaction between factors exists, an overall assessment cannot be made. In practice, it is suggested that possible confounding factors be identified and properly controlled at the planning stage of the studies. When significant interactions among factors are observed, subgroup analyses may be necessary for a careful evaluation of the treatment effect.

2.4. *Randomization*

Statistical inference on a parameter of interest of a population under study is usually derived under the probability structure of the parameter. The probability structure depends upon the randomization method employed in sampling. The failure of the randomization will have a negative impact on the validity of the probability structure. Consequently, the validity, accuracy, and reliability of the resulting statistical inference of the parameter are questionable. Therefore, it is suggested that randomization be performed using appropriate randomization method under a valid randomization model according the study design to ensure the validity, accuracy, and reliability of the derived statistical inference. Details regarding various randomization models and methods that are commonly employed in clinical research can be found in Chow and Liu.[12]

2.5. *Sample size determination/justification*

One of the major objectives of most studies during drug research and development is to determine whether the drug is effective and safe. During the planning stage of a study, the following questions are of particular interest to the pharmaceutical scientists: (i) how many subjects are needed in order to have a desired power for detecting a meaningful difference, (ii) what is the trade off if only a small number of subjects are available for the study due to limited budget and/or some scientific considerations. To address these questions, a statistical evaluation for sample size determination/justification is often employed. Sample size determination is usually referred to the calculation of sample size for some desired statistical properties such as power or precision, while sample size justification is to provide statistical justification for a selected sample size, which is often a small number.

For a given study, sample size can be determined/justified based on some criteria on type I error (a desired precision) or type II error (a desired power). The disadvantage for sample size determination/justification based on the criteria of precision is that it may have a small chance of detecting a true difference. As a result, sample size determination/justification based on the criteria of power becomes the most commonly used method. Sample size is selected to have a desired power for detection of a meaningful difference at a pre-specified level of significance.

In practice, however, it is not uncommon to observe discrepancies among study objective (hypotheses), study design, statistical analysis (test

statistic) and sample size calculation. These inconsistencies often result in (i) wrong test for right hypotheses, (ii) right test for wrong hypotheses, (iii) wrong test for wrong hypotheses, or (iv) right test for right hypotheses with insufficient power. Therefore, before the sample size can be determined, it is suggested that the following be carefully considered; (i) the study objective or the hypotheses of interest be clearly stated, (ii) a valid design with appropriate statistical tests be used, and (iii) sample size be determined based on the test for the hypotheses of interest.

Note that procedures for sample size calculation based on a pre-study power analysis for comparing means, proportions, time-to-event data, and variabilities can be found in Chow, Shao and Wang.[21]

2.6. *Statistical difference and scientific difference*

A statistical difference is defined as a difference that is unlikely to occur by chance alone, while a scientific difference is referred to as a difference that is considered to be of scientific importance. A statistical difference is also referred to as a statistically significant difference. The difference between the concepts of statistical difference and scientific difference is that statistical difference involves chance (probability), while scientific difference does not. When we claim there is a statistical difference, the difference is reproducible with a high probability.

When conducting a study, basically, there are four possible outcomes. The result may show that (i) the difference is both statistically and scientifically significant, (ii) there is a statistically significant difference yet the difference is not scientifically significant, (iii) the difference is of scientifically significant yet it is not statistically significant, and (iv) the difference is neither statistically significant nor scientifically significant.

If the difference is both statistically and scientifically significant or it is neither statistically or scientifically significant, then there is no confusion. However, in many cases, a statistically significant difference does not agree with the scientifically significant difference. This inconsistence has created confusion/arguments among pharmaceutical scientists and biostatisticians. The inconsistence may be due to large variability and/or insufficient sample size.

2.7. *One-sided test versus two-sided test*

For evaluation of drug product, the null hypothesis of interest is often the one of no difference. The alternative hypothesis is usually the one that there

is a difference. Statistical test for this setting is called a two-sided test. In some cases, the pharmaceutical scientist may test the null hypothesis of no difference against the alternative hypothesis that the drug is superior to the placebo. Statistical test for this setting is known as one-sided test.

For a given study, if a two-sided test is employed at the significance level of 5%, then the level of proof required is one out of 40. In other words, at the 5% level of significance, there is 2.5% chance (or one out of 40) that we may reject the null hypothesis of no difference in the positive direction and conclude the drug is effective at one side. On the other hand, if a one-sided test is used, the level of proof required is one out of 20. It turns out that one-sided test allows more ineffective drugs to be approved because of chance as compared to the two-sided test. It should be noted that when testing at the 5% level of significance with 80% power, the sample size required increases by 27% for a two-sided test as compared to a one-sided test. As a result, there is a substantial cost saving if a one-sided test is used.

However, there is no universal agreement among the regulatory, academia, and the pharmaceutical industry as to whether a one-sided test or a two-sided test should be used. The FDA tends to oppose the use of a one-sided test though several pharmaceutical companies on the Drug Efficacy Study Implementation (DESI) drugs at the Administrative Hearing have challenged this position. Dubey[26] pointed out that several viewpoints that favor the use of one-sided test were discussed in an administrative hearing. These points indicated that one-sided test is appropriate in the following situations of (i) where there is truly only concern with outcomes in one tail and (ii) where it is completely inconceivable that the results could go in the opposite direction.

2.8. *Good Statistics Practice*

Good Statistics Practice (GSP) is defined as a set of statistical principles for the best pharmaceutical practices in design and analysis of studies conducted at various stages of drug research and development.[8] The purpose of GSP is not only to minimize bias but also to minimize variability that may occur before, during, and after the conduct of the studies. More importantly, GSP provides a valid and fair assessment of the drug product under study. The concept of GSP can be seen in many guidelines and guidance that issued by the FDA and the International Conference on Harmonization (ICH) at various stages of drug research and development. These guidelines and guidances include Good Laboratory Practice

(GLP), Good Clinical Practice (GCP), current Good Manufacturing Practice (cGMP), and Good Regulatory Practice (GRP). Another example of GSP is the guideline on *Statistical Principles in Clinical Trials* recently issued by the ICH.[42] As a result, GSP can not only provide accuracy and reliability of the results derived from the studies but also assure the validity and integrity of the studies.

The implementation of GSP in pharmaceutical research and development is a teamwork, which requires mutual communication, confidence, respect, and cooperation between statistician, pharmaceutical scientists in the related areas, and regulatory agents. The implementation of GSP involves some key factors that have an impact on the success of GSP. These factors include (i) regulatory requirements for statistics, (ii) the dissemination of the concept of statistics, (iii) an appropriate use of statistics, (iv) an effective communication and flexibility, (v) statistical training. These factors are briefly described below.

In the pharmaceutical development and approval process, regulatory requirements for statistics are the key to the implementation of GSP. They not only enforce the use of statistics but also establish standards for statistical evaluation of the drug products under investigation. An unbiased statistical evaluation helps pharmaceutical scientists and regulatory agents in determining (i) whether the drug product has the claimed effectiveness and safety for the intended disease, and (ii) whether the drug product possesses good drug characteristics such as the proper identity, strength, quality, purity, and stability.

In addition to regulatory requirements, it is always helpful to disseminate the concept of statistical principles described above whenever possible. It is important for pharmaceutical scientists and regulatory agents to recognize that (i) a valid statistical inference is necessary to provide a fair assessment with certain assurance regarding the uncertainty of the drug product under investigation, (ii) an invalid design and analysis may result in a misleading or wrong conclusion about the drug product, (iii) a larger sample size is often required to increase statistical power and precision of the studies. The dissemination of the concept of statistics is critical to establish the pharmaceutical scientists and regulatory agents' brief in statistics for scientific excellence.

One of the commonly encountered problems in drug research and development is the misuse or sometimes the abuse of statistics in some studies. The misuse or abuse of statistics is critical which may result in either having the right question with the wrong answer or having the right

answer for the wrong question. For example, for a given study, suppose that a right set of hypotheses (the right question) is established to reflect the study objective. A misused statistical test may provide a misleading or wrong answer to the right question. On the other hand, in many clinical trials, point hypotheses for equality (the wrong question) are often wrongly used for establishment of equivalency. In this case, we have right answer (for equality) for the wrong question. As a result, it is recommended that appropriate statistical methods be chosen to reflect the design, which should be able to address the scientific or medical questions regarding the intended study objectives for implementation of GSP.

Communication and flexibility are important factors to the success of GSP. Inefficient communication between statisticians and pharmaceutical scientists or regulatory agents may result in a misunderstanding of the intended study objectives and consequently an invalid design and/or inappropriate statistical methods. Thus, effective communications among statisticians, pharmaceutical scientists and regulatory agents is essential for the implementation of GSP. In addition, in many studies, the assumption of a statistical design or model may not be met due to the nature of drug product under investigation, experimental environment, and/or other causes related/unrelated to the studies. In this case, the traditional approach of doing everything by the book does not help. In practice, since the concerns from a pharmaceutical scientist or the regulatory agent may translate into a constraint for a valid statistical design and appropriate statistical analysis, it is suggested that a flexible and yet innovative solution be developed under the constraints for the implementation of GSP.

Since regulatory requirements for the drug development and approval process vary from drug to drug and country to country, various designs and/or statistical methods are often required for a valid assessment of a drug product. Therefore, it is suggested that statistical continued/advanced education and training programs be routinely held for both statisticians and non-statisticians including pharmaceutical scientists and regulatory agents. The purpose of such continued/advanced education and/or training program is threefold. First, it enhances communications within the statistical community. Statisticians can certainly benefit from such a training and/or educational program by acquiring more practical experience and knowledge. In addition, it provides the opportunity to share/exchange information, ideas and/or concepts regarding drug development between professional societies. Finally, it identifies critical practical and/or regulatory issues that are commonly encountered in drug development and regulatory approval

process. A panel discussion from different disciplines may result in some consensus to resolve the issues, which helps in establishing standards of statistical principles for implementation of GSP.

3. Pharmaceutical Validation

3.1. *Assay validation*

When a new pharmaceutical compound is discovered, the FDA requires that an analytical method or test procedure for determination of the active ingredients of the compound be developed and validated before it can be applied to animal and/or human subjects. The cGMP requires that test methods, which are used for assessing compliance of pharmaceutical products with established specifications, must meet proper standards of accuracy and reliability. The USP/NF defines the validation of analytical methods as the process by which it is established, in laboratory studies, that performance characteristics of the methods meet the requirement for the intended analytical application.

The analytical application may be referred to as a drug potency which is usually based on gas chromatography (GC) or high performance liquid chromatography (HPLC) for potency and stability studies, immunoassays such as radioimmunoassay (RIA) for the *in vitro* activity of an antibody or antigen, or a biological assay for the *in vivo* activity such as median effective dose (ED_{50}). The performance characteristics include accuracy, precision, limit of detection (LOD), limit of quantitation (LOQ), selectivity (or specificity), linearity, range, and ruggedness, which are useful measures for assessment of accuracy and reliability of the assay results. Among these performance characteristics, accuracy, precision, and ruggedness are considered the primary parameters for the validation of an analytical method.

For the validation of an analytical method, whether the analytical method can generate true values is often of great concern. To address this question, one may measure how close the assay result obtained by the analytical method is to the true value. This performance characteristic is referred to as the accuracy of the assay result. In practice, one may consider the analytical method to be validated in terms of accuracy if the mean value is within $\pm 15\%$ of the actual value, except at LOQ, where it should not deviate by more than 20%.[60] In addition, the precision, which is defined as the degree of agreement among individual assay results when the assay method is applied repeatedly to multiple sampling of a homogenous sample can be measured based on measurement error of the assay. Similarly,

Shah *et al.*[60] indicated that one may claim that the analytical method is validated if the precision around the mean value does not exceed a 15% coefficient of variation (CV), except for LOQ, where it should not exceed 20% CV.

In many cases, different analysts and different laboratories under different operating circumstances such as different instruments, different lots of reagents, different elapse time, or different assay temperatures may perform a specific analytical method. Assay ruggedness is often used to assess the influence of uncontrollable factors or the degree of reproducibility on assay performance. One may conclude that the analytical method is validated in terms of reproducibility if its assay ruggedness is within 15% of the mean value.

Accuracy is typically assessed using multiple testing by linear regression. Precision can be assessed by testing the null hypothesis that the variability is less than an acceptable limit. Typical approaches for assessing assay ruggedness include the one-way nested random effects model and the two-way crossed-classification mixed model. For the assessment of assay ruggedness, it should be noted, however, that the classical analysis of variance method may produce negative estimates for the variance components and that the sum of best estimates of variance components may not be the best estimate of the total variability. In these situations, methods proposed by Chow and Shao[14] and Chow and Tse[23] are useful. In practice, the validation of an analytical method can be carried out by the following steps: First, it is important to develop a prospective protocol which clearly states the validation design, sampling procedure, acceptance criteria for the performance characteristics to be evaluated, and how the validation is to be carried out. Second, collect the data and document the experiment, including any violations from the protocol that may occur. The data should be audited to assure their quality. The collected data are then analyzed based on appropriate statistical methods. Appropriate statistical methods are referred to as those methods, which can reflect the validation design and meet the study objective. Finally, draw a conclusion regarding whether the analytical method is validated based on the statistical inference drawn about the accuracy, precision, and ruggedness of the assay results.

3.2. *Process validation*

The objective of the validation of a manufacturing process is to ensure that the manufacturing process does what it purports to do. A validated

process assures that the final product has a high probability of meeting the standards for identity, strength, quality, purity, and stability of the drug product. A manufacturing process is a continuous process, which usually involves a number of critical stages. For example, for the manufacturing of tablets, the process may include initial blending, mill, primary blending, final blending, compression, and coating stages. At each critical stage, some problems may occur. For example, the ingredients may not be uniformly mixed at the primary blending stage; the segregation may occur at the final blending stage, and the weight of tablets may not be suitably controlled during the compression stage. In practice, therefore, it is important to evaluate the performance of the manufacturing at each critical stage by testing in process and/or processed materials for potency, dosage uniformity, dissolution, and disintegration according to sampling plans and acceptance criteria stated in the USP/NF. These tests are usually referred to as the USP tests. For sampling plans of USP tests, the USP/NF requires that representative samples be drawn from the container.

A manufacturing process is considered to pass the USP/NF tests if each critical stage of the manufacturing process and the final product meet the required USP/NF specifications for the identity, strength, quality, and purity of the drug product. A manufacturing process is considered validated if at least three validation batches (or lots) pass all required USP/NF tests. Since manufacturing procedures vary from drug product to drug product and/or from site to site during the development of a validation protocol of manufacturing process, it is important to discuss the issues such as (i) critical stages, (ii) equipment to be used at each critical stage, (iii) possible problems, (iv) USP tests to be performed, (v) sampling plans, (vi) testing plans, (vii) acceptance criteria, (viii) pertinent information, (ix) test or specification to be used as reference, and (x) validation summary with project scientists to acquire a good understanding of the manufacturing process.

Process validation usually refers to as the establishment of documented evidence that a process does what it purports to do. Basically, there are four different types of manufacturing process validations in the pharmaceutical industry: prospective, concurrent, retrospective, and re-validation. Prospective validation establishes documented evidence that a process does what it purports to do based on a preplanned protocol. Prospective validation is usually performed in the situations where (i) historical data are not available or sufficient and in-process and end-product testing data are not adequate, (ii) new equipment or components are used, (iii) a new product

is reformulated from an existing product, or there are significant modifications or changes in the manufacturing process, and (iv) the manufacturing process is transferred from development laboratory to full-scale production. Retrospective validation provides documented evidence based on review and analysis of historical information, which is useful when there is a stable process with a larger historical database. One of the objectives of the retrospective validation is to support the confidence of the process. Concurrent validation evaluates the process based on information generated during actual implementation of the process. In some situations where (i) a step of the process is modified, (ii) the product is made infrequently, and (iii) a new raw material must be introduced, a concurrent validation is recommended. In practice, a well-established manufacturing process may need to be revalidated when there are changes in critical components (e.g. raw materials), changes/replacement of equipment, changes in facility/plant (e.g. location or size), and a significant increase and/or decrease in batch size.

For a validated process, there is no guarantee that if the test is performed again it will have a high probability of meeting the specification. Thus, it is of interest to conduct some in-house acceptance limits (specifications), which guarantee that future batches produced by the process will pass the USP test with a high probability. A common approach to process validation is to obtain a single sample and test the attributes of interest to see whether the USP/NF specifications are met. Bergum[4] proposed constructing acceptance limits that guarantee that future samples from a batch will meet a given product specification a given percentage of times. The idea is to consider a multiple stage test. If the criteria for the first stage are met, the test is passed. If the criteria for the first stage are not met, then additional stages of testing are done. If the criteria at any stages are met, the test is passed. Acceptance limits for a validation sample are them constructed based on sample mean and standard deviation of the test results to assure that a future sample will have at least a certain chance of passing a multiple stage test. More details can be found in Chow and Liu.[10]

4. Stability Studies

4.1. *Drug shelf-life*

For every drug product in the marketplace, the FDA requires that an expiration dating period (or shelf-life) must be indicated on the immediate container label. The shelf-life is defined as the time interval at which the characteristics of a drug product (e.g. strength) will remain within

the approved specifications after manufacture. Along this line, Shao and Chow[63] studied several statistical procedures for estimation of drug shelf-life. Before a shelf-life of a drug product can be granted by the FDA, the manufacturers (drug companies) need to demonstrate that the average drug characteristics can meet the approved specifications during the claimed shelf-life period through a stability study.

For determination of the shelf life of a drug product, both the FDA stability guideline and the stability guideline issued by the ICH requires that a long term stability study be conducted to characterize the degradation of the drug product over a time period under appropriate storage conditions. Both the FDA and ICH stability guidelines suggest that stability testing be performed at 3-month intervals during the first year, 6-month intervals during the second year, and annually thereafter. The degradation curve can then be used to establish an expiration dating period or shelf life applicable to all future batches of the drug product.

For a single batch, the FDA stability guideline indicates that an acceptable approach for drug products that are expected to decrease with time is to determine the time at which the 95% one-sided lower confidence bound for the mean degradation curve intersects the acceptable lower product specification limit, e.g. as specified in the USP/NF.[29]

4.2. *Statistical model*

Consider the case where the drug characteristic is expected to decrease with time. The other case can be treated similarly. Assume that drug characteristic decreases over time linearly (i.e. the degradation curve is a straight line). In this case, the slope of the straight line is considered as the rate of stability loss of the product. Let X_j be the jth sampling (testing) time point (i.e. 0 months, 3 months, etc.) and Y_{ij} be the corresponding testing result of the ith batch ($j = l, \ldots, n; i = l, \ldots, k$). Then

$$Y_{ij} = \alpha_i + \beta_i X_j + e_{ij} \tag{1}$$

where e_{ij} are assumed to be independent and identically distributed (i.i.d.) random errors with mean 0 and variance σ_e^2. The total number of observations is $N = kn$. The α_i (intercepts) and β_i (slopes) vary randomly from batch to batch. It is assumed that $\alpha_i (i = l, \ldots, k)$ are i.i.d. with mean a and variance σ_a^2, and that $\beta_i (i = l, \ldots, k)$ are i.i.d. with mean b and variance σ_b^2. The e_{ij}, α_i, and β_i are mutually independent.

If $\sigma_a^2 = 0$ (i.e. α_i are equal), then the above model has a common intercept. Similarly, if $\sigma_b^2 = 0$ (i.e. β_i are equal), then the above model has

a common slope. If both $\sigma_a^2 = 0$ and $\sigma_b^2 = 0$, then there is no batch-to-batch variation and the above model reduces to a simple linear regression. Under the above model, Chow and Shao[15] proposed several statistical tests for batch-to-batch variation.

4.3. *Statistical methods*

4.3.1. *Fixed batches approach*

If there is no batch-to-batch variation, a commonly used method for fitting the above model is the ordinary least squares (OLS) and a 95% lower confidence bound for $E(Y) = a + b\xi$, the expected drug characteristic at time ξ, can be obtained as

$$\hat{a} + \hat{b}\xi - t_{0.95}S(\xi)\,,$$

where $\hat{a}$ and $\hat{b}$ are the OLS estimators of a and b, respectively, $t_{0.95}$ is the one-sided 95th percentile of the t distribution with $N-2$ degrees of freedom, and

$$S^2(\xi) = MSE\left\{\frac{1}{N} + \frac{(\xi - \bar{X})^2}{k\sum_{j=1}^{n}(X_j - \bar{X})^2}\right\},$$

where

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

and

$$MSE = \frac{1}{N-2}\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \hat{a} - \hat{b}X_j)^2\,.$$

The estimated shelf-life can be obtained by solving the following equation

$$\eta = \hat{a} + \hat{b}\xi - t_{0.95}S(\xi)\,,$$

where η is a given approved lower specification limit.

When there is a batch-to-batch variation (i.e. there are different intercepts and different slopes), the FDA recommends the minimum approach be used for estimation of the shelf-life of a drug product. The minimum approach considers the minimum of the estimated shelf-lives of the individual batches. The minimum approach, however, has received considerable criticisms because it lacks of statistical justification. As an alternative, Ruberg and Hsu[58] proposed an approach using the concept of multiple comparisons

to derive some criteria for pooling batches with the worst batch. The idea is to pool the batches that have slopes similar to the worst degradation rate with respect to a pre-determined similarity (equivalence) limit.

4.3.2. *Random batches approach*

As indicated in the FDA guideline, the batches used in long-term stability studies for establishment of drug shelf-life should constitute a random sample from the population of future production batches. In addition, all estimated shelf-lives should be applicable to all future batches. As a result, statistical methods based on random effects model seem more appropriate. In recent years, several methods for determination of drug shelf-life with random batches have been considered.[7,15,16,49,61] Under the assumption that batch is a random variable, stability data can be described by a linear regression model with random coefficients. Consider the following model

$$Y_{ij} = X'_{ij}\beta_i + e_{ij}\,,$$

where Y_{ij} is the jth assay result (percent of label claim) for the ith batch, X_{ij} is a pxl vector of the jth value of the regressor for the ith batch and X'_{ij} is its transpose, β_i is a pxl vector of random effects for the ith batch, and e_{ij} is the random error in observing Y_{ij}. Note that $X'_{ij}\beta_i$ is the mean drug characteristic for the ith batch at X_{ij} (conditional on β_i). The primary assumptions for the model are similar to those for model (1). Since X_{ij} is usually chosen to be x_j for all i, where x_j is a pxl vector of nonrandom covariate which could be of the form $(1, t_j, t_j w_j)'$ or $(1, t_j, w_j, t_j w_j)'$, where t_j is the jth time point and w_j is the jth value of qxl vector of nonrandom covariate (e.g. package type and dosage strength). Denote $x_j = x(t_j, w_j)$, where $x(t, w)$ is a known function of t and w. If there is no batch-to-batch variation, the average drug characteristic at time t is $x(t)'b$ and the true shelf-life is equal to

$$\bar{t}_{true} = \inf\{t : x(t)'b \le \eta\}\,,$$

which is an unknown but nonrandom quantity. The shelf-life is then given by

$$\bar{t} = \inf\{t : L(t) \le \eta\}\,,$$

where

$$L(t) = x(t)'\hat{b} - t_{\alpha,nk-p}\left[\frac{x(t)'(X'X)^{-1}x(t)}{k(nk-p)}\mathrm{SSR}\right]^{1/2}\,,$$

in where SSR is the usual sum of squared residuals from the ordinary least squares regression.

When there is batch-to-batch variation, t_{true} is random since β_i is random. Chow and Shao[16] and Shao and Chow[61] proposed considering an $(1-\alpha)\times 100\%$ lower confidence bound of the εth quantile of t_{true} as the labeled shelf-life, where ε is a given small positive constant. That is,

$$P\{t_{label} \le t_\varepsilon\} \ge 1-\alpha\,,$$

where t_ε satisfies

$$P\{t_{true} \le t_\varepsilon\} = \varepsilon\,.$$

It follows that

$$t_\varepsilon = \inf\{t : x(t)'b - \eta = z_\varepsilon \sigma(t)\}\,,$$

where $z_\varepsilon = \Phi^{-1}(1-\varepsilon)$ and $\sigma(t)$ is the standard deviation of $x(t)'\beta_i$. As a result, the shelf-life is given by

$$\bar{t} = \inf\{t : x(t)'\bar{b} \le \bar{\eta}(t)\}\,,$$

where

$$\bar{\eta}(t) = \eta + c_\kappa(\varepsilon,\alpha) z_\varepsilon \sqrt{v(t)}\,,$$

$$c_\kappa(\varepsilon,\alpha) = \frac{1}{\sqrt{k} z_\varepsilon} t_{\alpha,K-1,\sqrt{k}z_\varepsilon}\,,$$

$$v(t) = \frac{1}{k-1} x(t)'(X'X)^{-1}X'SX(X'X)^{-1}x(t)\,.$$

Note that $t_{\alpha,K-1,\sqrt{k}z_\varepsilon}$ is the αth upper quantile of the noncentral t distribution with $(k-1)$ degrees of freedom and noncentrality parameter $\sqrt{k}z_\varepsilon$.

4.4. *Two-phase shelf-life estimation*

Unlike most drug products, some drug products are required to be stored at several temperatures such as -20°C, 5°C and 25°C (room temperature) in order to maintain stability until use.[47] The drug products of this kind are usually referred to as frozen drug products. Unlike the usual drug products, a typical shelf life statement for frozen drug products usually consists of multiple phases with different storage temperatures. For example, a commonly adopted shelf life statement for frozen products could be either (i) 24 months at -20°C followed by 2 weeks at 5°C or two days at 25°C or (ii) 24 months at -20°C followed by 2 weeks at 5°C and one days at 25°C.

As a result, the drug shelf life is determined based on a two-phase stability study. The first phase stability study is to determine drug shelf-life under frozen storage condition such as $-20°$C, while the second phase stability study is to estimate drug shelf-life under refrigerated or ambient conditions. A first phase stability study is usually referred to as a frozen study and a second phase stability study is known as a thawed study.

Since the stability study of a frozen drug product consists of frozen and thawed studies, the determination of the shelf-life involves a two-phase linear regression. The frozen study is usually conducted similar to a regular long term stability study except the drug is stored at frozen condition. In other words, stability testing will be normally conducted at 3-month intervals during the first year, 6-month intervals during the second year, and annually thereafter. Stability testing for the thawed study is conducted followed by the stability testing for the frozen study, which may be performed at 2-day intervals up to two weeks. It should be noted that the stability at the second phase (i.e. thawed study) might depend upon the stability at the first phase (i.e. frozen study). In other words, an estimated shelf-life from the thawed study followed stability testing at 3-month of the frozen study may be longer than that obtained from the thawed study followed the frozen study at 6-month. For simplicity, Mellon[47] suggested that stability from the frozen study and the thawed study be analyzed separately to obtain a combined shelf life for the drug product. As an alternative, Shao and Chow[62] consider the following method for determination of drug shelf lives for the two phases based on a similar concept proposed before.[16,61]

For the first phase shelf-life, we have stability data

$$Y_{ik} = \alpha + \beta t_i + \varepsilon_{ik}\,,$$

where $i = 1, \ldots, I \geq 2$ (typically $t_i = 0$, 3, 6, 9, 12, 18 months), $k = 1, \ldots, K_i \geq 1$, α and β are unknown parameters, and $\varepsilon_{ik,}$'s are i.i.d. random errors with mean 0 and variance $\sigma_1^2 > 0$. The total number of data for the first phase is $n_1 = \sum_i K_i$ $(= IK$ if $K_i = K$ for all $i)$.

At time $t_i, K_{ij} > I$ second phase stability data are collected at time intervals $t_{ij}, j = 1, \ldots, J \geq 2$. The total number of data for the second phase is $n_2 = \sum_i \sum_j K_{ij}$ $(= IJK$ if $K_{ij} = K$ for all i and $j)$. Data from two phases are independent. Typically, $t_{ij} = t_i + s_j$, where $s_j = 1$, 2, 3 days, etc.

Let $\alpha(t)$ and $\beta(t)$ be the intercept and slope of the second phase degradation line at time t. Since the degradation lines for the two phases intersect,

$$\alpha(t) = \alpha + \beta t\,.$$

Then, at time $t_i, i = 1, \ldots, I$, we have stability data

$$Y_{ijk} = \alpha + \beta t_i + \beta(t_i)s_j + e_{ijk}\,,$$

where $\beta(t)$ is an unknown function of t and e_{ijk}'s are i.i.d. random errors with mean 0 and variance $\sigma_{2i}^2 > 0$.

We assume that $\beta(t)$ is a polynomial in t. Typically,

$$\beta(t) = \beta_0 \text{ Common slope model}\,,$$

$$\beta(t) = \beta_0 + \beta_1 t \text{ Linear trend model}\,,$$

or

$$\beta(t) = \beta_0 + \beta_1 t + \beta_2 t^2 \text{ Quadratic trend model}\,.$$

In general,

$$\beta(t) = \sum_{h=0}^{H} \beta_h t^h\,,$$

where β_h's are unknown parameters and $H + 1 < \sum_j K_{ij}$ for all i, and $H < I$.

4.4.1. *First phase shelf-life*

The first phase shelf-life can be determined based on the first phase data $\{Y_{ik}\}$ as the time point at which the lower product specification limit intersects the 95% lower confidence bound of the mean degradation curve.[29,40] Let $\hat{\alpha}$ and $\hat{\beta}$ be the least squares estimators of α and β, based on the first phase data, and let

$$L(t) = \hat{\alpha} + \hat{\beta}t - t_{.05;n_1-2}\sqrt{v(t)}$$

be the 95% lower confidence bound for $\alpha + \beta t$, where $t_{.05;n_1-2}$ is the upper 0.05 quantile of the t-distribution with $(n_1 - 2)$ degrees of freedom,

$$v(t) = \hat{\sigma}_1^2 \left[\frac{nt^2 - (2\sum_{i,k} t_i)t + \sum_{i,k} t_i^2}{n\sum_{i,k} t_i^2 - (\sum_{i,k} t_i)^2} \right],$$

and

$$\hat{\sigma}_1^2 = \frac{1}{n_1 - 2} \sum_{i,k} (Y_{ik} - \hat{\sigma} - \hat{\beta}t_i)^2$$

is the usual error variance estimator based on residuals. Suppose that the lower limit for the drug characteristic is η (we assume that $\alpha+\beta t$ decreases as t qincreases). Then the first phase shelf-life is the first solution of $L(t) = \eta$, i.e.

$$\hat{t} = \inf\{t : L(t) \leq \eta\}\,.$$

Note that the first phase shelf-life is constructed so that

$$P\{\hat{t} \leq \text{the true first phase shelf-life}\} = 95\%$$

assuming that e_{ik}'s are normally distributed. Without the normality assumption, result approximately holds for large n_1.

4.4.2. *The case of equal second phase slopes*

To introduce the idea, we first consider the simple case where the slopes of the second phase degradation lines are the same. When $\beta(t) \equiv \beta_0$, the common slope β_0 can be estimated by the least squares estimator based on the second phase data:

$$\hat{\beta}_0 = \frac{\sum_{i,j,k}(s_j - \bar{s})Y_{ijk}}{\sum_{i,j,k}(s_j - \bar{s})^2}\,,$$

where s_j is the second phase time intervals and $\bar{s}$ is the average of s_j's. The variance of $\hat{\beta}_0$ is

$$V(\bar{\beta}_0) = \frac{\sigma_2^2}{\sum_{i,j,k}(s_j - \bar{s})^2}\,,$$

which can be estimated by

$$\hat{V}(\hat{\beta}_0) = \frac{\hat{\sigma}_2^2}{\sum_{i,j,k}(s_j - \bar{s})^2}\,,$$

where

$$\hat{\sigma}_2^2 = \frac{1}{n_2 - I(H+2)} \sum_{i,j,k}(Y_{ijk} - (\hat{\sigma} + \hat{\beta}t_i) - \hat{\beta}_0 s_j)^2\,.$$

For fixed t and s, let

$$v(t,s) = v(t) + \hat{V}(\hat{\beta}_0)s^2$$

and

$$L(t,s) = \hat{\sigma} + \hat{\beta}t + \hat{\beta}_0 s - t_{.95;n_1+n_2-2-I(H+2)}\sqrt{v(t,s)}\,.$$

For any fixed t less than the first phase true shelf-life, i.e. t satisfying $\alpha + \beta t > \eta$, the second phase shelf-life can be estimated as

$$\hat{s}(t) = \inf\{s \geq 0 : L(t,s) \leq \eta\}$$

(if $L(t,s) < \eta$ for all s, then $\hat{s}(t) = 0$). That is, if the drug product is taken out of the first phase storage condition at time t, then the estimated second phase shelf-life is $\hat{s}(t)$.

The justification for $\hat{s}(t)$ is that for any t satisfying $\alpha + \beta t > \eta$,

$$P\{\hat{s}(t) \leq \text{the true second phase shelf-life}\} = 95\%$$

assuming that e_{ik}'s and e_{ijk}'s are normally distributed. Without the normality assumption, the above result approximately holds for large n_1, and n_2.

In practice the time at which the drug product is taken out of the first phase storage condition is unknown. In such a case we may apply the following method to assess the second phase shelf-life. Select a set of time intervals $t_l < \hat{t}$, $l = 1, \ldots, L$, and construct a table (or a figure) for $(t_l, \hat{s}(t_l))$, $l = 1, \ldots, L$. If a drug product is taken out of the first phase storage condition at time t_o which is between t_l and t_{l+1}, then its second phase shelf-life is $\hat{s}(t_{l+1})$.

However, a single shelf-life label may be required. We propose the following method.

4.4.3. *Determination of a single two-phase shelf-life label*

In most cases, $L(\hat{t}, s)$ is less than η for all s, i.e. $\hat{s}(\hat{t}) = 0$. Hence, we propose to select a $\hat{t}_1 < \hat{t}$ such that $\hat{s}(\hat{t}) > 0$ and use $\hat{t}_1 + \hat{s}(\hat{t}_1)$ as the two phase shelf-life label. The justification for this two-phase shelf-life label is:

1. If the drug product is stored under the first phase storage condition until time $\hat{t}_1$, then

$$P\{\hat{t}_1 \leq \text{the true first phase shelf-life}\} \geq 95\%\,,$$

since $\hat{t}_1 < \hat{t}$.

2. If the drug product is taken out of the first phase storage condition at time $\hat{t}_1 < \hat{t}$, then its estimated second phase shelf-life is $\hat{s}(\hat{t})$, and

$$\begin{aligned} &P\{\hat{s}(\hat{t}_1) \leq \text{the true second phase shelf-life at time } t_0\} \\ &\quad \geq P\{\hat{s}(\hat{t}_0) \leq \text{the true second phase shelf-life at time } t_0\} \\ &\quad \geq 95\% \end{aligned}$$

However, this two-phase shelf-life label is very conservative if t_0 is much less than $\hat{t}_1$.

A general rule of choosing $\hat{t}_1$ is that $\hat{t}_1$ should be close to $\hat{t}$ while $\hat{s}(\hat{t})$ is reasonably large. For example, if the units of the first and second phase shelf lives are month and day, respectively, and if $\hat{t} = 24.5$, then we can choose $\hat{t}_1 = 24$; if $\hat{t} = 24$, then we choose $\hat{t}_1 = 23$. A table on $(t_l, \hat{s}(t_l)), l = 1, \ldots, L$, will be useful for the selection of $\hat{t}_1$.

4.4.4. *The general case of unequal second phase slopes*

In general, the slope of the second phase degradation line varies with time. Let $\bar{Y}_i$ be the average of Y_{ijk}'s with a fixed i, $Z_{ijk} = Y_{ijk} - \bar{Y}_i$, and $X_{hij} = (s_j - \bar{s})t_i^h$. Then the least squares estimator of $(\beta_0, \ldots, \beta_H)$ denoted by $(\hat{\beta}_0, \ldots, \hat{\beta}_H)$, is the least squares estimator of the following linear regression model:

$$Z_{ijk} = \sum_{h=0}^{H} \beta_h X_{hij} + \text{error}\,.$$

Let

$$\hat{\beta}(t) = \sum_{h=0}^{H} \hat{\beta}_h t^h$$

and

$$\hat{V}(\hat{\beta}(t)) = \hat{\sigma}_2^2 \mathbf{1}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{1}\,,$$

where $\mathbf{1}' = (1, t, t^2 \cdots t^H)$, $\mathbf{X}$ is the design matrix and

$$\hat{\sigma}_2^2 = \frac{1}{n_2 - (H+2)} \sum_{i,j,k} \left(Z_{ijk} - \sum_{h=0}^{H} \hat{\beta}_h X_{hij} \right)^2 .$$

The second phase shelf-life and the two-phase shelf-life label can be determined in the same way as described in the previous section with

$$L(t, s) = \hat{\alpha} + \hat{\beta}t + \hat{\beta}(t)s - t_{.05;n_1+n_2-(H+4)} \sqrt{v(t, s)}$$

and

$$v(t, s) = v(t) + \hat{V}(\hat{\beta}(t))s^2\,.$$

For the proposed method for two-phase shelf-life estimation, assume that the assay variabilities are the same across different phases. Detailed information regarding two-phase shelf-life estimation can be found.[17,62] In practice, the assay variability may vary from phase to phase. In this case, the proposed method is necessarily modified for determination of the expiration dating period of the drug product.

In practice, it is of interest to determine the allocation of sample size at each phase. For a fixed total of sample size, it is of interest to examine the relative efficiency for estimation of shelf lives using either more sampling time points in the first phase and less sampling time points in the second phase or less sampling time points in the first phase and more sampling time points in the second phase. The allocation of sampling time points at each phase then becomes an interesting research topic for two-phase shelf-life estimation. In addition, since the degradation at the second phase is highly correlated with the degradation at the first phase, it may be of interest to examine such correlation for future design planning.

4.5. *Practical issues*

4.5.1. *Matrixing and bracketing designs*

For a new drug product, stability studies are necessarily conducted not only to characterize the degradation of the compound over time but also to determine the expiration dating period (shelf-life). The estimated shelf-life should be applicable for all strengths and packages of the drug product. However, accelerated stability testing is required for 6 months and long-term stability testing is required for the length of shelf-life. The cost of the stability studies could be substantial. As a result, it is of interest to adopt a design where only a fraction of the total number of samples are tested but at the same still maintain the validity, accuracy and precision of the estimated shelf-life. For this consideration, matrixing and bracketing designs have become increasingly popular in drug research and development for stability. As indicated in the ICH stability guideline, bracketing design is defined as the design of a stability schedule so that at any time point only the samples on the extremes, for example, of container size and/or dosage strengths, are tested.[41] Matrix design is a design where only a fraction of the total number of samples is tested at any specified sampling point.[41,51] The matrixing design and bracketing design were evaluated by Pong and Raghavarao.[54]

Lin[43] indicated that a matrixing design might be applicable to strength if there is no change in proportion of active ingredients, container size, and immediate sampling time points. The application of a matrixing design to situations such as closure systems, orientation of container during storage, packaging form, manufacturing process, and batch size should be evaluated carefully. It is discouraged to apply a matrixing design to sampling times at two endpoints (i.e. the initial and the last) and at any time points beyond the desired expiration date. If the drug product is sensitive to temperature, humidity, and light, the matrixing design should be avoided.

4.5.2. *Bias and interval estimation of shelf-life*

As indicated in the FDA stability guideline, the estimated shelf-life of a drug shelf-life can be obtained at the time point at which the 95% one-sided lower confidence limit for the mean degradation curve intersects the acceptable lower specification limit. In practice, it is of interest to study the biasedness of the estimated shelf-life. If the bias is positive, the estimated shelf-life overestimates the true shelf-life. On the other hand, if there is a downward bias, the estimated shelf-life is said to underestimate the true shelf-life. In the interest of the safety of the drug product, the FDA might prefer a conservative approach, which is to underestimate rather than overestimate the true shelf-life. Sun *et al.*[66] studied distribution properties of the estimated shelf-life[16,61] for both cases with and without batch-to-batch variation. The result indicate that when there is no batch-to-batch variation (i.e. $\sigma_a^2 = \sigma_b^2 = 0$), there is a downward bias which is given by

$$\frac{t_\alpha \sigma_e}{b^2}\left[\frac{b^2}{n} + \frac{(b\hat{X} + a - \eta)^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\right]^{1/2},$$

where t_α, is the $(1-\alpha)$th quantile of the t distribution with $(k-1)$ degrees of freedom.

4.5.3. *Shelf-life estimation with multiple active components*

For the study of drug stability, the FDA guideline requires that all drug characteristics be evaluated. In most drug products, we obtain an estimated drug shelf-life based primarily on the study of the stability of the strength of the active ingredient. However, some drug products may contain more than one active ingredient. For example, Premarin (conjugated estrogens, USP) contains three active ingredients: estrone, equilin, and 17a-dihydroequilin.

The specification limits for each component are different. To ensure identity, strength, quality, and purity, it is suggested that each component be evaluated separately for determination of drug shelf-life. In this case, although a similar concept can be applied, the method suggested in the FDA stability guideline is necessarily modified. It should be noted that the assay values observed from each component might not add up to a fixed total, which is due to the possible assay variability for each component. The modified model should be able to account for these sources of variation. Pong and Raghavarao[55] proposed a statistical method for estimation of drug shelf-life for drug products with two components. The distributions of shelf-life for two components were evaluated by Pong and Raghavarao[56] under different designs.

4.5.4. *Stability analysis with discrete responses*

For solid oral dosage forms such as tablets and capsules, the FDA stability guideline indicates that following characteristics should be studied in stability studies: (i) Tablets — appearance, friability, hardness, color, odor, moisture, strength, and dissolution, and (ii) capsules — strength, moisture, color, appearance, shape brittleness and dissolution. Some of these characteristics are measured based on discrete rating scale. As a result, the usual methods for stability analysis may not be appropriate. Chow and Shao[18] proposed some statistical methods for estimation of drug shelf-life based on discrete responses following the concept as described in the FDA stability guideline. However, it may be of interest to consider a mixture of a continuous response variable (e.g. strength) and a discrete response variable (e.g. color or hardness) for estimation of drug shelf-life. This requires further research.

5. Bioequivalence and Bioavailability

In pharmaceutical research and development, *in vivo* bioequivalence testing is usually considered a surrogate for assessment of clinical efficacy and safety. This is based on the so-called *Fundamental Bioequivalence Assumption* that when two formulations of the same drug product or two drug products (e.g. a brand-name drug and its generic copy) are equivalent in the rate and extent of drug absorption, it is assumed that they will reach the same therapeutic effect or they are therapeutically equivalent.[13] Pharmacokinetic (PK) responses such as area under the blood or plasma concentration-time curve (AUC) and maximum concentration (C_{max},) are

usually considered to assess the rate and extent of drug absorption. The current regulation of the FDA requires that the evidence of bioequivalence in average bioavailabilities in terms of some primary PK responses such as AUC and C_{max}, between the two formulations of the same drug product or the two drug products be provided.[28,31] This type of bioequivalence is usually referred to as average bioequivalence (ABE). Under current ABE criterion, however, it is not clear whether we are able to demonstrate that the absorption profiles of a brand-name drug and its generic copies are similar; consequently, it is not clear whether the brand name drug and its generic copies will have the same therapeutic effect in terms of efficacy and safety and hence can be used interchangeably.

In medical community, as more generic drug products become available in the marketplace, it is of great concern whether a number of generic drug products of the same brand-name drug can be used safely and interchangeably. Basically drug interchangeability can be classified as drug prescribability or drug switchability. Drug prescribability is defined as the physician's choice for prescribing an appropriate drug product for his/her new patients between a brand-name drug product and a number of generic drug products of the brand-name drug product, which have been shown to be bioequivalent to the brand-name drug product. The underlying assumption of drug prescribability is that the brand-name drug product and its generic copies can be used interchangeably in terms of the efficacy and safety of the drug product. Under current practice, the FDA only requires evidence of equivalence in average bioavailabilities be provided, the bioequivalence assessment does not take into account equivalence in variability of bioavailability. A relatively large intrasubject variability of a test drug product (e.g. a generic drug product) as compared to that of the reference drug product (e.g. its brand-name drug product) may present a safety concern. To overcome this disadvantage, in addition to providing evidence of ABE, it is recommended that bioequivalence in variability of bioavailabilities between drug products be established. This type of bioequivalence is called *population bioequivalence* (PBE). In practice, although PBE is often considered for assessment of drug prescribability, it does not fully address drug switchability due to possible existence of the subject-by-formulation interaction.

Drug switchability is related to the switch from a drug product (e.g. a brand-name drug product) to an alternative drug product (e.g. a generic copy of the brand-name drug product) within the same subject whose concentration of the drug product has been titrated to a steady, efficacious, and safe level. As a result, drug switchability is considered more critical than

drug prescribability in the study of drug interchangeability for patients who have been on medication for a while. To assure drug switchability, it is recommended that bioequivalence be assessed within individual subjects. This type of bioequivalence is known as *individual bioequivalence* (IBE). The concept of IBE has attracted FDA's attention since introduced by Anderson and Hauck,[1] which has led to a significant change in regulator consideration for assessment of bioequivalence.[32] In what follows, we will focus on the review of guidance on *Statistical Approaches to Establishing Bioequivalence*, which was recently issued by the FDA.[32]

5.1. *Limitations of average bioequivalence*

Under current FDA regulation, two formulations of the same drug or two drug products are said to be bioequivalent if the ratio of means of the primary PK responses such as AUC and C_{max} between the two formulations of the same drug or the two drug products is within (80%, 125%) with 90% assurance.[28,31] A generic drug product can serve as the substitute of its brand-name drug product if it has been shown to be bioequivalent to the brand-name drug. The FDA, however, does not indicate that a generic drug can be substituted by another generic drug even though both of the generic drugs have been shown to be bioequivalent to the same brand-name drug. Bioequivalence among generic copies of the same brand-name drug is not required. As more generic drugs become available in the marketplace, it is very likely that a patient may switch from one generic drug to another. Therefore, an interesting question to the physicians and the patients is whether the brand-name drug and its generic copies can be used safely and interchangeably.

Chen[6] pointed out that current ABE approach for bioequivalence assessment has limitations for addressing drug interchangeability especially for drug switchability. These limitations include (i) ABE focuses only on the comparison of population average between the test and reference drug products, (ii) ABE does not provide independent estimated of the intrasubject variances of the drug products under study, and (iii) ABE ignores the subject-by-formulation interaction, which may have an impact on drug switchability. As a result, Chen[6] suggested that current regulation of ABE be switched to the approach of PBE and IBE to overcome these disadvantages.

Chow and Liu[11] proposed to perform a meta-analysis for an overview of ABE. The proposed meta-analysis provides an assessment of bioequivalence among generic copies of a brand-name drug that can be used as a tool to

monitoring the performance of the approved generic copies of the brand-name drug. In addition, it provides more accurate estimates of intersubject and intrasubject variabilities of the drug product.

5.2. *Drug interchangeability*

As indicated earlier, drug interchangeability can be classified as drug prescribability or drug switchability. It is recommended that PBE and IBE be used to assess drug prescribability and drug switchability, respectively. More specifically, the FDA guidance recommends that PBE be applied to new formulations, additional strength, or new dosage forms in NDAS, while IBE should be considered for ANDA (abbreviated new drug application) or AADA (abbreviated antibiotic drug application) for generic drugs. In what follows, we will only focus on the concept, decision rule, and statistical method of IBE for assessment of drug interchangeability.

5.2.1. *Individual bioequivalence*

The individual bioequivalence is motivated by the 75/75 rule which claims bioequivalence if at least 75% of individual subject ratios (i.e. relative individual bioavailability of the generic drug product to the innovator drug product) are within (75%, 125%) limits. Along this line, Anderson and Hauck[1] first proposed the concept of testing for individual equivalence ratios (TIER). The idea is to test individual bioequivalence based on the dichotomization of continuous PK metrics by calculating the p value for at least the observed number of subjects who fall within bioequivalence limits with the minimum proportion of the population in which the two drug products must be equivalent in order to claim individual bioequivalence.

It should be noted that no universal definition of IBE exists which is uniformly accepted by researchers from the regulatory agency, the academia and the pharmaceutical industry. For example, IBE may be established based on the comparison between distributions within each subject or it could be based on the distribution of the difference or ratio within each subject.[45] In addition to average bioavailability and variability of bioavailability, we may also consider assessment for the variability due to the subject by formulation interaction. In this case, IBE can be assessed by means of a union-intersection test approach, which concludes IBE if and only if all of the hypotheses are rejected at a pre-specified level of significance. Most current methods for assessment of IBE, however, are derived from the distribution of either difference or ratio within each subject. Under

this setting, IBE can be classified as probability-based and moment-based according to different criteria for bioequivalence.[1,27,39,59,64]

To address drug switchability, the FDA proposed the following aggregated, scaled moment based one-sided criterion:

$$IBC = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2)}{\max(\sigma_{WR}^2, \sigma_{W0}^2)} \leq \theta_I \,,$$

where σ_{WT}^2 and σ_{WR}^2 are the within subject variances for the test drug product and the reference drug product, respectively, σ_D^2 is the variance due to subject-by-formulation interaction, σ_{W0}^2 is a constant which can be adjusted to control the probability of passing IBE, θ_I and is the bioequivalence limit. The FDA 2001 guidance suggests that θ_I be chosen as follows

$$\theta_I = \frac{(\ln 1.25)^2 + \varepsilon_I}{\sigma_{W0}^2} \,,$$

where ε_I is the variance allowance factor which can be adjusted for control sample size. As indicated in the FDA 2001 guidance, ε_I may be fixed between 0.04 and 0.05. For the determination of σ_{W0}^2, the FDA 2001 guidance recommends the use of individual difference ratio (IDR), which is defined as

$$\begin{aligned} IDR &= \left[\frac{E(T-R)^2}{E(R-R')^2}\right]^{1/2} \\ &= \left[\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 + \sigma_{WR}^2)}{2\sigma_{WR}^2}\right]^{1/2} \\ &= \left[\frac{IBC}{2} + 2\right]^{1/2} . \end{aligned}$$

Therefore, assuming that the maximum allowable IDR is 1.25, substitution of $(\ln 1.25)^2/\sigma_{W0}^2$ for IBC without adjustment of the variance term approximately yields $\sigma_{W0} = 0.2$.

The FDA 2001 guidance suggests that a mixed effects model in conjunction with the restricted maximum likelihood (REML) method be used to estimate variance components of $\sigma_D^2, \sigma_{Wt}^2$ and σ_{WR}^2. An intuitive statistical test can then be obtained by simply replacing the unknown parameters with their corresponding estimates. However, exact statistical properties of the resultant test are unknown. The FDA 2001 guidance recommends that the small sample method proposed by Hyslop *et al.*[38] be used to obtain the confidence interval or confidence bound of the test. If the upper 95% confidence bound is less than θ_I, we conclude IBE.

5.3. *A review of the FDA guidance on population/individual bioequivalence*

As indicated earlier, the FDA 2001 guidance on *Statistical Approaches to Establishing Bioequivalence* is intended to address drug interchangeability. As a result, the guidance for assessment of PBE and IBE has a significant impact on pharmaceutical research and development. In what follows, we provide a comprehensive review of the FDA 2001 guidance on population and individual bioequivalence from both scientific/statistical and practical points of view. Without loss of generality, we will only focus on IBE.

5.3.1. *Aggregated criteria vs. disaggregated criteria*

The FDA 2001 guidance recommends aggregated criteria as described earlier for assessment of IBE. The IBE criterion takes into account for average of bioavailability, variability of bioavailability, and the variability due to subject-by-formulation interaction. Under the proposed aggregated criteria, however, it is not clear whether IBE criterion is superior to ABE criterion for assessment of drug interchangeability. In other words, it is not clear whether or not IBE implies ABE under aggregate criteria. Hence, the question of particular interest to pharmaceutical scientists is that *whether the proposed aggregated criterion can really address drug interchangeability*?

Liu and Chow[45] suggested disaggregated criteria be implemented for assessment of drug interchangeability. The concept of disaggregated criteria for assessment of IBE is described below. In addition to ABE, we may consider the following hypotheses testing for equivalence in variability of bioavailabilities, and variability due to subject-by-formulation interaction:

$$H_0 : \sigma_{WT}^2/\sigma_{WR}^2 \geq \Delta_v$$

$$\text{vs. } H_a : \sigma_{WT}^2/\sigma_{WR}^2 < \Delta_v$$

and

$$H_0 : \sigma_D^2 \geq \Delta_s$$

$$\text{vs. } H_a : \sigma_D^2 < \Delta_s$$

where Δ_v is bioequivalence limit for the ratio of intrasubject variabilities and Δ_s is an acceptable limit for variability due to subject-by-formulation interaction. We conclude IBE if both $100(1-\alpha)\%$ upper confidence limit for $\sigma_{WT}^2/\sigma_{WR}^2$ is less than Δ_v and $100(1-\alpha)\%$ upper confidence limit for

σ_D^2 is less than Δ_s. Under the above disaggregated criteria, it is clear that IBE implies ABE.

In practice, it is of interest to examine the relative merits and disadvantages between the FDA recommended aggregated criteria and the disaggregated criteria described above for assessment of drug interchangeability. In addition, it is also of interest to compare the aggregated and disaggregated criteria of IBE with the current ABE criterion in terms of the consistencies and inconsistencies in concluding bioequivalence for regulatory approval.

5.3.2. *Masking effect*

The goal for evaluation of bioequivalence is to assess the similarity of the distributions of the PK metrics obtained either from the population or from individuals in the population. However, under the aggregated criteria, different combinations of values for the components of the aggregated criterion can yield the same value. In other words, bioequivalence can be reached by two totally different distributions of PK metrics. This is another artifact of the aggregated criteria. For example, at the 1996 Advisory Committee meeting, it was reported that the data sets from the FDA's files showed that a 14% increase in the average (ABE only allow 80% to 125%) is offset by a 48% decrease in the variability and the test passes IBE but fails ABE.

5.3.3. *Power and sample size determination*

For the proposed aggregated criterion, it is desirable to have sufficient statistical power to declare IBE if the value of the aggregated criterion is small. On the other hand, we would not want to declare IBE if the value is large. In other words, a desirable property for assessment of bioequivalence is that the power function of the statistical procedure is a monotone decreasing function. However, since different combinations of values of the components in the aggregated criteria may reach the same value, the power function for any statistical procedure based on the proposed aggregated criteria is not a monotone decreasing function. The experience for implementing the aggregated criteria in regulatory approval of generic drugs is lacking.

Another major concern is how the proposed criteria for IBE will affect the sample size determination based on power analysis. Unlike ABE, there exists no closed form for the power function of the proposed statistical procedure for IBE. As a result, the sample size may be determined through a Monte Carlo simulation study. Chow and Shao[17] provided formulas (based

on normal approximation) for sample size calculation for assessment of PBE and IBE under a 2×4 replicated crossover design. Sample sizes calculated from the formulas were shown to be consistent with those obtained from simulation studies.

5.3.4. *Two-stage test procedure*

To apply the proposed criteria for assessment of IBE, the FDA 2001 guidance suggests the constant scale be used if the observed estimator of σ_{TR} or σ_{WR} is smaller than σ_{T0} or σ_{W0}. However, statistically, the observed estimator of σ_{TR} or σ_{WR} being smaller than σ_{T0} or σ_{W0} does not mean that σ_{TR} or σ_{WR} is smaller than σ_{T0} or σ_{W0}. A test on the null hypothesis that σ_{TR} or σ_{WR} is smaller than σ_{T0} or σ_{W0} is necessarily performed. As a result, the proposed statistical procedure for assessment of IBE becomes a two-stage test procedure. It is then recommended that the overall type I error rate and the calculation of power be adjusted accordingly.

5.3.5. *Study design*

The FDA 2001 guidance recommends a 2×4 replicated designs, i.e. (TRTR, RTRT) be used for assessment of IBE without any scientific and/or statistical justification. As an alternative to the 2×4 replicated design, the FDA 2001 guidance indicates that a 2×3 replicated crossover design, i.e. (TRT, RTR) may be considered. Several questions are raised. First, it is not clear whether the two replicated crossover designs the optimal design (in terms of power) among all 2×4 and 2×3 replicated crossover designs with respect to the aggregated criterion? Second, it is not clear what is the relative efficiency of the two designs if the total number of observations is fixed. Third, it is not clear how these two designs compare to other 2×4 and 2×3 replicated designs such as (TRRT, RTTR) and (TTRR, RRTT) designs and (TRR, RTT) and (TTR, RRT) designs. Finally, it may be of interest to study the relative merits and disadvantages of these two designs as compared to other designs such as Latin square designs and four sequence and four period designs.

Other issues regarding the proposed replicated designs include (i) it will take longer time to complete, (ii) subject's compliance may be a concern, (iii) it is likely to have a higher dropout rate and missing values especially in 2×4 designs, and (iv) there are little literature on statistical methods dealing with dropouts and missing values in a replicated crossover design setting.

Note that the FDA 2001 guidance provides detailed statistical procedures for assessment of PBE and IBE under the recommended 2×4 replicated design. However, no details regarding statistical procedures for assessment of PBE and IBE under the alternative 2×3 replicated design are given. Detailed statistical procedures for assessment of PBE and IBE are available.[19,20] In addition, Chow and Shao[17] pointed out that the statistical procedure for assessment of PBE under the recommended 2×4 replicated design as described in the FDA 2001 guidance was inappropriate due to the violation of the primary assumption of independence.

5.4. *Outlier detection*

The procedure suggested for detection of outliers is not appropriate for the standard 2×2, the 2×3 or the 2×4 replicated crossover designs because the observed PK metrics from the same subject are correlated. For a valid statistical assessment, the procedures proposed by Chow and Tse[22] and Liu and Weng[46] should be used. These proposed statistical procedures for outlier detection in bioequivalence studies were derived under crossover designs, which incorporate the correlations within the same subject. The FDA 2001 guidance provides little or no discussion regarding the treatment of identified outliers.

6. Statistical Principles for Good Clinical Practice

For approval of a drug product, the FDA requires that substantial evidence of the effectiveness and safety of the drug product be provided through the conduct of two adequate and well-controlled clinical studies. To assist the sponsors in preparation of final clinical reports for regulatory submission and review, the FDA developed guidelines for the format and content of a clinical report in 1988. In addition, in 1994, the Committee for Proprietary Medicinal Products (CPMP) Working Party on Efficacy on Medicinal Products of the European Community issued a similar guideline entitled A *Note for Guidance on Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products*. At the same time, the ICH also signed off on the step 4 final draft of the *Structure and Content of Clinical Study Reports* and recommended its adoption to the three regulatory authorities of the United States, European Community, and Japan. The ICH guidelines require that some critical statistical issues be addressed in the final clinical report. These critical issues include baseline comparability, adjustments for covariates, dropouts or missing values,

interim analyses and data monitoring, multicenter studies, multiplicity, efficacy subsets, active control trials, and subgroup analyses, which are briefly described below (see also, Pong and Chow.[53]).

6.1. *Baseline comparability*

Baseline measurements are those collected during the baseline periods as defined in the protocol. Baseline usually refers to at randomization and prior to treatment. Sometimes, measurements obtained at screening are used as baselines. Basically, the objectives for analysis of baseline data are threefold. First, the analysis of baseline data is to provide a description of patient characteristics of the targeted population to which statistical inference is made. In addition, the analysis of baseline data provided useful information regarding whether the patients enrolled in the study are a representative sample of the targeted population according to the inclusion and exclusion criteria of the trial. Second, since baseline data measure the initial patient disease status, they can serve as reference values for the assessment of the primary efficacy and safety clinical endpoints evaluated after the administration of the treatment. Finally, the comparability between treatment groups can be assessed based on baseline data to determine potential covariates for statistical evaluations of treatment effects. The ICH guideline requires that baseline data on demographic variables such as age, gender, or race and some disease factors such as specific entry criteria, duration, stage and severity of disease and other clinical classifications and subgroups in common usage or of known prognostic significance be collected and presented.

The commonly employed statistical tests for baseline comparability are Cochran-Mantel-Henzsel test for categorical data and analysis of variance for continuous variables. Preliminary investigation of baseline comparability helps identifying possible confounding and interaction effects between treatment and baseline characteristics.

6.2. *Adjustments for covariates*

For assessment of the efficacy and safety of a drug product, it is not uncommon that the primary clinical endpoints are affected by some factors (or covariates) such as demographic variables, patient characteristics, concomitant medications, and medical history. If these covariates are known to have an impact on the clinical outcomes, one may consider stratified randomization. In practice, however, one may collect information on some covariates, which may influential and yet unknown at the planning stage

of the trial. In this case, if patients are randomly assigned to receive treatments, the estimated treatment effect is asymptotically free of the accidental bias induced by these covariates. If the covariate were balanced, then the difference in simple treatment averages would be an unbiased estimate for the treatment effect. On the other hand, if the covariate is not balanced, then the difference in simple average between treatment groups will be biased for estimation of the treatment effect. In this case, it is suggested that the covariates be included in the statistical model such as an analysis of variance (or covariance) model for an unbiased estimate of the treatment effect. In the case where covariates are balanced between the treatment groups, it is still necessary to adjust for covariates for clinical endpoints in order to obtain valid inference of the treatment effect if the covariates are statistically significantly correlated with the clinical endpoints.

The ICH guidelines require that selection of and adjustments for demographic or baseline measurements, concomitant therapy, or any other covariate or prognostic factor should be explained. In addition, methods of adjustments, results of analyses, and supportive information should be included in the detailed documentation of statistical methods.

6.3. *Dropouts or missing values*

In clinical research, there are many possible causes for the occurrence of dropouts and missing values. These possible causes include the duration of the study, the nature of the disease, the efficacy and adverse effects of the drug under study, intercurrent illness, accidents, patient refusal or moving, or other administrative reasons. The ICH guidelines suggest that the reasons for the dropouts, the time to dropout, and the proportion of dropouts among treatment groups be analyzed to examine the effects of dropouts for evaluation of the efficacy and safety of the study drug. Little and Rubin[44] classified missing values into three different types based on the possible causes. If the causes of missing values are independent of the observed responses, then the missing values are said to be completely random. On the other hand, if the causes of missing values are dependent on the observed responses but are independent of the scheduled but unobserved responses, then missing values are said to be random. The missing values are said to be informative if the causes of missing values are dependent upon the scheduled but unobserved measurements.

If missing mechanism is either completely random or random, then statistical inference derived from the likelihood approaches based on patients who complete the study is still valid. However, the inference is

not as efficient as it supposes to be. If the missing values were informative, then the inference based on the completers would be biased. As a result, it is suggested that despite the difficulty, the possible effects of dropouts and missing values on magnitude and direction of bias be expressed as fully as possible.

6.4. *Interim analysis and data monitoring*

Interim analysis and data monitoring are commonly employed for clinical trials in treatment of life-threatening disease or severely debilitating illness with long-term follow-up and endpoints such as mortality or irreversible morbidity. Interim analyses based on the data monitoring can be classified into formal interim analysis and administrative analysis. The aim of a formal interim analysis is to determine whether a decision for early termination can be reached before the planned study completion due to compelling evidence of beneficial effectiveness or harmful side effects. The administrative interim analysis is usually carried out without any intentions of early termination because of the results of the interim analysis results. Since interim analyses, either formally or informally, can introduce bias and/or increase type I error, the ICH guidelines require that all interim analyses, formal or informal, pre-planned or ad hoc, by any study participant, sponsor staff member, or data monitoring group should be described in full, even if the treatment groups were not identified. Data monitoring without code-breaking should also be described, even if this kind of monitoring is considered to cause no increase in type I error.

6.5. *Multicenter studies*

A multicenter trial is often conducted to expedite the patient recruitment process. The objective of the analysis of clinical data from a multicenter trial is two-fold. It is not only to investigate whether a consistent treatment effect can be observed across centers but also to provide an estimate of the overall treatment effect. A set of four conditions under which evidence from a single multicenter trial would provide sufficient statistical evidence of efficacy is proposed.[50]

Although all of the centers in multicenter trials follow the same protocol, many practical issues are likely to occur. For example, some centers may be too small for a reliable interpretation of the results, while some centers may be too big which dominate the results. In addition, there may be a significant treatment-by-center interaction. As a result, a statistical test

for homogeneity across centers is necessarily performed for detection of possible quantitative or qualitative treatment-by-center interaction. Gail and Simon[34] indicated that the existence of a quantitative interaction between treatment and center dose not invalidate the analysis by pooling data across centers. However, if a qualitative interaction between treatment and center is observed, an overall or average summary statistic may be misleading and hence considered inadequate. In this case, treatment effect should be carefully evaluated by center.

6.6. *Multiplicity*

In clinical trials, multiplicity may occur depending upon the objective of the intended trial, the nature of the design, and statistical analysis. The causes of multiplicity are mainly due to the formulation of statistical hypotheses and the experiment-wise false positive rates in subsequent analyses of the data. The ICH guidelines require that the overall type I error rate be adjusted to reflect multiplicity. Basically, multiplicity in clinical trials can be classified as repeated interim analyses, multiple comparisons, multiple endpoints, and subgroup analyses.

In the interest of an overall type I error rate, the commonly employed approach is probably the application of the Bonferroni technique. The concept of Bonferroni's technique is to adjust p values for control of experiment-wise type I error rate for pairwise comparisons. Bonferroni's method does not require that the structure of the correlation among comparisons be specified. In addition, it allows an unequal number of patients in each treatment group. Bonferroni's method works well when the number of treatment groups is small. When the number of treatment groups increases, however, Bonferroni's adjustment for p values becomes very conservative and may lack adequate power for the alternative in which most or all efficacy endpoints are improved. In this situation, as an alternative, one may consider a modified procedure proposed by Hochberg (1988).[37] Hochberg's procedure is shown to be more powerful because it only requires one p value smaller than α to declare one statistically significant comparison.

6.7. *Efficacy subsets*

In clinical trials, despite the fact that there is a thoughtful study protocol, deviation from the protocol may be encountered during the course of the trial. In addition, it is very likely that patients will withdraw from the

study prematurely before the completion of the trial due to various reasons. Patients who complete the study might miss some scheduled visits. As a result, which patients should be included in the analysis for a valid and unbiased assessment of the efficacy and safety of the treatment is a legitimate question to ask.

To provide a fair and unbiased assessment of the treatment effect, the ICH guideline suggests that the primary analysis for the demonstration of the efficacy and safety of the drug product should be conducted based on the intention-to-treat sample. In addition to the intention-to-treat sample, some subsets of the intention-to-treat sample may be constructed for efficacy analysis. These subsets are usually referred to efficacy subsets. These efficacy subsets include (i) patients with any efficacy observations or with a certain minimum number of observations, (ii) patients who complete the study, (iii) patients with an observation during a particular time window, and (iv) patients with a specified degree of compliance. The ICH guidelines require that efficacy subsets be analyzed to examine the effects of dropping patients with available data from analyses because of poor compliance, missed visits, ineligibility, or any other reasons. Any substantial differences resulting from the analyses of the intention-to-treat sample and the efficacy subsets should be the subject of explicit discussion.

6.8. *Active control trials*

An active control trial is often considered an alternative to placebo control study for evaluation of the effectiveness and safety of a test drug with very ill patients or patients with severe or life-threatening diseases based on ethical considerations. The primary objective of an active control trial could be to establish the efficacy of the test drug, to show that the test drug is equivalent to an active control agent, or to demonstrate that the test drug is superior to the active control agent. Pledger and Hall[52] pointed out that active control trials offer no direct evidence of effectiveness of the test drug. The only trial that will yield direct evidence of effectiveness of the test drug is a placebo-controlled trial, which compares the test drug with a placebo. Temple[68] indicated that if we cannot be very certain that the active control agent in a study would have beaten a placebo group, the fundamental assumption of the active control study cannot be made and that design must be considered inappropriate.

ICH guidelines indicated that if an active control study is intended to show equivalence between the test drug and an active control, the analysis

should show the confidence interval for the comparison between the two agents for critical endpoints and the relation of that interval to the prespecified degree of inferiority that would be consider unacceptable.

7. Statistics in Diagnostic Imaging

The techniques for evaluation of the performance of diagnostic medical products are very different from therapeutic pharmaceuticals and nondiagnostic devices. However, medical imaging drugs are generally governed by the same regulations as other drug and biological products. Because of the medical imaging drugs have special characteristics that do not reflect from other drug and biological products. The purpose of this section will focus on the different considerations for designs in diagnostic studies.

7.1. *Introduction*

Medical imaging drug products are drugs used with medical imaging methods (such as radiography, computed tomography [CT], ultrasonography [US], and magnetic resonance imaging [MRI]) to provide information on anatomy, physiology and pathology. The term "images" can be used as films, likenesses or other renderings of the body, body parts, organ systems, body functions, or tissues. For example, an image of the heart obtained with a diagnostic radiopharmaceutical or ultrasound contrast agent may in some cases refer to a set of images acquired from different views of the heart. Similarly, an image obtained with an MRI contrast agent may refer to a set of images acquired with different pulse sequences and interpluse delay times. In other words, medical imaging uses advanced technology to "see" the structure and function of the living body. The intentions of a medical imaging drug have two-fold: (i) delineate nonanatomic structures such as tumors or abscesses (ii) detect disease or pathology within an anatomic structure. Therefore, the indications for medical imaging drugs

Table 1. Most common used contrast drug products in combination with medical imaging devices.

Modality	Contrast Drug Products
X-Ray and CT	Iodine agents (photon scattering)
MRI	Gadolinium, dysprosium, helium
Ultrasound	Liposomes, microbubbles
Suspensions Nuclear	Tc-99rn, TI-201, indium, samarium

may fall within the following general categories. However, they need not be mutually exclusive:

a. Structure delineation — normal or abnormal;
b. Functional, physiological, or biochemical assessment;
c. Disease or pathology detection or assessment;
d. Diagnostic or therapeutic patient management.

The details of drug regulations are shown in the draft guidance to INDs, NDAs, biologics license applications (BLAs), ANDAs, and supplements to NDAs or BLAs for the medical imaging drug and biological products. This guidance was issued by FDA for industry entitled *Development Medical Imaging Drugs and Biologics*.[30] Usually, images are created from computerized acquisition of digital signals. The medical imaging drugs can be classified into contrast drug products and diagnostic radiopharmaceuticals.

7.1.1. *Contrast drug product*

Contrast drug products are used to increase the relative difference of signal intensities and to provide the additional information in combination with an imaging device beyond by the device alone. In other words, imaging with the contrast drug product should add value when compared to imaging without the contrast drug product.

7.1.2. *Diagnostic radiopharmaceuticals*

Radiopharmaceuticals are used for a wide variety of diagnostic, monitoring, and therapeutic purposes. Diagnostic Radiopharmaceuticals are used to image or otherwise identify an internal structure or disease process. In other words, diagnostic Radiopharmaceuticals are radioactive drugs that contain a radioactive nuclide that may be linked to a legend and carrier. These products are used in planar imaging, single photon emission computed tomography (SPECT), positron emission tomography (PET), or with other radiation detection probes.

7.2. *Design of blinded-reader studies*

In order to demonstrate efficacy of a medical imaging drug, readers who are both independent and blinded should perform evaluation of images. These independent, blinded image evaluations are intended to limit possible bias that could be introduced into the images evaluation by non-independent or unblinded readers. This evaluation is conducted in controlled setting

with minimal clinical information provided to the reader. The definitions of "independent" and "blinded" are defined next.

The independent readers are defined as those who have not participated studies and who are not affiliated with the sponsor or with institutions at which the studies were conducted. The meaning of blinding differs from the common way the term used in therapeutic clinical trials. Blinding in this sense is a critical aspect of clinical trials of medical imaging agents. "Blinded readers" are those who are unaware (1) of treatment identity used to obtain a given image and (2) of patient-specific clinical information or study protocol. For example, blinded readers should not have the knowledge about which images were obtained prior to drug administration and which were obtained after drug administration, although this may be apparent upon viewing the images. In addition, blinded readers should not know the patients' final diagnoses and may have limited or no knowledge of the results of other diagnostic tests that were performed on the patients. In some cases, blinded readers should not be familiar with the inclusion and exclusion criteria for patient selection that were specified in the protocol.

7.2.1. *Assessing reader agreement*

As indicated in the draft guidance,[30] at least two independent, blinded readers (and preferably three or more) are recommended for each study that is intended to demonstrate efficacy. The purpose is to provide a better basis for the findings in the studies. Therefore, the determination of inter-reader agreement and variability is the typical design issue to blinded read studies.

According to the guidance, the consistency among readers should be measured quantitatively. The most commonly used statistical test to assess the inter-reader agreement is the κ (kappa) statistic. The Cohen's kappa coefficient,[24] is a measure of inter-reader agreement in terms of count data.

For a 2×2 table,

$$\kappa = \frac{P_0 - P_e}{1 - P_e},$$

where

$$P_0 = \sum_i pii = \text{proportion of observed agreement}$$

$$P_e = \sum_{i,j} pi.p.j = \text{proportion of expected agreement}$$

It assumes that two response variables are two independent ratings of the n subjects. It should be noted that the kappa coefficient equals +1 when there is complete agreement of the readers. When the observed agreement exceeds chance agreement, kappa is positive. Also, the magnitude of kappa statistics reflects the strength of agreement. In a very unusual practice, kappa could be negative when the observed agreement is less than chance agreement. The total range of kappa is between −1 and 1. The asymptotic variance of simple kappa coefficient can be estimated by the following, according to Fleiss *et al.*[33]:

$$\text{var}(\kappa) = \frac{A + B + C}{(1 - P_e)_n^2},$$

where

$$A = \sum_i pii[1 - (pi. + p.j)(1 - \hat{\kappa})]^2,$$

$$B = (1 - \hat{\kappa})^2 \sum_{i \neq j} \sum_{i,j} pij(pi. + p.j)^2,$$

$$C = [\hat{\kappa} - P_e(1 - \hat{\kappa})]^2.$$

For measuring the inter-reader agreement in continuous data, Snedecor and Cochran proposed the intra-class correlation.[65]

7.3. *Diagnostic accuracy*

To determine how well a diagnostic imaging agent can distinguish disease subjects and non-diseased subjects, the outcome may often be classified into one of the four groups depending on (i) whether disease is present and (ii) the results of the diagnostic test of interest (positive or negative). The terms "positive" and "negative" concern some particular disease status, which must be specified clearly. The categories can be defined in any meaningful way to the problem. For example, patients could be classified as having one or more tumors (positive) or no tumor (negative), malignant (positive) or benign/no tumor (negative).

It should be noted that the disease is often determined with a "truth" standard or "gold" standard. A "truth" standard or "gold" standard is an independent method of measuring the same variable being measured by the investigational drug that is known or believed to give the truth state of a patient or true value of a measurement. In other words, "truth" standards are used to demonstrate that the results obtained with the medical imaging

Table 2. The typical outcome table (2 × 2) in the evaluation of a diagnostic test.

		Disease Status	
		Present	Absent
Diagnostic Test	Positive	a(TP)	b(FP)
	Negative	c(FN)	d(TN)

drug are valid and reliable. For example, for a MRI contrast agent intended to visualize the number of lesions in liver or determine whether a mass is malignant, the truth standard might include results from the pathology or long-term clinical outcomes. In diagnostic imaging studies, "truth" or "gold" standard are usually called as standard of reference (SOR). Possible choices of SOR in an imaging trail are:

a. Histopathology;
b. Therapeutic response;
c. Clinical outcome;
d. Another valid imaging procedure (validated against a valid gold standard);
e. Autopsy.

TP, FP, FN, TN represent the true positive, false positive, false negative, and true negative, respectively. After completing a well-defined classification based on the disease status and diagnostic test of interest, the efficacy of imaging agent can be expressed as the diagnostic performance of the agent.

The simplest measure of diagnostic decision is the fraction of cases for which the physician is correct, which is often called "accuracy". In other words, the accuracy is defined as the proportion of cases, considering both positive and negative test results, for which the test results are correct. It also can be expressed in mathematics as following:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d}\,.$$

However, accuracy is of limited usefulness as an index of diagnostic performance because two diagnostic modalities can yield equal accuracies but perform differently with respect to the types of decisions. Also, it can be affected by the disease prevalence strongly. Due to the limitation of the accuracy index, the sensitivity and specificity are used in the evaluation scheme.

$$\text{Sensitivity} = \frac{\text{Number of TP decisions}}{\text{Number of actually positive cases}} = \frac{a}{a+c}$$

$$\text{Specificity} = \frac{\text{Number of TN decisions}}{\text{Number of actually negative cases}} = \frac{d}{b+d}$$

In effect, sensitivity and specificity represents two kinds of accuracy: the first is for actually positive cases and the second is for actually negative cases. However, very often a single pair of sensitivity and specificity measurements may provide a possibly misleading and even hazardous oversimplification of accuracy.[70] This is how the ROC (Receiver Operating Characteristic) curve comes into picture and is introduced in Sec. 7.4.1. It should be noted that the method for evaluating and comparing sensitivity and specificity for diagnostic tests is based on:

Assumption 1: Diagnostic tests are independent given the disease status;
Assumption 2: The gold standard is error free.

These two assumptions are not always valid. Several statistical methods have been considered.[2,3,57]

7.4. *Statistical analysis*

Most of the imaging trials are designed to provide dichotomous or ordered categorical outcomes. Therefore, the statistical tests for proportions and rates are commonly used, and the methods based on ranks are often applied to ordinal data. The analyses based on odds ratios and the Mantel-Haenszel procedures are useful for data analysis. In addition, the use of model-based techniques, such as logistic regression models for binomial data, proportional odds models for ordinal data, and log-linear models for normal outcome variables are usually applied.

The diagnostic validity can be assessed in many ways. For example, the pre- and post-images can be compared to the gold standard, and the sensitivity and specificity of the pre-image compared to the post-image. Similarly, the same approaches can be used for two different active agents. The common methods used to test for differences in diagnosis are the McNemar test and Stuart-Maxwell test. The confidence intervals for sensitivity and specificity, and other measures can be also provided in the analysis.

Recently, the Receiver Operating Characteristic (ROC) analyses are becoming increasing important. Not only because it is recommended in the FDA draft guidance,[30] but also its advantage over more traditional measures of diagnostic performance.[48]

7.4.1. *Receiver operating characteristic (ROC) analyses*

In the use of most diagnostic test, test data do not necessarily fall into one of two obviously defined categories. Imaging studies usually require some confidence threshold be established in the mind of the decision maker. For example, if an image suggests the possibility of disease, how strong the suspicion is in order for the image to be called positive? Therefore, the decision maker chooses between positive and negative diagnosis by comparing his/her confidence concerning with an arbitrary confidence threshold. Figure 1 is an example of the model that underlies ROC analysis. The bell-shaped curves represent the probability density distributions of a decision maker's confidence in a positive diagnosis that arise from actually positive patients and actually negative patients.

The true positive fraction (TPF) is represented by the area under the left-hand distribution to the threshold. Similarly, the false positive fraction (FPF) is represented by the area under the left-hand distribution to the threshold. These imply that the sensitivity and specificity vary inversely as the confidence threshold is changed. In other words, TPF and FPF will increase or decrease together as the confidence threshold is changed.

If we change the decision threshold several times, we will obtain several different pairs of TPF and FPF. These pairs can be plotted as points on a graph, such as that in Fig. 2. This curve is called the ROC curve for

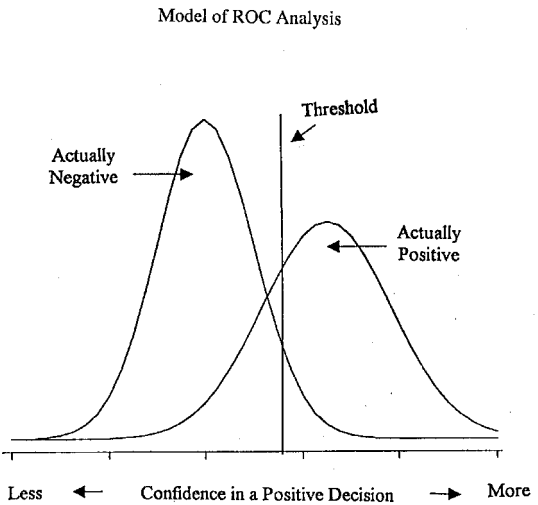


Fig. 1. Model of ROC Analysis.

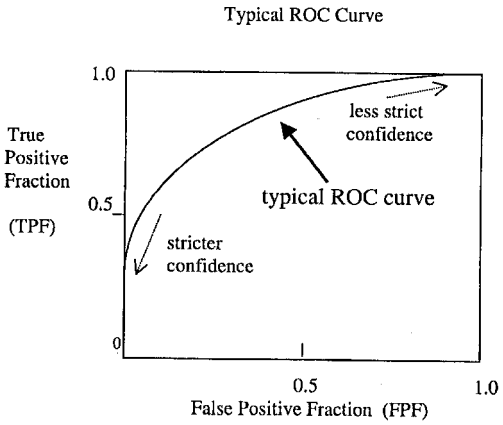


Fig. 2. Typical ROC Curve.

diagnostic test. Then, we may conclude that better performance is indicated by an ROC curve that is higher to the left in the ROC space.

A practical technique for generating response data that can be used to plot a ROC curve is called the *rating method*. This method requires the decision maker select a value from a continuous scale, such as definitely negative, probably negative, questionable, probabyly positive or definitely positive.

The advantages of the ROC curves are it is simple and graphical. Also, it is independent of prevalence and it provides a direct visual comparison between tests on a common scale. However, the drawbacks of the ROC curves are the decision thresholds and the numbers of subjects are usually not displayed on the graph. In addition, the appropriate software may not be widely available.

The ROC curve provides more information than just a single sensitivity and specificity pair to describe the accuracy of a diagnostic test. The curve depicts sensitivity and specificity levels over the entire range of decision thresholds. However, it would be helpful if the performance of a diagnostic test could be assessed by a single number. One such measurement that can be derived from the ROC curve is the area under the curve (AUC). If a diagnostic test that discriminates almost perfect, then its ROC curve passes near the upper left corner. This makes an AUC approaching 1. On the other hand, if the curve of a test that discriminates almost randomly, then the curve would lie near the 45 degree diagonal line. This would turn an AUC close to 0.5. The AUC range is between 0.5 and 1.

The AUC is calculated by summing the area of the trapezoids formed between the graph and the horizontal axis. This nonparametric method of calculation makes no assumptions regarding the underlying distributions of the diseased and non-diseased status. The meaning of AUC has been proved mathematically to be the probability that a random pair of positive/diseased and negative/non-diseased individuals would be identified correctly by the diagnostic test.[35] Also, it had been shown that the statistical properties of the Mann-Whitney-Wilcoxon statistics could be used to predict the statistical properties of AUC.[36] For comparing corrected ROC curves, Delong *et al.*[25] suggested a nonparametric approach for comparing the AUCs. For the parametric approach, Swets and Pickett[67] proposed a more exact method using the maximum likelihood estimation to estimate the AUC and its standard error. A comparison of nonparametric and binomial parametric areas can be found in Center and Schwartz.[5]

References

1. Anderson, S. and Hauck, W. W. (1990). Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **18**: 259–273.
2. Baker, S. G., Cannor, R. J. and Kessler, L. G. (1998). The partial testing design: A less costly way to test equivalence for sensitivity and specificity. *Statistics in Medicine* **17**: 2219–2232.
3. Baker, S. G. (1990). A simple EM algorithm for capture-recapture data with categorical covariates. *Biometrics* **46**: 1193–1200.
4. Bergum, J. S. (1990). Constructing acceptance limits for multiple stage tests. *Drug Development Industrial Pharmaceutical* **16**: 2153–2166.
5. Center, R. M. and Schwartz, J. S. (1985). An evaluation of methods for estimating the area under the receiver operating characteristics (ROC) curve. *Medical Decision Making* **5**: 149–156.
6. Chen, M. L. (1997). Individual bioequivalence — A regulatory update. *Journal of Biopharmaceutical Statistics* **7**: 5–11.
7. Chow, S. C. (1992). Statistical design and analysis of stability studies. Presented at the *48th Conferences of Applied Statistics*, Atlantic City, N.J., December.
8. Chow, S. C. (1997). Good statistics practice in the drug development and regulatory approval process. *Drug Information Journal* **31**: 1157–1166.
9. Chow, S. C. (2000). *Encyclopedia of Biopharmaceutical Statistics*. Marcel Dekker, Inc., New York.
10. Chow, S. C. and Liu, J. P. (1995). *Statistical Design and Analysis in Pharmaceutical Science*, Marcel Dekker, Inc., New York.
11. Chow, S. C. and Liu, J. P. (1997). Meta-analysis for bioequivalence review. *Journal of Biopharmaceutical Statistics* **7**: 97–111.

12. Chow, S. C. and Liu, J. P. (1998). *Design and Analysis of Clinical Trials*, Wiley, New York.
13. Chow, S. C. and Liu, J. P. (2000). *Design and Analysis of Bioavailability and Bioequivalence*, Marcel Dekker, Inc., New York.
14. Chow, S. C. and Shao, J. (1988). A new procedure for estimation of variance components. *Statistics and Probability Letters* **6**: 349–355.
15. Chow, S. C. and Shao, J. (1989). Test for batch-to-batch variation in stability analysis. *Statistics in Medicine* **8**: 883–890.
16. Chow, S. C. and Shao, J. (1991). Estimating drug shelf-life with random batches. *Biometrics* **47**: 1071–1079.
17. Chow, S. C. and Shao, J. (2002). *Statistics in Drug Research — Methodologies and Recent Development*. Marcel Dekker, Inc., New York.
18. Chow, S. C. and Shao, J. (2003). Stability analysis with discrete response. *Journal of Biopharmaceutical Statistics*. To appear.
19. Chow, S. C., Shao, J. and Wang, H. (2002a). Statistical tests for population bioequivalence. *Statistica Sinica*, In press.
20. Chow, S. C., Shao, J. and Wang, H. (2002b). Individual bioequivalence testing under 2×3 designs. *Statistics in Medicine* **21**: 629–648.
21. Chow, S. C., Shao, J. and Wang, H. (2003). *Sample Size Calculation in Clinical Research*. Marcel Dekker, Inc., New York, In press.
22. Chow, S. C. and Tse, S. K. (1990). Outliers detection in bioavailability/ bioequivalence, *Statistics in Medicine* **9**: 549–558.
23. Chow, S. C. and Tse, S. K. (1991). On variance estimation in assay validation. *Statistics in Medicine* **10**: 1543–1553.
24. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**: 37–46.
25. DeLong, E. R. and DeLong, D. M. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**: 837–845.
26. Dubey, S. D. (1991). Some thoughts on the one-sided and two-sided test. *Journal of Biopharmaceutical Statistics* **1**: 139–150.
27. Esinhart, J. D. and Chinchilli, V. M. (1994). Extension to the use of tolerance intervals for assessment of individual bioequivalence. *Journal of Biopharmaceutical Statistics* **4**: 39–52.
28. FDA (1992). Guidance on *Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*. Division of Bioequivalence, Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.
29. FDA (1993). Guideline for *Submitting Documentation for the Stability of Human Drugs and Biologics*. Center for Drugs and Biologics, Office of Drug Research and Review, Food and Drug Administration, Rockville, Maryland.
30. FDA (2000a). Draft Guidance on *Developing Medical Imaging Drugs and Biologics. Division of Medical Imaging and Radiopharmaceutical Drug Product*. Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

31. FDA (2000b). Guidance for industry: *Bioavailability and Bioequivalence Studies for Orally Administrated Drug Products — General Considerations*. Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.
32. FDA (2001). Guidance for industry: *Statistical Approaches to Establishing Bioequivalence*. Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.
33. Fleiss, J. L., Cohen, J. and Everitt, B. S. (1969). Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin* **72**: 323–327.
34. Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatments and patient subsets. *Biometrics* **41**: 361–372.
35. Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley, New York.
36. Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Diagnostic Radiology* **143**: 29–36.
37. Hochberg, Y. (1988). A sharper Bonferronis procedure for multiple test of significance. *Biometrika* **75**: 800–803.
38. Hyslop, T., Hsuan, F. and Holder, D. J. (2000). A small-sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* **19**: 2885–2897.
39. Holder, D. J. and Hsuan, F. (1993). Moment-based criteria for determining bioequivalence. *Biometrika* **80**: 835–846.
40. ICH (1993). *Stability Testing of New Drug Substances and Products*. Tripartite International Conference on Harmonization Guideline.
41. ICH (1994). ICH QIA. Guideline for industry. *Stability Testing Guideline of New Drug Substances and Products*. September.
42. ICH (1997). *Statistical Principles in Clinical Trials*. International Conference on Harmonization.
43. Lin, T. Y. D. (1994). Applicability of matrixing and bracketing approach to stability study design. Presented at the *4th ICSA Applied Statistics Symp.*, Rockville, Maryland.
44. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Values*. John Wiley and Sons, New York.
45. Liu, J. P. and Chow, S. C. (1997). Some thoughts on individual bioequivalence, *Journal of Biopharmaceutical Statistics* **7**: 41–48.
46. Liu, J. P. and Weng, C. S. (1992). Detection of outlying data in bioavailability/bioequivalence studies. *Statistics in Medicine* **10**: 1375–1389.
47. Mellon, J. I. (1991). Design and analysis aspects of drug stability studies when the product is stored at several temperatures. Presented at the *12th Annual Midwest Statistical Workshop*, Muncie, Indiana.
48. Metz, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology* **21**: 720–733.
49. Murphy, J. R. and Weisman, D. (1990). Using random slopes for estimating shelf-life. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 196–203.

50. Nevins, S. E. (1988). Assessment of evidence from a single multicenter trial. *Proceedings of the Biopharmaceutical Section of the American Statistical Association,* 43–45.
51. Nordbrock, E. (2000). Stability matrix design. In *Encyclopedia of Biopharmaceutical Statistics*. Ed. S. Chow, Marcel Dekker, Inc., New York.
52. Pledger, G. W. and Hall, D. (1986). Active control trials: Do they address the efficacy issue? *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 1–7.
53. Pong, A. and Chow, S. C. (1997). Statistical/practical issues in clinical trials. *Drug Information Journal* **31**: 1167–1174.
54. Pong, A. and Raghavarao, D. (2000). Comparison of bracketing and matrixing designs for a two-year stability study. *Journal of Biopharmaceutical Statistics* **10**: 217–228.
55. Pong, A. and Raghavarao, D. (2001). Shelf life estimation for drug products with two components. *Proceedings Biopharmaceutical Section of the American Statistical Association*.
56. Pong, A. and Raghavarao, D. (2002). Comparing distributions of drug shelf lives for two components under different designs. *Journal of Biopharmaceutical Statistics* **12**(3): 277–293.
57. Qu, Y. and Hadgu A. (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect test. *Journal of the American Statistical Association* **93**: 920–928.
58. Ruberg, S. and Hsu, J. (1992). Multiple comparison procedures for pooling batches in stability studies. *Technometrics* **34**: 465–472.
59. Schall, R. and Luus, R. E. (1993). On population and individual bioequivalence. *Statistics in Medicine* **12**: 1109–1124.
60. Shah, V. P., Midha, K. K., Dighe, S., McGilveray, I. J., Skelly, J. P., Yacobi, A., Layoff, T., Viswanathan, C. T., Cook, C. E., McDowall, R. D., Pittman, K. A. and Spector, S. (1992). Analytical methods validation: bioavailability, bioequivalence and pharmaceutical studies. *Pharmaceutical Research* **9**: 588–592.
61. Shao, J. and Chow, S. C. (1994). Statistical inference in stability analysis. *Biometrics* **50**: 753–763.
62. Shao, J. and Chow, S. C. (2001a). Two-phase shelf-life estimation. *Statistics in Medicine* **20**: 1239–1248.
63. Shao, J. and Chow, S. C. (2001b). Drug shelf life estimation. *Statistica Sinica* **11**: 737–745.
64. Sheiner, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine* **11**: 1777–1788.
65. Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th edn. Ames, The Iowa State University Press.
66. Sun, Y., Chow, S. C., Li, G. and Chen, K. W. (1999). Assessing distributions of estimated drug shelf lives in stability analysis. *Biometrics* **55**: 896–899.
67. Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic System*. Academic Press, NY.

68. Temple, R. (1983). Difficulties in evaluating positive control trials. *Proceedings of the Biopharmacentical Section of the American Statistical Association*, 1–7.
69. USP/NF (2000). *The United States Pharmacopeia 24 and the National Formulary 19*. United States Pharmacopeia Convention, Inc., Rockville, Maryland, USA.
70. Zweig, M. I. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**(4): 561–577.

About of Author

Shein-Chung Chow is President of the US Operations of StatPlus, Inc. directing the fuction/activity of technical consulting and biostatistics and data management for drug research and development services. Dr. Chow also provides technical supervision and guidance to project teams on statistical issues and presentations before clients, regulatory agencies or scientific bodies, defending the appropriateness of statistical methods used in clinical trial design or data analyses or the validity of reported statistical inferences. He identifies the best statistical and data management practices, organizes and leads working parties in the development of statistical design, analyses and presentation applications, and participates on Data Safety Monitoring Boards. He is also an Adjunct Professor at Temple University, Philadelphia, PA and an Adjunct Professor at National Cheng-Kung University, Taiwan. His professional activities include playing key roles in many professional organizations such as officer, Board of Directors member, Advisory Committee member, and Executive Committee member. He has served as Program-chair, session-chair/moderator, panelist and instructor/faculty at many professional conferences, symposia, workshops, tutorials and short courses. He is currently on the Editorial Boards of the *Journal of Biopharmaceutical Statistics*, *Statistica Sinica*, the *Journal of Food and Drug Analysis.* Dr. Chow is the Editor of Biostatistics Book series at Marcel Dekker, Inc., New York, NY. He was elected Fellow of the American Statistical Association in 1995 and an elected member of the ISI (International Statistical Institute) in 1999. He was the recipient of the DIA Outstanding Service Award (1996), ICSA Extraordinary Achievement Award (1996), and Chapter Service Recognition Award of the American Statistical Association (1998). Dr. Chow was appointed Pharmaceutical Scientific Advisor to the Bureau of Pharmaceutical Affairs, Department of Health,

Republic of China in 1999. Dr. Chow is past President of the International Chinese Statistical Association, Chair of the Advisory Committee on Chinese Pharmaceutical Affairs and a member of the Advisory Committee on Statistics of the DIA.

Dr. Chow is the author or co-author of over 120 professional papers, and co-author of the following books: *Advanced Linear Models*, *Design and Analysis of Bioavailability and Bioequivalence Studies* (1st and 2nd editions), *Statistical Design and Analysis in Pharmaceutical Science*, *Design and Analysis of Clinical Trials*, *Design and Analysis of Animal Studies in Pharmaceutical Development*, *and Encyclopedia of Biopharmaceutical Statistics*.

Dr. Chow received his BS in Mathematics from National Taiwan University, Taiwan, and PhD in Statistics from the University of Wisconsin, Madison, Wisconsin.

CHAPTER 13

STATISTICS IN TOXICOLOGY

JAMES J. CHEN

Division of Biometry and Risk Assessment,
National Center for Toxological Research,
US Food and Drug Administration,
Jefferson, AR 72079, USA
Tel: 870-543-7007; jchen@nctr.fda.gov

Toxicology is the study of the adverse effects of chemical substances on biological systems. Toxicological research is typically directed toward providing scientific information for the hazard potential of drugs and chemicals used by humans. Human epidemiology and animal toxicology are two major sources of scientific information for evaluation of toxic chemicals or drugs. Epidemiological studies, which attempt to associate disease or other adverse outcomes with an exposure, have the advantage of directly measuring an effect in humans at exposure conditions. Main limitations on the epidemiological studies are the lack of comprehensive data associated with unintentional or complex exposures, such as quantifying the actual dose concentration and no safety data for new drug or chemical products. Safety evaluation of the use of drug and chemicals are primarily based on animal studies in which animals are considered as surrogates for humans. *In Vitro* mutagenicity studies and structure-activity relationships may be used to support the interpretation of the information from the animal or human studies. In this chapter, we focus on two major toxicological studies: long-term carcinogenicity testing and reproductive testing.

Statistical analyses of various endpoints have been of two aspects: qualitative testing and quantitative estimation for risk assessment. The qualitative testing is to determine if the chemical cause an adverse health effect (if there is a statistically significant difference between treated and control groups. Statistical analysis discussed in this section focuses on the qualitative testing with respect to carcinogenic and reproductive endpoints. Statistical modeling for quantitative risk estimation is given in Chapter 11.

1. Animal Carcinogenicity Experiments

Long-term rodent bioassays have been the government's primary means of screening chemicals to assess carcinogenic potential to human risk. The United States Food and Drug Administration (FDA) and other countries require that new drugs and certain medical devices must be approved for safety and effectiveness for their intended use before being marketed. As a part of the drug approval process, the FDA requires that the sponsor submit the results of a rodent tumorigenicity bioassay to assess the carcinogenic potential of a drug for chronic use of humans. In the last 25 years the National Toxicology Program (NTP) has conducted about 500 long-term animal carcinogenesis bioassays for safety assessment of environmental compounds, and Food and Drug Administration (FDA) has reviewed hundreds of such studies of pharmaceuticals conducted by drug companies. Data from these studies have been a major database for safety assessment of compounds in the environment and industry.

A standard carcinogenic study is conducted in both sexes of two rodent species, typically rats and mice. A carcinogenicity experiment consists of a control and several dose groups. The *maximum tolerated dose* (MTD) has been used as the high-dose level. The MTD is defined as the dose that causes no more than a 10% body weight decrement, as compared to the appropriate control groups. The MTD is often estimated from the results of subchronic studies (generally three months of duration). Typically, dosage is measured in mg/kg body weight per day. The number of dose groups and allocation of animals among the dose groups depend on the objective of the study. A typical NTP carcinogenicity experiment consists of a control and three dose levels (0, 1/4 NTD, 1/2 MTD, MTD) with 50 animals per group. Animals are assigned randomly to dose groups or cages. As an example, consider a situation of 200 animals to be assigned to four groups of 50 with four animals from the same group caged together. Thus, 52 cages are used for the 200 animals. Each animal, first, is given to a number according to their order of presentation. A random number sequence of 52 cages numbers each with 4 replicates is, then, generated for placing animals in cages. For example, a sequence may be

Animal number	1	2	3	4	5	6	7	8	9	10	..
Random cage number	42	7	8	13	9	11	18	7	22	38	..

The animal #1 would be placed in cage #42, and animal #2 in cage #7, and so on. After randomization of the animals to cages (and into experimental

dose groups), the cage position may need to be rotated during the course of the experiment in order to balance the environmental effects.

The animals are given the test substance for a major portion of their lifespan. The test substance may be given in the diet or administered by other routes, such as inhalation, skin paints, or oral gavage. The experiment is terminated according to a predetermined stopping time, for example, 78–104 weeks for mice and 104 weeks for rats. Animal body weights and food consumption are measured weekly, the weeks of death of animals are recorded. Animals which die or are sacrificed are necropsied. Tissues taken from different organs and sites are examined microscopically for the presence of tumors for an evidence of carcinogenic effects.

One main objective of a long-term carcinogenicity experiment is to compare control and dose groups of animals with respect to tumor development. Statistical analysis of tumor responses includes the comparisons between dosed and control groups as well as a test for dose-related trend for each tumor site/organ. A typical experiment investigates approximate 20–50 tumor sites routinely. Because a large number of statistical tests are performed, the chance of false positive findings could increase. For example, the false positive rate is about 0.64 ($\approx 1-(1-0.05)^{20}$) for tests of 20 independent tumor types (sites/organs) all at the 0.05 significance level. For a particular tumor type, the primary response variable (endpoint) for comparison is the incidence of *first* tumors. One factor that affects the performance of methods is the animal survival time. A high degree of animal mortality will cause a significant censoring of the tumor response. Comparisons should be adjusted for the survival time because the crude incidence rate can be biased by the differential mortality (across groups). Another complication is that most tumor types are occult and therefore detectable only after the animal has died; that is, the time to the (first) tumor onset is not directly observable. This section will describe the commonly used statistical procedures for the analysis of animal tumor response data.

1.1. *Time-to-tumor model*

Kodell and Nelson[1] presented a tumor-death model which uses survival/sacrifice data to describe the sequence of events comprised by histological appearance of a tumor followed by death from that tumor. Three random variables can be used to describe the model:

X: The potential time to tumor onset, transition time from the normal state (N) to the tumor-bearing state (T).

T: The potential time from tumor onset to death, transition time from tumor state (T) to the death from the tumor state (D_T).
Z: The potential time until death from a competing cause, transition time from the normal state (N) or the tumor state (D_T) to the death from competing risk (D_{NT}).

Sacrificed animals are considered to be dead from a competing risk. The three random variables X, T, Z completely determined the fate of each animal. The two random variables Y and Z are the survival time of an animal, where $Y = X + T$ is the potential time until death from tumor. Note that X is not observable for the occult tumors.

A survival-adjusted method, that has been widely accepted, is to require that pathologists assign a "context of observation" (cause-of-death) to each tumor.[2] Tumors can be classified as "incidental", "fatal", and "mortality-independent (or observable)". Tumors that do not alter an animal's risk of death and are observed only as the result of a death from an unrelated cause are classified as an incidental context. Tumors that affect mortality by either directly causing death or indirectly increasing the risk of death are classified as a fatal context. Tumors, such as skin tumors, whose detection occurs at times other than when the animal dies are classified as a mortality-independent (or observable) context. It should be noted that the validity of context of observation is under the assumption: tumor-bearing and tumor-free animals of the same age have identical hazard functions for death unrelated to tumor.

In the context of observation, one of the four events will be observed on each animal:

A. Appearance of a visible tumor (mortality-independent context, X is observable).
B. Animal died from the tumor of interest (fatal context, $Y < Z$).
C. Animal had a tumor and died from competing cause (incidental context, $X \leq Z < Y$).
D. Animal did not have a tumor and died from a competing cause ($Z \leq X$).

Let $t_1, t_2, \ldots, t_m$ be the distinct times at which the above events are observed, and a_k, b_k, c_k, and d_k, $k = 1, \ldots, m$, are the number of events of A, B, C, and D at time t_k, respectively. Define the tumor resistance (survival) functions for X and Y as $S_X(t) = \Pr(X \geq t)$, and $S_Y(t) = \Pr(Y \geq t)$. Let $f_X(t)$ and $f_Y(t)$ be the density function of X and Y, respectively. For the tumors observed in a mortality-independent context, the likelihood function

is given as

$$L_a = \prod f_X(t_k)^{a_k} S_X(t_k)^{d_k} .$$

The likelihood function for the tumors observed in a fatal context is

$$L_b = \prod f_Y(t_k)^{b_k} S_Y(t_k)^{d_k} .$$

The likelihood functions L_a and L_b are essentially the same. The likelihood function for the tumors observed in an incidental context is

$$L_c = \prod [1 - S_X(t_k)]^{c_k} S_X(t_k)^{d_k} .$$

In the general case, when a tumor is observed in a fatal cases for some animals and is also observed in an incidental context for other animals, the likelihood function is

$$L_d = \prod [f_Y(t_k)]^{b_k} [S_Y(t_k) - S_X(t_k)]^{c_k} [S_X(t_k)]^{d_k} .$$

Kodell *et al.*[3] showed that

$$S_Y(t) - S_X(t) = [1 - Q(t)] S_Y(t) ,$$

where $Q(t) = S_X(t)/S_Y(t)$ is the conditional probability of tumor onset after time t, given tumor-free survival through time t. It follows that

$$L_d = \prod [f_Y(t_k)]^{b_k} [S_Y(t_k)]^{c_k + d_k} [1 - Q(t_k)]^{c_k} Q(t_k)^{d_k} .$$

That is, L_d can be expressed as the product of the two likelihood functions

$$L_d^b = \prod [f_Y(t_k)]^{b_k} [S_Y(t_k)]^{c_k + d_k} ,$$

and

$$L_d^c = \prod [1 - Q(t_k)]^{c_k} Q(t_k)^{d_k} .$$

The L_d^b and L_d^c represent the contributions of the fatal and incidental tumors, respectively.

1.2. *Estimation*

An important first step in the evaluation of animal carcinogenicity data is to estimate the animal survival curve for the assessment of any effects of exposure to the test compound on mortality. The survival curve for each dose group is calculated by the Kaplan-Meier method.[4] In this calculation, the weeks of death for animals killed accidentally or sacrificed are considered as censored observations.[5] For a given group, suppose that the death time of the animals are observed at $t_k, k = 1, \ldots, m$. Let n_k denote the number

of animals that died at or after t_k (the number of animals at risk), and x_k denote the number of deaths (out of n_k). The Kaplan-Meier estimate of the conditional probability of survival beyond t_k given survival beyond $t_{(k-1)}$ is $(n_k - x_k)/n_k$. The estimated survival function is

$$\hat{S}(t) = \prod_{t_k \le t} \frac{n_k - x_k}{n_k}, \quad t_k \le t < t_{(k+1)}.$$

The variance of the $\hat{S}(t)$ is calculated by Greenwood's formula

$$V[\hat{S}(t)] = \hat{S}^2(t) \sum_{t_k \le t} \frac{x_k}{n_k(n_k - x_k)} t_k \le t \le t_{(k+1)}.$$

The estimation of the tumor survival functions $S_X(t)$ and $S_Y(t)$ depends on the context of observation. When all tumors are observed in a mortality-independent context or fatal context, the Kaplan-Meier method can be used to estimate the tumor survival function. The calculation is the same as estimating animal survival function. But, the n_k represents the number of animals at risk (have not developed a tumor), and x_k is the number of tumor observed (or death caused by the tumor).

For the tumors observed in an incidental context, these tumors are only discovered at necropsy, either after sacrifice or after has died from the cause unrelated to the presence of tumor. Let the experiment period be partitioned into J sub-interval such that $(t_{j-1}, t_j], j = 1, \ldots, J$, where $t_0 = 0$ and t_J denotes the time at which the terminal sacrifice is scheduled. Let c_j and d_j denote the number of animals that died in the jth time interval for which the tumor is present or absent, respectively. The total number of deaths in the jth time interval is $(c_j + d_j)$. Hoel and Walburg[6] proposed the tumor prevalence estimate as

$$\hat{R}(t) = \frac{c_j}{c_j + d_j}, \quad t_{(j-1)} < t \le t_j.$$

The prevalence method requires to partition the experimental period into several time intervals. The following three partitions have been used in practice: (1) (0, 50], (51, 80], (81, 104], interim sacrifice (if any), and terminal sacrifice; (2) (0, 52], (53, 78], (79, 92], (93, 104], interim sacrifice (if any), and terminal sacrifice; and (3) the Peto *ad hoc* interval determined by the tumor prevalence data based on the assumption of non-decreasing prevalence function.[2] The maximum likelihood estimate of $R(t)$ is estimated by the "pooling adjacent violators" method.

For the tumors observed in both incidental and fatal contexts, the maximum likelihood estimate of $S_X(t)$ can be obtained by estimating $S_Y(t)$ and

$Q(t)$ separately, provided that $Q(t)$ is monotonically nonincreasing. Consequently, the $S_Y(t)$ and $Q(t)$ are estimated by the Kaplan-Meier estimator and Hoel-Walburg estimator described above. The estimate of tumor onset distribution[3] is

$$\hat{S}_X(t) = \left[\prod_{t_k \le t} \frac{n_k - b_k}{n_k}\right] \left(\frac{c_j}{c_j + d_j}\right), \quad t_j - 1 \le t < t_j \,.$$

The variance and the variance of $\hat{S}_X(t)$ is obtained using a first-order Taylor series,

$$\text{var}[\hat{S}_X(t)] \simeq [\hat{S}_X(t)]^2 \left[\sum_{t_k \le t} \frac{x_k}{n_k(n_k - x_k)} + \frac{c_j}{(c_j + d_j)d_j}\right].$$

1.3. *Testing*

First, consider testing for difference in animal survival functions, or difference in the incidence of tumors observed in a mortality-independent context or fatal context. The logrank test or the death-rate method[2] is the most widely used procedure for testing the age-specific differences among groups.

Consider a carcinogenicity experiment with g groups $(d_1, \ldots, d_g)$. Taking all dose groups together as one, suppose that the death times are observed at t_k, a time point at which tumors are found in any group, $k = 1, \ldots, m$. Let n_{ik} denote the number of animals in the ith group that died at or after t_{ik} (at risk), and x_{ik} denote the number of deaths (out of n_{ik}). Animal death-tumor data at time t_k can be summarized in a $2 \times g$ table as

Summary of animal tumor-death data at t_k.

Dose	d_1	d_2	$\cdots$	d_g	Total
# With Tumors	x_{1k}	x_{2k}	$\cdots$	x_{gk}	$x_{.k}$
# At Risk	n_{1k}	n_{2k}	$\cdots$	n_{gk}	$n_{.k}$

The expected number of tumors in the ith group at time t_k is $e_{ik} = x_{.k} f_{ik}$, where $f_{ik} = n_{ik}/n_{.k}$, $i = 1, \ldots, g$. Thus, the observed and expected numbers of tumors in the ith group over the entire experiment are $O_i = \sum_{k=1}^{m} x_{ik}$ and $E_i = \sum_{k=1}^{m} e_{ik}$, respectively. Define

$$D_i = O_i - E_i = \sum_{k=1}^{m} (x_{ik} - e_{ik})$$

and

$$V_{rs} = \sum_{k=1}^{m} \frac{x_{.k}(n_{.k} - x_{.k}) f_{rk}(\delta_{rs} - f_{sk})}{n_{.k} - 1}$$

where δ_{rs} is defined as 1 if $r = s$ and 0 otherwise. Let $\boldsymbol{D_a} = (D_1, \ldots, D_g)'$ and $\boldsymbol{V_a}$ be the $g \times g$ matrix with the (r, s) entry V_{rs}. Then

$$X_a = \boldsymbol{D_a'} \boldsymbol{V_a^-} \boldsymbol{D_a}$$

can serve as a test for heterogeneity among the g groups, where $\boldsymbol{V_a^-}$ is a generalized inverse of $\boldsymbol{V_a}$. Under the null hypothesis, X_H is asymptotically distributed as χ^2 distribution with $g - 1$ degrees of freedom. Also, a dose-related trend test can be considered by using

$$Z_a = \boldsymbol{l'D_a} / \sqrt{\boldsymbol{l'V_a l}}\,,$$

where $\boldsymbol{l} = (d_1, \ldots, d_g)'$.

For the tumors observed in an incidental context, the Mantel and Haenszel[7] test or the prevalence method[2] can be used for comparing the prevalence rates among groups. The prevalence method used is very similar to the death-rate method, except that each interval defines the tumor-death time as described in the estimation. The vector of the differences of observed and expected values $\boldsymbol{D_b}$ is calculated the same way as described for the fatal tumors, and the corresponding covariance matrix V_b is computed. The χ^2 test statistics for heterogeneity and trend can be calculated similarly.

For the tumors observed in both fatal and incidental and contexts, the data for the fatal tumors and for the incidental tumors are analyzed separately by the death-rate and prevalence methods, respectively. The test for the difference in the time of tumor onset is based on the pooled vector $\boldsymbol{D} = \boldsymbol{D_a} + \boldsymbol{D_b}$, with covariance matrix $\boldsymbol{V} = \boldsymbol{V_a} + \boldsymbol{V_b}$.[2] The test statistic for heterogeneity is given by

$$X = (\boldsymbol{D_a} + \boldsymbol{D_b})'(\boldsymbol{V_a} + \boldsymbol{V_b^-}(\boldsymbol{D_a} + \boldsymbol{D_b}))\,.$$

The trend test is given by

$$Z = \boldsymbol{l'}(\boldsymbol{D_a} + \boldsymbol{D_b}) / \sqrt{\boldsymbol{l'}(\boldsymbol{V_a} + \boldsymbol{V_b})\boldsymbol{l}}\,.$$

The choice of time intervals for calculating the incidental tumor component of the test of Peto *et al.*[2] is an important consideration. The use of the *ad hoc* time intervals can be problematic.[8] Moreover, the procedures described above for the analysis of occult tumor all require the information of tumor lethality or cause of death. Some argue that the determination

whether a tumor causes an animal's death is a rather complicate and subjective process. It is often difficult for a pathologist to classify accurately and objectively a tumor type as straight causing or not causing animal's death. Furthermore, the validity of context of observation relies on the assumption that tumor-bearing and tumor-free animals of the same age have identical hazard functions for death unrelated to tumor. Chen and Moore[9] showed the Peto test performs poorly when there is a large reduction in survival times in the dosed groups.

1.4. *Other methods*

Dinse and Lagakos[10] proposed a logistic regression model as an alternative prevalence test for the incidental tumors,

$$\exp(\mu + \tau t + \theta d_i)/[1 + \exp(\mu + \tau t + \theta d_i)]\,. \tag{1}$$

They derived the likelihood score test of $\theta = 0$. The logistic regression analysis assumes that tumor prevalence is a smooth function of ages and it does not require the choice of time intervals. In addition, the logistic regression model can easily incorporate other covariates, and the software is ready available. The logistic regression has been adopted by the NTP as a standard analysis for dose-related trend test.

Bailer and Portier[11] proposed an alternative survival-adjusted approach that do not require the cause of death information. The approach modifies the Cochran-Armitage test to account for the survival times of those animals that die prior to study termination without tumor presence. The Bailer and Portier approach, has been referred to as the Poly-κ test, can be used to replace the Peto's procedure when the cause of death information is not available.

Let y_{ij} be a binary response indicating presence or absence of a tumor type of the jth animal in the ith group who dies at time t_{ij}, $i = 1, \ldots, g$, $j = 1, \ldots, n_i$. If all animals survive during the whole experiment period, the probability of developing a tumor for a animal in the ith group, say, μ_i can be modelled by the linear-logistic model

$$\text{logit}(\mu_i) = \alpha + \beta d_i\,.$$

Let T denote the length of the experiment and t_{ij} denote the death time of the ijth animal. Bailer and Portier[11] defined a weight equal to 1 if a tumor is present at death, and a weight equal to $\delta_{ij} = (t_{ij}/T)^{\kappa}$ if the animal dies without tumor presence. The parameter δ_{ij} reflects a less-than-whole

animal contribution, and κ depends on the tumor type/site. The Poly-κ trend test for the null hypothesis $\mathrm{H}_0 : \beta = 0$ against $H_1 : \beta > 0$, is given as

$$z = \frac{\sum_i y_{i.} d_i - p'_{..} \sum_i n'_{i.} d_i}{p'_{..}(1 - p'_{..})[\sum_i n'_{i.} d_i^2 - (\sum_i n'_{i.} d_i)^2 / \sum_i n'_{i.}]}$$

where $y_{j.} = \sum_j y_{i.}$, $n'_{ik} = \sum_j \delta_{ij}$, and $p'_{..} = \sum_{ij} y_{ij} / \sum_{ij} \delta_{ij}$. Under the null hypothesis, z is asymptotically standard normally distributed. Bieler and Williams[12] proposed a modification to account for the random variation due to δ_{ijk}.

$$z = \frac{\sum_i a_i p'_{i.} d_i - (\sum_i a_i d_i)(\sum_i a_i p'_{i.}) / \sum_i a_i}{\{C[\sum_i a_i d_i^2 - (\sum_i a_i d_i)^2 / \sum_i a_i]\}^{1/2}},$$

where $C = \sum_{ij} (r_{ij} - \bar{r}_i)^2 / (N - g)$, $a_i = (n'_i)^2 / n_i$, $p'_{i.} = y_{i.} / n'_i$, $r_{ij} = y_{ij} - p'_{..} \delta_{ij}$, $\bar{r}_i = \sum_j r_{ij} / n_i$, $n'_i = \sum_j \delta_{ij}$, $p'_{..} = \sum_i y_{i.} / \sum_i n'_i$, $y_{i.} = \sum_j y_{ij}$, and n_i is the total number of animals in the ith group, and $N_k = \sum_i n_i$. Under the null hypothesis, z_k is asymptotically standard normally distributed. The values for κ are between 1 to 6 from the examination of the NTP historical data.[11] Recently, the NTP has adopted the modified Poly-3 test ($\kappa = 3$) as a standard test for trend and compares the results against the Poly-1.5 and Poly-6 tests.

1.5. *Example*

Stallard and Whitehead[13] presented the results of a carcinogenicity experiment with four dose groups in male mice. The control, low, medium groups contained 60 animals and the high dose group contained 59 animals. The experiment lasted for 105 weeks. The tumor data are shown in Table 1. Since the data contained mixture of fatal and incidental tumors, the Peto cause-of-death test was used. The fatal tumors and incidental tumors were analyzed separately by the logrank test and the prevalence method, respectively. The vector for the difference between the observed and expected numbers of tumors is $D_a = (1.4976, -6.7991, 7.1938, 12.4952)$ for the fatal tumor and is $D_b = (-0.1489, -8.0581, -8.8191, 17.0532)$ for the incidental tumor with the variance-covariance matrices

$$V_a = \begin{bmatrix} 6.4575 & -2.7921 & -2.2296 & -1.4358 \\ -2.7921 & 6.2768 & -2.1184 & -1.3663 \\ -2.2296 & -2.1184 & 5.4485 & -1.1005 \\ -1.4358 & -1.3663 & -1.1005 & 3.9026 \end{bmatrix}$$

Table 1. Tumor data from a carcinogencity experiment presented by Stallard and Whitehead.[13]

Dose	Deaths without Tumors (frequency in parentheses)	Deaths with Tumors (frequency in parentheses)
Control	15, 62, 90, 92, 96, 97, 101, 105(22)	56, 65, 66, 76, 77, 80, 81, 86*, 87, 89, 93, 95, 97, 98(2), 103, 104, 105*(14)
Low	24, 27, 53, 64, 68, 47, 82, 83, 94, 96, 97, 99, 102, 105*(6) 102(2), 103, 104, 105(27)	63, 75, 78, 84, 85, 95, 96, 97, 98, 101,
Medium	5, 7, 39, 65, 70, 75, 76, 80, 82, 83, 87, 91(2), 92, 96(2), 97, 98(2), 99, 100(2), 102, 105(23)	47, 52, 65, 69, 70, 88, 91, 95, 99, 100, 104, 105*(3)
High	16, 18, 49, 55, 59, 77, 85(2), 105	57*, 60, 66, 70(2), 74(2), 76, 78, 83(2), 84(3), 85, 88, 89, 92, 93(2), 94, 95*, 95, 96, 97, 98*, 98(2), 99, 100, 101, 102*, 102, 103, 104, 105*(15)

and

$$V_b = \begin{bmatrix} 7.8714 & -2.8425 & -2.9518 & -2.0772 \\ -2.8425 & 8.7858 & -3.4885 & -2.4549 \\ -2.9518 & -3.4885 & 8.9895 & -2.5493 \\ -2.0772 & -2.4549 & -2.5493 & 7.0813 \end{bmatrix},$$

respectively. Hence, the pooled vector for the fatal and incidental tumors combined is $D = (1.3487, -14.8842, -16.0130, 29.5484)$ with the variance-covariance matrix

$$V = \begin{bmatrix} 14.3289 & -5.6346 & -5.1814 & -3.5130 \\ -5.6346 & 15.0626 & -5.6069 & -3.8212 \\ -5.1814 & -5.6069 & 14.4380 & -3.6498 \\ -3.5130 & -3.8212 & -3.6498 & 10.9839 \end{bmatrix}.$$

The χ^2_3 test statistic for heterogeneity among the 4 groups is $X_H = 88.42$ and the Z statistic for dose-response trend is $Z = 4.599$, where the dose metric is $l = (0, 1, 2, 3)$. Both the heterogeneity test and trend test show statistically significant. The z-score from the Poly-3 test is 4.4328.

1.6. *Exact trend tests for incidental tumors*

In animal carcinogenicity bioassay experiments, the number of animals developing certain tumor types of interest is often small. The methods

described previously use the asymptotic normal approximation. In general, the mortality patterns, number of intervals used in the partition, and numbers and patterns of tumor occurrence in each interval can have effects on the accuracy of an asymptotic test. When the total number of tumor occurrences is small, the normal approximation may not be reliable.[14] The exact permutation test is recommended. The following will describe an exact permutation trend test for tumors observed in an incidental context. The test is a generalization of the Fisher's exact test to the $2 \times g$ table. The tumors observed in a mortality-independent or fatal context can be tested in a similar way. However, Fairweather *et al.*[15] discussed limitations of applying exact methods to fatal tumors.

The data for the tumors observed in the jth interval can be summarized as

Summary of animal prevalence data at interval $(t_{(j-1)}, t_j]$.

Dose	d_1	d_2	$\cdots$	d_g	Total
# with Tumors	r_{1j}	r_{2j}	$\cdots$	r_{gj}	$r_{.j}$
# deaths	n_{1j}	n_{2j}	$\cdots$	n_{gj}	$n_{.j}$

where n_{ij}, here, is the total number of animals from the ith group that died in the jth time interval, r_{ij} is the number of animals (out of n_{ij}) found to have the tumor of interest, and $r_{.j}$ and $n_{.j}$ are the row marginal totals which are fixed for all $j = 1, \ldots, J$.

Conditional on $r_{.j}$ and $n_{.j}$, under the null hypothesis of no difference among groups, the conditional distribution of $(r_{1j}, r_{2j}, \ldots, r_{gj})$ is the multivariate hypergeometric distribution

$$P(r_{1j}, r_{2j}, \ldots, r_{gj}) = \frac{\binom{n_{1j}}{r_{1j}}\binom{n_{2j}}{r_{2j}} \cdots \binom{n_{gj}}{r_{gj}}}{\binom{n_{.j}}{r_{.j}}} .$$

Assuming independence among the J tables, the joint probability of $(r_{11}, \ldots, r_{21}, \ldots, r_{gJ})$ for a random permutation outcome is

$$P(r_{11}, \ldots, r_{21}, \ldots, r_{gJ}) = \prod_{j=1}^{J} P(r_{1j}, r_{2j}, \ldots, r_{gj}) .$$

The trend score associated with $(r_{1j}, r_{2j}, \ldots, r_{gj})$ in the j-interval is defined by

$$S_j = \sum_{i=1}^{g} d_i r_{ij} .$$

The probability distribution for the trend score statistic S_j is

$$P(S_j = s_j) = \sum_{\Omega_j} P(r_{1j}, r_{2j}, \ldots, r_{gj}) ,$$

where Ω_i consists of all possible permutations of r_{ij} such that $\sum_{i=1}^{g} r_{ij} = r_{\cdot j}$ and $\sum_{i=1}^{g} d_i r_{ij} = s_j$. The trend score associated with a random permutation $(r_{11}, \ldots, r_{21}, \ldots, r_{gJ})$ is

$$S = \sum_{j=1}^{J} S_j = \sum_{j=1}^{J} \sum_{i=1}^{g} d_i r_{ij} .$$

The probability distribution for the trend score S is

$$P(S = s) = \sum_{\Omega} P(S_1 = s_1) \cdots P(S_2 = s_2) \cdots P(S_J = s_J) ,$$

where Ω consists of all possible permutations $(r_{11}, \ldots, r_{21}, \ldots, r_{gJ})$ such that $\sum_{j=1}^{J} s_j = s$. Let s^* denote the trend score associated with the observed outcome. The exact one-tailed p-value for a positive trend is

$$p\text{-value} = \sum_{s \geq s^*} P(S = s) .$$

Note that when $k = 1$ and $g = 2$, this procedure becomes the Fisher exact test.

Traditionally, the definition of the p-value of an exact test is the cumulative sum of the probability of the observed outcome and the probabilities of all more extreme outcomes. The trend p-value described above is the sum of the probabilities of all permutations whose *trend scores* are greater than or equal to the trend score of the observed outcome s^*. Thus, every permutation with a trend score equal to s^* is included in the p-value computation irrespective of the magnitude of its probability of occurrence. Chen *et al.*[16] argued that those permutations whose probabilities are greater (less extreme) than the probability of the observed outcome should not be included. A less conservative exact p-value is

$$p\text{-value} = \sum_{s > s^*} P(S = s) + \sum_{\omega} P(S_1 = s_1) \cdot P(S_2 = s_2) \cdots P(S_J = s_J) ,$$

where ω consists of all permutations such that $\sum_{j=1}^{k} s_j = s^*$ and $P(r_{11}, \ldots, r_{21}, \ldots, r_{gJ}) \leq P^*$, and P^* denotes the probability of the observed outcome.

The lung adenoma data observed in female mice from a two-year feeding study of phenylephrine hydrochloride conducted by the National Toxicology

Table 2. The incidence of lung adenoma in female mice in the two-year feeding study of phenylephrine.

Weeks		0 ppm	1250 ppm	2500 ppm	Total
0–52	r_{1j}	0	0	0	0
	n_{1j}	1	0	3	4
53–78	r_{2j}	0	0	0	0
	n_{2j}	1	3	3	7
79–92	r_{3j}	1	0	0	1
	n_{3j}	9	5	2	16
93–104	r_{4j}	0	3	5	8
	n_{4j}	39	42	42	123

Table 3. Computations of the p-value for the exact trend test.

79–92			93–104			Combined	
Pattern	S_3	P_3	Pattern	S_4	P_4	$S_3 + S_4$	Prob.
						$> 16,250$	0.01987
1, 0, 0	0	0.5625	1, 1, 6	16,250	0.008343	16,250	0.00469
	0	0.5625	0, 3, 5	16,250	0.009482	16,250	0.00533
0, 1, 0	1250	0.3125	2, 0, 6	15,000	0.003774	16,250	0.00118
	1250	0.3125	0, 4, 4	15,000	0.012165	16,250	0.00380
	1250	0.3125	2, 0, 6	15,000	0.027736	16,250	0.00867
0, 0, 1	2500	0.1250	0, 5, 3	13,750	0.009482	16,250	0.00118
	2500	0.1250	2, 1, 5	13,750	0.025707	16,250	0.00321
	2500	0.1250	1, 3, 4	13,750	0.048660	16,250	0.00608

Program (NTP, 1987) is analyzed for illustration. This experiment contained a control and two dose groups, 1250 and 2500 ppm. In the analysis of tumor incidence data, NTP generally groups the animals into the following four time intervals to adjust for intercurrent mortality: 0–52, 53–78, 79–92, 93–104 weeks. Table 2 shows the number of lung adenomas and the number of deaths occurring in the four time intervals.

The computations of the p-value for the exact trend test are shown in Table 3.

Each row in the table corresponds to a permutation for which the score is 16,250 from the sum of Columns S_3 and S_4. The exact p-value ($= 0.0540$) is the sum of the probabilities of the right most column, and the exact p-value ($= 0.0393$) by the Chen *et al.*[16] is obtained by excluding two probability values 0.00867 and 0.00608 as they are greater than 0.00533, the probability

of the data observed. The MH asymptotic trend test gives a p-value of 0.0347. For a significance level of 5%, the Chen *et al.*[16] adjustment would indicate a statistically significant result, in agreement with the MH trend test.

1.7. *Discussion*

A typical carcinogenicity experiment examines approximately 20–50 tumor types/sites, statistical tests often perform for 10–30 types/sites routinely. Performing several tests without appropriately accounting for the multiplicity effect can inflate the overall Type I error rate or familywise error rate (FWE). Statistical Methods for the analysis of multiple tumor types/sites have been proposed by several authors.[17–20] Haseman[21] presented a rule rejecting a hypothesis for a rare tumor (spontaneous rate at most 0.01) when $p \leq 0.05$ and for a common tumor (spontaneous rate greater than 0.01) with $p \leq 0.01$. The Center for Drug Evaluation and Research (CDER) of FDA has adopted the "Haseman rule" in comparing tumor incidence rates between the control and dose groups in its evaluation of tumorigenicity studies of new drugs, and has recently recommended a new rejection rule for a positive dose-related *trend* test with $p \leq 0.025$ for rare rumors and $p \leq 0.005$ for common tumors.[22]

Interpreting results of carcinogenicity experiments is a complex process, and there are risks of both false negative and false positive results. The relatively small number of animals used, and the low tumor incidence rates can cause carcinogenicity of a compound not to be detected (i.e. a false negative error is committed). Because of the large number of comparisons involved, there is also a great potential or finding statistically significant positive trends or differences in some tumor types that are due to chance alone (i.e. a false positive error is committed). The inflated false-positive rate can invalidate the use of animal carcinogenicity data. Controlling both false positive and false negative rates should be the central issue in the statistical analysis of animal carcinogenicity experiments from the safety assessment viewpoint.

2. Reproductive Studies

Reproductive studies are conducted to assess reproductive risk to mature adults and to the developing individual from the exposure to drugs and environmental compounds. Adverse reproductive and developmental effects include effects on male and female fecundity, spontaneous abortion, infant

and child death, congenital malformations, growth retardation, and mental retardation. Three segments of study are required in preclinical animal testing for each new drug depending on how women might be exposed to the drug.[23] These are referred to as Segment I (fertility and general reproductive performance), Segment II (developmental effects), and Segment III (prenatal and postnatal evaluations).

The Segment I study is aimed at providing an overall evaluation of the effects of drugs on fertility in both sexes, the course of gestation, early and late stages of the development of the embryo and fetus, and postnatal development. The studies may be conducted by treating animals of only one sex and mating with untreated animals of the opposite sex, or by treating both male and female animals. Segment II is primarily aimed at detecting teratogenic effects. The drug is given to the pregnant females during the period of organogenesis, e.g. days 6–15 for rats and mice, and days 6–18 for rabbits. The offspring are removed one or two days before term, and corpora lutea, resorption sites, and live and dead fetuses are examined. Fetuses are weighed and examined for anomalies. Segment III is aimed at the evaluation the effects of drugs on the late stages of gestation and on parturition and lactation. The drug is given to pregnant females in the final one-third of gestation and continued throughout lactation to weaning, e.g. gestation day 15 to postnatal day 21 for rats or mice. The effects on duration of gestation is determined. Pup birth and developmental data including litter size, weight, and postnatal growth and mortality, along with impaired maternal behavior are recorded and measured.

Mice, rats, and rabbits are the most commonly used species for reproductive and developmental studies. The experimental design is very similar to the carcinogenicity experiment consisting an untreated control and three dose groups. The US regulatory guidelines generally recommend about 20 pregnant rodents and 15 nonrodent animals per dosage group. The ICH guideline recommends 16 to 20 pregnant animals per group. All adult animals are necropsied at terminal sacrifice.[24,25]

Regulatory requirements specify that a wide range of endpoints must be measured, recorded, and analyzed. The endpoints can be divided into two categories: parental and embryonic/fetal endpoints. Since the test compound is administered to the adult animal, the effect of the test compound occurs in the female that receives the compound, or that is mated to a male that receives the compound, the treatment affects the fetuses indirectly via the dam. The fetal responses from the same dam are expected to be more alike than responses from different dams. This phenomenon is referred

to as the "litter effect". In the analysis of embryonic/fetal endpoints, the experimental unit should be the entire litter rather than an individual fetus. Failure to account for the intra-litter correlations by using each fetus as the experimental unit will inflate the Type I error and will reduce the validity of the test. The classical approach to the analysis of reproductive data is based on the litter mean. However, this approach does account for differences in litter sizes. An alternative approach is to model the fetal endpoints in a litter as correlated outcomes (clustered data). These two approaches will described in this section. Statistical methods described will be according to three measures, continuous, binary, and count.

2.1. *Per-litter based analysis*

Consider a typical experiment of g groups, a control and $g-1$ dose groups. Assume that the ith group contains m_i female animals. Let y_{ijk} be the response from a fetus out of n_{ij} examined or tested for a particular developmental outcome, $1 \leq i \leq g$, $1 \leq j \leq m_i$, and $1 \leq k \leq n_{ij}$. Note that y_{ijk} may be an indicator variable representing the presence or absence of a particular malformation type or a continuous variable representing a fetal weight or postnatal performance measurement. Depending on the endpoint of interest, n_{ij} may represent the number of viable fetuses, number of implants, or number of measurements. The litter-based analysis is based on the per-litter response $y_{ij} = \sum_k y_{ijk}/n_{ij}$. For a continuous response, y_{ij} will represent the mean fetus response; for a discrete variable, it will represent the sum of the fetal responses. The y_{ij} can be viewed and analyzed as a maternal endpoint.

2.1.1. *Continuous data*

Continuous data such as body weights, organ weights, or behavioral measurements conducted on offspring following birth are measured on a continuous scale. The continuous endpoints are measured either at the litter level in an adult animal (e.g. maternal body weights) y_{ij}, or at the individual fetus level (fetal body weights) y_{ijk}. Analysis of variance (ANOVA) is the most commonly used procedure for the analysis of continuous data.[25] The ANOVA method assumes that data are independently and normally distributed with homogeneous variance. Transformations such as the logarithmic, square-root and arc-sine are often applied to satisfy the normality assumption and stabilize the variance. A simple one-way ANOVA analysis is the comparison of maternal endpoints among groups. Developmental

endpoints are analyzed similarly but in terms of the average within each litter as described.[26] Nonparametric methods are used when the assumption of normality fails. The nonparametric analysis is initiated by ranking all observations of the combined groups. The repeated measures ANOVA is often used for the analysis of postnatal behavioral data.[27]

2.1.2. *Binary data*

Binary endpoints can also be measured either at the parent level, such as success or failure of pregnancy, or at the individual fetal level, such as presence or absence of a particular malformation type. Statistical methods for the analysis of the prenatal and fetal responses are different. Asymptotic chi-square test is often used for the analysis of prenatal binary endpoints for comparing the incidence rates among several groups.[28] The Cochran-Armitage test is used to test for trend.[29,30] The Fisher exact test is the best known permutation test for comparing two groups. General computational algorithms and software to perform all possible permutations are given by Mehta *et al.*[31] The general approach to the analysis of fetal binary endpoints is to consider the proportion per-litter such as the proportion of live fetuses with a certain type of malformation. Typically, the proportions are transformed by an arc-sine transformation, and then the parametric ANOVA methods are used. The litter-based approach does not use the data effectively since it does not account for the litter size. For example, one out of two is treated as the same as five out of ten.

2.1.3. *Count Data*

A number of primary reproductive endpoints are measured in counts. In a dominant lethal assay, male mice are treated with a suspect mutagen, and then are mated with females. The numbers of corpora lutea, implantations, lives, and dead conceptuses are counted to assess reproductive effects of the test compound. Count data are often normalized by the square root transformation, the transformed data are then analyzed as continuous data using the parametric ANOVA methods. Count data can also be analyzed by nonparemetric methods.

2.1.4. *Example*

A study of the effects of maternal exposure to diethylhexyl phthalate (DEHP) in rats is presented as an example. Table 4 contains a summary of

Table 4. Reproductive parameters after exposure of pregnant Fisher 344 rats to diethylhexyl phahalate in the feed on getational days 0–20.

	Diethylhexyl phahalate % in feed				
Endpoint	0.0	0.5	1.0	1.5	2.0
No. pregnant dams	24	23	22	24	25
Maternal wgt gd 0	173.60 (3.25)	175.37 (3.37)	172.42 (3.12)	171.77 (2.80)	173.33 (2.95)
Maternal wgt gd 20	248.30 (3.64)[a]	246.00 (2.90)	232.73 (3.25)*	217.76 (3.25)*	184.15 (4.28)*
Maternal wgt gain	74.69 (1.71)[a]	70.63 (2.48)	60.31 (1.44)*	45.99 (1.78)*	10.82 (3.00)*
Gravid uterine wgt	49.79 (1.39)[a]	44.83 (2.49)	46.81 (1.31)	43.86 (0.96)	19.08 (3.71)*
Maternal liver wgt	9.75 (0.12)[a]	11.72 (0.17)*	12.06 (0.17)*	12.21 (0.19)*	11.11 (0.15)*
Corpora lutea	10.91 (0.33)	10.96 (0.22)	11.18 (0.30)	11.17 (0.18)	10.52 (0.61)
Implantation sites	10.92 (0.31)	9.83 (0.54)	10.59 (0.24)	10.58 (0.22)	10.40 (0.41)
% viable	10.46 (0.32)[a]	9.39 (0.53)	10.05 (0.26)	10.08 (0.25)	8.00 (0.70)*
% male	53.56 (3.65)	45.77 (4.51)	46.39 (3.31)	54.44 (3.00)	50.66 (6.10)
% malformation	1.27 (0.71)[a]	0.00 (0.00)	1.92 (1.11)	3.13 (1.06)	2.87 (1.64)
Fetal weight	3.022 (0.029)[a]	3.143 (0.035)*	2.852 (0.053)*	2.557 (0.034)*	2.266 (0.041)*

[a]Significance in linear trend

*Significantly different from control

the analysis for selected endpoints. For the detailed design and analysis the reader is referred to Tyl *et al.*[32] As described the ANOVA was used for the analysis of continuous endpoints and per-litter proportion data. Prior to analysis, the arc-sine square root transformation was performed to all maternal or per-litter proportion data. When a significant ($p < 0.05$) dose effect occurred, Duncan's multiple range test was used for pairwise comparisons between control and each dose group. A test of linear dose-response trend was performed using contrast tests. A one-sided test was used for pairwise comparisons except for the maternal and fetal body weights and percentage of males per litter.

2.2. *Likelihood and quasi-likelihood/ generalized-estimating-equations approaches*

As discussed, the fetal responses with a litter are not independent. The proper experimental unit in the analysis should be the litter with the fetal responses representing multiple observations from a single experimental unit. The likelihood-based and generalized estimating equations (or quasi-likelihood) are the two commonly used approaches to modeling the correlated data.

2.2.1. *Modeling continuous data*

A general approach to modeling fetal data can be carried out in terms of a mixed-effects model. Dempster *et al.*[33] proposed a normal mixed-effects model with two levels of variance, in which litter effect is modeled by a nested random factor and dose by a fixed factor. The response y_{ijk} in a litter is expressed as a mixed effects model with two sources of variations: the between litter γ_{ij} and within litter variations e_{ijk},

$$y_{ijk} = \mu_{ij} + \gamma_{ij} + e_{ijk}\,.$$

The random components γ_{ij} and e_{ijk} are independently normally distributed with $\mathrm{E}(\gamma_{ij}) = 0$, $\mathrm{var}(\gamma_{ij}) = \sigma_a^2$, and $\mathrm{E}(e_{ijk}) = 0$, $\mathrm{var}(e_{ijk}) = \sigma^2$. Thus, the mean and variance of y_{ijk} are $\mathrm{E}(y_{ijk}) = \mu_i$ and $\mathrm{var}(y_{ijk}) = \sigma_a^2 + \sigma^2$. The intra-litter correlation between y_{ijk} and $y_{ijk'}$, for $k \neq k'$ is $\phi = \sigma_a^2/(\sigma^2 + \sigma_a^2)$. The mean parameter μ_{ij} is often modeled as a linear function of dose $\mu_{ij} = \beta_0 + \beta_1 d$ for trend test ($H_o : \beta_1 = 0$).

Computation techniques for the maximum likelihood estimation of the parameters of linear mixed effects models have been proposed by many authors.[34–36] The estimates can be obtained using the PROC MIXED

procedure of SAS.[37] Alternatively, the generalized estimating equations (GEE) approach[38] can be used to estimate the fixed effects parameters β. Under the normal model, the likelihood-based and GEE approaches have the same estimating equations for the mean parameters, but unlike the likelihood approach, the GEE uses the method of moments to estimate the variance component parameters.

2.2.2. *Modeling binary data*

Let y_{ijk} denote the presence or absence of a response. Assume that the mean and variance of y_{ijk} are $\mathrm{E}(y_{ijk}) = \mu_i$, and $\mathrm{var}(y_{ijk}) = \mu_i(1-\mu_i)$ and the correlation is $\mathrm{corr}(y_{ijk}, y_{ijk'}) = \phi_i$, where $k, k' = 1, \ldots, n_{ij}$, and $k \neq k'$. The parameter μ_i is the probability of a developmental effect of the ith group, and ϕ_i is the intra-litter correlation coefficient. The binary responses $y_{ij1}, \ldots, y_{ijn_{ij}}$ within each litter are assumed exchangeable, that is, if $\{k_1, \ldots, k_l\}$ is a subset of $\{1, \ldots, n_{ij}\}$, then

$$\Pr(y_{ij1} = 1, \ldots, y_{ijl} = 1) = \Pr(y_{ij_{k_1}} = 1, \ldots, y_{ij_{k_l}} = 1)\,,$$

for all $l = 1, \ldots, n_{ij}$.

Let $y_{ij} = (y_{ij1} + \cdots + y_{ijn_{ij}})$, then the mean and variance of y_{ij} are $\mathrm{E}(y_{ij}|n_{ij}) = n_{ij}\mu_i$ and $\mathrm{var}(y_{ij}|n_{ij}) = n_{ij}\mu_i(1-\mu_i)[\phi_i(n_{ij}-1)+1]$. The intra-litter correlation coefficient generally is positive ($\phi_i > 0$). Thus, the variance $n_{ij}\mu_i(1-\mu_i)[\phi_i(n_{ij}-1)+1]$ is greater than the nominal binomial variance $n_{ij}\mu_i(1-\mu_i)$. The distribution of y_{ij} is known as an extra-binomial variate. Note that if $\phi_i = 0$, then all y_{ijk}'s are independent binary random variables and y_{ij} is a binomial,

$$P(y_{ij}) = \binom{n_{ij}}{y_{ij}} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{n_{ij}-y_{ij}}\,.$$

In a binomial or extra-binomial model, the mean function is often modeled by a logit function, $\mathrm{logit}(\mu_i|\boldsymbol{z}_{ij}) = \boldsymbol{\beta}'\boldsymbol{z}_{ij}$. The dose response model for trend test is given by

$$\mu_i = \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)}\,.$$

The beta-binomial model is the commonly known distribution used for modeling the extra-binomial variation data.[39] The beta-binomial model assumes that responses within the same litter occur according to a binomial

distribution and the probability of responses is assumed to vary among litters according to a beta distribution:

$$P(y_{ij}) = \binom{n_{ij}}{y_{ij}} \frac{\mathrm{B}(a_i + y_{ij}, b_i + n_{ij} - y_{ij})}{\mathrm{B}(a_i, b_i)}$$

where $\mathrm{B}(a_i, b_i) = \Gamma(a_i)\Gamma(b_i)/\Gamma(a_i + b_i)$, where $\Gamma(\cdot)$ is the gamma function, $a_i > 0$, and $b_i > 0$. Under the reparameterization $\mu_i = a_i/(a_i + b_i)$ and $\phi_i = (a_i + b_i + 1)^{-1}$, the parameters μ_i and ϕ_i are, respectively, the mean and the intra-litter correlation parameters in the ith group. That is, the mean of y_{ij} is $\mathrm{E}(y_{ij}) = n_{ij}\mu_i$, and the intra-litter correlation is $\mathrm{corr}(y_{ijk}, y_{ijk'}) = \phi_i$. The variance of y_{ij} is $\mathrm{var}(y_{ij}) = [\phi_i(n_{ij} - 1) + 1][n_{ij}\mu_i(1 - \mu_i)]$. When $\phi_i = 0$, then y_{ij} becomes a binomial variable. The parameters can be estimated by the maximum likelihood method. The likelihood ratio or Wald test is often used to test for the significance of parameters. An advantage of the use of the likelihood-based beta-binomial model is that parameters $\boldsymbol{\beta}$ as well as the intra-litter correlations ϕ's can be tested directly. One problem with the beta-binomial model is the bias and instability of the MLE's of $\boldsymbol{\beta}$ as discussed by Williams.[40] Williams[41] proposed using the quasi-likelihood method as an alternative approach to the beta-binomial model.

In the quasi-likelihood approach, only assumptions on the mean and variance are required: $\mathrm{E}(y_{ij}|n_{ij}) = n_{ij}\mu_i$ and $\mathrm{var}(y_{ij}|n_{ij}) = n_{ij}\mu_i(1 - \mu_i)$ $[\phi(n_{ij} - 1) + 1]$. Note that in the quasi-likelihood estimation, the intra-litter correlations typically are modeled as constant across groups. The coefficients of the β's can be obtained by solving the quasi-likelihood estimating (score) equations:

$$S(\beta_l) = \sum_{i=1}^{g} \sum_{j=1}^{m_i} z_{ijl} \frac{y_{ij} - n_{ij}\mu_i}{[1 + (n_{ij} - 1)\phi]} = 0\,, \quad l = 1, \ldots, 2\,.$$

The intra-litter correlation coefficient is calculated by equating with the mean of Pearson chi-square statistics, i.e.

$$\phi = \frac{1}{N - 2} \sum_{i=1}^{g} \sum_{j=1}^{m_i} \frac{(y_{ij} - n_{ij}\mu_i)^2}{n_{ij}\mu_i(1 - \mu_i)[\phi^{-1} + (n_{ij} - 1)]}\,,$$

where $N = \sum_i^g \sum_j^{m_i} n_{ij}$. The parameters β's and ϕ are estimated by solving $S(\beta_k) = 0$ and ϕ alternatively until convergence. The Wald test is often used to test for the significance of β's.

2.2.3. *Modeling count data*

Count data are generally modeled by a Poisson distribution. Let n_{ij} be an observed count from the jth animal in the ith group $1 < i < g$ and $1 < j < m_i$. If n_{ij} has the Poisson distribution

$$p(n_{ij}) = \frac{\mu_i^{n_{ij}} e^{-\mu_i}}{n_{ij}!}\,, \quad n_{ij} = 0, 1, 2, \ldots\,,$$

the mean and variance of n_{ij} are $\mathrm{E}(n_{ij}) = \mathrm{var}(n_{ij}) = \mu_i$. In the Poisson model, the mean function is often modeled by a log-linear function, $\log(\mu_i|\boldsymbol{z}_{ij}) = \boldsymbol{z}_{ij}\boldsymbol{\beta}$. The dose-response model for trend test is $\mu_i = \exp(\beta_0 + \beta_1 d_i)$.

A common complication in the analysis of count data is that the observed variation often exceeds or falls behind the variation that is predicted from a Poisson model. The classical approach is to assume that the mean of the Poisson has a gamma distribution which leads to a negative binomial (gamma-Poisson) distribution for the observed data,

$$p(n_{ij}) = \frac{\Gamma(n_{ij} + \phi_i^{-1})}{\Gamma(n_{ij} + 1)\Gamma(\phi_i^{-1})} \left(\frac{\phi_i \mu_i}{1 + \phi_i \mu_i}\right)^{n_{ij}} \left(\frac{1}{1 + \phi_i \mu_i}\right)^{\frac{1}{\phi_i}}$$

where $\phi_i > 0$. The maximum likelihood estimation of the negative binomial model was described in details by Lawless.[42] The significance of the parameters can be tested using either the likelihood ratio test or Wald test. Like the beta-binomial model, the negative binomial model can be applied to testing for the extra-Poisson variation.

A limitation of the parametric approach is in its restriction on $\phi \geq 0$. In applications, for example, the number of litter implant or the number of corpora lutea may exhibit a sub-Poisson variation. The quasi-likelihood approach[43] provides a method to model sub-Poisson variation data. The quasi-likelihood approach assumes the mean and variance of count data are of a negative binomial form, $\mathrm{E}(n_{ij}) = \mu_i$ and $\mathrm{var}(n_{ij}) = \mu_i(1 + \phi_i \mu_i)$. The coefficients of the β's can be obtained by solving the score equations

$$S(\beta_l) = \sum_{i=1}^{g} \sum_{j=1}^{m_i} z_{ijl} \frac{n_{ij} - \mu_i}{\mu_i + \phi \mu_i^2} = 0\,.$$

The parameter ϕ is calculated by equating with the mean of Pearson chi-square statistics,

$$\phi = \frac{1}{N - 2} \sum_{i=1}^{g} \sum_{j=1}^{m_i} \frac{(n_{ij} - \mu_i)^2}{\mu_i \phi^{-1} + \mu_i^2}\,,$$

where $N = \sum_i^g \sum_j^{m_i} n_{ij}$. The parameters β's and ϕ are estimated by solving $S(\beta_k) = 0$ and ϕ alternatively until convergence. Quasi-likelihood approaches to estimating and testing the mean of various mixed Poisson models are given by Chen and Ahn.[44] Examples of analyses of fetal response toxicity data using the likelihood and the quasi-likelihood/GEE approaches are given in Chen.[45]

2.3. *Multiple developmental outcomes*

The standard approach for assessment of developmental risks of a compound has been based on the analysis of each developmental endpoint separately. It has been suggested that the developmental toxicity outcomes (i.e. death/resorption, malformation, growth retardation, etc.) may represent different degrees of responses to a toxic insult and occur in a dose-related manner.[45,46] These developmental outcomes are likely to be correlated. Therefore, a joint analysis of multiple developmental outcomes can have some advantages: it can increase the power of detecting effects if the multiple outcomes are manifestations of some common biological effects, and it allows investigations of associations among the multiple outcomes if they are the results of different biological mechanisms. Various multivariate models have been developed for simultaneous analysis of multiple endpoints.[47–50]

References

1. Kodell, R. L. and Nelson, C. J. (1980). An illness-death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics* **36**: 267.
2. Peto, R., Pike, M. C., Day, N. E. *et al.* (1980). Guidelines for simple sensitive significance tests for carcinogenic effects in long-term animals experiments. In *Long-term and Short-term Screening Assays for Carcinogens*: *A Critical Appraisal*, Annex to Supplement 2. Lyon, International Agency for Research on Cancer, 311.
3. Kodell, R. L., Shaw, G. W. and Johnson, A. M. (1982). Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments. *Biometrics* **38**: 43.
4. Kaplan, E. L. and Meier, P. (1982). Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association* **53**: 457. Experiments. *Biometrics* **38**: 43.
5. Gart, J. J., Krewski, D., Lee, P. N. *et al.* (1986). Statistical methods in cancer research, *The Design and Analysis of Long-Term Animal Experiments*, Vol. 3, Lyon, International Agency for Research on Cancer.
6. Hoel, D. and Walburg, H. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute* **49**: 361.

7. Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**: 719.
8. Kodell, R. L., Chen, J. J. and Moore, G. E. (1994). Comparing distributions for time to onset of disease in animal tumorigenicity experiments. *Communications in Statistics — Theory and Methods* **23**: 959.
9. Chen, J. J. and Moore, G. E. (1994). Impact of surviving time on tests for carcinogenicity. *Communications in Statistics — Theory and Methods* **23**: 1375.
10. Dinse, G. E. and Lagakos, S. W. (1983). Regression analysis of tumor prevalence data. *Applied Statistics* **32**: 236.
11. Bailer, A. J. and Portier, C. J. (1988). Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics* **44**: 417.
12. Bieler, G. S. and Williams, R. L. (1993). Ratio estimates, the delta methods, and quantal response tests for increased carcinogenicity. *Biometrics* **49**: 793.
13. Stallard, N. and Whitehead, A. (1999). Modified Weibull multi-state models for the analysis of animal carcinogenicity data. *Environmental and Ecological Statistics* **6**: xx.
14. Chen, J. J. and Gaylor, D. W. (1986). The upper percentiles of the distribution of the logrank statistics for small numbers of tumors. *Communications in Statistics — Theory and Methods* **15**: 991.
15. Fairweather, W. R., Bhattacharyya, P. P., Ceuppens, G. *et al.* (1998). Biostatistical methodology in carcinogenicity studies. *Drug Information Journal* **32**: 401.
16. Chen, J. J., Kodell, R. L. and Pearce, B. A. (1997). Significance levels of randomization trend tests in the event of rare occurrences. *Biometrical Journal* **39**: 327.
17. Heyse, J. F. and Rom, D. (1980). Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical Journal* **30**: 883.
18. Westfall, P. H. and Young, S. S. (1989). p-value adjustment for multiple tests in multivariate binomial models. *Journal of American Statistical Association* **84**: 780.
19. Chen, J. J. (1996). Global tests for analysis of multiple tumor data from animal carcinogenicity experiments. *Statistics in Medicine* **15**: 1217.
20. Chen, J. J., Lin, K. K., Huque, M. *et al.* (2000). Weighted p-value adjustments for animal carcinogenicity. *Trend Test Biometrics* **56**.
21. Haseman, J. K., Huff, J. and Boorman, G. A. (1984). Use of historical control data in carcinogenicity studies in rodents. *Toxicology and Pathology* **12**: 126.
22. Lin, K. K. and Rahman, M. A. (1998). Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies in new drugs. *Journal of Biopharmaceutical Statistics* **8**: 1.
23. US Food and Drug Administration. (1996) International Conference on Homonisation. *Guideline on Detection of Reproduction for Medicinal Products,* FDA, Rockvill, MD.

24. Christian, M. S. and Hoberman, A. M. (1996). Perspectives on the US, EEC, and Japanese developmental toxicity testing guidelines. In *Handbook of Developmental Toxicology*, ed. R. D. Hood, CRC Press, NY, 551.
25. Tyl, R. W. and Marr, M. C. (1996). Developmental toxicity testing — methodology. In *Handbook of Developmental Toxicology*, ed. R. D. Hood, CRC Press, NY, 175.
26. Healy, M. J. R. (1972). Animal litters as experimental units. *Applied Statistics* **21**: 155.
27. Karpinski, K. F. (1991). In *Statistics in Toxicology*, eds. D. Krewski and C. Frankin, Gordon and Breach Science, NY, 393.
28. Agresti, A. (1990). *Categorical Data Analysis*, John Wiley and Sons, NY.
29. Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10**: 417.
30. Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**: 375.
31. Mehta, C. R., Patel, N. R. and Senchaudhuri, P. (1992). *Journal of Computation and Graph Statistics* **1**: 21.
32. Tyl, R. W., Price, C. J., Marr, M. C. *et al.* (1983). *Teratologic Evaluation of Diethylhexyl Phthalate in Fisher 344 Rats*, Research Triangle Park, NC.
33. Dempster, A. P., Selwyn, M. R., Patel, C. M. *et al.* (1984). Statistical and computational aspects of mixed model analysis. *Applied Statistics* **33**: 203.
34. Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**: 320.
35. Laird, M. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**: 963.
36. Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association* **83**: 1014.
37. SAS Institute Inc. (1994). *Getting Started with PROC MIXED.* SAS Institute Inc., Cary, NC.
38. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13.
39. Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**: 949.
40. Williams, D. A. (1988). Estimation bias using beta-binomial distribution in teratology. *Biometrics* **44**: 305.
41. Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**: 144.
42. Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15**: 209.
43. Breslow, N. E. (1984). Extra-Poisson variation in log-linear model. *Applied Statistics* **33**: 38.
44. Chen, J. J. and Ahn, H. (1996). Fitting mixed Poisson regression models using quasi-likelihood methods. *Biometrical Journal* **38**: 81.

45. Chen, J. J. (1998). Analysis of reproductive and developmental studies. In *Design and Analysis of Animal Studies in Pharmaceutical Development*, eds. S. C. Chow and J. P. Liu, Marcel Dekker, NY, 309.
46. Kimmel, C. A. and Gaylor, D. W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Analysis* **8**: 15.
47. Chen, J. J. and Gaylor, D. W. (1992). Correlations of developmental endpoints observed after 2, 4, 5-trichlorophenoxyacetic acid exposure in mice. *Teratology* **45**: 241.
48. Lefkopoulou, M., Moore, D. and Ryan, L. M. (1989). The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *Journal of the American Statistical Association* **84**: 810.
49. Chen, J. J., Kodell, R. L., Howe, R. B. *et al.* (1991). Analysis of trinomial responses from reproductive and developmental toxicity experiments. *Biometrics* **47**: 1049.
50. Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of American Statistical Association* **87**: 651.

About the Author

James Chen is Senior Mathematical Statistician in the Division of Biometry and Risk Assessment at the National Center for Toxicological Research, US Food and Drug Administration. He received his BS degree from Taiwan Tsing-Hua University, M.A. degree from University of Pittsburgh, and PhD from Iowa State University. Dr. Chen is a Fellow of the American Statistical Association. He has over 100 scientific publications in peer-reviewed journals and numerous invited subject review articles. Dr. Chen has served on the FDA, EPA, and interagency committees and workshop that directed at developing scientific and regulatory issues and guidelines, and has provided consultations to FDA, and EPA scientists on the statistical analysis of toxicological data and on risk assessment procedures.

This page intentionally left blank

CHAPTER 14

SOME STATISTICAL ISSUES OF RELEVANCE TO CONFIRMATORY TRIALS

GEORGE Y. H. CHI*, KUN JIN and GANG CHEN†

*Division of Biometrics I, US Food and Drug Administration, HFD-710 W0C2, Room 2033 1451 Rockville Pike Rockville, MD 20852, USA Tel: 301-827-1515; *chig@cder.fda.gov*

LU CUI‡

Division of Biometrics and Data Management, Aventis Pharceuticals, Inc.

1. Introduction

1.1. *Overview of drug development*

From its initial discovery to final marketing, a new drug in the United States has to go through various stages of development. Typically, preclinical animal studies are needed to determine the toxicity and carcinogenicity potential of a new compound. If these studies offer no suggestion of toxicity and evidence of carcinogenicity, then small studies on human volunteers are conducted in Phase 1 to determine the metabolism and pharmacologic actions of the drug in humans, dose dependent side effects, and if possible evidence of effectiveness. The information collected will permit the design of well-controlled, scientifically valid studies in Phase 2. The clinical studies in Phase 2 are conducted to evaluate the effectiveness of the drug for a particular indication or indications in patients with the disease or condition under study and to determine the common short term side effects and risks associated with the drug. In addition, for serious diseases such as cancer, a drug may be approved conditionally based on acceptable surrogate efficacy

†The reviews expressed in this chapter are those of the authors and not necessarily those of the Food and Drug Administration.

‡The contribution by this author was made while he was still with the Division.

endpoints in Phase 2. The information gathered in these early phases is used to guide the drug sponsor in the planning of Phase 3 clinical trials. Phase 3 studies are expanded controlled and uncontrolled studies carried out to obtain additional effectiveness and safety information that is needed to evaluate the benefit-risk relationship of the drug and to provide adequate basis for physician labeling [21 CFR Sec. 312.20].

One of the critical requirements for pre-marketing approval of a new drug in the United States is the demonstration of the effectiveness of the drug through Phase 3 clinical trials. The United States Code of Federal Regulations [21 CFR Sec. 314.126(a)] requires that to establish the efficacy of a new drug, the sponsor must produce reports of adequate and well-controlled clinical trials that demonstrate its effectiveness. Generally, the evidence of effectiveness is based on at least two adequate and well-controlled studies. Some of the characteristics of an adequate and well-controlled study are described in the regulation [21 CFR Sec. 314.126(b)]. Another critical requirement for pre-marketing approval of a new drug is safety. The safety of the drug is evaluated on the basis of the entire database obtained from all three phases of drug development prior to approval, and continues to be evaluated through Phase 4 clinical trials, if required as a condition for approval. To be approved, the drug sponsor must demonstrate that there is sufficient information to show that the drug is safe for use under the conditions prescribed, recommended, or suggested in its proposed labeling. Once approved, the safety of the new drug continues to be monitored through post-marketing adverse reaction reports.

A drug that is not yet approved for marketing may be used for treatment in patients with a serious or immediately life-threatening disease condition and for whom no comparable or satisfactory alternative drug or other therapy is available. The regulations allow such treatment to be carried out under a special protocol or treatment IND [21 CFR Secs. 312.34, 312.83].

For treating serious or life-threatening illnesses, certain new drug products may qualify under accelerated approval [21 CFR Sec. 314,500]. For treating a serious or life-threatening illness, a new drug that has demonstrated, based on adequate and well-controlled trials, an effect on a surrogate endpoint, or on some clinical benefits other than survival or irreversible morbidity, may be approved for marketing. The drug should provide meaningful therapeutic benefit to patients over existing treatments. For example, it may be effective in treating patients who are unresponsive to available therapy, or it may improve patient response over available therapy. The surrogate endpoint should have been reasonably validated

through epidemiologic, therapeutic, patho-physiologic, or other evidence as predictive of clinical benefit.

Accelerated approval is granted under the condition that the drug sponsor will conduct further study to verify and describe its clinical benefit in relation to the surrogate endpoint, or the ultimate clinical outcome of primary interest.

Some publications[14,61,71] including the US Code of Federal Regulations [21 CFR Secs. 321–314 (2001)], and the US Food and Drug Administration International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), E1–E10 documents (1998–1999), provide a good overview and source of reference for the entire drug development procss and the United States drug regulatory requirements.

This chapter focuses primarily on some concepts, principles and issues related to establishing the evidence of drug efficacy in a confirmatory trial. The discussion of selected topics on certain aspects of the design, conduct and analysis of a clinical trial generally revolves around two fundamental statistical principles that are particularly critical to a confirmatory trial.

1.2. *Confirmatory trial*

In drug development, a confirmatory trial is a clinical trial that is prospectively designed to provide the primary source of evidence necessary to support the efficacy claim of the drug under investigation. The evidence expected of a confirmatory trial must be compelling, especially for mortality or serious morbidity trials where ethical consideration often precludes the possibility of conducting a second trial. This will also be critical for future active control trials where the current drug may be used as the control. The strength of evidence is measured among other things by the overall quality of the trial, internal and external consistency of the trial results. The internal and external consistencies of the trial results are the outcomes of the trial, whereas the overall quality of the trial is to a large extent under the control of the experimenter. The overall quality of a trial should be evaluated with respect to the following aspects of the study. These aspects should include the appropriateness of the design, acceptability of the study conduct, quality of the data collected, adequacy of the power, maintenance of the probability of the overall type I error, control of bias and confounding, proper method of analysis, and correct interpretation of trial results.

The accepted standard for the design of a confirmatory trial includes proper randomization, desired level of blinding, absence of confounding and choice of an appropriate control. In addition, other fundamental statistical principles should be closely adhered to in a confirmatory trial. These principles include minimization of the potential for bias and control of the probability of the overall type I error at the desired α-level. These two principles are essential to a Phases 2 or 3 confirmatory trial. A confirmatory trial should have assay sensitivity, that is, the capability of showing a difference when the study treatment is effective, a concept defined in the ICH-E10 (1999) document on the choice of control.

1.3. *Two fundamental statistical principles of clinical trials*

Minimization of the potential for bias and control of the probability of the overall type I error, or the probability of making a false positive conclusion, are two fundamental statistical principles that are central to the quality of a trial. Both principles are intimately related to the design, conduct and analysis of the trial, and the final interpretation of the trial results. Deviation from these two principles has the potential of weakening the strength of evidence, rendering the trial results uninterpretable, or sometimes, even invalidating the entire trial results. Proper attention to these two principles throughout the trial from design and conduct to the final analysis and interpretation will strengthen the evidence and improve the likelihood of success of a trial.

Minimizing the potential for bias will be discussed first in the next section. In Sec. 3, the control of the probability of the overall type I error will be discussed through issues and problems related to multiple testing. Clinical decision rule will be introduced as an intuitive and natural way of handling the multiple testing problems. The concept of decision structure for a clinical trial incorporates the clinical, statistical and regulatory perspectives on the issue of multiplicity. In Sec. 4, interim analysis and design modification will be discussed through the framework of decision structure. The problems and issues related to the important topic of active control non-inferiority trial of current interest will be discussed in the last section from the perspective of validity of inference and interpretation.

2. Minimizing the Potential for Bias

In a clinical trial, bias refers to the consequence of any design, property of the study treatment or characteristics of the disease, intentional or

unintentional conduct, and decision that results in a systematic exaggeration of the treatment difference either in favor or against the study treatment in a show-difference trial. It also refers to the consequence of a systematic dampening of treatment difference in favor of the study treatment in an active control non-superiority (non-inferiority or equivalence) trial. Bias affects the estimate of the true treatment effect and may lead to drawing incorrect conclusion regarding the overall effect of the study treatment. This is important for instance in an active control non-superiority trial where it is crucial to obtain an unbiased estimate of the effect of the active control.

2.1. *Potential sources for bias in a clinical trial*

There are many potential sources for bias in a randomized controlled clinical trial. These sources include confounding, operational bias during trial execution, evaluation bias in outcome measurement, structural bias in the trial design, and statistical bias in the method of analysis. In any given situation, bias could come from one or more or even a combination of these sources. Some of these are discussed below.

2.1.1. *Confounding*

The primary objective of a clinical trial is to demonstrate that the study treatment is effective as expected. Therefore, the trial must allow the experimenter to attribute any observed effect to the study treatment, and to the study treatment alone by ruling out all other potential explanations. Confounding occurs when one cannot attribute the entire observed treatment difference to the study treatment. The consequence of confounding here is bias, a bias that may be for or against the study treatment. A common source of confounding is an imbalance between the treatment arms in some important baseline covariates or prognostic factors that may have direct influence on the clinical outcomes of interest. A standard technique for avoiding confounding is randomization. Randomization may provide some assurance that the treatment arms are relatively balanced with respect to all known or unknown factors that may affect a patient's response to the treatment. Randomization, if carried out properly, can minimize the chances for such imbalances to occur at baseline.

Confounding may occur simply due to failure of the randomization to achieve necessary ***balance at baseline*** between the treatment and the control relative to some important and relevant known or unknown

demographic or prognostic factors. When there are imbalances in known baseline covariates or prognostic factors, then one will be less assured about absence of imbalances in important but unknown factors. This type of imbalance can often occur when the sample size is modest or when the randomization scheme is flawed. It is important for a confirmatory trial to have proper randomization and to stratify relative to some of the most important factors to avoid such problems and concerns.

However, even proper randomization may not be able to protect against bias resulting from differential treatment of the experimental and the control arms. Such differential treatment of the two arms may occur through operational bias introduced as a result of unblinding, through inherent structural bias as to be illustrated by the example on duodenal ulcer prevention trial in Sec. 2.1.4, or through statistical bias in the method of analysis.

2.1.2. *Operational bias*

If a randomized controlled trial is open label, that is, if there is no blinding, then operational bias can easily be introduced. For instance, in a study of a new treatment for headache, the investigator may consciously or unconsciously allow subjects on study treatment to have concurrent use of aspirins. This type of conduct can introduce bias into the trial favoring the study treatment. ***Evaluation bias*** can also easily find its way into such trial. For example, if the evaluation involves adjudication of certain event, then the adjudicator may assess the event differently depending upon whether the subject is on study treatment or control.

For certain clinical trials, it is simply impossible to maintain blinding at all. In trials comparing different surgical procedures, it is frequently not possible to have blinding. In some oncology trials, it may not be possible to maintain blinding because the trials may involve comparing treatments with different toxicity profiles and delivery systems, e.g. pill vs. intravenous injection.

In some randomized trials, even though the trials may be blinded, they can still potentially become partially unblinded. For example, it is known that β-blocks, a class of drugs for treating certain heart diseases, can lower heart rate significantly. The mean heart rate of patients treated with β-blockers can be about 12 beats/minute lower than that of patients given placebo. Thus, knowing the specific properties of the study treatment, one can correctly guess the patient treatment assignment with a high likelihood of success, resulting in some degree of unblinding.

2.1.2.1. Levels of blinding

Blinding is one of the basic techniques used to control the potential introduction of bias in a randomized controlled trial. It is generally recommended to maintain a level of blinding permissible by the study.

When the individual patients are blinded to their own treatment assignments, the trial is called **single blind**. The blinding is generally achieved by giving patients medications that are identical in appearance but may contain either the study treatment, the active comparator, or an inactive ingredient (placebo) depending upon the type of design.

When the investigators, evaluators, raters, or anyone who can influence the course of the trial, are also blinded to the patient treatment assignments during the entire course of the experiment, the trial is called **double blind**. In a double-blind trial, provisions are made so that the investigators will be able to break the blind in individual patient during an emergency, or when it is determined that the risk to the patient requires specific cares to be taken by the clinician to protect the patient's safety. In a double blind trial, all study personnel are blinded to the patient treatment codes and should have no knowledge of or access to the results of any interim treatment comparative analysis.

2.1.2.2. Minimizing the likelihood of unblinding

In a blinded trial, how easy is it for the individual patient to unblind his/her own treatment? Can the study treatment and the control be easily distinguished through appearance, taste, shape, size, route of administration, regimen, frequency of administration, and side effects? Special blinding techniques may be needed in a trial in order to minimize the likelihood that the individual patient will be able to unblind his/her own treatment codes. Of course, the investigator and other trial personnel should be instructed not to provide any information to the patient that may help the patient to unblind his/her own treatment code.

How easy is it for the investigator, evaluator, or rater to have access to the patient treatment codes or to unblind the patient treatment assignments? Can the investigator unblind or partially unblind the patient treatment assignments through patient's reported side effects, laboratory, physiological, and clinical parameters? For example, for the congestive heart failure trial involving exercise tolerance test, measures need to be taken to ensure that the administrator who conducts the treadmill test has no access to patient treatment experiences and baseline information so that the

likelihood of unblinding the patient treatment assignments can be minimized. Or in a trial that requires event adjudication, the adjudicator should not have access to information that may help reveal the patient treatment assignments. However, for the investigator, available blinding techniques may be more limited because the investigator has access to all of the patient data. The attention here should be more on reducing the impact of potential bias in the event of unblinding by the investigator.

How easy is it for the trial personnel to have access to treatment codes, to perform treatment comparative analysis, or to receive treatment comparative analysis results? How easy is it for the trial personnel to make changes to the trial, e.g. patient enrollment criteria, addition or deletion of centers, increasing sample size, changing endpoints, etc. For example, if there is an independent Data Monitoring Committee involved, then how easy is it for the trial personnel to have access to treatment comparative analysis results through members of the Data Monitoring Committee? If the treatment comparative analysis is actually done by the trial personnel, then what safeguard is there to protect the integrity of the trial from changes done to the trial based on knowledge of the interim analysis results? Here, the only available tool to minimize unblinding is to establish a sound clinical trial infrastructure, clear designation of roles and responsibilities, and strict standard operating procedures governing the interaction, communication and dissemination of results among various operating units. These operating units include units within the trial organization such as the safety monitoring group, the data management group, the data analysis group, and the custodian of patient treatment codes. They also include units outside the sponsor's organization such as outside consultants, Contract Research Organization, and Data Monitoring Committee, if there is one. When interim analysis is planned, the standard operating procedures should address some issues including the following. Who is doing the interim analysis? How the interim analysis is being carried out? To whom the results should be reported? Who can have access to the results of the interim analysis? How confidentiality of the results can be maintained? For guidance on these issues, one may refer to the Guidance on the Establishment and Operation of Clinical Trial Data Monitoring Committees (2001).

These questions and issues regarding blinding and how to avoid or minimize the potential opportunity for unblinding should be discussed and addressed at the design stage of a confirmatory trial. Appropriate blinding techniques tailored to each trial by taking into consideration the trial design, special properties of the study treatment, and characteristics of the disease

should be described in the study protocol. Trial infrastructure and standard operating procedures should be designed in relation to these considerations.

2.1.2.3. Assessing and reducing the impact of unblinding

Of course, the mere occurrence of unblinding will not necessarily result in bias. It is the subsequent conduct on the part of some study personnel intentionally trying to influence or change the outcome either in favor or against the study treatment that results in bias. Thus, in addition to minimize the likelihood of unblinding, one should also consider reducing the severity and impact of potential bias introduced in the event of unblinding.

Generally, the severity and impact of bias introduced subsequent to unblinding depends on the level at which the unblinding occurs, the ease with which the patient response can be altered, the quanitfiability of the bias, and the magnitude of the bias relative to the overall study treatment effect size.

When some individual patients are unblinded to their own treatment assignments, then these patients may introduce biases into their own individual responses. These biases introduced by the individual patients may not be consistently in the same direction, nor systematic. So the net effect of all the biases may be much less severe. The impact of the bias depends upon the number of individual patients involved and the magnitude of the net bias relative to the overall study population size and the overall treatment effect size. Occasionally, biases introduced by one or two patients may change the overall conclusion. However, in most cases, these individual biases may not be easily quantified. Various strategies may be taken at the analysis stage to examine the impact of these biases when known. Such strategies may range from excluding some or all of these affected patients from the analysis to imputing conservative values for these patients' responses.

When unblinding occurs at the investigator level, then the potential bias introduced will affect the outcomes from the particular investigator site or center. Such bias tends to be more severe because within each center, the bias introduced tends to be more systematic and consistent, and affects the entire outcome from that particular center. In a trial with only a few centers, or when the centers involved account for a significant proportion of the total patients, the result can be devastating. If such biases are known and quantifiable, then appropriate measures may be taken to account for them at the analysis stage. Otherwise, available analysis strategies for examining

the impact of such biases may be limited only to dropping the affected centers from the analysis. In such event, randomization stratified within center may become an important issue.

Finally, when unblinding occurs at the study level, treatment comparisons may be made based on the entire available data. Knowledge of such treatment comparative analysis can lead to subtle changes in patient demographics, addition or deletion of sites, and sometimes even to early trial termination. This type of changes made to the trial may result in bias favoring the study treatment. The impact of this kind of bias may be quite significant because such changes are based on the interim treatment comparative data. The consequences of such bias may include declaring an ineffective study treatment to be effective, prescribing the study treatment for a more general patient population than warranted, and incorrect labeling of the study treatment. It may be difficult to detect this kind of bias as a result of unblinding at the study level, and even when detected, it may be difficult to assess its impact. When unblinding occurs at the study level, the entire study results may be voided.

The potential impact of unblinding may also depend upon the type of efficacy endpoints involved. There are generally two types of endpoints, the objective and the subjective. In general, for subjective endpoints such as scores based on investigator or patient evaluation, concerns for potential unblinding are understandable. In such cases, an investigator or patient can easily assign certain scores, or components of the scores, in a manner that is favoring the study treatment. Subjective endpoints arise frequently in clinical trials. For objective endpoints, the outcomes are less vulnerable to such alterations. However, some forms of operational bias can still be introduced. For example, in many clinical trials, besides the study treatment and the control, concomitant use of other effective drugs are allowed for all patients. In such cases, differential usage of concomitant medications between the study treatment and the control may potentially lead to bias.

Objective endpoints also may vary in degree of objectivity. For objective endpoints such as mortality or serious morbidity, it is difficult for anyone to alter their outcomes. For some other objective endpoints, operational bias can still be introduced. For example, in congestive heart failure trials, exercise tolerance test is often used to measure a patient's symptomatic improvement. In such trials, patients are asked to walk on treadmill. The administrator of the treadmill test would ask patients to walk until exhaustion. The total walking time will be recorded. Here the exercise tolerance test is an objective endpoint, but it may be affected by how hard the administrator

pushes the patient to exhaustion. If the administrator is unblinded or partially unblinded to the patient treatment assignments, then differential handling of patients depending upon their treatment assignments will introduce operational bias. So in such cases, the administrator should not have access to the patient data and the investigator should not communicate to the patient his/her treatment assignment and treatment information.

It is clear that for a confirmatory trial, one should require double blinding. In addition, measures for minimizing the likelihood of unblinding and steps for reducing the impact of bias in the event of unblinding should be described clearly in the protocol. These efforts should include a sound infrastructure for the trial, clear designation of roles and responsibilities, and strict standard operating procedures that are tailored for the trial at hand and that take into account the type of design and specific trial requirements.

2.1.3. *Structural bias*

In a controlled clinical trial, even randomization and blinding may not fully protect against ***structural bias*** resulting from flaws in the design. Design flaws can occur, and not infrequently. Thus, one should be on guard against such structural bias. When it occurs, it has the potential to invalidate the results of the entire trial. There is usually no satisfactory remedy for structural bias as illustrated in Examples 1 and 2.

Example 1. In the late 1970s and early 1980s, it is customary for duodenal ulcer prevention trials to consider a double-blind placebo-controlled design with patients whose duodenal ulcers were recently healed on acute treatment randomized to either the same treatment at half the dose or placebo. Patients were scheduled at regular intervals for endoscopic examination to determine the presence or absence of duodenal ulcers. The intervals are usually of 3-, 4- or 6-months duration, and the entire trial usually lasts 12 months. At the request of the physician, patients could also be endoscoped upon presentation of symptoms such as pain. Patients found to have recurrent duodenal ulcers were discontinued from the study; the remaining patients continued in the trial until symptomatic recurrence, the next scheduled endoscopy, or the end of the trial. So these trials met the basic requirements of being randomized, double blind and placebo-controlled. What is the problem? For H2-receptor antagonists such as ranitidine and cimetidine, it was generally recognized that they provide symptom relief and promote healing of duodenal ulcers in short term acute treatment trials.

In the prevention trial, patients with symptomatic recurrent ulcers are discontinued from the trials. Therefore, if the test drug merely relieves symptoms without actually preventing recurrences, then simply by the present design, more symptomatic recurrences would be observed among patients on placebo. To further compound the problem, regular endoscopies are scheduled 3, 4 or 6 months apart. It is also known that relatively short time, 4–8 weeks, is required for duodenal ulcers to heal either with or without treatment in an acute trial. Thus it is entirely possible that during the intervals between successive endoscopies, symptomatic ulcers may have recurred among patients who remain in the trial, healed before the next scheduled endoscopy, and hence escaped detection. It is clear that there is a bias favoring the H2-receptor antagonist as a result of the properties of the drug and the design. These and related issues were extensively discussed at a FDA Gastrointestinal Advisory Committee meeting.[52] A renowned gastroenterologist at the time questioned whether duodenal ulcers could recur under maintenance treatment. However, it was demonstrated in a subsequent trial that duodenal ulcers could recur under maintenance treatment.[8,46,62] A more detailed discussion of this example and the related design issues can be found in Chi,[10] and Elashoff *et al.*[18] This interesting example illustrates a combination of design flaws, evaluation bias due to the analgesic property of the H2-blockers, and the spontaneous healing of duodenal ulcers over time.

2.1.4. *Statistical bias*

A randomized controlled trial that is blinded and has no design flaw can still have ***statistical bias*** introduced at the final analysis stage. Statistical bias can arise as a result of the method of analysis, the manner in which patients or data are excluded from the data set, or the manner in which missing data are being handled. This kind of statistical bias can sometimes be fairly subtle.

An important principle in the analysis of clinical trial data is the so-called intent-to-treat principle,[20,28,47] see ICH-E9 (1999). The intent-to-treat principle simply espouses the view that the primary analysis should be performed on the outcome measures from all of the randomized patients. When there is no other source of bias, such as design flaw, then the intent-to-treat analysis should provide an unbiased estimate of the treatment effect when there are no patient exclusions and no missing data. When the outcome measures are not available from all of the randomized patients, then

efficacy subset analysis is likely to provide biased estimate of the treatment effect. Little and Rubin[53] and Little[54] defined three types of missing data mechanisms, missing completely at random (MCAR), missing at random (MAR) and informative missingness. There are various statistical models proposed to handle clinical trial data with these types of missing mechanisms. However, missing data mechanisms in clinical trials are difficult to verify whether they are MCAR, MAR or informative missingness. Generally, clinicians and biostatisticians in the field believe that most missing data mechanisms in clinical trials are likely to be informative missingness for which statistical approaches are less well developed. One imputation method that has often been used is the so-called Last Observation Carried Forward (LOCF) analysis where the last available observation is substituted for all the subsequent missing observations. The LOCF analysis implicitly assumes that the last observation is an unbiased representation of what the missing observation would have been had the patient been followed, which is also an unverifiable assumption. Various issues related to the LOCF analysis are discussed.[6,28,47,59,66,103]

The handling of missing data is a difficult problem, and the best strategy is to adhere to the intent-to-treat principle by minimizing the likelihood of missing data. The ICH E-8 Guidance on General Considerations for Clinical Trials (1997) recommends that the study protocol should specify procedures for the follow-up of patients who stop treatment prematurely. Furthermore, the ICH E-9 Guidance on Statistical Principles for Clinical Trials (1998) states that compliance with the intent-to-treat principle requires complete follow-up of all randomized patients for the primary outcome measures. For a confirmatory trial, procedures for such complete follow-up of all randomized patients should be carefully spelled out in the protocol and diligently carried out during the trial. The intent-to-treat principle should be followed with robustness or sensitivity analyses performed if needed.

The following example illustrates an interesting case of structural bias combined with analysis bias in a study for a serious progressive disease.

Example 2. In this example, the investigational new drug is manufactured only in limited quantity; hence only a small percent of the patients can be given the new drug. The study protocol has an unusual design. After a patient satisfies some eligibility criteria, he/she enters a pool and becomes eligible for random selection for the study treatment according to the fixed percentage. If a patient is selected, then the study treatment will be given. On the other hand, if a patient is not selected at a given pool of eligible,

then that patient remains unselected until the next selection. If that patient dies before the next selection, then he/she will be counted as an event in the unselected group. Otherwise, at the next selection, this patient enters the pool along with the newly eligible. Again patients in this second pool will have the same probability of being selected for the study treatment. This process continues with selections conducted about every month for over a year. Generally, for the selected patients, there is a delay of about three months between the time of selection and the actual time of administration of the study treatment. There is a one-year follow-up. The primary endpoint of interest is mortality. The length of survival of a patient is measured from the time of first eligibility to subsequent death, lost-to-follow-up, or to the end of the study. The comparison of the survival experience between the selected group and the unselected group from all ten selections is evaluated by the log-rank test statistic.

There are two kinds of bias in this example. The first kind is a structural bias, and the second is an analysis bias. The structural bias can be seen as follows. The definition of selected and unselected groups are outcome (survival) dependent. This is because a patient who was not selected in any given selection could still be selected in a subsequent selection provided that that patient was able to survive till that selection. A patient who was not selected and died before a selection would automatically remain in the unselected group. This design would enrich the selected group with patients that have better survival experience up to the time of selection, and enrich the unselected group with patients who died before a selection.

There is also a bias in the survival analysis. If a patient who is selected after a given selection, then his/her survival time is measured from the time of first eligibility to either death, lost-to-follow-up, or end of the study. But for such patient, the time from first eligibility to the time of selection is actually not under the study treatment. In fact, in general, there is an additional delay of about three months before the study treatment is actually given to a selected patient. Thus a large proportion of his/her survival experience would not be under the study treatment, but would be attributed to the study treatment. Such bias favors the study treatment. However, it is not advisable to consider the survival time as measured from the time of selection or the actual time the selected patient is given the study treatment, because the selected group and the unselected group are outcome dependent and hence there is no valid randomization.

As it turns out, in this example, a majority of the patients were considered for eligibility in the second selection. To maintain randomization

and to minimize bias, it was recommended that patients from the second selection be considered for analysis. Thus, the selected patients consist of those who were selected in the second selection, and the unselected patients consist of those who were not selected at the second selection. Now, for these unselected patients, some of them were actually selected at a subsequent selection. So, in the survival analysis, the survival time for the unselected patients are measured from the time of first eligibility to either death, lost-to-follow-up, or censored at time of subsequent selection, actual time of study treatment administration, or the end of the study.

2.2. *Some measures to minimize potential bias*

In view of the various potential sources of bias in a clinical trial, it is important for a confirmatory trial to consider adopting appropriate measures at the design stage to minimize the impact of potential biases.

Randomization is the standard technique used in clinical trials to achieve balance in both known and unknown important baseline covariates, prognostic and demographic factors between the treatment arm and the control. It is important to use proper randomization scheme that will not lead to unblinding to achieve the necessary balance.

Operational bias is difficult to control. Blinding is the most important technique for controlling operational bias. A confirmatory trial should be blinded at the study level. If necessary, special blinding techniques should be considered at the individual patient level and the investigator level. The aim is to minimize the likelihood of unblinding by the individual patients, the investigators, evaluators or raters, or the study personnel and to minimize its impact in the event of actual unblinding. In order to minimize the impact of bias due to unblinding, randomization should be centralized and stratified within each center, provided that the randomization procedure itself does not lead to potential unblinding. All blinding techniques should be clearly documented, described, standardized and operationalized. At the study level, the only blinding technique that can be implemented is through proper infrastructure, clear delineation of roles and responsibilities, and strict standard operating procedures. For instance, it should be made clear who has custody of the patient treatment codes, circumstances under which the patient treatment codes can be accessed, the parties that may have the authority to access them, and the proper procedure for such accession. These considerations should also include outside consultants, contract research organizations, and data monitoring committees. This is especially

critical when the trial involves planned interim analyses and a data monitoring committee. When there is a data monitoring committee, there should be standard operating procedures that define the responsibility and the proper span of authority of this committee, and govern the conduct and communication between the trial sponsor and the committee, whether the committee is independent or not. For example, when a data monitoring committee requests the interim results from the trial sponsor, who in the sponsor's organization is doing the interim analysis, and how can the interim analysis results be kept confidential from the sponsor's study personnel?

When the independent Data Monitoring Committee (DMC) recommends to the trial sponsor that the study be terminated early, certainly the trial sponsor should be provided with the interim analysis results and the reason for the recommendation. The sponsor's study personnel that receive such recommendation and the interim analysis results would have the interim comparative results. Thus the sponsor's study personnel would now have access to the comparative interim results. If the sponsor decides to continue the trial, then there is a potential opportunity for operational bias to be introduced by the study personnel at this point. How can one prevent such potential opportunity for operational bias to be introduced? The recent FDA draft document on the Establishment and Operation of Clinical Trial Data Monitoring Committee (2001) provides some relevant guidance on these issues.

To avoid structural bias, properties of the study treatment, type of treatment administration, the nature of the disease, the objective of the trial and other pertinent information should be well understood to insure that the design of the trial is not flawed.

To reduce the problem with missing data, it is recommended that a confirmatory trial should attempt complete follow-up on all missing primary response data from dropouts or others, and better documentation of reasons for dropping out and missing primary response data. This will minimize the impact of bias and may provide the basis for determining the proper method of handling the missing data.

To properly assess the strength of evidence in a trial, one needs to know whether the blind has been broken. If the blind has been broken, what is the extent of unblinding? Whether and how such information was used to bias the trial outcome? Whether the bias can be quantified? What analysis strategies can be applied to examine the impact of the bias, and whether the trial can still provide the strength of evidence sought, if at all? These are important issues to be addressed in a confirmatory trial.

3. Inflation of the Overall Type I Error Rate

The second fundamental statistical principles in clinical trial is to control the overall probability of type I error, i.e. the probability of declaring the study treatment to be effective when it is not, at a desired significance level of α. In a given experiment, this probability can frequently be inflated in various ways. For example, when the test statistic for the null hypothesis involves a nuisance parameter that is estimated from the data, then this may inflate this probability. This section discusses some frequently encountered situations in which this probability can be inflated.

3.1. *Multiple testing*

Multiple testing surfaces frequently in clinical trials. It manifests itself in various forms. It often appears in multiple comparisons, repeated testing, multiple endpoints, multiple indications, and subgroup analyses, or combinations thereof. The basic problem with multiple testing in a clinical drug trial is that it increases the probability of the overall type I error, i.e. the probability of declaring the drug or certain dose of the drug to be effective for the desired treatment indication when in fact it is not. This inflation can be illustrated as follows in the context of multiple comparisons.

3.2. *Multiple comparisons*

In clinical drug trials, it is often of interest to design a parallel placebo-controlled study with multiple treatments, or multiple doses of a test drug. The primary objective of such a trial is to demonstrate that one or more of the treatments, or doses of the drug works, and a secondary objective is to compare different treatments, or to characterize the drug's dose response relationship. For such a study, how should one most efficiently evaluate the effect of the treatments or the effect of the drug? How should the primary hypothesis or hypotheses be defined? And how should the test or tests be carried out?

For purposes of discussion, consider N treatments or doses of a drug and a placebo. Let $\mu_i, i = 0, 1, 2, \ldots, N$ represent the mean changes from baselines for placebo and the N treatments or doses of the drug respectively. Let $H_{o123\cdots N} : \Delta\mu_1 = \Delta\mu_2 = \cdots = \Delta\mu_N = 0$ represents the global null hypothesis. Let $H_{oi} : \Delta\mu_i = 0, i = 1, 2, \ldots, N$ denote the individual null hypotheses of no difference between the ith-treatment or the ith-dose of the drug and placebo. Let us suppose that we have suitable test statistics

to test each one of these individual null hypotheses, $H_{oi} : \Delta\mu_i = 0, i = 1, 2, \ldots, N$, and let $p_{oi}, i = 1, 2, \ldots, N$ denote the p-values associated with the corresponding tests. Now can we reject an individual null hypothesis, $H_{oi} : \Delta\mu_i = 0$, if the p-value $p_{oi} < \alpha$? The answer is not necessarily so, because by testing the individual null hypothesis H_{oi} at the same α level does not control the probability of the overall type I error, the probability of rejecting at least one individual null hypothesis when the global null hypothesis is true, at the α level. This probability will be inflated as a result of the multiple testing. The inflation of the overall type I error rate can be seen as follows. If there are N individual and independent comparisons with each individual comparisons done at a nominal significance level of α_i, then

The probability of the overall type I error

$$
\begin{aligned}
&= P(\text{Reject } H_{o123\cdots N} | H_{o123\cdots N}) \\
&= P(\text{Reject at lease one } H_{oi} : \Delta\mu_i = 0 | H_{o123\cdots N}) \\
&= 1 - P(\text{Fail to reject all } H_{oi} : \Delta\mu_i = 0 | H_{o123\cdots N}) \\
&= 1 - \Pi P(\text{Fail to reject } H_{oi} : \Delta\mu_i = 0 | H_{oi}), \text{ assuming independence} \\
&\geq 1 - \Pi(1 - \alpha_i) . \qquad (1)
\end{aligned}
$$

If the individual comparisons are made at the same significance level of $\alpha_i = \alpha$, then from Eq. (1), it follows that under independence assumption,

$$\text{The probability of the overall type I error} > 1 - (1 - \alpha)^N > \alpha . \qquad (2)$$

The following values illustrate the increase in the probability of the overall type I error rate as a result of an increase in the number of comparisons each performed at $\alpha = 0.05$ level. The probability of the overall type I error

$$= 0.05, 0.098, 0.143, 0.226, 0.401$$

corresponding to $N = 1$, 2, 3, 5 and 10 comparisons, respectively.

From these values, Eq. (1), and inequality (2), it is clear that in order for the probability of the overall type I error to be controlled at a given α level, it is necessary to test the individual null hypothesis H_{oi} at a significance level $\alpha_i < \alpha, i = 1, 2, \ldots, N$, in such a way that the probability of the overall type I error does not exceed α. Various multiple comparison procedures have been proposed to accomplish this.

3.2.1. *Some p-value based multiple comparisons procedures*

The most commonly used multiple comparison procedures are those that can be described as p-value based procedures, so-named because these procedures are all defined in terms of the individual p-values, p_{oi}. These general procedures include the simple Bonferroni procedure, the Holm's[35] sequential rejective procedure, the Hochberg[33] procedure, the Hommel[37] procedure, and the more general Simes procedure. Except for the Simes procedure, most of these procedures are not powerful when the number of comparisons increases. However, they control the probability of the overall type I error at the desired α level, and they have enjoyed great popularity due to their relative simplicity. For a more detailed discussion of these and related procedures, the reader is referred to Hochberg and Tamhane,[34] Samuel-Cahn,[81] Chi,[11] and Sarkar.[83]

These p-value based procedures are all inferences based on the individual p-values, $p_{o(i)}, i = 1, 2, \ldots, N$, associated with the individual null hypotheses, $H_{o(i)}, i = 1, 2, \ldots, N$. These p-value based procedures adjust the individual α_i in such a manner that the probability of the overall type I error is controlled at the desired α level. If these p-value based procedures reject any one of the individual null hypotheses at the α_i level, then the global null hypothesis, $H_{o123\cdots N} : \Delta\mu_1 = \Delta\mu_2 = \Delta\mu_3 = \cdots = \Delta\mu_N = 0$, will be rejected for sure, where $\Delta\mu_i$ is the difference between the ith-treatment or the ith-dose of the test drug and placebo. These p-value based procedures are sometimes also referred to as the step-up procedures. They generally do not make use of the correlation that may exist between the various test statistics and tend to be conservative. Furthermore, it should be noted that most of the p-value procedures implicitly treat all comparisons as of equal importance.

3.2.2. *The closure principle and the step-down procedures*

In contrast, a test that rejects the global null hypothesis, $H_{o123\cdots N}$, at a significance level of α permits one to conclude that overall the drug is superior to placebo. However, the rejection of this global null hypothesis does not shed any light as to which specific treatment(s) or dose(s) actually works. Certainly, if the test fails to reject the global null hypothesis, then one can only conclude that the study fails to detect a difference among various treatments and placebo, or among different doses of the drug and placebo.

The US regulations [21 CFR Sec. 314.50(d)(5)(v)] also require that evidence be provided to support the dosage and administration section of the labeling, including support for the dosage and dose interval recommended. Thus, when a test rejects the global null hypothesis, it is also of interest to know which one or more of these treatments or doses of the drug is actually superior to placebo. Can one then simply use the p-values, $p_{o(i)}$, obtained from the individual comparisons of $H_{o(i)} : \Delta\mu_i = 0, i = 1, 2, \ldots, N$ and compare each at the same nominal significance level say, α, and then conclude that that treatment or dose is or is not superior to placebo? The answer is no, not in general, because suppose that the true state of nature is as follows:

$$\Delta\mu_1 = \Delta\mu_2 = \cdots = \Delta\mu_{N-1} = 0 \quad \text{and} \quad \Delta\mu_N \neq 0\,. \tag{3}$$

Now suppose the global null is rejected, and then one tests each of the individual null hypothesis, $H_{o(i)}, i = 1, 2, \ldots, N$ at the α level. It follows from the series of equations in (1) that the probability of rejecting at least one of the individual null hypotheses, $H_{o(i)}, i = 1, 2, \ldots, N-1$ assuming independence, is $> 1 - (1-\alpha)^{N-1} > \alpha$ given the true state of nature according to Eq. (3). Thus, the probability is greater than α that one will erroneously reject one of the individual null hypotheses, $H_{o(i)}, i = 1, 2, \ldots, N-1$, and conclude that one of the first $(N-1)$ treatments or doses is effective.

Equation (3) is only one possible true state of nature. The following lists out all the remaining true states of nature.

Level 0 : $\Delta\mu_1 = \Delta\mu_2 = \cdots = \Delta\mu_N = 0$

Level 1 : $\Delta\mu_1 = \Delta\mu_2 = \cdots = \Delta\mu_{i-1} = \Delta\mu_{i+1} = \cdots = \Delta\mu_N = 0\,,$
$\Delta\mu_i \neq 0,\ i = 1, 2, \ldots, N$

Level 2 : $\Delta\mu_1 = \Delta\mu_2 = \cdots = \Delta\mu_{i-1} = \Delta\mu_{i+1} = \cdots = \Delta\mu_{j-1} =$
$\Delta\mu_{j+1} = \cdots \Delta\mu_N = 0,\ \Delta\mu_i \neq 0,\ \Delta\mu_j \neq 0,\ i = 1, 2, \ldots,$
$N-1,$ all $j > i\,.$

. . . . . .

Level $N-1$: $\Delta\mu_i = 0,\ i = 1, 2, \ldots, N,$ and $\Delta\mu_j \neq 0,$ all $j \neq i\,.$

Since the above argument is also applicable to all the other possible true states of nature, it follows that controlling the probability of the type I error associated with the global null hypothesis, $H_{o123\cdots N}$, at the α level, and the probability of the type I error associated with the individual null hypotheses $H_{o(i)}, i = 1, 2, \ldots, N$ each at the α level, is not sufficient to

protect the probability of the overall type I error at the α level. That is, in this case, the probability of erroneously rejecting one of the individual null hypotheses when it is true will be greater than α. In other words, simply because an individual p-value $p_{o(i)} < \alpha$ it does not imply that one can reject the corresponding individual null hypothesis $H_{o(i)}$, even when the global null hypothesis has been rejected. However, there are some classical multiple comparisons procedures within the framework of analysis of variance that permit such conclusions. For example, the Scheffé test, the Dunnett procedure, and the Tukey procedure. An interesting example of a global test within the context of a multi-factorial combination drug trial is discussed in Hung, Chi and Lipicky.[38]

This suggests that perhaps one should consider controlling the probability of the type I error associated with all the possible true states of nature. The closure principle proposed[60] to a general class of closed testing procedure is a step-down procedure that begins with testing the global null hypothesis (corresponding to the state of nature in Level 0) at the α level. If this test fails to reject the global null hypothesis, then the procedure stops. If this test rejects the global null hypothesis, then the procedure steps down to test each of the partial null hypotheses at the next lower level (corresponding to the states of nature in Level 1) at the same α level. The procedure stops when none of the partial null hypotheses is rejected. When one or more of the partial null hypotheses are rejected, then the procedure steps down to the next lower level (corresponding to the states of nature in Level 2). It tests each of the partial null hypotheses that imply the previously rejected partial null hypotheses at the α level (A hypothesis H_1 implies another hypothesis H_2, if the rejection of H_1 implies the rejection of H_2). The process either stops at some level for failing to reject any of the implying partial null hypotheses at that level, or continues to the last level (corresponding to the true states of nature in Level $N-1$) and tests each of the implying individual null hypotheses at α level.

The closed testing procedure guarantees that the probability of the overall type I error will be maintained at the α level. However, it does not guarantee that in a given application, it will necessarily be able to continue to the last level of testing all the implying individual null hypotheses. Even when it does reach the last level, it may fail to reject any of the implying individual null hypotheses. In other words, the closed testing procedure may fail to identify any treatment or dose that is superior to the control. A closed testing procedure may be able to can take advantage of the correlation between the various test statistics. This feature makes the

closed testing procedure somewhat more attractive when one is confronted with multiple endpoints which is the next multiplicity issue to be discussed. Again it should be noted that the closed testing procedure also implicitly treats all comparisons as of equal importance.

3.3. *Multiple endpoints*

In medicine, a disease is characterized by multiple clinical endpoints that may be correlated. These clinical endpoints may describe the stage and severity of the disease, and the signs and symptoms associated with the disease, etc. Some clinical endpoints provide direct information regarding the state of the disease; they are capable of revealing whether the disease has been modified, such as healing of ulcer as determined by endoscopic examination. Some clinical endpoints provide indirect information regarding the state of the disease; these clinical endpoints include signs and symptoms known to be associated with the disease. In a disease where death is a potential outcome, mortality or survival is considered as a very special and unique clinical endpoint; it is an objective endpoint, and due to its seriousness, it supersedes all other endpoints in importance. On the other end of the spectrum are the so-called surrogate endpoints or surrogate markers, or biomarkers. They are endpoints, usually not clinical in nature, but may be associated or correlated with the clinical endpoints of interests.

The effect of a drug may manifest itself in a number of endpoints, and each endpoint may provide a measure of drug effect. From a regulatory perspective (The 1962 Amendment to the 1938 Food, Drug and Cosmetic Act requires drug to show clinical benefit in adequate and well-controlled studies), only endpoints that can provide a measure of relevant clinical benefits that may lead to potential efficacy claims are of interests. This requirement essentially restricts the endpoints to clinical endpoints. But clinical endpoints have varying importance and reliability, and not every clinical endpoint has the potential of leading to an efficacy claim. In a few special instances, surrogate endpoints have been used directly to support drug efficacy claim, as in blood pressure for hypertension trials and CD4 counts in AIDS trials. In some disease areas, due to the lack of good or available treatments, a surrogate endpoint may be used to support drug efficacy claim conditional on subsequent demonstration of meaningful and real clinical benefit under the accelerated approval program (See Mathieu[61] and 21 CFR Sec. 314.500 Subpart H, 2001).

Therefore, in the design of a clinical drug trial, the investigator should first determine what constitutes a clinical benefit and how best to measure

it. The ICH-E9 Guideline (1998) calls for the designation of a single clinical endpoint as the "primary endpoint". The effect of the drug will then be measured relative to this "primary endpoint". This recommendation may be reasonable when there is only one clinically relevant endpoint for the disease under study, or when there is only one clinical endpoint that is of primary interest. Often in practice, this recommendation is followed because the investigator is not willing to make the necessary adjustment for multiplicity testing or to consider applying a closed testing procedure that would preserve the probability of the overall type I error at α. In either case, the investigator usually would designate one among several endpoints as the "primary endpoint", and the remaining endpoints as "secondary endpoints". Such practice often leaves ambiguity as to what to do with those "secondary endpoints" that may lead to the efficacy claim themselves. This often leads to difficult situations and undesirable consequences as illustrated by the following scenario.

After a trial fails to reject the null hypothesis defined by the designated "primary endpoint" at the significance level α, the investigator then turns to the "secondary endpoints". Often, some of the observed p-values for these endpoints show "apparent significance", i.e. with p-values less than α. The question is whether one can claim that the trial has demonstrated that the drug is effective based on those "secondary endpoints" that show "apparent significance". This type of situation often results in controversy, particularly when the "secondary endpoint" turns out to be mortality or some serious irreversible morbidity endpoints. For an interesting recent discussion of such an example, one may refer to the articles by Fisher[21,22] and Moyé.[63–65] See also Chi[10] for the SOLVD prevention trial.

From a statistical and regulatory standpoint, when the primary hypothesis fails to be rejected at the significance level α, then the trial has failed to provide the expected strength of evidence for the efficacy of the drug. Additional testing of hypotheses can only result in an inflation of the probability of the overall type I error over and beyond α. The probability of the overall type I error can be controlled only if all desired hypotheses to be tested have been pre-specified with a proper allocation of α. A recent proposal by Moyé[64,65] and related commentaries[16,70] considers testing the secondary hypotheses at some significance level $\alpha*$ between 0.05 and 0.10 to accommodate the so called "surprise finding". This proposal is clearly inflating the overall type I error rate beyond the desired level of $\alpha = 0.05$ and is not desirable from this perspective.

In our view, for a confirmatory trial, one should maintain the fundamental principle of controlling the probability of the overall type I error at

the desired level α. Thus, designating only one of several clinical endpoints as "primary" may not necessarily be the best option.

3.3.1. *Current clinical trial practice*

When there are multiple clinical endpoints, each of which can lead to the same efficacy claim, the current practice usually considers one of two general approaches. The first approach is to designate all clinical endpoints that can lead to the same claim as "co-primary endpoints". Appropriate hypotheses are then defined in terms of these endpoints. A method for testing these hypotheses that accounts for multiple testing can then be applied. The second approach is a conditional or sequential approach. One designates only one "primary endpoint" to be tested at the significance level of α as recommended in the ICH-E9 Guidance document. The remaining endpoints are then designated as "secondary endpoints". If the primary null hypothesis fails to be rejected, then no efficacy claim can be made. When the primary null hypothesis is rejected, then the efficacy claim is made. Additionally, those "secondary endpoints" that are closely correlated with the "primary endpoint" can be tested at the nominal significance level of α; but these tests will not lead to additional efficacy claims, but are meant to provide additional information to describe the efficacy finding based on the "primary endpoint". For those "secondary endpoints" that are not closely related to the "primary endpoint", they may be tested at an overall nominal significance level of α appropriately adjusted for multiplicity. Those "secondary endpoints" that reach nominal significance after adjustment may be described in the labeling of the drug.

In the above two approaches, to adjust for multiple "co-primary endpoints" or to adjust for multiple "secondary endpoints" after rejection of the primary null hypothesis, one may consider various multiplicity adjustment procedures such as the p-valued based procedures and the closed testing procedures that were discussed in Secs. 3.2.1 and 3.2.2. These procedures may be applied here if one assumes that the "co-primary endpoints" are of equal importance and have equal likelihood of demonstrating the proposed efficacy claim, or that the "secondary endpoints" are of equal clinical relevance and importance in terms of allowing their appearance in the labeling of the drug. A few examples of multiple testing has been discussed by many authors.[2,17,27,51,67,69,76,80,82,87,88,105,106]

The p-value based procedures and the closed testing procedures are only formal statistical procedures and they lack the proper clinical and

regulatory perspectives. They are being applied as if all the endpoints are of equal importance and have equal likelihood of achieving the desired efficacy claim. However, in most practical situations, such assumption is not appropriate. The fact is that most of the times, these clinical endpoints have varying or differential clinical relevance and importance, and different likelihood of demonstrating the efficacy claim. Furthermore, from a clinical and regulatory perspective, some of these endpoints may be used together rather than individually in a more complex way in assessing the drug effect. The proper approach, in our view, is to define *a priori* a ***decision set*** and a ***clinical decision rule*** for assessing the drug efficacy claim. A decision set is simply a set of clinical endpoints and a clinical decision rule is simply a decision tree consisting of several decision paths or branches. Each decision path is defined by one or more of the endpoints from the decision set and it may lead to the desired efficacy claim. The null hypothesis at each decision path is tested by a test statistic that reflects this decision path. If the null hypothesis at each decision path is being rejected at certain significance level α_i, then the trial would have demonstrated efficacy. The significance levels α_i are allocated across the various decision paths in a manner that will maintain the overall probability of type I error. A clinical decision rule should have proper statistical support structure to provide valid statistical inference. The concept of clinical decision rule is more formally discussed and illustrated by examples in the next section.[13,42]

3.4. *Clinical decision rules*

It is important to understand that in current clinical trial practice, the term, "primary endpoint", is used to identify a clinical endpoint that is being used in a trial for demonstrating the efficacy of the study treatment. It is an endpoint that has the burden of providing the primary evidence for the desired efficacy claim. When there are more than one "primary endpoint", then they are "co-primary" relative to one another. The "secondary endpoints" and "tertiary endpoints" do not have such a role. But the mere fact that a clinical endpoint is placed in the category of "secondary endpoint" does not imply that it has lesser clinical importance or lacks the ability to independently provide the evidence for an efficacy claim. This is one reason why controversies frequently arise in clinical trials as in the case of carvedilol discussed earlier.

A clinical decision rule is a natural and meaningful way of assessing drug efficacy. It reflects the clinical and regulatory perspectives in handling the

multiple testing issue. How a clinical decision rule is defined actually varies from disease to disease. It depends on the current state of clinical knowledge about the disease, and the degree of acceptability of the endpoints considered. In order to define clinical decision rule in a clear and clinically meaningful way, some general definitions of clinical endpoints are needed.

3.4.1. *A hierarchical definition of clinical endpoints*

The following hierarchical definitions of clinical endpoints are proposed with a regulatory perspective.

Definition 1. A ***clinical endpoint*** is a clinical variable which either directly or indirectly reflects the condition of an underlying disease.

For example, death, disease progression, tumor size, pain intensity, signs/symptoms, hospitalization are clinical endpoints.

Definition 2. A clinical endpoint is a ***primary endpoint*** if it satisfies the following conditions:

- It can provide a measure of clinical benefit realized in the patient that is acceptable by the clinicians in the field as a meaningful measure of the drug effect for the disease under treatment.
- It is an endpoint such that a positive finding in this endpoint may result in an efficacy claim.

Examples of primary endpoints are time to death, time to disease progression, tumor response, diastolic blood pressure, presence or absence of ulcer.

Definition 3. A primary endpoint is a ***principal*** primary endpoint if a positive finding in this endpoint alone is sufficient to result in the efficacy claim.

Principal primary endpoints are usually endpoints that are **objective** and can **directly** demonstrate that the underlying disease has been modified. For example, mortality, absence of ulcers, and objective measures of disease progression. The regulatory perspective here is that a positive finding on a principal primary endpoint ***alone*** is sufficient for proof of efficacy. For many diseases, there may not be a principal primary endpoint. In general, the choice for principal primary endpoint may be limited.

Definition 4. A primary endpoint is a ***co-primary*** endpoint if a positive finding in this endpoint is sufficient to result in an efficacy claim, provided other primary endpoints considered do not show an inconsistent effect.

Co-primary endpoints are endpoints that may be more **subjective** in nature, less well defined, or may provide only **indirect** measure of the condition of the underlying disease such as signs and symptoms. The regulatory perspective here is that a positive finding on a co-primary endpoint alone is sufficient for proof of efficacy **only if other primary endpoints considered do not show an inconsistent effect**.

Some examples of co-primary endpoints include all cause hospitalization, cause specific hospitalization, time to disease progression, tumor response, disease related signs/symptoms, etc. From a clinical perspective, co-primary endpoints are not as important as principal primary endpoints because they may not be as objective, or as well defined, and may not provide direct measures of disease modification. On the other hand, improvement as measured by a co-primary endpoint may support a claim provided other primary endpoints considered do not show contradictory findings.

It is important to point out that by our definitions, both principal primary endpoints and co-primary endpoints are primary endpoints. Their nature can not be changed simply because they are placed in the "secondary endpoint" category, a common term used in current practice.

Definition 5. A clinical endpoint is ***secondary*** if it satisfies the following conditions:

- It is a clinical endpoint that provides a measure of clinical benefit realized in the patient that is acceptable by the clinicians in the field as a meaningful measure of the drug effect for the disease under treatment.
- It is an endpoint such that a positive finding in this endpoint alone is ***not*** sufficient to result in the drug's efficacy claim.

It should be pointed out that secondary endpoint as we define here has a different meaning than that used in current practice. In current practice, "secondary endpoints" may include both primary and secondary endpoints as we define above.

It should be noted that according to our definition, a positive finding in a secondary endpoint alone could not result in an efficacy claim. However, positive findings in several secondary endpoints may together provide sufficient evidence to support a claim. This is illustrated in Example 6. Evidence from secondary endpoints may or may not be needed to support the evidence provided by a primary endpoint.

There are several reasons for the above hierarchical definition of clinical endpoints. The first reason is that in each disease, there are usually multiple relevant clinical endpoints of varying importance. Secondly, efficacy

assessment depends on the nature of the clinical endpoints and the relevance of the clinical benefits that are measured by these clinical endpoints to the indication sought. Thirdly, the hierarchical nature of these endpoints would become important in defining the clinical decision rule, to be defined below, which essentially operationalizes these endpoints in the assessment of efficacy.

Definition 6. A ***decision set*** relative to a disease is a set of relevant endpoints determined by the clinicians that will be used to assess the effectiveness of the treatment for the disease in a clinical trial. A decision set may contain principal primary endpoints, co-primary endpoints, and secondary endpoints.

Ideally, the decision set for a disease should be determined by the consensus of the medical experts in the field. However, for diseases that are not well understood, consensus may be hard to come by. But it is exactly in this kind of disease where multiple endpoints become a difficult issue. This is because, the experts themselves may not know exactly which clinical endpoints are most relevant and important, and they often may not even agree.

A decision set may consist of principal primary, co-primary, and secondary endpoints, and it may even contain surrogate endpoints. A decision set may change over time as better understanding of a disease may modify the way efficacy should be assessed. It usually should not contain too many endpoints. It should contain sufficient number of endpoints to allow all potentially acceptable ways of assessing efficacy of the drug.

The hierarchical nature of the endpoints is operationalized in a clinical decision rule.

Definition 7. A ***clinical decision rule*** is simply a decision tree defined relative to a decision set for the purpose of assessing the effectiveness of a drug in treating a given disease. Each decision path or branch is defined in terms of one or more endpoints from the decision set. In each path, a decision is made based on the outcomes of all endpoints involved.

It is important to realize that the clinical decision rule as defined here simply represents the clinical and regulatory way of expressing how evidence of efficacy of the drug can be assessed based on the endpoints in a decision set. It is independent of any statistical considerations. It simply reflects the various alternative hypotheses of interest. For instance, no consideration is given to how the hypotheses are defined, control of the probability of the overall type I error, etc.

Therefore, a clinical decision rule should have a proper statistical support structure. A proper statistical support structure should include appropriate null and alternative hypotheses reflecting the entire clinical decision rule. It should have proper test statistic at each decision point, appropriate allocation of α to each decision point so as to maintain the probability of the overall type I error for the clinical decision rule at α, and sufficient sample size or power for the entire clinical decision rule.

Definition 8. A ***decision structure*** is a clinical decision rule with a proper statistical support structure.

A decision structure requires proper allocation of α among the various decision paths of a clinical decision rule. If one wishes to allocate all the α to one particular decision path, then this can easily be accommodated by defining all the other decision paths as sub-paths of this particular decision path, or simply defines the clinical decision rule with only this path. The sub-paths will be tested only when the primary decision path has demonstrated the efficacy of the drug.

Decision structure generalizes the concept of "primary endpoints" and "secondary endpoints" used in current practice by *decision paths*, and *sub-paths* in a conditional decision path of a clinical decision rule.

The earlier discussion of labeling consideration for "secondary endpoints" can also be transferred to the present decision structure framework by considering clinical decision rule with sequential or conditional paths that allow secondary outcomes based on sub-paths to enter the labeling of the drug.

Ideally, for a given disease, the clinical decision rule should be based on a consensus of the clinical experts in the field. For example, the Food and Drug Administration has advisory committees for various disease areas. The members of the advisory committees include medical experts in various diseases. These advisory committees often discuss issues regarding proper choices of endpoints. The consensus regarding decision set and clinical decision rule in a given disease may be reached among the corresponding committee members. For some diseases, there may be general consensus; but for many other diseases, there may be no agreement. This is especially true in disease areas where the state of knowledge is still evolving. For such diseases, the clinical trial sponsor and the regulatory agency to which the submission will be submitted should reach agreement on the decision set and the clinical decision rule that are acceptable to both parties prior to commencement of the trial. It is the responsibility of the statistician to provide proper statistical support structure for the clinical decision rule.

In a clinical trial with only a single primary endpoint, T_1, the decision structure is simple. Its statistical support structure includes the simple null hypothesis, its alternative, and an appropriate test statistic all defined in terms of this primary endpoint. The sample size and power are calculated accordingly. It can be expressed by the symbolic notation ($T_1 > 0$) which is to be interpreted as follows. The new treatment will be declared as superior to the control, if the test statistic rejects the null hypothesis of no difference between the new treatment and control at the significance level of α, and if the difference is favoring the new treatment.

Many trials lack a well-defined clinical decision rule. Among those that do, many do not have proper statistical supports struture as illustrated by the following two examples.

Example 3. In a recent study of the drug carvedilol for the treatment of congestive heart failure, there was an "unexpected" mortality finding (observed p-value < 0.001).[16] However, in this study, mortality was not even stated in the protocol as an endpoint. There was a lengthy debate as to whether carvedilol has demonstrated a mortality benefit in this trial. If it were accepted as a positive demonstration of a mortality benefit, then this would imply that the final decision rule has been altered from the one originally proposed in the protocol. This would certainly lead to an increase in the probability of the overall type I error. This example illustrates a trial with a decision rule that is not exhaustive, in the sense that the decision set is not complete, since mortality is not in the decision set. In addition, the post-hoc attempt to alter the clinical decision rule is without proper statistical support.

Example 4. According to its label, VASOTEC is indicated "for stable asymptomatic left ventricular dysfunction: it decreases the rate of development of overt heart failure and decreases the incidence of hospitalization for heart failure".[73] This claim is essentially based on the SOLVD Prevention Trial. The SOLVD Prevention Trial has all cause mortality as the only primary endpoint, and hospitalization for CHF, development of CHF, and incidence of MI as three, among several, secondary endpoints. The trial results show that mortality is not significant ($p = 0.60$), but hospitalization for CHF ($p < 0.002$) and development for CHF ($p < 0.002$), and incidence of MI ($p < 0.024$) show nominally significant p-values. So VASOTEC was approved for the above indication based on the "apparent" significance of the "secondary endpoints", hospitalization for CHF and development of CHF. This time, the decision rule used to assess the efficacy of the drug is

again altered from the one declared in the protocol. It leaves unanswered the potential inflation in the probability of the overall type I error.

These two examples describe clinical trials with multiple primary endpoints where the decision rules are not complete and lack proper statistical support structures. They led to post-hoc attempts to alter the decision rules by attaching "statistical significance" to findings in endpoints that are selected in a retrospective manner. A clinical decision rule should in principle and in practice exhaust all possible ways of assessing the efficacy of the drug.

The following example illustrates a clinical decision rule that is exhaustive.

Example 5. In a clinical trial, the decision set consists of three co-primary endpoints, $\{T_1, T_2, T3\}$ without a principal primary endpoint. The clinical decision rule is ($T_1 > 0$, or $T_2 > 0$, or $T_3 > 0$). That is, the new drug would have demonstrated efficacy if it can be shown to be superior to placebo in any one of the three co-primary endpoints and no inconsistent trend in the other co-primary endpoints. This clinical decision rule considers all three co-primary endpoints as of equal importance and to have equal likelihood of demonstrating drug efficacy, and hence the application of either a p-value based procedure or a closed testing procedure would be appropriate.

The following example illustrates a different situation where the primary endpoints have varying importance.

Example 6. In a clinical trial, the decision set consists of three primary endpoints, $\{T_1, T_2, T_3\}$, where T_1 is a principal primary endpoint, T_2 and T_3 are two co-primary endpoints. The clinical decision rule is defined as ($T_1 > 0$ at α_1 or $T_2 > 0$ at α_2 or $T_3 > 0$ at α_3), where α_1, α_2 and α_3 are chosen to preserve the probability of the overall type I error at α.

This clinical decision rule is different from the preceding example in that one specifies a different allocation of α among the three separate hypotheses. So a closed testing procedure or a p-value based procedure may not be directly applicable. The trial sponsor usually determines the α allocation. In addition, the drug efficacy is established if the principal primary endpoint T_1 can be shown to be positive regardless of the outcomes of the other two co-primary endpoints. For example, one may think of T_1 as mortality.

The following example shows that even if individual secondary endpoint may not support an efficacy claim, several secondary endpoints together may be sufficient to support an efficacy claim.

Example 7. In a study of Alzheimer patients, T_1 = Alzheimer's Disease Assessment Scale — Cognitive Subscale, and T_2 = Clinician's Interview Based Impression of Change, are two secondary endpoints. The clinical decision rule is ($T_1 > 0$ at $\alpha = 0.05$ and $T_2 > 0$ at $\alpha = 0.05$).

The decision to use a combination of two different types of outcome measurements to evaluate the efficacy of anti-dementia drugs was made with the support of a number of experts working in the field of dementia at FDA's Anti-dementia Assessment Symposium in 1989. This decision rule was recently reaffirmed by the expert panel of the Peripheral and Central Nervous System Drugs Advisory Committee in meetings on issues concerning mild cognitive impairment and cardiovascular dementia. The Alzheimer's Disease Assessment Scale — Cognitive Subscale, a performance based assessment instrument, ensures that the effect of the drug is on the "core" phenomena of dementia. The Clinician's Interview Based Impression of Change, a global assessment, ensures that the effect of the drug is clinically meaningful.

It has been argued that the clinical decision rule in the above example is too conservative because the probability of the overall type I error is less than 0.05. This argument is valid only if both T_1 and T_2 are primary endpoints. The reason why this should not be considered as conservative is because both T_1 and T_2 are secondary endpoints, and by our definition either one alone is not sufficient to support a claim. This is one important reason why a hierarchical definition for the various clinical endpoints is proposed. The next example illustrates the importance of hierarchy in clinical decision rule.

Example 8. In a clinical trial, the decision set $\{T_1, T_2\}$ consists of a principal primary endpoint T_1, and a co-primary endpoint, T_2. The clinical decision rule is ($T_1 > 0$ at α_1) or ($T_2 > 0$ at $\alpha_2 | T_1 \not< 0$ at α_1), where α_1 and α_2 are chosen to preserve the probability of the overall type I error at α.

This example illustrates the difference between a principal primary endpoint and a co-primary endpoint as follows. The new drug can claim efficacy if it is significantly better than placebo relative to the principal primary endpoint T_1 at a significance level of α_1. On the other hand, the new drug can also claim efficacy if it is superior to placebo relative to the co-primary endpoint T_2 at a significance level of α_2, provided it is not worse than placebo relative to the principal primary endpoint T_1 at the significance level of α_1. The condition ($T_1 \not< 0$ at α_1) may need further elaboration in each case. This same condition is implicit in the preceding examples. However, in current practice, this condition is usually not clearly mentioned.

The next example illustrates a clinical decision rule without proper statistical structures.

Example 9. In an epilepsy trial, the decision set consists of three clinical endpoints, $\{A, B, C\}$, where A is a primary endpoint and B and C are secondary endpoints. The clinical decision rule adopted is that either ($A > 0$ at α) or ($B > 0$ at α and $C > 0$ at α). Let Δ_A, Δ_B and Δ_C denote the parameters corresponding to the endpoints A, B and C respectively.

This decision rule reflects the following complex null and alternative hypotheses:

$$H_{oc} : \Delta_A = 0 \quad \text{and} \quad (\Delta_B = 0 \text{ or } \Delta_C = 0)\,,$$

$$H_{ac} : \Delta_A \neq 0 \quad \text{or} \quad (\Delta_B \neq 0 \text{ or } \Delta_C \neq 0)\,.$$

The sponsor proposes to test instead the restricted null hypothesis,

$$H_{or} : \Delta_A = 0 \quad \text{and} \quad \Delta_B = 0 \quad \text{and} \quad \Delta_C = 0\,,$$

for no treatment effect. Note that H_{oc} consists of two axes, ($\Delta_A = 0$ and $\Delta_B = 0$) and ($\Delta_A = 0$ and $\Delta_C = 0$), while H_{or} consists of only the origin. It is shown by Jin and Chi[42] that the probability of the overall type I error for the proposed clinical decision rule under H_{oc} should be

$$\begin{aligned}\alpha_C &= \sup_{(\Delta_A,\Delta_B,\Delta_C)0H_{oc}} P_{(\Delta_A,\Delta_B,\Delta_C)}\{|Z_A| > c_A \text{ or } (|Z_B| > c_B \text{ and } |Z_C| > c_C)\} \\ &= \max\{P_{(\Delta_A=0,\Delta_B=0)}(|Z_A| > c_A \text{ or } |Z_B| > c_B)\,, \\ &\quad P_{(\Delta_A=0,\Delta_c=0)}(|Z_A| > c_A \text{ or } |Z_C| > c_C)\}\,,\end{aligned}$$

where Z_A, Z_B, and Z_C are the corresponding test statistics.

The probability of the overall type I error corresponding to the restricted null hypothesis is

$$\alpha_r = \sup_{(\Delta_A,\Delta_B,\Delta_C)0H_{or}} P_{(\Delta_A,\Delta_B,\Delta_C)}\{|Z_A| > c_A \text{ or } (|Z_B| > c_B \text{ and } |Z_C| > c_C)\}\,.$$

It is easy to see that

$$\alpha_C \geq \alpha_r\,,$$

with strict inequality holds true in general.

Therefore, if the rejection region, or critical values c_A, c_B and c_C, are defined to preserve α_r at the desired level, say, $\alpha = 0.05$, then the probability of the overall type I error, α_C will be inflated beyond $\alpha = 0.05$. Therefore, the p-value calculated with the restricted null hypothesis is smaller than that calculated with the complete null hypothesis.

This example illustrates that it is necessary to clearly specify the statistical support structure for a complex clinical decision rule. Otherwise, it can easily weaken the strength of evidence or compromise the validity of the statistical inference, for example, through an improper calculation of the p-value.

As noted earlier, clinical decision rule is defined relative to multiple endpoints. However, one can easily extend the concept to include other efficacy assessment schemes involving multiple comparisons, repeated testing, multiple indication etc. Furthermore, it is critical that a clinical decision rule should have a proper statistical support structure. A confirmatory trial should have a well-defined decision structure.

The p-value based procedures and the closed testing procedures previously discussed are formal statistical procedures developed for testing hypotheses in a multiple testing situation. As pointed out earlier, these procedures do not make reference to any clinical or regulatory considerations. For example, no distinction is made with respect to the nature, meaning and relative importance of the clinical endpoints, and no attempt is made to take into consideration any regulatory perspective on how these endpoints should be used in assessing efficacy. They essentially assume that all multiple endpoints or multiple comparisons are of equal importance for purpose of efficacy assessment. Therefore, in any given clinical situation, a direct application of these procedures to assess the efficacy of a new treatment without proper clinical and regulatory considerations may be quite inappropriate. The following two examples illustrate this problem.

Example 10. In a trial seeking an indication for treatment of transient insomnia, data on latency to persistent sleep (LPS) are collected at night at baseline and on four post-baseline nights. The trial sponsor proposes that the drug would have demonstrated efficacy if it can show superiority to placebo at any one of the four nights relative to LPS. Since there are multiple testing, the sponsor proposes to use a p-valued based procedure. From a purely statistical perspective, this proposal is acceptable. But this kind of decision rule lacks clinical perspective and is not appropriate for this indication. The reason is that according to this decision rule, the drug could be approved for the treatment of transient insomnia if the trial shows that the drug is superior to placebo at any one of the four nights. Thus, a possible winning scenario is for the drug to show superiority on any one of the nights except the first night. Unfortunately, this kind of outcome can not support the desired indication, because it is expected that a drug

for treating transient insomnia should show effect on the first night. The treatment effects on the subsequent nights are used mainly for assessing its tolerance profile, i.e. whether the drug effect diminishes over time. Clinically, showing an effect on the first night is a necessary condition for approval. Therefore, the application of any p-valued based procedure or closed testing procedure would not be appropriate. One possible approach is to apply a conditional testing procedure: starting with the first night, test LPS with a significance level of α, and continue testing for the subsequent night as long as the current night shows a treatment effect significant at the same α level.

In developing statistical methodology for handling multiple endpoints in clinical trials, one should also be mindful of the interpretability of the results. Clinical interpretation of the results of statistical analysis of multiple endpoints often presents formidable challenge to both statisticians and clinicians. An interesting way of reducing the problem with multiple endpoints is to define in some manner a composite endpoint, an index or a global statistic. The following examples show different ways composite endpoint can arise and the problem of interpretation that may accompany such composite endpoints.

The handling of mortality data in some central nervous system drug trials provides a good illustration.

Example 11. In trials treating acute stroke or ALS, some patients in either drug or placebo group die during the trials. The primary endpoints in these trials are often some neurological scores, such as NIH Stroke Scale, Modified Rankin Scale, Barthel Index, Glasgow Outcome Scale in stroke trials, or Appel score and vital capacity measurement in ALS trials. All these scores, except for the Modified Rankin Scale, do not include death as a part of the measurement. To analyze the complex data consisting of neurological scores and death, a composite endpoint is sometimes proposed as a way to compare treatment effect. To illustrate a difficulty in interpretation of the results of such composite endpoint, consider the following proposal. The proposed method will rank the deaths from the earliest to the last, then rank the patients who are alive according to their neurological scores with the worst ranked next to the rank of the last death. The new ranking scores will then be used in an ordinary ANOVA analysis. The results from such an analysis will be difficult to implement in a regulatory setting, particularly the meaning of the assigned rank scores for death and neurological functioning measures will be hard to interpret. For example, the ranks for the

last death and the worst neurological score only differ by one. These two very different clinical outcomes are essentially treated as the same outcome in the analysis. The comment is not meant to criticize the statistical merits of the composite endpoint analysis, but merely to point out a problem in interpreting the results based on this kind of composite endpoint. Any other global test statistics will likely face similar difficulty since it is attempting to evaluate drug effect by using a single measure to summarize two very different outcomes.

Example 12. The objective of this stroke trial is to demonstrate whether a 24-hour continuous infusion of a new drug combined with t-PA when t-PA treatment is initiated within three hours of stroke onset offers superior outcome at day 90 than t-PA treatment alone. The primary endpoints consist of the NIH Stroke Scale, the Barthel Index, Modified Rankin Scale and the Glasgow Outcome Scale. The proposed analysis is a modification of the Wald-like global test statistic used in Tilley *et al.*[93] If the global test statistic demonstrates a "significant" difference, then the individual component scores will be examined. The problem with this global test statistic is that it can show a "significant" difference even when the individual component scores may show inconsistent results. For example, one possible scenario could be that the Barthel Index shows "significance", but all the other three scores show a negative trend. Now, there is no adjustment made for multiple testing after the global test shows "significance". A closed testing procedure should be applied in order to maintain the probability of the overall type I error. This same issue existed in Tilley *et al.*[93] However, the t-PA trial was fortunate to have a highly effective treatment so that all individual component scores showed significant and consistent findings so that the application of a closed testing procedure would produce similar conclusion. But in the current proposed trial, such consistent and significant outcomes in the individual component scores may be unlikely, in view of all the recent failed stroke trials.

In clinical drug trials, resolutions of subtle issues like the ones discussed above require a close interaction between the clinicians and the statisticians. Any attempt to develop a statistical methodology without careful consideration of the interpretive issue is hazardous.

From all of the preceding examples, it becomes very clear that one should try to avoid the following types of problems. The first type of problem is that of asserting drug efficacy claim based on clinical decision rule that is not pre-specified, but defined retrospectively. This practice

inflates the probability of the overall type I error rate. The second type of problem refers to clinical decision rules that do not provide the kind of evidence appropriate for the indication sought and the strength of evidence needed. The third type of problem is a clinical decision rule that lacks proper statistical support structure. The fourth type of problem is the use of a statistical procedure inappropriate for a given multiple testing situation. Finally, one should avoid defining composite endpoints or global test that would render the result of the analysis difficult to interpret or clinically unacceptable.

The following general procedure for defining a decision structure should be prospectively described in the study protocol for each trial, especially a confirmatory trial.

For the disease under study, define clearly the indication desired. Understand clearly the efficacy criteria needed for the new treatment to satisfy in order to get this indication. Based on the criteria, identify a decision set of clinically relevant endpoints. These endpoints may be called principal primary, co-primary and secondary as defined earlier. These definitions include a hierarchical order based on their clinical importance, objectivity, etc. and are defined from both clinical and regulatory perspectives. From this decision set, one should define the clinical decision rule for assessing the efficacy of the drug. The clinical decision rule is a decision tree consisting of most if not all decision branches or paths that can lead to a drug's efficacy claim. Each decision branch or path is defined in terms of one or more endpoints from the decision set. Some decision branches or paths may be sequential or conditional, thus forming sub-branches or sub-paths. At each decision point, decision will be made based on the outcome of the endpoints used in that particular decision path or branch. Then, one should provide the necessary and appropriate statistical support structure for this clinical decision rule. In defining a decision structure, one should also consider the level and strength of the evidence required of the study.

The statistical support structure should include:

- The appropriate hypotheses to be tested; the hypotheses include the proper null and alternative hypotheses reflecting the clinical decision rule. Unlike a simple null hypothesis that usually consists of a single parameter zero, the null hypothesis for a complex clinical decision rule is a region in a multidimensional parameter space. Failure to clarify this null hypothesis region could jeopardize the final statistical analysis, and weaken the strength of evidence or even invalidate the trial results.

- The appropriate test statistic to be used at each decision branch or path, and the associated statistical analysis plan.
- The desired allocation of α across the decision branches or paths so that the probability of the overall type I error is preserved at the desired α level.
- Sufficient sample size and an optimal overall power relative to the entire clinical decision rule.

So far, the discussion of clinical decision rule is focused on a given time point in the trial, usually at the end of the trial. In the next section, we shall discuss clinical decision rule over some indices typically time or information fraction.

4. Interim Analysis and Sequential Clinical Decision Rule

In the 1970s, clinical trial investigators began to question on ethical grounds whether a trial, with mortality or serious irreversible morbidity as the primary outcomes of interest, should be continued to its intended end when interim analysis based on accumulating data shows that the treatment is effective. In those early days, interim analyses were routinely done by the investigators without any regard to issues of multiple testing and its consequent inflation in the probability of the overall type I error. So in this kind of trials, the decision rule, which is usually defined in terms of one primary endpoint such as mortality, is repeatedly tested at various times in the course of the trial. The problem at that time is that there was no appropriate statistical support structure for the decision process. The desire for stopping the trial early for efficacy has been the driving force behind the development of group sequential procedures in the subsequent decades.

A *group sequential procedure* is a statistical procedure that provides for a series of test statistics based on the accumulating data. The simple null hypothesis is tested by these statistics. An early termination rule is implemented through a stopping boundary defined by the critical values or in terms of an increasing sequence of nominal significance levels for each test. This boundary is defined so that the probability of the overall type I error will be maintained at a desired α-level for a two-sided test.

Group sequential procedures that allow for interim analyses and early termination had been proposed by many authors.[23,68,74,75,94,108] These group sequential procedures are used for a type of analysis customarily referred to as *formal interim analysis*, and they are usually implemented for mortality and irreversible morbidity trials.

A group sequential procedure with formal interim analyses is an example of a *sequential clinical decision rule*.

Definition 8. A sequential clinical decision rule is a clinical decision rule, repeatedly applied over some indices typically time or information fraction.

A **sequential decision structure** is a sequential clinical decision rule with a proper statistical support structure.

Note: alternatively, one can define each decision (path) that can result in an efficacy assessment as a clinical decision rule, and then consider a family of such clinical decision rules along with its statistical support structure.

The group sequential procedures referred to above that allow for interim analyses and early termination are simple sequential decision structures. They are simple because the clinical decision rule at each time point involves one and the same primary endpoint such as mortality or a composite endpoint consisting of several types of event endpoints.

In a trial with a sequential clinical decision rule, the decision rule itself can be more complex when multiple endpoints and/or multiple doses are involved. Proper statistical support structure is needed for these more complex sequential clinical decision rules.

4.1. *Interim analysis and design modifications*

The principal components of a confirmatory trial include the target patient population, the study design, the decision structure which includes a decision set of key clinical endpoints, a clinical decision rule with its statistical support structure for assessing efficacy, and possibly an interim analysis plan, etc. There is always an interest on the part of the trial sponsor to make changes to the trial based on interim treatment comparative analysis of accumulating data prior to the intended end of the trial. These changes may involve changes in the patient population, in the decision structure including the clinical decision rule, the test statistics, the interim analysis plan, and the sample size. For instance, after an interim analysis, it may be of interest to drop one or more treatment arms, to change or drop one or more endpoints, or in a group sequential trial, to change the interim analysis schedule, to change the stopping boundary, the α-spending function, or to increase the sample size. Since such proposed changes are based on the interim treatment comparative analysis, they may introduce serious bias into the study and inflate the probability of the overall type I error.

Cui *et al.*[15] showed that in a group sequential trial, sample size re-estimation based on observed interim treatment difference can inflate the probability of the overall type I error, essentially because the sample size estimate is interim outcome dependent. A methodology was developed based on weighted test statistics that would permit sample size re-estimation at any one of the pre-scheduled interim analysis times without inflating the probability of the overall type I error. The methodology is quite flexible and allows one to retain the original stopping boundary. Other design modification schemes with or without sample size re-estimation are currently being investigated.

Except for changes in the characteristics of the patients yet to be enrolled, one may view most of these design modification strategies as modifications of certain aspects of the decision structure. For example, these modifications may involve deleting one or more decision branches or paths, changing the test statistic at a decision point, re-allocating the α's and re-estimating the sample size. How such modifications impact on the validity of the statistical inference needs to be investigated. For instance, in a group sequential trial with an interim analysis plan, if such modification affects the underlying Brownian motion process, then one needs to be able to develop valid statistical test procedure in the absence of, say, the property of independent increments. Generally, one should pre-specify the desired modifications, and describe how decisions to modify or not to modify the decision structure are made conditional on the interim outcome data. Finally, one should ascertain that the modified decision structure can still permit valid and meaningful interpretation of the study results.

4.2. *Adaptive two-stage designs*

Design modification of a clinical trial can also be prospectively built into a two- or multi-stage randomized trial. In a two-stage design, modifications may be considered at the end of the first stage. Bauer and Köhne[3] considered the problem of sample size re-estimation at the end of the first stage. The two-stage combination test is based on Fisher's product test and the assumption that the samples from the two stages are independent. Proschan and Hunsberger[77] generalized the method of Bauer and Köhne through the concept of conditional error function. Liu and Chi[56] further considered allowance for stopping at the end of the first stage for futility, and the problem of defining a unique overall p-value by proposing the use of a class of generalized conditional error functions. All

three papers assumed independence of the samples from the two stages.[5,105] A general theory of adaptive two- or multi-stage design that would permit modifications other than sample size re-estimation, and for dependent samples have recently been considered.[4,5,57] Again as in the case of group sequential trials, it would be of interest to develop a general theory of adaptive two- or multi-stage design that would permit modification of the decision structure for the general case with dependent samples. Such theory should be developed within the framework of pre-specification of the desired modifications and how the decision to modify or not to modify the decision structure is based on the interim outcome data. In addition, the distributions of the adapted statistics should be derived, and the overall probability of type I error should be maintained. Furthermore, the uniqueness of the overall p-value should be demonstrated to confirm the validity of the statistical inference. Such theory would be extremely important as it forms the basis for various practical applications in clinical drug trials as discussed.[49] For example, the adaptive two-stage design would be a natural framework for combining a traditional Phase 2 study with a Phase 3 study, or for accelerated approval of potentially life-saving drugs in diseases that do not have available treatment.[4]

4.3. *Interim analysis and data safety monitoring committee*

Most mortality or serious irreversible morbidity trials have formal planned interim analysis and stopping rules. There is also an independent Data Safety and Monitoring Committee (DMC). The primary responsibility of a DMC is to recommend to the trial sponsor early termination of the trial for either efficacy or safety reason. The requirement of being independent is to maintain the integrity of the trial. It also frequently happens that the DMC also makes recommendation for changes to the design and conduct of the trial. This kind of recommendations often raises concern. The following are a few important points to consider for a DMC. For general guidance, one may refer to the FDA's pending *Guidance on the Establishment and Operation of Clinical Trial Data Monitoring Committees*[101] *and Data Monitoring Committees in Clinical Trials.*[112]

- The trial protocol should have the prior approval of the DMC. The DMC cannot arbitrarily modify the protocol design. Even though DMC may be independent, its independence refers to the lack of direct interest in the outcome of the trial. It does not imply that the DMC can recommend

design modifications arbitrarily. Any design modification recommended by the DMC also has the potential for inflation in the probability of the overall type I error. This is because the proposed modification by the DMC is likely to be based on the interim treatment comparative data. The guideline should clearly define the span of authority of the DMC.

- The DMC should have standard operating procedures governing its conduct. It has obligation and responsibility to maintain confidentiality of the interim outcome of the trial. It should not communicate the interim results to anyone outside the DMC, unless for purpose of early termination according to the pre-specified termination rule or for safety reason.
- There should be clear standard operating procedures governing the communication between the trial sponsor and the DMC. There should be clear guidelines regarding the documentation of minutes of meetings, decisions reached, and written communications.

The ability to make design modifications is obviously very attractive. However, one should be mindful of the real potential for bias to be introduced as a result of access to the interim unblinded treatment comparative data. As previously noted, if the characteristics of the patients enrolled subsequent to an interim analysis have changed, then this may seriously bias the outcome, unless the drug label can accurately describe the patient population for whom the drug may be prescribed. It is desirable to have an independent third party that is responsible for conducting interim analysis either in a group sequential trial setting or a two- or multi-stage design trial. There should be clear and sound guidelines and standard operating procedures governing the role, responsibility and conduct of the independent third party.

In general, any desired design modification or change should be prespecified at the design stage. Furthermore, one should describe how decisions on modifications are made conditional on the interim outcome data, the distribution of the adapted test statistic, and the overall p-value for assessing the significance of the trial finding. The proposal should also include methods for addressing any adverse impacts such modification or change may introduce, such as, bias, changes in patient population, inflation in the probability of the overall type I error, and proof of the validity of the statistical inference based on the adapted test statistic. Generally, the clinical decision rule should be as complete as possible. The design modification, other than sample size adjustment, should be restricted to deletion

of decision branches or paths. The addition of new decision branches or paths should be discouraged.

4.4. *Two-stage design in a randomized trial and accelerated approval*

Accelerated approval of potentially life saving drugs in diseases that do not have available treatment may be done through an accepted surrogate (or surrogates) whose validity has been established. Validity of the surrogate means it is likely to predict clinical benefit of primary interest. The idea of accelerated approval is to first conditionally approve the use of the drug based on a positive outcome on the surrogate endpoint, and then subsequently confirm its effectiveness through the clinical endpoints of primary interests.

FDA's *Oncology Initiatives*[96] recognize that the predictive value of partial responses may still be a matter of discussion for all types of cancer. But for refractory malignant disease or for diseases that have no adequate alternative, clear evidence of anti-tumor activity is a reasonable basis for approving the drug. In these cases, studies confirming a clinical benefit may appropriately be completed after the conditional approval.

So in essence, the *Oncology Initiative* has gone a step further in permitting a surrogate to be used in certain situations even though the surrogate has not been fully validated. One may refer to the example of gemtuzumab ozogamicin in relapsed acute myeloid leukemia as discussed in Bross *et al.*[9] It is for this reason that a two-stage design may be well suited for accelerated approval. The drug may be tested at the end of the first stage for conditional approval based on the surrogate, and then confirmed at the end of the second stage by the primary clinical endpoints.

A two-stage design in a randomized trial offers the following advantages. First, it places the evaluation of the surrogate and the clinical outcomes of primary interest in the same trial, and hence would be able to provide some checks on the validity of the surrogate in relation to the clinical benefit of interest.

Secondly, a more appropriate way for conditional approval is to first identify a decision set of clinical endpoints of primary interest that will be used at the end of the second stage to evaluate the efficacy of the drug. Appropriate allocation of α should be considered for multiple endpoints, including the surrogate as well as interim analysis at the end of the first stage, in order to maintain the probability of the overall type I error at

α. At the end of the first stage, the efficacy will be evaluated relative to the surrogate endpoint as well as all the clinical endpoints in the decision set. If it fails to demonstrate efficacy relative to the surrogate and all the clinical endpoints, then the trial stops and the drug fails to gain accelerated approval. If any of the clinical endpoints in the decision set demonstrates efficacy, then the trial can be terminated at the end of the first stage. Otherwise, if the test demonstrates efficacy relative to the surrogate only, then one proceeds to calculate the conditional power of showing a positive outcome at the end of the second stage given the current interim outcomes for each of the clinical endpoint in the decision set. If the conditional powers show that the likelihood of such an outcome is very low for all the clinical endpoints, and a sample size increase would be unacceptably large, then the accelerated approval probably should be withheld. Thus, accelerated approval should be granted if at least one such conditional power is sufficiently high, or if sample size can be increased so that at least one clinical endpoint will have sufficient conditional power to suggest a positive outcome at the end of the second stage. If the trial continues to the second stage and the final results show that there is no clinical benefit, then the drug may need to be withdrawn from the market.

5. Active Control Trials

The recent fifth revision of the World Medical Association Helsinki Declaration (2000) has generated a great deal of discussion[19,92,102] and a renewed interest in active control trials. Historically, for trials with mortality or serious morbidity outcome, delaying or withholding available treatments would increase the mortality or irreversible morbidity outcome, the use of a placebo is considered unethical. Thus, active control comparative trials have been proposed.[24,25] An active control superiority trial allows for a direct comparison of a new study treatment against a standard therapy or standard of care. It establishes the efficacy of a new study treatment by demonstrating that the new treatment is superior to the active control.[49,50,90] For example, in oncology for cancers that have standard therapies, a new study treatment must be compared to and show superiority to a standard therapy in two randomized controlled clinical trials. Unless the new treatment represents a new advance in the treatment of the disease, it would generally be more difficult to show that the new treatment is better than an effective active control. Thus, it is not surprising to find that when a new treatment fails to show superiority to the active control, the sponsor or investigator

would attempt to assert that the new treatment is either equivalent to, or no worse than the active control. But it is well known that failing to reject the null hypothesis of equality between the new treatment and the active control does not imply that they are equivalent.[25,91] As White[107] aptly puts it, "the lack of evidence of a difference can not be interpreted as evidence of a lack of difference".

It has been suggested by various authors[7,39] that if the intention is to show that the new intervention is equivalent or non-inferior to the active control, then one needs to define the null hypothesis and the corresponding alternative hypothesis appropriately. More specifically, the alternative hypothesis should reflect the hypothesis of interest, namely that of equivalence, or non-inferiority. To do that, these authors suggested that a certain equivalence or non-inferiority margin should be pre-specified in the hypothesis. For example, in bioequivalence studies, an equivalence margin is generally set at $\pm 20\%$ of the control effect. In many active control non-inferiority trials, an arbitrary fixed threshold is pre-specified — for example, a threshold of 1.25 for the hazard ratio of the study treatment relative to the active-control. If the upper limit of the 95% CI for the study treatment effect relative to the active control lies beneath this threshold, then non-inferiority is inferred.

The major concern with using an arbitrary fixed threshold is that it is unrelated to the active control effect defined as the difference between the control response and the non-existing placebo response. When the active control effect is relatively small, this may lead to a loss of all of the active control effect plus more (or a loss of too great a percent of the active control effect). In other words, if the demonstration of non-inferiority is based on an arbitrary fixed margin, the new treatment may be approved even though it may be less efficacious than a placebo.

Having recognized this problem, it has been proposed that the margin should be set at half the lower limit of the 95% confidence interval for the estimate of the active control effect to ensure that the new drug is better than placebo. A similar approach was used by the FDA Center for Biologics Evaluation and Research in a thrombolytic trial where instead of 95%, they used 90% confidence interval. The null hypothesis of sufficient inferiority will be rejected if the upper limit of the 95% confidence interval for the hazard ratio of the new treatment relative to the active control is below this cutoff. This method has also been called the "two-95% confidence intervals approach". While this fixed cutoff is not arbitrary and is linked to an estimate of the control effect, it has been criticized as being too

conservative because it compares two "statistically worst" cases. Success is when the new treatment demonstrates a retention of at least 50% of the statistically worst active control effect.

Various authors[30,31,36,44,79,86] have proposed not to directly pre-specify a fixed margin, but rather define simply the percent of the control effect one wishes to retain. Non-inferiority is then demonstrated by the new treatment, if it can be shown that the new treatment retains at least the desired percent of the active control effect. The active control effect may be estimated from non-concurrent placebo or standard control studies using mixed effects model.

5.1. *Retention of certain percent of active control effect — hypothesis for a non-inferiority trial*

Let T, C and P denote the treatment, the control and the placebo respectively. The hypothesis for a non-inferiority trial can be formulated in the following two different ways.

1) If one specifies an arbitrary fixed margin, say $\delta = 20\%$, then the hypothesis can be formulated as follows:

$$H_o : T - C \leq -\delta \quad \text{vs.} \quad H_a : T - C > -\delta\,.$$

A fixed margin may be appropriate if one believes the effect can only be attributed to the treatment. For example, in cancer trials, tumor shrinkage may be attributed only to the treatment. Otherwise, the use of an arbitrary fixed margin is highly questionable since the active control effect is not accounted for in the margin. Particularly, if an active control effect size is relatively small, using an arbitrary fixed margin may lead to the approval of a study treatment that is actually inferior to the placebo.

2) If the objective of a non-inferiority trial is to demonstrate that the new treatment retains a certain percent of the active control effect, then the null and alternative hypotheses can be written as follow:

$$H_o : (T - P_1)/(C_1 - P_1) \leq \pi \quad \text{vs.} \quad H_a : (T - P_1)/C_1 - P_1) > \pi \tag{4}$$

or

$$\begin{aligned} &H_o : ((T - C_1) - (C_1 - P_1))/(C_1 - P_1) \leq \pi - 1 \quad \text{vs.} \\ &H_a : ((T - C_1) - (C_1 - P_1))/(C_1 - P_1) > \pi - 1 \end{aligned} \tag{5}$$

where π is the percent retention desired, C_1 represents the current active control, P_1 the current placebo if placebo were to be present, and $(C_1 - P_1)$ is the current active control effect assumes to be positive.

The hypotheses in (5) can be written as

$$\begin{aligned} H_o &: (T - C_1) \leq -(1-\pi)(C_1 - P_1) \quad \text{vs.} \\ H_a &: (T - C_1) > -(1-\pi)(C_1 - P_1)\,. \end{aligned} \tag{6}$$

Now since in the current trial, there is no placebo, the active control effect on the right side of the hypotheses cannot be estimated from current active control trial. The assessment of the assumption of $C_1 - P_1 > 0$ is based on non-concurrent, randomized, placebo controlled trials. If there are adequate and well-controlled non-concurrent placebo or standard control studies that can provide consistent estimates of the active control effect and if the current active control study is as similar as possible to these non-concurrent studies in terms of design, patient populations and therapeutic settings, then perhaps the assumption that the current control effect $(C_1 - P_1)$ is a certain fraction, θ, of the historical active control effect, $(C_0 - P_0)$, may be considered reasonable. That is, $(C_1 - P_1) = \theta(C_0 - P_0)$, with $0 \leq \theta \leq 1$. Under this assumption, (6) can be written as,

$$\begin{aligned} H_o &: (T - C_1) \leq -(1-\pi)\theta(C_0 - P_0) \quad \text{vs.} \\ H_a &: (T - C_1) > -(1-\pi)\theta(C_0 - P_0)\,. \end{aligned} \tag{7}$$

The right side in the hypotheses can be viewed as the margin. For example, by setting $\pi = 1/2$, then the margin is defined as a 50% preservation of the active control effect, $\theta(C_0 - P_0)$. By setting $\pi = 0$, the non-inferiority trial becomes a superiority over "placebo" trial, where "placebo" is represented by $C_1 - \theta(C_0 - P_0)$. By setting $\pi = 1$, the non-inferiority trial becomes an active control superiority trial.

Test statistic defined in terms of the estimate of $(T - C_1)$ from the current active control trial and the estimate of $(C_0 - P_0)$ from non-concurrent placebo or standard control trials can be used to test the null hypothesis in (7). Holmgren (1999) derived a test statistic involving relative risks. Rothmann *et al.*[79] derived a test statistic involving hazard ratios from mortality trials, and derived some important properties of the corresponding test statistic. In each case, the hypotheses in (4)–(7) need to be rewritten to reflect the corresponding efficacy measure. Rothmann *et al.*[79] showed that the two-95% confidence intervals approach is conservative in the sense that it controls the probability of the type I error associated with hypotheses (7) at the 0.003 level when using survival endpoint. It was further shown that if one controls the probability of the type I error for the one-sided test at the 0.025 level, then a unique γ% confidence interval for the active control estimate can be derived. The use of a test statistic for a retention of certain

percent of the active control effect is conditionally equivalent to an asymmetric two-confidence interval approach. The asymmetric two-confidence interval approach is analogous to the two-95% confidence intervals approach, except by replacing the 95% confidence interval for the active control effect estimate by the shorter γ% confidence interval. It follows from this that the use of point estimate for the active control effect always inflates the probability of the type I error. The shorter confidence interval would lead to a higher threshold or margin. A detailed discussion of the general methodology and issues related to the design, analysis and interpretation of such active control non-inferiority trials is given in Rothmann *et al.*[79]

Implicit in the formulation of the hypotheses in (4), and its modification in (7) are the following fundamental assumptions:

First, there are adequate non-concurrent placebo or standard control trials that consistently demonstrated the active control effect over various patient populations studied. Such consistency provided the necessary confidence regarding the existence of the active control effect in the current patient population and clinical trial setting. The false positive rate of the assessment of the control effect should be taken into the consideration of the current non-inferiority trial. In many cases, active control non-inferiority trial would probably not be possible to do.

Second, it is assumed that the active control effect exists in the current active control trial. Even though there are non-concurrent placebo or standard control trials on the basis of which the active control effect is estimated, it does not follow that in the current active control trial, the active control effect necessarily exists. This is because many causes may lead to this. For example, the patient samples in the non-concurrent trials and the patient sample in the current trial may not be representative of the same patient population. Take an extreme case. Suppose that the active control works mainly in female patients. So suppose in the non-concurrent trials, the patient samples are mostly females, while the patient sample in the current trial is mainly males. Thus, in the patient sample in the current trial, the active control will not show much of an effect.

Third, it is assumed that the active control effect in the current active control trial is a certain fraction, θ, of the active control effect determined in the non-concurrent placebo or standard control trials. This assumption requires information that may or may not be available that would permit its determination. Rothmann *et al.*[79] show that the method is applicable if one knows what this fraction, θ, is. In fact, because the patient samples in the non-concurrent studies and the patient sample in the current study may not be representative of the same patient population, one generally

cannot assume that the active control effect is constant between the non-concurrent studies and the current study. Hence, some kind of adjustment factor θ may be needed. In other circumstances, such adjustment may also be necessary. For example, if the standard of care has improved over time, then the active control effect will be somewhat reduced. Similarly, in anti-infective area, if the bacteria have become resistant to the current antibiotics, then the active control effect will be reduced. Allowing the use of new and effective concomitant medications also reduces the active control effect. A more subtle situation is when the non-concurrent placebo or standard control studies were stopped early based on the interim analyses results. It is known that estimates of the active control effects may be biased upward based on interim results.[109,110] Proper adjustment is needed for such active control effect estimates.[55]

It should be noted that the hypothesis in (4), which is appropriate only if the assumption $C_1 - P_1 > 0$ holds, reflects the following clinical philosophy. If a new treatment can offer some additional clinical benefits, e.g. a better toxicity profile or ease of administration, then this new treatment may still be beneficial even when some efficacy is lost as compared to the active control. From this perspective, then one objective of an active control non-inferiority trial is to rule out all differences of "clinical importance" between the new treatment and the active control. This would permit one to conclude that the new treatment is effective even though one has not established that it is non-inferior to the active control. In such trials, the term "active control non-inferiority trial" is a misnomer because the trial objective is not to show non-inferiority of the new treatment to the active control, but that the new treatment is simply effective. To show the new treatment is non-inferior or equivalent to the active control, one will need to specify a much more stringent criteria than that in (4), that is one should demand a retention much greater than 50% of the active control effect, perhaps at least 85% based on preliminary research results.

For active control non-inferiority trial, the problem regarding bias towards no difference is a crucial issue because the null hypothesis would be easier to be rejected when there is bias towards no difference. This kind of bias can be introduced without having to unblind the treatment codes. When the primary efficacy endpoint is mortality or irreversible morbidity event, then perhaps such bias would be of less concern.

Example 13. In clinical trials involving colorectal cancer, it is considered unethical to use placebo. In a recent study of xeloda for the treatment of colorectal cancer, the objective is to demonstrate that the new treatment

is effective in an active control trial. Since xeloda has better side effects profile and ease of administration. The clinicians felt that a retention by xeloda of at least 50% of the active control effect represents an acceptable level of efficacy in view of its better side effects profile and ease of administration. There are a number of non-concurrent standard control studies involving the active control. These studies demonstrated fairly consistent active control effect. Based on the data from these non-concurrent studies, estimates for the log hazard ratio of the active control effect and its standard error are 0.234 and 0.075 respectively. This is equivalent to a hazard ratio estimate of 1.264 with a 95% confidence interval, (1.091, 1.464). For a 50% retention in the active control effect, the largest estimate of the active control effect that is allowed with type I error controlled at 0.025 is 1.228 which corresponds to the lower limit of a 30% confidence interval. Thus, 50% retention of this active control effect produces a cutoff of 1.114. The clinicians believe that the active control effect should not have diminished over time. Thus, in this analysis, $\theta = 1$. The current active control trial produces a hazard ratio estimate of 0.92 with a 97.5% confidence interval upper limit of 1.09 that is below the cutoff of 1.114 (For comparison, the original protocol proposed to consider a fixed cutoff of 1.20). Thus, this trial demonstrates that it rules out a loss of more than 50% of the active control effect. In fact, the new treatment is likely to retain at least 61% of the active control effect. The conclusion that one may draw from this trial is that the new treatment is ***effective***. It retains at least 61% of the active control effect. However, one cannot claim that this new treatment is non-inferior to the active control. For such a claim, one would require that the new treatment should retain at least a certain percent of the active control effect that is much greater than 50% specified here. But for such a trial, the sample size required to maintain reasonable power would become prohibitively large.

This example illustrates that clinical decision rule is sometimes fairly subtle. It is not simply a matter of defining a set of primary and secondary endpoints, desired comparisons among several treatment arms, and an allocation of α. It involves a deeper understanding of what kind of evidence will be presented when the trial is finished, whether the evidence is appropriate for the indication sought, and whether the evidence would be sufficient for approval.

In summary, if for ethical reason, one has to use an active control, then an active control superiority trial can always be done. The problem is that

it may not be easy to demonstrate that the new treatment is superior to the active control, unless the active control happens to be ineffective in the current study. But in the latter case, if one were to claim that the new drug is superior to the active control, then it would be misleading. Thus, even in an active control superiority trial, perhaps one can only conclude that the new treatment is effective, unless one is certain that the active control is effective in the current trial.

On the other hand, due to all the critical assumptions needed in doing an active control non-inferiority trial, most of the times, such trials may not be possible and are not recommended because these assumptions are not verifiable. One of the basic concerns is that the active control may not work in the current patient population or trial setting. When this is true, then the assumption that the active control is effective would inflate the type I error rate. Another concern is that one may end up demonstrating a drug that is actually inferior to a placebo through such a trial. Therefore, such trials may be contemplated if there are other studies or information that can help to alleviate these concerns which can not be verified within the active control trial itself. Even then, the issue regarding bias towards no difference should be properly addressed.

In the actual design of an active control non-inferiority trial, the active control effect, the proportion of control effect to be preserved, the control of the probability of the type I error should be properly determined in light of the objective. One should also pay attention to issues such as multiplicity testing, interim analysis and design modification. These issues may take on a different complexity due to the special nature of an active control non-inferiority trial. Proper standard operating procedures should be designed to minimize the introduction of bias towards no difference.

References

1. Armitage, P. and Berry, G. (1994). *Statistical Methods in Medical Research.*
2. Bauer, P. (1991). Multiple testings in clinical trials. *Statist. Med.* **10**: 871–890.
3. Bauer, P. and Köhne, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**: 1029–1041.
4. Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statist. Med.* **18**: 1833–1848.
5. Bauer, P., Brannath, W. and Posch, M. (2001). Flexible two-stage designs: An overview. *Meth. Infor. Med.* **40**: 117–121.

6. Begg, C. B. (2000). COMMENTARY: Ruminations on the intent-to-treat. *Controlled Clin. Trials* **21**: 241–243.
7. Blackwelder, W. C. (1982). Proving the null hypothesis in clinical trials. *Controlled Clin. Trials* **3**: 345–353.
8. Boyd, E. J. S., Penston, J. G., Johnston, D. A. and Wormsley, K. G. (1988). Does maintenance therapy keep duodenal ulcer healed? *Lancet*, 1324–1327.
9. Bross, P. F., Beitz, J., Chen, G., Chen, X. H., Duffy, E., Kieffer, L., Roy, S., Sridhara, R., Rahman, A., Williams, G. and Pazdur, R. (2001). Approval Summary: Gemtuzumab ozogamicin in relapsed acute myeloid leukemia. Report from FDA. *Clin. Cancer Res.* **7**: 1490–1496.
10. Chi, G. Y. H. (1985). A design problem in ulcer prevention trials. Proceedings of the Biopharmaceutical Subsection of the *Am. Statist. Assoc.*, 100–105.
11. Chi, G. Y. H. (1998). Multiple testings: Multiple comparisons and multiple endpoints. *Drug Infor. J.* **32** (Suppl.): 1347s–1362s.
12. Chi, G. Y. H. and Liu, Q. (1999). The attractiveness of the concept of a prospectively designed two-stage clinical trial. *J. Biopharmaceutical. Statist.* **9**: 537–547.
13. Chi, G. Y. H. (2000). Clinical decision rules and multiple endpoints: A regulatory perspective. Presented at the *2nd Inter. Conf. Multiple Comparisons* held at the Humboldt University/Charite, Berlin, Germany, June 25–28.
14. Chow, S. C. and Liu, J. P. (1998). *Design and Analysis of Clinical Trials*: *Concept and Methodologies*, John Wiley and Sons, Inc.
15. Cui, L., Hung, H. M. J and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**: 853–857.
16. D'Agostino, R. B., Sr. (2000). Controlling alpha in a clinical trial: The case for secondary endpoints. *Statist. Med.* **19**: 763–766.
17. Dunnett, C. W. and Tamhane, A. C. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics* **51**: 217–227.
18. Elashoff, J. D., Koch, G. G. and Chi, G. Y. H. (1988). Designing a clinical trial to demonstrate prevention of ulcer recurrence: Modelling simulation approaches. *Statist. Med.* **7**: 877–888.
19. Ellenberg, S. S. and Temple, R. (2000). Placebo-controlled trials and active control trials in the evaluation of new treatments, Part II: Practical issues and specific cases. *Ann. Int. Med.* **133**(6): 464–470.
20. Fisher, L. D., Dixon, D. O., Herson, J., Frankowski, R. F., Hearron, M. S. and Peace, K. E. (1990). *Intention-to-Treat in Clinical Trials*, Chapter 7. Statistical issues in drug research and development, ed. E. Karl, Peace, Marcel Dekker, Inc.
21. Fisher, L. D. and Moyé, L. A. (1999a). Carvedilol and the Food and Drug Administration approval process: An introduction. *Controlled Clin. Trials* **20**: 1–15.
22. Fisher, L. D. (1999b). Carvedilol and the Food and Drug Administration (FDA) approval process: The FDA paradigm and reflections on hypothesis testing. *Controlled Clin. Trials* **20**: 16–39.
23. Fleming, T. R., Harrington, D. P. and O'Brien, P. C. (1984). Designs for group sequential tests. *Controlled Clin. Trials* **5**: 348–361.

24. Fleming, T. R. (1987). Treatment evaluation in active control studies. *Cancer Treatment Reports* **71**(11): 1061–1065.
25. Fleming, T. R. (1990). Evaluation of active control trials in AIDS. *J. AIDS* **3** (Suppl. 2): S82–S87.
26. Fleming, T. R. (2000). Design and interpretation of equivalence trials. *Am. Heart J.* **139**: s171–s176.
27. Follman, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Statist. Med.* **14**: 1163–1175.
28. Gillings, D. and Koch, G. (1991). The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Infor. J.* **25**: 411–424.
29. Goodman, S. (1999). Toward evidence-based medical statistics 1: The p value fallacy. *Ann. Intern. Med.* **130**: 995–1004.
30. Hassalblad, V. and Kong, D. F. (2001). Statistical methods for comparison to placebo in active-control trials. *Drug Infor. J.* **32**(2): 435–450.
31. Hauck, W. W. and Anderson, S. (1999). Some issues in the design and analysis of equivalence trials. *Drug Infor. J.* **33**: 109–118.
32. Hauschke, D., Schafer, J. and Pigeot, I. (2001). Statistical approaches for the choice of delta. Presented at the *37th Ann. Meet. of the Drug Infor. Assoc.*, July 8–12, Denver, Colorado.
33. Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**: 800–802.
34. Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, Wiley, New York.
35. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian. *J. Statist.* **6**: 65–70.
36. Holmgren, E. B. (2001). Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *J. Biopharmaceut. Statist.* **9**(4): 651–659.
37. Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**: 383–386.
38. Hung, H. M. J., Chi, G. Y. H and Lipicky, R. J. (1993). Testing for the existence of a desirable dose combination. *Biometrics* **49**: 85–94.
39. Huque, M., Dubey, S. and Fredd, S. (1989). Establishing therapeutic equivalence with clinical endpoints. *Proc. Biopharmaceut. Sec.*, 46–52.
40. Huque, M. F. and Sankoh, A. J. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *J. Biopharmaceut. Statist.* **7**: 545–564.
41. Jin, K. and Chi, G. Y. H. (1997). Application of bootstrap in handling multiple endpoints. *Proc. Biopharmaceut. Sec. Am. Statist. Assoc.*, 150–155.
42. Jin, K. and Chi, G. Y. H. (1998). Clinical decision rules and statistical support structures — a novel approach to handling the multiple endpoints problem. *Proc. Biopharmaceut. Sec. Am. Statist. Assoc.*, 56–62.
43. Kieser, M., Bauer, P. and Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometric. J.* **41**(3): 261–277.

44. Koch, G. G. and Tangen, C. M. (1999). Nonparametric analysis of covariance and its role in noninferiority clinical trials. *Drug Infor. J.* **33**: 1145–1159.
45. Koch, G. G. (2000). Discussion for Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statist. Med.* **19**: 781–784.
46. Kurata, J. H. and Koch, G. G. (1988). Response to H2-receptor antagonists and duodenal ulcer recurrence. *Am. J. Gastroenterol.* **83**(12): 1427–1428.
47. Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clin. Trials* **21**: 167–189.
48. Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**(3): 659–663.
49. Leber, P. (1986). The placebo control in clinical trials — A view from the FDA. *Psychopharmacol. Bull.* **22**(1): 30–32.
50. Leber, P. (1989). Hazards of inference: The active control investigation. *Epilepsia* **30** (Suppl. 1): s57–s63.
51. Lehmacher, W., Wassmer, G. and Reitmer, P. (1991). Procedures for two sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47**: 511–521.
52. Lewis, J. H. (1985). Summary of the Gastrointestical Drug Advisory Committee Meeting, March 21 and 22. *Am. J. Gastroenterol.* **80**: 581–583.
53. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc.
54. Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Assoc.* **90**(431): 1112–1121.
55. Liu, A. and Hall, W. J. (1999). Unbiased estimation following a group sequential test. *Biometrika* **86**: 71–78.
56. Liu, Q. and Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57**: 172–177.
57. Liu, Q. (2001). On general two-stage adaptive designs with dependent data. Personal communication, 1–22.
58. Lavori, P. W., Dawson, R. and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statist. Med.* **14**: 1913–1925.
59. Malinckrodt, C., Clark, W. and David, S. (2001). Accounting for dropout bias using mixed-effects models. *J. Biopharmaceut. Statist.* **11**(1&2): 9–12.
60. Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* **63**: 655–660.
61. Mathieu, M. (1997). *New Drug Development*: *A Regulatory Overview*, Parexel International Corporation, Waltham, Massachusetts.
62. McManus, J. and Wormsley, K. G. (1989). A new perspective on what maintenance therapy of duodenal ulcer achieves — Selected Summaries. *Gastroenterology* **96**(4): 1218–1220.
63. Moyé, L. A. (1999). End-point interpretation in clinical trials: The case for discipline. *Controlled Clin. Trials* **20**: 40–49.

64. Moyé, L. A. (2000a). Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statist. Med.* **19**: 767–779.
65. Moyé, L. A (2000b). Response to commentaries on Alpha calculus in clinical trials: Considerations and commentary for the new millennium. *Statist. Med.* **19**: 795–799.
66. Myers, W. R. (2000). Handling missing data in clinical trials: An overview. *Drug Infor. J.* **34**: 525–533.
67. Neuhäuser, M., Steinijans, V. W. and Bretz, F. (1999). The evaluation of multiple clinical endpoints, with application to asthma. *Drug Infor. J.* **33**: 471–477.
68. O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**: 549–556.
69. O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**: 1079–1087.
70. O'Neill, R. T. (2000). Commentary on Alpha calculus in clinical trials: Considerations and commentary for the new millenium. *Statist. Med.* **19**: 785–793.
71. Peace, K. E. (1988). *Biopharmaceutical Statistics for Drug Development*, Marcel Dekker, Inc.
72. Peace, K. E. (1990). *Statistical Issues in Drug Research and Development*, Marcel Dekker, Inc.
73. Physicians' Desk Reference (1999). *Physicians' Desk Reference*, Medical Economics Company, Inc., Montvale, NJ 07645-1742.
74. Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**(2): 191–199.
75. Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* **38**: 153–162.
76. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**: 487–498.
77. Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**: 1315–1324.
78. Riis, P. (2000). Perspectives on the 5th revision of the Declaration of Helsinki. *J. Am. Med. Assoc.* **284**(23): 3045–3046.
79. Rothmann, M., Chen, G., Li, N. and Chi, G. Y. H. (2001). Design and analysis of non-inferiority mortality trials in oncology. *DIA 37th Annual Meeting*, July 8–12, Denver, Colorado. To appear in *Stat. in Med. 2003.*
80. Rüger, B. (1978). Das maximale signifikanzniveau des tests. Lehno Ho ab, wenn k under n gegebenen tests zur Ablehnung fuhren. *Metrika* **25**: 171-178.
81. Samuel-Cahn, E. (1996). Is the Simes improved Bonferroni procedure conservative? *Biometrika* **83**(4): 928–933.
82. Sankoh, A. J., Huque, M. F. and Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statist. Med.* **16**: 2529–2542.
83. Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the Simes Conjecture. *Ann. Statist.* **26**(2): 494–504.

84. Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**: 190–197.
85. Siegel, J. P. (2000). Equivalence and non-inferiority trials. *Am. Heart J.* **139**: s166–s170.
86. Simon, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics* **55**: 484–487.
87. Tamhane, A. J., Hochberg, Y. and Dunnett, C. W. (1996). Multiple test procedures for dose finding. *Biometrics* **28**: 519–531.
88. Tang, D. I., Geller, N. L. and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**: 23–30.
89. Tang, D. I., Gnecco, C. and Geller, N. L. (1989). An approximate likelihood ratio test for a normal mean vector with non-negative components with application to clinical trials. *Biometrika* **76**: 577–583.
90. Temple, R. (1983). Difficulties in evaluating positive control trials. *Proc. Am. Statist. Assoc.* (*Biopharmaceut. Sec.*), 1–7.
91. Temple, R. (1996). Problems in interpreting active control equivalence trials. *Account. Res.* **4**: 267–275.
92. Temple, R. and Ellenberg, S. S. (2000). Placebo-controlled trials and active control trials in the evaluation of new treatments, Part I: Ethical and scientific issues. *Ann. Int. Med.* **133**(6): 455–463.
93. Tilley, B. C., Maler, J., Geller, N. L., Lu, M., Legler, J., Brott, T., Lyden, P. and Grotta, J. (1996). Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. *Stroke* **27**(11): 2136–2142.
94. Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Statist. Assoc.* **77**(380): 855–861.
95. US Code of Federal Regulations 21: Parts 300–499 (2001). US Government Printing Office.
96. US Food and Drug Administration (1996). Oncology Initiatives. http://www.fda.gov/opacom/backgrounders/cancerbg/html.
97. US Food and Drug Administration (1998–2000). International Conference on Harmonization Requirements for Registration of Pharmaceuticals for Human Use (ICH) E-1-E-11. http://www.FDA.gov/CDER/guidance/ index.html.
98. US Food and Drug Administration (1997). International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). E-8: Guidance on General Considerations for Clinical Trials. Federal Register 62, 242, December 17, 1997/Notices, 66113–66119.
99. US Food and Drug Administration (1998). International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). E-9: Guidance on Statistical Principles for Clinical Trials. Federal Register 63, 179, September 16, 1998/Notices, 49583–49598.

100. US Food and Drug Administration (1999). International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). E-10: Guidance on Choice of Control Group in Clinical Trials. Federal Register 64, 185, September 24, 1999/Notices, 51767–51780.
101. US Food and Drug Administration (2001). On the *Establishment and Operation of Clinical Trial Data Monitoring Committees* — Draft Guidance. Federal Register: Pending.
102. Vastag, B. (2000). Helsinki discord? A controversial declaration. *J. Am. Med. Assoc.* **284**(23): 2983–2985.
103. Verbeke, G., Molenberghs, G., Bijnens, L. and Shaw, D. (1997). *Linear Mixed Models in Practice*. New York: Springer.
104. Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics***54**: 696–705.
105. Wassmer, G., Reitmeir, P., Kieser, M. and Lehmacher, W. (1999). Procedures for testing multiple endpoints in clinical trials: An overview. *J. Statist. Plan. Inference* **82**: 69–81.
106. Westfall, P. H. and Young, S. S. (1992). *Resampling-Based Multiple Testing*, Wiley, New York.
107. White, H. D. (1998). Thrombolytic therapy and equivalence trials — Editorial Comment. *J. Am. Coll. Cardiol.* **31**(3): 494–496.
108. Whitehead, J. (1983). *The Design and Analysis of Sequential Clinical Trials*. New York: Ellis Horwood.
109. Whitehead, J. (1986a). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**: 573–581.
110. Whitehead, J. (1986b). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics* **42**: 461–471.
111. World Medical Association Declaration of Helsinki — Ethical principles for medical research involving human subjects. *J. Am. Med. Assoc.* **284**(23): 3043–3045.
112. Ellenberg, S. S., Fleming, T. R. and DeMets, D. L. (2002). *Data Monitoring Committees in Clinical Trials*, Wiley, England.

About the Author

Dr. Chi is currently the Director, Division of Biometrics I, Center for Drug Evaluation and Research, Food and Drug Administration(FDA), Rockville, Maryland.

He received his doctorate degree in Mathematics (1970) from the Carnegie-Mellon University. He has taught at the University of Pittsburgh, University of Florida, and visited the University of Bucharest as a Senior Fulbright Scholar. Prior to joining FDA in 1983, Dr. Chi has been associated with the LIPID project at the University of North Carolina as a NIH research fellow.

Dr. Chi has an active interest in regulatory research. Recent research publications on factorial trials (*Biometrics*, March 1993, December 1995 *and Communication in Statistics* **23**, 1994) dealt with the development of new test statistics for evaluating the effectiveness of antihypertensive combination drugs. Current activities include the analysis of ambulatory blood pressure data, analysis of multiple endpoints, prospectively designed two-state trials (*Biometrics* 2001), design modifications, clinical decision rules and multiplicity, the design and analysis of active control non-inferiroity trials, the preparation of interim analysis and multiple endpoints guidelines, and guidance documents on how to review NDA/INDs. He is currently a member of the Statisticsal Policy Coordinating Committee.

Section 3
Statistical Methods in Epidemiology

This page intentionally left blank

CHAPTER 15

STATISTICS IN GENETICS

ZHAOHAI LI

George Washington University, Department of Statistics,
315 Funger Hall, Washington, DC 20052, USA
Tel: 202-994-6357; zli@research.circ.gwu.edu

MINYU XIE

Central China Normal University, Wuhan, China
and George Washington University, Washington DC, USA

1. Introduction

Recent advances in molecular genetics have provided opportunities for genetic studies of complex human traits. Many human diseases such as cystic fibrosis, insulin dependent diabetes mellitus, hypertension, and schizophrenia, are considered as having some genetic component. Locating the genes that affect susceptibilities to these diseases is important in understanding the etiology of the diseases and may result in better treatment. In this chapter, we give an overview of segregation and linkage analysis of genetic data. We start with an introduction of basic genetic concepts and relevant terminology.

1.1. *Genetic terminology*

Each individual has 23 pairs of *chromosomes*. One of the 23 pairs is formed by two *sex chromosomes*, X and Y. Every woman carries XX chromosomes and every man carries XY. The other 22 pairs are called *autosomal chromosomes*. We focus on autosomal chromosomes in this chapter. The human genome can be imagined as two parallel straight lines. A given location in the genome (like a segment or point on the straight line) is called a *locus*.

The different forms of genetic variants are termed as *alleles*. In molecular genetics or biology, the term, *gene*, is used to refer to both an allele or a locus. Alleles are often denoted by letters or numbers, such as A, a, B, b, 1, 2, 3, etc. The proportion of a specific allele at a given locus in the population is called the *allele frequency* or *allele probability*. For example, $p = P(A) = 0.3$, implies that 30% of alleles at a given locus are "A." The allele frequency can be estimated from genetic data. At any given locus, each individual has two alleles, such as AA, Aa, or aa. The pair of alleles at a locus is referred to as the *genotype*. The order of the two alleles in the genotype is not relevant. Thus, "Aa" and "aA" are considered to be the same genotype. If an individual has two identical alleles (e.g. AA or aa) at a given locus, then the individual is said to be *homozygous* at that locus. If the two alleles are different (e.g. Aa), then the individual is said to be *heterozygous* at that locus. Usually, the determination of the genotype of an individual at a given locus requires laboratory work. However, some of characteristics are readily observable, such as an individual's eye color and height. An observable characteristic is called the *phenotype*. The relationship between the genotype and the phenotype is not necessarily one to one. It is possible that several different genotypes correspond to one phenotype. If genotype AA and genotype Aa both have the same phenotype (characteristic), but different from that of genotype aa, then allele A is said to be *dominant* to allele a, or allele a is said to be *recessive* to allele A. In this case, the phenotype or characteristic corresponding to AA is called dominant, and the phenotype corresponding to aa is called recessive. If the phenotype associated with the genotype Aa is different from those of both AA and aa, then alleles A and a are said to be *codominant*. For example, there are three alleles, A, B, and O, at the blood group locus. Genotypes AA and AO have the same phenotype, blood type A, and genotypes BB and BO have the same phenotype, blood type B. Hence, allele A is dominant to allele O, and allele B is also dominant to allele O. The genotype OO has a recessive phenotype, blood type O, and the genotype AB has a codominant phenotype, blood type AB.

When we consider more than one locus simultaneously, the alleles (at different loci) received by an individual from one parent are called a *haplotype*. A pair of haplotypes is a multilocus genotype. Suppose that there are two loci, locus one with two alleles A and a, and locus two with two alleles B and b. Figure 1 presents a hypothetical family with two parents and a son and a daughter in which the son received haplotype ab from the mother and haplotype Ab from the father, and the daughter received

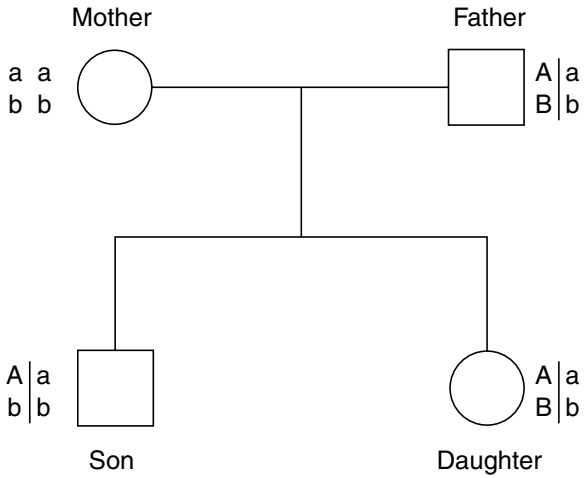


Fig. 1.

haplotypes AB from the father and ab from the mother. The squares and circles in Fig. 1 indicate males and females, respectively. Notice that there is a vertical bar between the two haplotypes for some individuals in Fig. 1. The bar notation represents a *phase known* genotype. That is, alleles on the same side of the bar are from the same parent, or equivalently the maternal and paternal origin for each allele is known. The haplotype that the son received from the father is Ab which is different from the two original haplotypes of the father, AB and ab. Therefore, the alleles at the two loci have been *recombined* during the process of transmitting from the father to the son. This phenomenon is referred to as *crossing-over*. Only an odd number of cross-overs between two loci are observable. The probability of an odd number of cross-overs between two loci is called the *recombination fraction*, denoted by θ.

An individual receives one of the two alleles from the genotype of each parent with equal probability. Suppose the genotype of a parent is Aa, then the above principle implies that $P\{\rightarrow A|Aa\} = P\{\rightarrow a|Aa\} = \frac{1}{2}$, where $\{\rightarrow A|Aa\}$ denotes the event that a parent transmits allele A to the offspring given that the parent has genotype Aa. This is frequently referred to as *Mendel's first law* or *the principle of independent segregation*. If the loci of a two-locus genotype are on different chromosomes, then the transmission of the alleles at one locus is independent of the transmission of the alleles at the other locus. For example, suppose a parent has genotype $AaBb$

(two-locus genotype), and the two loci are on the different chromosomes, then

$$P\{\rightarrow AB|AaBb\} = P\{\rightarrow A|Aa\} \times P\{\rightarrow B|Bb\} = \frac{1}{4}\,,$$

$$P\{\rightarrow Ab|AaBb\} = P\{\rightarrow A|Aa\} \times P\{\rightarrow b|Bb\} \ = \frac{1}{4}\,,$$

$$P\{\rightarrow aB|AaBb\} = P\{\rightarrow a|Aa\} \times P\{\rightarrow B|Bb\} \ = \frac{1}{4}\,,$$

$$P\{\rightarrow ab|AaBb\} = P\{\rightarrow a|Aa\} \times P\{\rightarrow b|Bb\} \ = \frac{1}{4}\,.$$

This principle of *independent assortment* is often called *Mendel's second law.*

1.2. *The Hardy–Weinberg equilibrium*

Random mating is defined as: any female is equally likely to mate with any male. That is, the probability of the mating type is the product of the probabilities of the genotypes of the female and male mates, for example, $P\{AA \times Aa\} = P(AA)P(Aa)$. Notation $AA \times Aa$ indicates the mating type resulting from an AA individual mating with an Aa individual.

Next, we consider a locus with two alleles, A and a. Assume that the allele frequencies (or probabilities) for the two alleles are

$$P(A) = p\,, \quad P(a) = q\,,$$

where $p+q=1$. A population is said to be in *equilibrium* if the proportions (or probabilities) of the three genotypes of the current generation in the population are

$$P(AA) = p^2\,, \quad P(Aa) = 2pq\,, \quad P(aa) = q^2\,. \tag{1}$$

Under random mating, the allele and genotype probabilities for next generation are the same as the current generation, i.e.

$$P(A) = p\,, \quad P(a) = q\,,$$

and

$$P(AA) = p^2\,, \quad P(Aa) = 2pq\,, \quad P(aa) = q^2\,.$$

The genotypic and allelic probabilities stay the same from generation to generation for an equilibrium population. This is known as the Hardy–Weinberg law.[19,67]

If the genotypic probabilities of the current generation do not satisfy condition (1), then equilibrium will be reached after one generation of random mating.[68] For details of the derivation of the Hardy–Weinberg law, readers are referred to an excellent book on population genetics by C. C. Li.[29]

1.3. *Linkage and linkage equilibrium*

From Mendel's second law, if two genetic loci are on different chromosomes, then the transmission of alleles (segregation) at one locus is independent of that at the other locus; there, the recombination fraction is $\theta = \frac{1}{2}$. If the two genetic loci are close together, the alleles that are paternal (or maternal) in origin tend to transmit together (*cosegregation*) to an offspring. This phenomenon is known as *linkage*. The closer the two loci are, the smaller the probability for crossing over; thus, the recombination fraction for the two linked loci is smaller than $\frac{1}{2}$.

Next, we consider two genetic loci in more detail. Suppose that the first locus has two alleles, A and a; and the second locus has two alleles, B and b. The respective allele probabilities are

$$P(A) = p\,, \quad P(a) = q\,, \quad P(B) = u\,, \quad P(b) = v\,,$$

where $p + q = 1$ and $u + v = 1$. There are nine two-locus joint genotypes. Under the assumptions of random mating and no linkage ($\theta = \frac{1}{2}$), if the genotypic probabilities are:

$AABB$	$AABb$	$AAbb$	$AaBB$	$AaBb$	$Aabb$	$aaBB$	$aaBb$	$aabb$
p^2u^2	$2p^2uv$	p^2v^2	$2pqu^2$	$4pquv$	$2pqv^2$	q^2u^2	$2q^2uv$	q^2v^2

then, the genotypic probabilities of the next generation will be the same as the current generation. Hence, if each locus is in equilibrium separately in a population; and the two loci are not linked, then the two loci are jointly in equilibrium. For a population not in equilibrium jointly, the joint equilibrium will not be reached after a single generation of random mating. The joint equilibrium is approached as the number of generations $n \to \infty$. The speed of approaching joint equilibrium depends on the recombination fraction, θ (see the following Eq. (3)).

If alleles at two loci are in random association (independent), the two loci are said to be in a state of *linkage equilibrium*. To illustrate this concept, consider two diallelic loci with alleles A and a for locus one, and alleles B

and b for locus two. If these loci are in a state of linkage equilibrium, then the haplotype probabilities satisfy:

$$\begin{aligned} P(AB) &= P(A)P(B)\,, \quad P(Ab) = P(A)P(b)\,, \\ P(aB) &= P(a)P(B)\,, \quad P(ab) = P(a)P(b)\,. \end{aligned} \tag{2}$$

Under linkage equilibrium, the joint probabilities of each two locus haplotype are equal to the products of the corresponding single-locus allele probabilities. If alleles at two loci are not in random association (dependent), this situation is referred to as *allelic association* or *linkage disequilibrium*. For two diallelic loci, under linkage disequilibrium, we have

$$\delta = P(AB) - P(A)P(B) \neq 0\,,$$

where δ is the departure from equilibrium; it is the linkage disequilibrium parameter. It can be shown that

$$\begin{aligned} P(AB) &= P(A)P(B) + \delta\,, \quad P(Ab) = P(A)P(b) - \delta\,, \\ P(aB) &= P(a)P(B) - \delta\,, \quad P(ab) = P(a)P(b) + \delta\,. \end{aligned}$$

If linkage disequilibrium exists ($\delta \neq 0$) initially for a population, under random mating, then the linkage disequilibrium parameter will approach zero as the number of generations, $n \to \infty$. Specifically, let δ_0 be the initial disequilibrium parameter and θ be the recombination fraction between two loci, then after n generations,

$$\delta_n = (1-\theta)^n \delta_0\,, \tag{3}$$

where δ_n is the linkage disequilibrium parameter of the nth generation. The linkage disequilibrium will decrease quickly over generations when the linkage is weak, i.e. the recombination fraction θ is large (close to $\frac{1}{2}$). If the two loci are unliked, $\theta = \frac{1}{2}$ and equilibrium is reached very quickly; if the two loci are tightly linked, $\theta \simeq 0$, and disequilibrium will continue for many generations. This is the basis for fine mapping using disequilibrium. Therefore, a large linkage disequilibrium is often considered to be evidence of linkage (small θ).

For readers who would like to know more about linkage and population genetics in general, they are referred to the books by Li[29] and Hartl and Clark.[20] Most books on statistical methods in genetics provide an introductory chapter on basic genetic concepts and terminology.[13,27,40,54] The book by Watson *et al.*[65] gives a comprehensive description of the molecular biology of genes. Olson *et al.*[38] wrote a tutorial on genetic mapping of complex traits. For fundamentals of genetic epidemiology, readers are referred to Khoury *et al.*[23] and Thompson.[64]

2. Segregation Analysis

Mendel proposed that two factors (alleles) segregate from one another in the process of forming gamates (sperm or ovum) that constitute the genetic makeup of the next generation. The Mendelian *segregation ratio* is the conditional probability of the genotype in the offspring given the mating type, for example,

$$P\{aa|Aa \times Aa\} = \frac{1}{4}\,.$$

These ratios are fixed under the Mendelian inheritance model. Hence, one task of *segregation analysis* is to test whether or not the observed phenotype data among the offspring is consistent with Mendelian inheritance. In general, segregation analysis tests models of inheritance with family data.

2.1. *Estimating allele probability (Gene frequency)*

Suppose that a random sample of individuals is selected to observe the phenotypes at a given autosomal locus. If all alleles are codominant, one can estimate the frequency of an allele at the locus by counting the number of alleles in the sample and then dividing by the total number of genes (alleles) in the sample. There are twice as many alleles as individuals because each individual has two alleles at a given locus. Consider a given locus which has two codominant alleles A and a. Suppose that a random sample of n individuals is ascertained to study the locus, with n_{AA}, n_{Aa}, n_{aa} corresponding to the counts of the three genotypes, AA, Aa, aa respectively ($n_{AA} + n_{Aa} + n_{aa} = n$). Then frequency p_A of allele A is estimated by

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n}\,.$$

Likewise

$$\hat{p}_a = \frac{2n_{aa} + n_{Aa}}{2n}\,,$$

with $\hat{p}_A + \hat{p}_a = 1$. The variances of the estimated allele frequencies are

$$\mathrm{var}(\hat{p}_A) = \frac{2np_A(1-p_A)}{(2n)^2} = \frac{p_A(1-p_A)}{2n}\,, \quad \mathrm{var}(\hat{p}_a) = \frac{p_a(1-p_a)}{2n}\,.$$

Example: The human MN blood type locus is a codominant locus with two alleles M and N. Li[29] cites the results of L. Ride (1935; cf. Haldance, 1938) on MN blood type data. More than one thousand Chinese residents in Hong Kong were tested for the MN blood type. The following results were obtained:

Blood Types	MM	MN	NN	Total
Numbers	342	500	187	1029

$$\hat{p}_M = \frac{2 \times 342 + 500}{2 \times 1029} = \frac{1184}{2058} = 0.5753\,.$$

Suppose that a given locus has k codominant alleles with n_i alleles of type i in a random sample of n individuals, where $n_1 + n_2 + \cdots + n_k = 2n$. Then, the allele frequency p_i of allele type i is estimated by

$$\hat{p}_i = \frac{n_i}{2n}\,.$$

It is easy to see that the allele counts $(n_1, n_2, \ldots, n_k)$ has a multinomial distribution with parameters $(p_1, p_2, \ldots, p_k)$. Thus,

$$E(\hat{p}_i) = p_i\,, \quad \text{var}(\hat{p}_i) = \frac{p_i(1-p_i)}{2n}\,, \quad \text{cov}(\hat{p}_i, \hat{p}_j) = -\frac{p_i p_j}{2n}\,, \quad \text{for } i \neq j\,.$$

In fact, $\hat{p}_i$ is a maximum likelihood estimate of p_i. Both maximum likelihood estimates (MLE) and likelihood ratio tests play very important roles in genetic analysis.

Consider a locus with two alleles A and a. If allele A is dominant, then the genotype AA and Aa each have the same phenotype. Let n_A be the number of individuals with either genotype AA or Aa, and let n_a be the number of individuals with genotype aa, where $n_A + n_a = n$. Under the random mating assumption and by Hardy–Weinberg equilibrium, the probability of the recessive genotype equals the squared probability of allele a, i.e. $P(aa) = p_a^2$. Therefore

$$\hat{p_a^2} = \frac{n_a}{n}\,, \quad \text{and} \quad \hat{p}_a = \sqrt{\frac{n_a}{n}}\,.$$

Note n_a has a binomial distribution with parameter p_a^2. Hence, $\hat{p}_a^2$ is the MLE for p_a^2. By the invariance principle of MLE, $\hat{p}_a$ is the MLE for p_a. By using the δ-method[47] and the fact that

$$\text{var}(\hat{p_a^2}) = \frac{p_a^2(1-p_a^2)}{n}\,,$$

we could derive

$$\text{var}(\hat{p}_a) = \frac{1-p_a^2}{4n}\,.$$

For a dominant locus with more than two alleles, such as the ABO blood locus, estimating allele frequencies is a little more complex. The EM algorithm can be used to obtain the MLEs of the allele frequencies.[9,27,32,39,56]

2.2. *Testing Hardy–Weinberg equilibrium*

Hardy–Weinberg equilibrium is commonly assumed in genetic analyses. The validity of this assumption can be tested by the Pearson χ^2 test:

$$\chi^2 = \sum \frac{(O-E)^2}{E}\,.$$

For a two allele codominant locus, under Hardy–Weinberg equilibrium (H_0), the expected numbers of individuals with various genotypes in a random sample of n individuals are

AA	Aa	aa
np_A^2	$2np_Ap_a$	np_a^2

Let n_{AA}, n_{Aa}, n_{aa} be the observed numbers of individuals with genotypes AA, Aa, aa, respectively. Then,

$$\hat{p}_A = \frac{2n_{AA}+n_{Aa}}{2n}\,, \quad \hat{p}_a = \frac{2n_{aa}+n_{Aa}}{2n}\,.$$

The expected numbers of the individuals with various genotypes can be estimated by:

$$\hat{E}_{AA} = n\hat{p}_A^2\,, \quad \hat{E}_{Aa} = 2n\hat{p}_A\hat{p}_a\,, \quad \hat{E}_{aa} = n\hat{p}_a^2\,.$$

The one degree of freedom Pearson χ^2 statistic is

$$\chi^2 = \frac{(n_{AA}-\hat{E}_{AA})^2}{\hat{E}_{AA}} + \frac{(n_{Aa}-\hat{E}_{Aa})^2}{\hat{E}_{Aa}} + \frac{(n_{aa}-\hat{E}_{aa})^2}{\hat{E}_{aa}}\,.$$

Reject the null hypothesis (H_0) that the population is in Hardy–Weinberg equilibrium when the χ^2 value is large, larger than $\chi^2_{1-\alpha}(1)$, where α is the level of significance.

Example: The following data were obtained from a hypertension genetic study. A random sample of 197 individuals were genotyped for the angiotensin-converting enzyme (ACE) locus with

AA	Aa	aa
26	93	78

From this data set, we have

$$\hat{p}_A = 0.3680\,, \quad \hat{p}_a = 0.6320\,, \quad \hat{E}_{AA} = 197\times(0.3680)^2 = 26.68\,,$$

$$\hat{E}_{Aa} = 2\times 197\times 0.3680\times 0.6320 = 91.63\,,$$

$$\hat{E}_{aa} = 197\times(0.6320)^2 = 78.69\,,$$

and

$$\chi^2 = \frac{(26 - 26.68)^2}{26.68} + \frac{(93 - 91.63)^2}{91.63} + \frac{(78 - 78.69)^2}{78.69} = 0.0439\,.$$

This test fails to reject the null hypothesis that the population is in Hardy–Weinberg equilibrium at the significance level $\alpha = 0.05$, since $\chi^2_{0.95}(1) = 3.841$.

2.3. *Segregation analysis of dominant loci*

One method to show the genetic basis of single locus inheritance of a given disease is to demonstrate the Mendelian segregation ratio. Consider a rare dominant two allele disease locus with allele A and a. Assume that the allele A causes the disease with frequency $P(A) = p \simeq 0$. The individuals with genotypes AA and Aa will have the disease, while individuals with genotype aa will be unaffected. The observable phenotypes are affected or unaffected. The Mendelian segregation ratios for six mating types under random mating are presented in Table 1.

Table 1. Mendelian segregation ratios for 6 mating types under random mating.

		Genotype			Phenotype	
Mating	P(Mating)	AA	Aa	aa	Affected	Unaffected
$AA \times AA$	p^4	1	0	0	1	0
$AA \times Aa$	$4p^3q$	$\frac{1}{2}$	$\frac{1}{2}$	0	1	0
$AA \times aa$	$2p^2q^2$	0	1	0	1	0
$Aa \times Aa$	$4p^2q^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{4}$
$Aa \times aa$	$4pq^3$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$aa \times aa$	q^4	0	0	1	0	1

Among the five possible mating types which produce affected offspring, the mating type $Aa \times aa$ is the most likely to occur according to the above probabilities of mating types and the fact that $p = P(A) \simeq 0$.

The most informative ascertainment procedure or sampling scheme is to select families with one affected parent and another unaffected parent. The mating of such families is commonly assumed to be $Aa \times aa$. Let $\tau = P\{Aa|Aa \times aa\} \simeq P\{\text{Affected}|Aa \times aa\}$ be the Mendelian ratio parameter, and let X be the random variable for the number of affected offspring for

an ascertained family. Then, X has a binomial distribution with probability function

$$L(r|\tau) = P\{X = r\} = \binom{n}{r} \tau^r (1-\tau)^{n-r}\,, \tag{4}$$

where n is the number of offspring. It is worth noting that the genotypes of the offspring are conditionally independent given the parental mating type. To demonstrate the Mendelian segregation ratio, we test $H_0 : \tau = \frac{1}{2}$.

2.4. *Approximated χ^2 test*

Suppose that k families with parental mating type $Aa \times aa$ were ascertained, and each family has r_i affected offspring among a total of n_i offspring $(n_1 + n_2 + \cdots + n_k = n)$. Let $X = X_1 + X_2 + \cdots + X_k$, then

$$E(X_i) = n_i\tau\,, \quad \mathrm{var}(X_i) = n_i\tau(1-\tau)\,, \quad E(X) = n\tau\,, \quad \mathrm{var}(X) = n\tau(1-\tau).$$

By the central limit theorem, we have

$$\frac{X - n\tau}{\sqrt{n\tau(1-\tau)}} \xrightarrow{D} N(0,1)\,.$$

Hence,

$$\frac{(X - n\tau)^2}{n\tau(1-\tau)} \xrightarrow{D} \chi_1^2\,,$$

where $\xrightarrow{D}$ denotes covergence in distribution. Under $H_0 : \tau = \frac{1}{2}$, we calculate the test statistic

$$\chi^2 = \frac{(\sum_{i=1}^k r_i - \frac{n}{2})^2}{\frac{n}{4}}\,.$$

Reject H_0 when χ^2 is large.

2.5. *Likelihood ratio test*

With the above family data structure and formula (4), the likelihood function of τ is

$$L(\tau|r_1, \ldots, r_k) = \prod_{i=1}^k L(\tau|r_i) = \left[\prod_{i=1}^k \binom{n_i}{r_i}\right] \tau^r (1-\tau)^{n-r}\,,$$

where $r = \sum_{i=1}^k r_i$. The MLE of τ is $\hat{\tau} = \frac{r}{n}$. The likelihood ratio statistic is

$$\chi^2 = \frac{L(\hat{\tau})}{L(\frac{1}{2})}\,.$$

2.6. *Segregation analysis of recessive loci*

Consider a rare recessive two allele disease locus. For an individual to be affected, the genotype of the individual has to be aa. There are three mating types, $Aa \times Aa$, $Aa \times aa$, $aa \times aa$, which could produce affected offspring. Let $P(A) = p$ and $P(a) = q = 1 - p$. Then, it can be shown that the conditional probabilities of mating types, given that one offspring is affected, are:

$$P\{Aa \times Aa|\text{Affected}\} = p^2\,,$$

$$P\{Aa \times aa|\text{Affected}\} = 2pq\,,$$

$$P\{aa \times aa|\text{Affected}\} = q^2\,.$$

Since the disease is rare recessive, i.e. $P(a) = q \simeq 0$, then $Aa \times Aa$ is the most likely mating type given that one child is affected. Therefore, the ascertainment procedure for a rare recessive disease is to select families with at least one affected child and then assume the mating type is $Aa \times Aa$ for analysis. The Mendelian segregation ratio for mating type $Aa \times Aa$ with a recessive disease is

$$\tau = P\{aa|Aa \times Aa\} = \frac{1}{4}\,.$$

We can estimate and test the segregation ratio τ.

It is possible that some families with at least one affected child are not included in the study sample due to chance. Fisher[15] in his classical paper recognized the need for correcting for the incomplete selection in segregation analysis and proposed methods to take the ascertainment procedure into account in the analysis. When families are selected on the basis of having at least one affected offspring, the affected individuals initially identified are called *probands*. It is possible that one ascertained family has more than one proband. The probability that an affected individual is a proband is called the ascertainment probability, and it is denoted by

$$\pi = P\{\text{Proband}|\text{Affected}\}\,.$$

The probability that a family, with r affected offspring, is not ascertained is

$$P\{\text{Not Ascertained}|r \text{ Affected}\} = (1 - \pi)^r\,.$$

Therefore, the probability that a family with r affected offspring is ascertained is

$$P\{\text{Ascertained}|r \text{ Affected}\} = 1 - (1 - \pi)^r\,.$$

If the ascertainment probability is one, i.e. $\pi = 1$, then all families with at least one affected offspring are ascertained. This situation is referred to as *complete ascertainment*. The situation where $\pi < 1$ is referred to as *incomplete ascertainment*. When the ascertainment probability is very small, that is, $\pi \simeq 0$, then $1-(1-\pi)^r \simeq r\pi$, and the phrase *single ascertainment* is used. Thus, for the single ascertainment procedure, the probability that an affected family is ascertained is proportional to the number of affected offspring. Under the single ascertainment procedure, almost all ascertained families have a single proband.

2.7. *Segregation analysis with complete ascertainment*

Under the complete ascertainment procedure, the number of affected offspring for each ascertained family has a truncated binomial distribution[15] with the likelihood function

$$L(\tau|r_i, s_i) = \frac{\binom{s_i}{r_i}\tau^{r_i}(1-\tau)^{s_i-r_i}}{1-(1-\tau)^{s_i}}, \quad r_i = 1, \ldots, s_i, i = 1, \ldots, n, \qquad (5)$$

where $\tau = P\{aa|Aa \times Aa\}$ is the Mendelian segregation ratio, s_i is the number of offspring for the ith family, r_i is the number of affected offspring, and n is the number of families. The goal of segregation analysis is to estimate and test the Mendelian segregation ratio τ.

Assume that all ascertained families have the same number of offspring, $s_i = s$ for all i. Let a_r be the number of families with r affected offspring $(r = 1, 2, \ldots, s)$ and n_s be the total number of ascertained families, then $\sum_{r=1}^{s} a_r = n_s$, and $\sum_{r=1}^{s} r a_r = A$ is the total number of affected offspring. From (5), the likelihood function based on n_s ascertained families is

$$L(\tau) = \prod_{i=1}^{n_s} L(\tau|r_i, s_i) = \prod_{r=1}^{s} \left[\frac{\binom{s}{r}\tau^r(1-\tau)^{s-r}}{1-(1-\tau)^s}\right]^{a_r}.$$

The maximum likelihood estimate for τ is the solution of the score equation $\frac{\partial L(\tau)}{\partial \tau} = 0$, which is equivalent to

$$\frac{s\tau}{1-(1-\tau)^s} = \frac{A}{n_s} = \bar{r}. \qquad (6)$$

There is no closed form solution to Eq. (6). The solution can be obtained by interative algorithms such as the: Newton–Raphson, Fisher scoring, and

EM algorithms. The Fisher information for τ is

$$I(\tau) = E\left(-\frac{\partial^2 L(\tau)}{\partial \tau^2}\right) = \frac{sn_s}{1-(1-\tau)^s}\left\{\frac{1-(1-\tau^s-s\tau(1-\tau)^{s-1}}{\tau(1-\tau)[1-(1-\tau)^s]}\right\}.$$

The variance of the MLE $\hat{\tau}$ can be estimated by $\frac{1}{I(\hat{\tau})}$.

2.8. *Segregation analysis with incomplete ascertainment*

In order for an individual selected at random, from families with parental mating type $Aa\times Aa$, to become a proband, the individual has to be affected and ascertained. Thus,

$$\begin{aligned} P\{\text{Proband}\} &= P\{\text{Affected and Selected}\} \\ &= P\{\text{Selected}|\text{Affected}\}P\{\text{Affected}\} = \pi\tau\,, \end{aligned}$$

where π is the ascertainment probability, and τ is the Mendelian segregation ratio. A family is considered to be segregating if the family has at least one affected offspring. The probability that a family with s offspring is a segregating family and not ascertained is $(1-\pi\tau)^s$. Thus, the probability that a family with s offspring is ascertained is $1-(1-\pi\tau)^s$. Let B be the random variable denoting the number of probands in a family. Then, the fact that a family is ascertained is equivalent to $B>0$ with $P(B>0) = 1-(1-\pi\tau)^s$. The likelihood function for an ascertained family with r affected offspring with sibship size s is[36]

$$\begin{aligned} L(\pi,\tau) &= P\{X=r|B>0;\ s,\pi,\tau\} \\ &= \frac{[1-(1-\pi)^r]\binom{s}{r}\tau^r(1-\tau)^{s-r}}{1-(1-\pi\tau)^s}\,. \end{aligned} \tag{7}$$

This likelihood function can be used to estimate and test ascertainment probability, π, and the Mendelian segregation ratio, τ. Since the analysis takes the ascertainment procedure into account, it is often referred to as an ascertainment bias corrected analysis.

There are two special cases to the likelihood function (7). When the ascertainment probability $\pi=1$, that is, complete ascertainment, the likelihood function (7) reduces to (5). If the ascertainment probability is very small, i.e. single ascertainment ($\pi\simeq 0$), then

$$(1-\pi)^r \simeq 1-r\pi\,, \quad (1-\pi\tau)^s \simeq 1-s\pi\tau\,,$$

and the likelihood function (7) is approximately

$$L(\pi,\tau) = P\{X = r|B > 0;\ s,\pi,\tau\} = \binom{s-1}{r-1}\tau^{r-1}(1-\tau)^{s-r}\,.$$

In the case when the exact number of probands for each ascertained family is known, the likelihood function is given by

$$\begin{aligned} L(\pi,\tau) &= P\{X = r, B = b|B > 0;\ s,\pi,\tau\} \\ &= \frac{\binom{r}{b}\pi^b(1-\pi)^{r-b}\binom{s}{r}\tau^r(1-\tau)^{s-r}}{1-(1-\pi\tau)^s}\,. \end{aligned}$$

In addition to the likelihood based inference for ascertainment probability and the Mendelian segregation ratio for a recessive locus under incomplete ascertainment, there are two other simple methods: the proband method and the singles method.

The proband method was initiated by Weinberg and described in detail by Fisher.[15] Suppose that n segregating families are ascertained and s_i, r_i, b_i are the sibship size, the number of affected offspring, and the number of probands for the ith family, respectively. The segregation ratio and the ascertainment probability are estimated by the proband method as

$$\hat{\tau} = \frac{\sum_{i=1}^n b_i(r_i-1)}{\sum_{i=1}^n b_i(s_i-1)} \quad \text{and} \quad \hat{\pi} = \frac{\sum_{i=1}^n b_i(b_i-1)}{\sum_{i=1}^n b_i(r_i-1)}\,.$$

If a proband is the only proband within an ascertained family, such a proband is referred to as a *single*. Let d be the number of singles in a sample of ascertained families. The segregation ratio and the ascertainment probability are estimated by the singles methods as

$$\hat{\tau} = \frac{\sum_{i=1}^n r_i - d}{\sum_{i=1}^n s_i - d} \quad \text{and} \quad \hat{\pi} = \frac{\sum_{i=1}^n b_i - d}{\sum_{i=1}^n r_i - d}\,.$$

The segregation analysis methods described above are for qualitative traits (i.e. affected or unaffected). Methods for segregation analysis of quantitative traits for single-locus and polygenic control were developed by Morton and MacLean.[37] This "mixed model" method is based on a likelihood approach. Several programs implemented the "mixed model," such as Pedigree Analysis Package (PAP),[22] SEGPATH,[44] and POINTER.[26] Bonney[5] developed a family of regressive models for segregation analysis of quantitative traits that allowed simultaneous adjustment for covariates and estimation of the parameters of Mendelian models. These models were implemented in the Statistical Analysis for Genetic Epidemiology (S.A.G.E.)

computer program.[11] Terwilliger and Ott[63] provided a list of computer programs for genetic analysis.

3. Linkage Analysis

Linkage analysis investigates whether or not two loci are physically located near one another on the same chromosome. Alleles from two linked loci (physically close) tend to segregate together, that is, they are passed from parent to child as a single unit. This phenomenon deviates from Mendel's second law of independent assortment. Linkage analysis is one of the methods used to localize disease traits in the human genome. Evidence of linkage between a known marker system and a putative gene for a disease is considered to be the highest level of statistical evidence that the disease is due to a genetic mechanism. Linkage analysis localizes a gene solely on the basis of its location, without regard to its biochemical function. This approach is called "positional cloning."

Alleles cosegregating due to linkage between two loci in one family may be different from alleles in another family. The cosegregating phenomenon due to linkage is only observable within families. Therefore, family data or data from biologically related subjects are necessary for detecting linkage. However, cosegregating caused by allelic association (linkage disequilibrium) can be detected by general population studies. Allelic association is a property of alleles, while linkage is a property of loci. They are two different but related concepts.

The measurement for linkage between two loci is the recombination fraction, θ. The closer the two loci, the less likely that a cross-over will occur between them and the smaller the recombination fraction, θ. Two extreme cases are: (1) $\theta = \frac{1}{2}$, the two loci are far apart and segregate independently (Mendel's second law of independent assortment); (2) $\theta = 0$, the two loci are identical and actually are one locus. The range of the recombination fraction is $0 \leq \theta \leq \frac{1}{2}$. Through map functions,[17,25,34,47] the recombination fraction, θ, between the two loci can be translated to the genetic distance between the two loci. Genetic distance is correlated with physical distance, but they are different. Linkage analysis estimates and tests the recombination fraction, θ. There are two types of statistical methods for linkage analysis: model-based and model free.

3.1. *The LOD score method*

The LOD (log-odds) score method is based on the maximum likelihood ratio test. Haldane and Smith[18] used the maximum likelihood approach to

linkage analysis. The LOD score method was widely used after Morton[35] published tables of log-odds (or LOD) scores that could be used in the analysis of family data. The LOD score method is considered to be a model-based procedure. Usually, the mode of inheritance, the number of alleles, and the penetrances of each genotype are assumed to be known for the LOD score method. The only unknown parameter is the recombination fraction θ. The conditional probability of observing the corresponding phenotype, say affection status, given the specified genotype, is referred to as *penetrance*. Ott[41] described the LOD score method in detail.

Assume that the likelihood function for a given family is $L(\theta)$, and $\hat{\theta}$ is the maximum likelihood estimate of θ. The LOD score for testing $H_0 : \theta = \frac{1}{2}$ vs. $\theta < \frac{1}{2}$ is

$$Z(\hat{\theta}) = \log_{10} \frac{L(\hat{\theta})}{L(\theta = \frac{1}{2})} .$$

Traditionally, reject H_0 and claim linkage if $Z(\hat{\theta}) > 3$.[35] The p-value corresponds to $Z(\hat{\theta}) > 3$ is less than 10^{-4}.

3.1.1. *Example of phase known data*

Figure 2 depicts a three generation hypothetical family with a binary phenotype (affected or unaffected) and marker genotype data for each individual. The dark symbols indicate that a subject is affected with an autosomal rare dominant disorder. We further assume that the disease locus has two alleles D and d with full penetrance. Hence, the penetrances are

$$P\{\text{Affected}|DD\} = P\{\text{Affected}|Dd\} = 1 , \quad P\{\text{Affected}|dd\} = 0 .$$

From the phenotype and marker locus genotype data in Fig. 2, the joint genotype (i.e. two locus genotype) of marker and trait loci and phase information can be inferred and are given in Fig. 3. Both grandmother and mother are homozygous with genotype dd at the disease locus because they are normal. Since the grandfather is affected, his genotype at the disease locus is either DD or Dd. It is reasonable to assume that his genotype is Dd since the disease is rare. This assumption has no impact on the linkage analysis because the grandparental genotypes are only used to determine the phase of the father. There are two possible phases for the grandfather and either one of them will give rise to the same phase for the father. The father is affected, so he must have at least one D allele. He must also receive one d allele from his mother. Therefore, the genotype of the father is Dd at the disease locus. The fact that the father received a haplotype

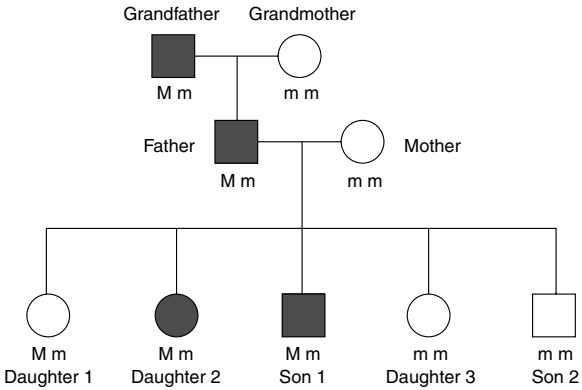


Fig. 2.

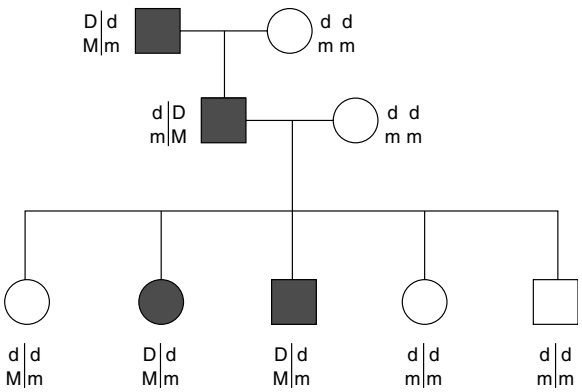


Fig. 3.

dm from his mother determines his phase. Hence, his genotype is dm/DM as shown in Fig. 3. Since the mother is double homozygous, each of the five children must receive a haplotype dm from her. The two locus genotypes and phases for the five children are inferred from the parental information and shown in Fig. 3. The mother is not informative for recombination. The father produced one recombinant gamate and four non-recombinant gamates. The likelihood function is

$$L(\theta) = \theta^r (1-\theta)^{N-r} = \theta(1-\theta)^4 ,$$

where r is the number of recombinants, and N is the number of gamates. The maximum likelihood estimate of θ is $\hat{\theta} = \frac{r}{N} = 0.2$, and the LOD score

is

$$Z(\hat{\theta}) = \log_{10} \frac{L(\hat{\theta})}{L(\theta = 0.5)} = \log_{10} \frac{0.20 \times 0.8^4}{0.5^5} = 0.4185\,.$$

The evidence is not strong enough to support linkage.

3.1.2. *Example of phase unknown data*

Suppose that the grandparents are missing in Figs. 2 and 3. Then, the phase of the father cannot be determined with certainty. With the given genotype, the father's genotype could have two possible phases, dm/DM (phase I) (Fig. 4) and dM/Dm (phase II) (Fig. 5). Under the phase dm/DM, there are one recombinant and four non-recombinants. There are one non-recombinant and four recombinants under the phase dM/Dm. Each of these two phases has equally likely probability ($\frac{1}{2}$) of being correct. The likelihood function becomes:

$$\begin{aligned} L(\theta) &= P\{\text{Data}\} \\ &= P\{\text{Data}|\text{Phase I}\}P\{\text{Phase I}\} + P\{\text{Data}|\text{Phase II}\}P\{\text{Phase II}\} \\ &= \frac{1}{2}\theta(1-\theta)^4 + \frac{1}{2}\theta^4(1-\theta)\,. \end{aligned}$$

The MLE for θ no longer has a closed form solution in this case. Many numerical procedures can be employed to obtain the MLE, $\hat{\theta}$. Using the LINKAGE program,[28] we find $\theta = 0.21$ (approximately), with LOD score $Z(\hat{\theta}) = 0.1249$. Terwilliger and Ott[63] provided detailed descriptions and guidance about the LINKAGE program.

If the data consists of more than one family, then the joint likelihood function is the product of the likelihood functions of each family under the

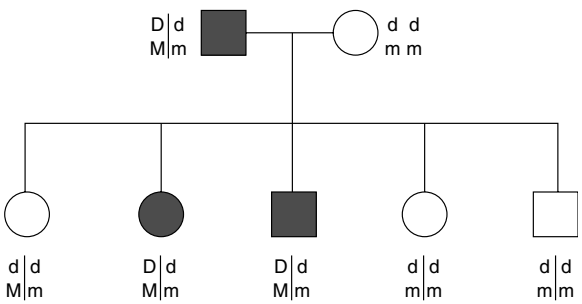


Fig. 4.

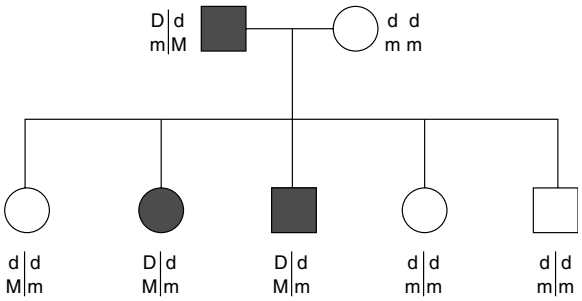


Fig. 5.

assumption that the families are independent of each other. The LOD score linkage analysis can be carried out similarly.

3.2. *The affected sib pair (ASP) method*

Sib pair linkage studies are commonly used for the investigation of genetic components involved in complex traits because a sib pair is the simplest family unit and easy to ascertain. Penrose[42,43] initiated the method based on the idea that sib pairs with similar phenotypes should have an excess of allele sharing while sib pairs with dissimilar phenotypes should have a deficit of allele sharing. The method was further developed to give rise to the affected sib pair method (ASP). The ASP method ascertains sib pairs where both sibs are affected. Hence, they should have an excess of allele sharing. One measurement of allele sharing is the number of alleles shared identical-by-descent (IBD). Two alleles that were transmitted from a common ancestor are referred to as being IBD. For example, suppose that the parental mating type is $Aa \times aa$ and the genotypes of both sibs are Aa, then allele "A" of the sib pair is IBD, but allele "a" of the sib pair may or may not be IBD.

Let I be a random variable denoting the number of alleles shared IBD by a sib pair. It can be shown that the distribution of I for a sib pair is:

$$P\{I = 0\} = \frac{1}{4}, \quad P\{I = 1\} = \frac{1}{2}, \quad P\{I = 2\} = \frac{1}{4}.$$

Hence,

$$E(I) = 1, \quad \text{var}(I) = \frac{1}{2}.$$

The conditional distribution of I, given the disease status of the members of a pair, provides the theoretical basis for the ASP method. It

was derived by Suarez *et al.*[61] under the one locus with two alleles assumption. Risch[48] generalized these results to other relatives and to multilocus models based on the recurrence risks of the disease. We introduce some notation and concepts before we describe the ASP methods.

Assume that there are two alleles, T and t, at the trait locus with probabilities $P\{T\} = p$ and $P\{t\} = q$. Let Y be a binary random variable indicating whether or not an individual is affected by a given disorder, i.e.

$$Y = \begin{cases} 1\,, & \text{Affected} \\ 0\,, & \text{Unaffected}\,. \end{cases}$$

The three penetrances are denoted as

$$f_1 = P\{Y = 1|TT\} = P\{\text{Affected}|TT\}\,,$$

$$f_2 = P\{Y = 1|Tt\}\,,$$

$$f_3 = P\{Y = 1|tt\}\,.$$

The prevalence rate of the disorder in the population is derived under the Hardy–Weinberg equilibrium as

$$K_P = P\{Y = 1\} = P\{Y = 1|TT\}P\{TT\} + P\{Y = 1|Tt\}P\{Tt\}$$

$$+ P\{Y = 1|tt\}P\{tt\} = p^2 f_1 + 2pqf_2 + q^2 f_3\,.$$

The genetic variance for the binary trait is

$$V_G = \text{var}(E(Y|G)) = V_A + V_D\,,$$

where G is a random variable representing the three genotypes, $V_A = 2pq[p(f_2 - f_1) + q(f_3 - f_2)]^2$ is the additive variance, and $V_D = p^2q^2(f_1 - 2f_2 + f_3)^2$ is the dominance variance. The additive model corresponds to $f_2 = \frac{f_1+f_3}{2}$; the dominance model corresponds to $f_2 = f_1$; and the recessive model corresponds to $f_2 = f_3$.

Let I_M be the number of alleles shared IBD at the marker locus, X be the random variable denoting the number of affected sibs in a sib pair, and θ be the recombination fraction between the trait and marker loci. Suarez *et al.*[61] derived the distribution $P\{I_M = j|X = k\}, j = 0, 1, 2, k = 0, 1, 2$, which is given in Table 2.

Given a set of parameters (K_P, V_A, V_D, θ), the deviation of the conditional distribution of I_M given $X = 2$, under the alternative hypothesis of linkage, from the expected distribution $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, under the null hypothesis of no linkage, is the largest among the three cases $(X = 2, X = 1, X = 0)$.

Table 2. Distribution $P\{I_M = j|X = k\}$ by Suarez *et al.*[61]

	$j=2$	$j=1$	$j=0$
$k=2$	$\frac{1}{4}+\frac{(\Psi-\frac{1}{2})V_A+(\Psi^2-\frac{1}{4})V_D}{d_2}$	$\frac{1}{2}-\frac{2(\Psi^2-\Psi+\frac{1}{4})V_D}{d_2}$	$\frac{1}{4}-\frac{(\Psi-\frac{1}{2})V_A+(2\Psi-\Psi^2-\frac{3}{4})V_D}{d_2}$
$k=1$	$\frac{1}{4}-\frac{(2\Psi-1)V_A+(2\Psi^2-\frac{1}{2})V_D}{d_1}$	$\frac{1}{2}+\frac{2(2\Psi^2-2\Psi+\frac{1}{2})V_D}{d_1}$	$\frac{1}{4}+\frac{(2\Psi-1)V_A+(4\Psi-2\Psi^2-\frac{3}{2})V_D}{d_1}$
$k=0$	$\frac{1}{4}+\frac{(\Psi-\frac{1}{2})V_A+(\Psi^2-\frac{1}{4})V_D}{d_0}$	$\frac{1}{2}-\frac{2(\Psi^2-\Psi+\frac{1}{4})V_D}{d_0}$	$\frac{1}{4}-\frac{(\Psi-\frac{1}{2})V_A+(2\Psi-\Psi^2-\frac{3}{4})V_D}{d_0}$

where $d_2 = 4(K_P^2 + V_A/2 + V_D/4)$, $d_1 = 4(2K_P - V_A - V_D/2 - 2K_P^2)$, $d_0 = 4(1 - 2K_P + K_P^2 + V_A/2 + V_D/4)$, and $\Psi = \theta^2 + (1-\theta)^2$.

Hence, the ASP design will provide more information for detecting linkage. There are many test statistics for detecting linkage based on the ASP design. The most popular one is the "mean test." It is defined as

$$T = \frac{(n_2 + \frac{1}{2}n_1) - \frac{n}{2}}{\sqrt{\frac{n}{8}}},$$

where n is the total number of affected sib pairs, n_2 is the number of affected sib pairs with $I_M = 2$, and n_1 is the number of affected sib pairs with $I_M = 1$. Under the null hypothesis of no linkage, T has an asymptotically standard normal distribution. The power and sample size calculations can be carried out with given values of (K_P, V_A, V_D, θ) by using the formulas in the above table. The power of the mean test has been investigated by Knapp *et al.*,[24] Suarez and Eerdewegh,[62] and Blackwelder and Elston.[3] It performs adequately under various conditions.

Risch[48] formulated the ASP method based on recurrence risk. Recurrence risk of a sib pair is defined as the conditional probability that one sib is affected given that the other sib is affected, i.e. $K_R = P\{Y_2 = 1|Y_1 = 1\}$, where Y_1 and Y_2 are random variables indicating affection status of the two sibs. The recurrence risk ratio of a sib pair is the ratio of the sibling recurrence risk relative to the population prevalence rate, i.e. $\lambda_S = \frac{K_R}{K_P}$. Similarly, the recurrence risk ratio between parent and offspring can be defined by replacing the sibling recurrence risk with the parent-offspring recurrence risk, denoted by λ_O. The conditional distribution of I_M given that two sibs are affected is trinomial with probabilities:

$$z_0 = \frac{1}{4} - \frac{1}{4\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2(1 - \Psi)(\lambda_S - \lambda_O)],$$

$$z_1 = \frac{1}{2} - \frac{1}{2\lambda_S}(2\Psi - 1)^2(\lambda_S - \lambda_O),$$

$$z_2 = \frac{1}{4} + \frac{1}{4\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2\Psi(\lambda_S - \lambda_O)],$$

where $z_i = P\{I_M = i|\text{Two sibs affected}\}$ and $\Psi = \theta^2 + (1 - \theta)^2$. This parameterization is different from that of Suarez *et al.*[61] The mean test statistic for $H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta < \frac{1}{2}$ is the same as before. The only difference is that the power and sample size calculations are in terms of the parameters $(\lambda_S, \lambda_O, \theta)$. A likelihood ratio test can be performed based on the trinomial distribution with parameters $(\lambda_S, \lambda_O, \theta)$. The affected sib pair method has been generalized to affected relative pairs[48] and to affected relative-sets or affected-pedigree-member (APM) methods.[66]

3.3. *The Haseman–Elston procedure*

The LOD score and ASP methods described above handle the linkage analysis of qualitative traits. However, many complex traits are quantitative, i.e. those measured on a continuous scale, instead of a discrete scale. Having a complex disease is often determined by applying a threshold to a quantitative measurement, such as hypertension defined by blood pressure or obesity defined by body mass index. However, it is important to develop statistical methods to study linkage between a marker locus and a locus underlying a quantitative trait.

Let x_{1j} and x_{2j} denote the continuous trait values for two sibs in a sib pair, respectively. We assume the trait values have the following structure:

$$x_{1j} = \mu + g_{1j} + e_{1j}\,,$$

$$x_{2j} = \mu + g_{2j} + e_{2j}\,,$$

where μ is the overall mean, g_{1j} and g_{2j} represent genetic contributions to the trait values, and e_{1j} and e_{2j} are residuals.

We consider a single locus with two alleles A_1 and A_2. The allele frequencies of A_1 and A_2 are p and $q = 1 - p$, respectively. The mean trait values of individuals, with the three possible genotypes, are defined as follows:

A_2A_2	A_2A_1	A_1A_1
$-a$	d	a

Then, the additive genetic variance is $\sigma_a^2 = 2pq[a + (q - p)d]^2$, and the dominance variance is $\sigma_d^2 = (2pqd)^2$. The total genetic variance is the sum of the additive and dominance genetic variances, that is, $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$. Let $\sigma_e^2 = E(e_{1j} - e_{2j})^2$ be the residual variance for each genotype and ρ be the residual correlation coefficient between the two sibs. Throughout this section, we assume that there is no dominance, i.e. $\sigma_d^2 = 0$.

Haseman and Elston[21] showed that the expectation of the squared difference in the observed trait values of a sib pair, conditional on the proportion of alleles shared IBD at the trait locus, satisfies the regression equation

$$E(y_j|\pi_j) = \sigma_e^2 + 2\sigma_g^2 - 2\sigma_g^2\pi_j = \beta_0 + \beta_1\pi_j\,, \tag{8}$$

where $y_j = (x_{1j} - x_{2j})^2$, $\beta_0 = \sigma_e^2 + 2\sigma_g^2$ and $\beta_1 = -2\sigma_g^2$.

With the candidate gene approach, we collect sib pair data as $(y_1, \pi_1), \ldots, (y_n, \pi_n)$ and perform regression analysis to obtain estimates $\hat{\sigma}_e^2$

and $\hat{\sigma}_g^2$. Then, the estimate of heritability is obtained from:

$$H = \frac{\sigma_g^2}{\sigma_e^2 + \sigma_g^2}$$

by substituting $\hat{\sigma}_e^2$ and $\hat{\sigma}_g^2$. Such an approach has been extended to multiple trait loci with multiple alleles.[60]

When the proportion of alleles shared IBD at the marker locus, π_{jm}, instead of trait locus, is available, the regression equation becomes[21]

$$E(y_j|\pi_{jm}) = \sigma_e^2 + 2\sigma_g^2\Psi - 2(1-2\theta)^2\sigma_g^2\pi_{jm} = \gamma_0 + \gamma_1\pi_{jm}\,, \tag{9}$$

where $y_j = (x_{1j} - x_{2j})^2$, $\gamma_0 = \sigma_e^2 + 2\sigma_g^2\Psi$, $\Psi = \theta^2 + (1-\theta)^2$, $\gamma_1 = -2(1-2\theta)^2\sigma_g^2$, and θ is the recombination fraction between the trait and marker loci. Recall that $\theta \in [0, \frac{1}{2}]$, and π_{jm}, the proportion of marker alleles shared IBD for the jth sib pair, assumes values $0, \frac{1}{2}, 1$. The proportion of marker alleles shared IBD is often estimated. The regression equation in terms of the estimated proportion $\hat{\pi}_{jm}$ becomes

$$E(y_j|\hat{\pi}_{jm}) = \gamma_0 + \gamma_1\hat{\pi}_{jm}\,, \tag{10}$$

where γ_0 and γ_1 are the same as those in (9), and $\hat{\pi}_{jm}$ takes values $\frac{i}{4}$, $i = 0, 1, 2, 3, 4$.[21] Notice that, $\gamma_1 = 0$ implies $\theta = \frac{1}{2}$ when $\sigma_g^2 > 0$. The least squares estimate $\hat{\gamma}_1$ is obtained from sib pair genetic data $(y_1, \hat{\pi}_{1m}), \ldots, (y_n, \hat{\pi}_{nm})$ by performing regression analysis based on formula (10). The least squares estimate $\hat{\gamma}_1$ can be used to test $H_0 : \theta = \frac{1}{2}$ (No linkage) vs. $H_1 : \theta < \frac{1}{2}$ (Linkage). A significantly negative $\hat{\gamma}_1$ indicates $\theta < \frac{1}{2}$. Hence, reject H_0 if $\hat{\gamma}_1$ is less than the appropriate $C < 0$ at a level α test.

3.4. *The ED and EC sib pair design*

Haseman–Elston[21] model is based on randomly sampled sib pairs. Certain sampling schemes that select sib pairs based on their trait values have greater power.[3,6,10,49] Risch and Zhang[49] concluded that three types of sib pairs, selected on the basis of trait values, provide the most power to detect linkage for a quantitative trait locus (QTL): (1) extremely discordant (ED) sib pairs where one has a high and the other a low trait value; (2) extremely concordant (EC) for high trait values; (3) extremely concordant for low trait values.[49,50,69,70] They investigated the power of these three sib pair designs under different genetic models and concluded that the extremely discordant sib pair design has the greatest power. Hence their recommendation is that the extremely discordant sib pair design be used for linkage studies of QTLs

in humans. Eaves and Meyer[10] also obtained the power of ED sibpairs by simulation.

Suppose that N ED sib pairs are selected for genotyping in a linkage study. Let n_0, n_1, n_2 be the number of sib pairs with $IBD = 0, 1, 2$, respectively. If the marker locus is linked to the trait locus, more ED sib pairs should have $IBD = 0$ due to the selection process. Hence, if n_0 is significantly larger than n_2, then linkage is indicated. Under H_0 (no linkage) (n_0, n_1, n_2) has a trinomial distribution with parameters $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. A test statistic can be based on $n_0 - n_2$. We have

$$E_{H_0}(n_0 - n_2) = 0\,, \quad \mathrm{var}_{H_0}(n_0 - n_2) = \frac{N}{2}\,,$$

where $N = n_0 + n_1 + n_2$.

Define a test statistic

$$T_{\mathrm{ED}} = \frac{n_0 - n_2}{\sqrt{\frac{N}{2}}}\,.$$

This statistic has an asymptotically standard normal distribution. Reject H_0 and declare linkage when T_{ED} is large. Sample size and power formulas are given in Risch and Zhang.[49] For the EC sib pair design, the test statistic is

$$T_{\mathrm{EC}} = \frac{n_2 - n_0}{\sqrt{\frac{N}{2}}}\,.$$

Several existing procedures that combine ED and EC sib pairs into one test give each ED (EC) pair the same weight.[16,30] Rao[46] noticed that these methods can be improved by exploiting the quantitative variability in the tail distribution of the trait. Li and Gastwirth[31] developed a test giving greater weight to the more discordant (concordant) ED (EC) pairs.

3.5. *The transmission/disequilibrium test (TDT)*

Assume that there are two alleles, D_1 and D_2, at a disease locus, and two marker alleles, M_1 and M_2, at a marker locus. Suppose that n affected children are ascertained. From these families, there will be $4n$ parental marker alleles, $2n$ of which are transmitted and $2n$ of which are not transmitted. If the marker locus is in the neighborhood of the disease locus and the disease allele is due to a recent mutation, then a specific marker allele associated with the disease allele will have higher frequency among diseased individuals compared to normal individuals. The imbalanced transmission of one allele relative to the other suggests that linkage exists between the marker and disease loci.

Table 3. Numbers a, b, c, and d of the transmitted and non-transmitted marker alleles M_1 and M_2 among $2n$ parents of n affected children.[59]

Transmitted allele	Non-transmitted allele		
	M_1	M_2	Total
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

Spielman *et al.*,[59] summarizes the number of alleles transmitted and not transmitted to the n affected children of the $2n$ parents, as presented in Table 3.

Notice that entry b in the above 2×2 table represents the number of parents who are M_1M_2 at the marker locus with one transmitted allele, M_1, and one non-transmitted allele, M_2. Since each parent of an affected offspring contributes exactly one transmitted and one non-transmitted allele in the above 2×2 table, the transmission/disequilibrium test (TDT) proposed by Spielman *et al.*[57,59] is the McNemar test resulting from a matched case-control design. The one degree of freedom χ^2 test statistic is

$$\chi^2_{TD} = \frac{(b-c)^2}{b+c}\,.$$

The McNemar test is based on the standard normal approximation to a binomial distribution.

The theoretical background for the TDT is given in Table 4, from Curnow *et al.*[7]

In Table 4, $m = P(M_1)$ and $p = P(D_1)$ are the allele frequencies, θ is the recombination fraction between the marker and trait loci, $\delta = P(M_1D_1) - P(M_1)P(D_1)$ is the linkage disequilibrium (association) parameter, and

$$B = \frac{p[p(f_{11}-f_{12})+(1-p)(f_{12}-f_{22})]}{p^2 f_{11} + 2p(1-p)f_{12} + (1-p)^2 f_{22}}\,,$$

where

$$f_{11} = P(\text{Affected}|D_1D_1)\,,$$

$$f_{12} = P(\text{Affected}|D_1D_2)\,,$$

$$f_{22} = P(\text{Affected}|D_2D_2)$$

Table 4. Probabilities of combination of transmitted and non-transmitted marker alleles M_1 and M_2 among $2n$ parents of n affected children.[7]

Transmitted allele	Non-transmitted allele	
	M_1	M_2
M_1	$m^2 + \frac{Bm\delta}{p}$	$m(1-m) + \frac{B(1-\theta-m)\delta}{p}$
M_2	$m(1-m) + \frac{B(\theta-m)\delta}{p}$	$(1-m)^2 - \frac{B(1-m)\delta}{p}$

are the three penetrances, i.e. the probabilities that individuals with disease genotype D_1D_1, D_1D_2, and D_2D_2 have the disease. The entry in the upper-right of Table 4, $m(1-m) + \frac{B(1-\theta-m)\delta}{p}$, is the conditional probability that a parent with marker genotype M_1M_2 transmits allele M_1 given that the child is affected, that is

$$\begin{aligned} p_{12} &= P\{\text{Parent} = M_1M_2 \to M_1 | \text{Child Affected}\} \\ &= m(1-m) + \frac{B(1-\theta-m)\delta}{p}\,. \end{aligned}$$

Similarly, in the lower-left of Table 4,

$$\begin{aligned} p_{21} &= P\{\text{Parent} = M_1M_2 \to M_2 | \text{Child Affected}\} \\ &= m(1-m) + \frac{B(\theta-m)\delta}{p}\,. \end{aligned}$$

If $\delta \neq 0$ (association), then $H_0 : \theta = \frac{1}{2}$ is equivalent to $H_0 : p_{12} = p_{21}$. Hence, the TDT is a joint test for linkage and association. The diagonal terms in Table 4 are independent of the recombination fraction θ. That is, as expected, homozygous parents have no information about linkage which is the reason why the TDT statistic χ^2_{TD} only involves b and c and not a or d in Table 3.

The TDT is related to the concept of haplotype relative risk (HRR).[14,40] The TDT has the advantage that it only requires parent and child data from families with one affected child. It does not require multiple sibs such as ASP. The disadvantage of the TDT is that it can detect linkage only if association is present. The TDT has been generalized to multiallelic markers,[2,51,55,57] to families without parental genotype information,[4,8,33,52,58] and to quantitative traits.[1,45,53]

4. Discussion

The statistical methods discussed in this chapter are only a small selection from problems arsing from the molecular data that are becoming available to search for disease genes. The properties of currently used statistical methods still need to be investigated, and new statistical methods for genome-wide inference need to be developed. For an excellent review on major contribution of statistics to genetics over the last century, and current and future research problems, readers are referred to Elston and Thompson.[12]

Acknowledgments

This was partly supported by NIH grants CA 64363 and EY 14478 and China Nature Science grant. We would like to thank Professor Chin Long Chiang and Mr. Dennis Buckman for their excellent comments.

References

1. Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *American Journal of Human Genetics* **60**: 676–690.
2. Bickeböller, H. and Clerget-Darpoux, F. (1995). Statistical properties of the allelic and genotypic transmission/disequilibrium test for multi-allelic markers. *Genetic Epidemiology* **12**: 865–870.
3. Blackwelder, W. C. and Elston, R. C. (1982). Power and robustness of sib-pair linkage test and extension to larger sibships. *Communication Statistical Theory Methods* **11**: 449–484.
4. Boehnke, M. and Langefeld, C. D. (1998). Genetic association mapping based on discordant sib pairs: The discordant-alleles test. *American Journal of Human Genetics* **62**: 950–961.
5. Bonney, G. E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: Regressive models. *American Journal of Medical Genetics* **18**: 731–749.
6. Carey, G. and Williamson, J. (1991). Linkage analysis of quantitative traits: Increased power by using selected samples. *American Journal of Human Genetics* **49**: 786–796.
7. Curnow, R. N., Morris, A. P. and Whittaker, J. C. (1998). Locating genes involved in human disease. *Applied Statistics* **47**: 63–76.
8. Curtis, D. (1997). Use of siblings as controls in case-control association studies. *Annals of Human Genetics* **61**: 319–333.
9. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* **B39**: 1–38.
10. Eaves, L. and Meyer, J. (1994). Locating human quantitative trait loci: Guidelines for the selection of sibling pairs for genotyping. *Behaviour Genetics* **24**: 443–455.

11. Elston, R. C., Bailey-Watson, J. E., Bonney, G. E. *et al.* (1986). A package of computer programs to perform statistical analysis for genetic epidemiology. Presented at the *7th International Congress of Human Genetics*, Berlin.
12. Elston, R. C. and Thompson, E. A. (2000). A century of biometrical genetics. *Biometrics* **56**: 659–666.
13. Falconer, D. S. (1989). *Introduction to Quantitative Genetics*, Longman Scientific and Technical with John Wiley and Sons, Inc. New York.
14. Falk, C. T. and Rubinstein, P. (1987) Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* **51**: 227–233.
15. Fisher, R. A. (1952). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* **6**: 13–25.
16. Gu, C., Todorov, A. and Rao, D. C. (1996). Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of QTLs. *Genetic Epidemiology* **13**: 513–533.
17. Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics* **8**: 299–309.
18. Haldane, J. B. S. and Smith, C. A. B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* **14**: 10–31.
19. Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science* **28**: 49–50.
20. Hartl, D. L. and Clark, A. G. (1997). *Principles of Population Genetics*, Sinauer Associates, Inc. Sunderland, Massachusetts.
21. Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behaviour Genetics* **2**: 3–19.
22. Hasstedt, S. J. and Carwright, P. E. (1981). PAP-pedigree analysis package, University of Utah, Department of Medical Biophysics and Computing, Technical Report No. 13. Salt Lake City, Utah.
23. Khoury, M. J., Beaty, T. H. and Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology*, Oxford University Press, New York and Oxford.
24. Knapp, M., Wassmer, G. and Baur, M. P. (1995). Linkage analysis in nuclear families, I. Optimality criteria for affected sib-pair tests. *Human Heredity* **44**: 37–43.
25. Kosambi, D. D. (1994). The estimation of map distances from recombination values. *Annals Eugenics* **12**: 172–175.
26. Lalouel, J. M. and Yee, S. (1980). POINTER: A computer program for complex segregation analysis with pointers. Technical Report, Population Genetics Laboratory, University of Hawaii, Honolulu.
27. Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*, Springer-Verlag, New York.
28. Lathrop, G. M., Lalouel, J. M., Juier, C. *et al.* (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of National Academic Sciences USA* **81**: 3443–3446.

29. Li, C. C. (1988). *First Course in Population Genetics*, The Boxwood Press, Pacific Grove, California.
30. Li, Z. and Zhang, H. (2000). Mapping quantitative trait loci in humans using both extreme discordant and concordant sib pairs: A unified approach for meta-analysis. *Communication Statistical Theory Methods* **29**: 1115–1127.
31. Li, Z. and Gastwirth, J. L. (2001). A weighted test using both extreme discordant and concordant sibpairs for detecting linkage. *Genetic Epidemiology* **20**: 34–43.
32. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.
33. Monks, S. A., Kaplan, N. L. and Weir, B. S. (1998). A comparative study of sibship tests of linkage and/or association. *American Journal of Human Genetics* **63**: 1507–1516.
34. Morgan, T. H. (1928). *The Theory of Genesics*, Yale University Press, New Haven, Conn.
35. Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7**: 277–318.
36. Morton, N. E. (1959). Genetic tests under incomplete ascertainment. *American Journal of Human Genetics* **11**: 1–16.
37. Morton, N. E. and MacLean, C. J. (1974). Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. *American Journal of Human Genetics* **26**: 489–503.
38. Olson, J. M., Witte, J. S. and Elston, R. C. (1999). Tutorial in biostatistics: Genetic mapping of complex traits. *Statistics in Medicine* **18**: 2961–2981.
39. Ott, J. (1977). Counting methods (EM algorithm) in human pedigree analysis: Linkage and segregation analysis. *Annals of Human Genetics* **40**: 443–454.
40. Ott, J. (1989). Statistical properties of the haplotype relative risk. *Genetic Epidemiology* **6**: 127–130.
41. Ott, J. (1999). *Analysis of Human Genetic Linkage*, The Johns Hopkins University Press, Baltimore and London.
42. Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs brothers and sisters of unspecified parentage. *Annals of Eugenics* **6**: 133–138.
43. Penrose, L. S. (1953). The general purpose sib-pair linkage test. *Annals of Eugenics* **18**: 120–124.
44. Province, M. A. and Rao, D. C. (1995). General purpose model and a computer program for combined segregation and path analysis (SEGPATH): Automatically creating computer programs from symbolic language model specifications. *Genetic Epidemiology* **12**: 203–219.
45. Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Human Heredity* **47**: 342–350.
46. Rao, D. C. (1998). CAT scans, PET scans, and genomic scans. *Genetic Epidemiology* **15**: 1–18.

47. Rao, D. C., Keats, B. J. B., Lalouel, J. M., Morton, N. E. and Yee, S. (1979). A maximum likelihood map of chromosome 1. *American Journal of Human Genetics* **31**: 680–696.
48. Risch, N. (1990). Linkage strategies for genetically complex traits II. The power of affected relative pairs. *American Journal of Human Genetics* **46**: 229–241.
49. Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**: 1584–1589.
50. Risch, N. and Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: Sample size considerations. *American Journal of Human Genetics* **58**: 836–843.
51. Schaid, D. J. (1996). General score tests for associations of genetic markers with disease using cases and parents. *Genetic Epidemiology* **13**: 423–449.
52. Schaid, D. J. and Rowland, C. (1998). The use of parents, sibs, and unrelated controls to detection of association between genetic markers and diseases. *American Journal of Human Genetics* **63**: 1492–1506.
53. Schaid, D. J. and Rowland, C. M. (1999). Quantitative trait transmission disequilibrium test: Allowance for missing parents. *Genetic Epidemiology* **17**: S307–S312.
54. Sham, P. (1998). *Statistics in Human Genetics*, Arnold, London and New York.
55. Sham, P. C. and Curtis, D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics* **59**: 323–336.
56. Smith, C. A. B. (1957). Counting methods in genetical statistics. *Annals of Human Genetics* **21**: 254–276.
57. Spielman, R. S. and Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59**: 983–989.
58. Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**: 450–458.
59. Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**: 506–516.
60. Stoesz, M. R., Cohen, J. C., Mooser, V., Marcovina, S. and Guerra, R. (1997). Extension of the Haseman–Elston method to multiple alleles and multiple loci: Theory and practice for candidate genes. *Annals of Human Genetics* **61**: 263–274.
61. Suarez, B. K., Rice, J. and Reich, T. (1978). The generalized sib pair IBD distribution: Its use in the detection of linkage. *Annals of Human Genetics* **42**: 87–94.
62. Suarez, B. K. and van Eerdewegh, P. (1984). A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *American Journal of Medical Genetics* **18**: 135–146.

63. Terwilliger, J. D. and Ott, J. (1994). *Handbook of Human Genetic Linkage*, The John Hopkins University Press, Baltimore and London.
64. Thompson, E. A. (1986). Genetic epidemiology: A review of the statistical basis. *Statistics in Medicine* **5**: 291–302.
65. Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A. and Weiner, A. M. (1987). *Molecular biology of the Gene*, The Benjamin/Cummings Publishing Company, Inc., Menlo Park, California.
66. Weeks, D. E. and Lange, K. (1988). The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics* **42**: 315–326.
67. Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. Jahresh. Verein f. vaterl. Naturk. in *Württemberg* **64**: 368–382.
68. Wentworth, E. N. and Remick, B. L. (1916). Some breeding properties of the generalized Mendelian population. *Genetics* **1**: 608–616.
69. Zhang, H. and Risch, N. (1996). Mapping quantitative trait loci in humans using extreme concordant sib pair: Selected sampling by parental phenotypes. *American Journal of Human Genetics* **59**: 951–957
70. Zhao, H., Zhang, H. and Rotter, J. I. (1997). Cost-effective sib-pair designs in the mapping of quantitative-trait loci. *American Journal of Human Genetics* **60**: 1211–1221.

About the Author

Zhaohai Li is an Associate Professor of Statistics and Biostatistics at the Biostatistics Center, Statistics Department, George Washington University, Washington, DC. He obtained, his MS degree in Mathematical Statistics from Center China Normal University and PhD in Statistics from Columbia University. His research interests include statistical methods for genetic epidemiology and empirical Bayes methods for meta-analysis.

This page intentionally left blank

CHAPTER 16

DOSE-RESPONSE MODELING IN HEALTH RISK ASSESSMENT

YILIANG ZHU

Department of Epidemiology and Biostatistics,
College of Public Health, University of South Florida,
13201 Bruce B Downs Blvd, Tampa, FL 33612-2805, USA
Tel: (813)974-6674; yzhu@hsc.usf.edu

Human health protection against environmental exposure to chemical hazards is of fundamental public health importance, and thus of intensive research interests among industries, governments, and academics. As the scientific basis of health protection, risk assessment employs a wide range of statistical tools. Among them, dose-response modeling plays a central role in assessing exposure-related health risk and deriving safety exposure levels for environmental regulation. This article illustrates the use of dose-response modeling in risk assessment through examples of carcinogenicity, developmental toxicity, and neurotoxicity. Data from these examples include binary, clustered categorical, and longitudinal measurements, and require careful consideration for effective and innovative use of statistical methods such as generalized estimating equations and nonlinear mixed effects models. We also discuss the problem of benchmark dose estimation and some open statistical issues encountered in risk assessment. Whereas the examples and applications are directly related to environmental health, the methods illustrated in this article are widely applicable to many problems in medicine and biological studies.

1. Introduction

There are sufficient scientific evidences that link chemical exposure to various adverse health effects, including carcinogenicity, developmental toxicity, mutangenicity, immunotoxicity and neurotoxicity.[33] The United States Environmental Protection Agency's (US EPA) TSCA inventory currently registers more than 65,000 chemical substances that are in active use in the USA, and this number is increasing. However, only about 10% of

these substances have been tested for carcinogenicity, and an even smaller number have been tested for non-cancer effects. For example, it is estimated that about 3% to 28% of all chemicals are neurotoxicants.[33] A large number of the more than 500 registered pesticide ingredients are estimated to affect the nervous system of the targeted species to varying degrees. Of the 588 chemicals listed by the American Conference of Governmental Industrial Hygienists, 167 affected the nervous system or behavior at some exposure level.[1] It is further estimated that of the approximately 200 chemicals to which one million or more workers are exposed in the USA, more than one third may have adverse effects on the nervous system if sufficient exposure occurs.[2] Thus, there are increasing scientific and regulatory interests in estimating the risk of exposure to chemicals of various toxic potentials with regard to their overall impact on human health.

1.1. *Health risk assessment*

Health risk assessment is composed of some or all of the following components: hazard identification; dose-response assessment; exposure assessment; and risk characterization. Hazard identification involves the detection of exposure-induced adverse health effects, with respect to dose, route, timing and duration of exposure, through either case report or designed human/animal studies. Dose-response analysis is typically done through designed animal experiments because controlled exposure to humans is generally not feasible. The purpose of dose-response assessment is to determine an exposure range in which exposure-induced risk is of a controllable magnitude. This estimated range of exposure in turn provides a quantitative base for risk characterization. The third component, exposure assessment, is to identify exposed or potentially exposed population, describe its size and composition, and present the types, magnitudes, frequencies, and duration of the exposure. As an integration of hazard identification, dose-response assessment and exposure assessment, a statement of the consequence of exposure then comes as a result of risk characterization.

In the past decade or so, dose-response assessment in the context of non-cancer risk assessment has largely focused on determining a no-observed-adverse-effect-level (NOAEL) or lowest-observed-adverse-effect-level (LOAEL), largely ignoring the shape of the underlying dose-response relationship. A NOAEL is the highest experimental dose at which the increase in adverse effects relative to the control group is not significant. A LOAEL, on the other hand, is the lowest experimental dose at which

there is a significant increase in risk. A reference dose (RfD) or reference concentration (RfC) is then calculated by dividing a NOAEL, or a LOAEL when a NOAEL is not determined, by uncertainty factors to account for interspecies difference in response, different exposure routes and other study variations.[3] The RfD is an estimate, with uncertainty of an order of magnitude, of a daily exposure to the human population that is likely to be without appreciable risks of deleterious health effects during a lifetime.[3] It provides a quantitative basis for setting up regulatory levels of the chemical under testing.

Since a NOAEL/LOAEL is limited to only the experimental doses, its determination ignores the shape of the dose-response relationship, and the risk associated with the NOAEL/LOAEL varies substantially between experiments.[12,21,24] Recognizing the inconsistency in the NOAEL/LOAEL approach, the US EPA and several other international agencies have recommended the benchmark dose method (BMD)[12] as an alternative or supplementary approach in risk assessment of developmental toxicity as well as other non-cancer effects.[4,39–41] It is expected that, through sound statistical methods, the BMD approach can overcome the weaknesses of the NOAEL/LOAEL approach and provide a more consistent quantification of risk.

1.2. *Benchmark Doses*

Befaore we describe the benchmark dose method, it helps to first discuss measures of risk. Consider an outcome measure Y of health effects. Assume that in the absence of exposure, Y has a cumulative distribution function $F_0(y) = Pr(Y < y)$. Chemical exposure at level d may cause adverse effects to the exposed population, resulting in an altered distribution $F_d(y) = Pr(Y < y|d)$ relative to that of the control population, $F_0(y) = Pr(Y < y|d = 0)$. The alteration may occur in many ways, including such common phenomenon as mean shifting and change in variance. Characterization of a dose-response relationship in risk is essentially to characterize any alterations in the distribution as a function of exposure. Statistically, the problem often reduces to identifying a quantity that reflects the changes to $F_0(y)$. We can consider, for example, a set A of particular value of Y, and use the probability $\pi(d = 0) = Pr(Y \in A|d = 0)$ as a reference level for the changes in $\pi(d) = Pr(Y \in A|d)$. Significant changes from $\pi(0)$ to $\pi(d)$ signify a risk as well as a dose-response relationship. However, the probability $\pi(d)$ may not be a direct measure of risk

since A is not necessarily a region of adverse values. Since this approach assesses overall risk in terms of changes to the population, not necessarily identifying adverse effects in individuals, we call it population-level risk characterization.

Although the value of Y may not always lead to a definitive diagnosis of adverse effects especially when Y is a continuous measure, special cases do arise where A is a region of adversity on grounds of toxicological criteria. These include the case of Y being a binary or dichotomized measure, so that $Y = 1$ indicates the presence of adverse effects. As a result $A = \{1\}$, and $\pi(d)$ is a direct measure of risk. Under these circumstances, it is feasible to assess adverse effects on an individual basis. We call this approach individual-level risk characterization.

While the principle of population-level risk assessment is clear, the possibility of distributional changes to the population is almost infinite. Let us consider the special case of the location-scale family of distributions where $F_d(y) = F((y-\mu_d)/\sigma_d)$ with mean μ_d and scale σ_d. Obviously, normal distributions are members of this family. For convenience, we make $\mu_0 = 0$, and $\sigma_0 = 1$ for the control distribution $F(y)$. If toxic effects are manifested as a mean shifting, and a shifting of c units is toxicologically or clinical meaningful, we can then choose $A = \{Y > c\}$. It follows that $\pi(0) = 1 - F(c)$, and $\pi(d) = 1 - F((c-\mu_d)/\sigma_d)$. The case of $A = \{Y < c\}$ can be implemented analogously. Here, c can be a percentage of the control mean, variance, or data range. If toxic effects are manifested as c units change of variance relative to the control variance, we can choose $A = \{|Y| > c\}$. It follows that $\pi(0) = 1 - F(c) + F(-c)$, and $\pi(d) = 1 - F(c/\sigma_d) + F(-c/\sigma_d)$. In the case of both a mean shifting and variance change, $A = \{|Y| > c\}$ remains applicable. If changes to $F_0(y)$ are such that $F_d(y)$ is no longer in the same family, it is still possible to quantify the changes using preferably non-parametric methods.[5]

Given the dose-response model $\pi(d) = Pr(Y \in A|d)$, we can choose a benchmark response (BMR) level γ, say 10%, and identify the corresponding level of exposure, i.e. benchmark dose (BMD) that induces the specified BMR. Specifically, the BMD can be defined by the multiplicative excess risk

$$\frac{\pi(\mathrm{BMD}_\gamma) - \pi(0)}{1 - \pi(0)} = \gamma \tag{1}$$

or alternatively the additive excess risk,

$$\pi(\mathrm{BMD}_\gamma) - \pi(0) = \gamma\,. \tag{2}$$

We then divide the BMD_γ by certain safety factors to obtain a RfD.

The estimation of BMDs involves a few steps. First, we determine a reference "risk" $\pi(0) = Pr(A|d = 0)$ by characterizing the control distribution with regard to certain aspects that are sensitive to potential changes caused by exposure. Second, we use the reference probability $\pi(d) = Pr(A|d)$ to establish a dose-response relationship. Third, we estimate BMDs based on the fitted dose-response models using Eqs. (1) or (2). Since BMD is a point estimate, a lower confidence limit is often computed as a more conservative measure. Gaylor *et al.*[16] give a detailed account of procedures for computing BMDs with different types of data. While the procedures are clear in principle, many technical challenges remain. Because of the complex nature of data arising from environmental studies, it is important to employ effective, and is challenging to develop innovative methods to handle statistical issues in risk assessment. This article mainly illustrates dose-response modeling using advanced statistical methods such as generalized estimating equations (GEEs) and mixed effects models. We discuss binary data for cancer risk assessment in Sec. 2, clustered multinomial data of developmental toxicity in Sec. 3, and longitudinal data from neurobehavioral toxicity screening studies in Sec. 4. We conclude with some open issues in Sec. 5.

2. Binary Data: Carcinogenicity

Binary outcome data indicate the presence or absence of adverse effects in each individual. In cancer studies this means that subjects can be classified as having or not having a tumor in a target organ at some specified time following exposure to a test substance. It is often assumed that the effects observed in a subject is independent of that observed in others. This assumption needs to be checked, and is not the case for example in reproductive and developmental experiments where littermates may respond more similarly than others from different litters. This situation is addressed in Sec. 3.

Table 1 is a summary of adenoma/carcinoma incidence in liver from an experiment on Japanese Medaka (Oryzias latipes) exposed to N-nitrosodiethylamine (DEN), a known carcinogen. This dataset is the 4th replication in a study reported by Brown-Patterson *et al.*[6] The primary purpose of this study is to test for non-linear dose-response relationships below the 1% BMR level. This is an issue of important implications in regulatory policy. Currently, EPA's testing protocol requires studies to include a lowest dose level near the 1% BMR level. A linear dose-response relationship is assumed from below the 1% BMR level, and linear extrapolation

Table 1. Frequency of adenoma/carcinoma in the liver of medaka exposed to DEN.

Dose (ppm)	0	0.075	0.15	0.3	0.6	1.5	3.0
Cases	5	6	11	12	19	27	41
%	0.36	0.43	0.77	0.86	1.36	4.20	6.48
Sample Size	1387	1385	1427	1400	1393	643	633

Table 2. Fitted probit model for the incidence of adenoma/carcinoma.

Coefficient	Estimate	Std. Err.	t-value
Intercept	−2.632	0.072	−36.752
DEN	0.831	0.148	5.625
DEN^2	−0.153	0.046	−3.333

towards 0 is used to derive a RfD at the 0.1% BMR level. (The terminology extrapolation in risk assessment context excludes the control as an exposure level.) As a result, extrapolation would lead to an over-estimation (or under-estimation) of risk if the true dose-response is curved upwards (or downwards).

From Table 1 we can see that the incidence of either adenoma or carcinoma or both is elevated by exposure to DEN and the 1% BMR is somewhere between 0.3 and 0.6 ppm. Since Medaka were housed in water tanks in group during both exposure and growing-out periods, similar living condition shared by Medaka in the same tank may result in clustering effects, i.e. data dependence. We checked the full study data by comparing the sample variance with the binomial variance, and did not find any indication of clustering. Therefore, we used a probit link function in conjunction with a binomial distribution, within the framework of generalized linear models,[28] to model the mean response probability of an adenoma or carcinoma. This led to a probit dose-response model $\pi(X\beta) = \Phi(X\beta)$, where Φ is the cumulative distribution function of the standard normal, X is a set of covariates including dose, and possibly other covariates or interaction terms, and β are a vector of the regression coefficients. A simple quadratic model of dose was fit to the data, yielding the results given in Table 2. The residual deviance for the model was 0.984 on 4 degrees of freedom, indicating a reasonable fit to the data. From this model, the 1% BMR level was estimated to be about 0.396 ppm. While the quadratic term is statistically significant, signaling some degree of non-linearity below the 1% BMR level, it is the impact of the non-linearity on risk quantification that we were

really interested in. To this end, we decided to estimate the curvature

$$|\Phi''|/(1+(\Phi')^2)^{3/2}$$

of the dose-response curve, and then explored the relationship between the curvature and risk.

To see the potential bias in risk estimation resulted from linear extrapolation, we compared the risk based on the fitted model and that based on linear extrapolation below the 1% BMR level. Assuming that the spontaneous risk of the control population is 0.0042 as predicted from the fitted model, the extrapolated risk was

$$\pi_{lin}(d) = 0.0145\dot{d} + 0.0042\,.$$

We plotted the relative bias $(\pi_{lin} - \Phi)/\Phi$ against the estimates of curvature of the fitted model (Fig. 1). We can see that the largest bias is about 6.6% of the estimated risk, occurring at $d = 0.17$ ppm with curvature $=$ 0.0212. At the 0.1% excessive risk level, the relative bias is about 5.5%, and the curvature there is 0.0202. These are clear evidences of a significant non-linearity in the true dose-response curve. However, the impact of non-linearity on risk estimation is of a small magnitude. On the other hand,

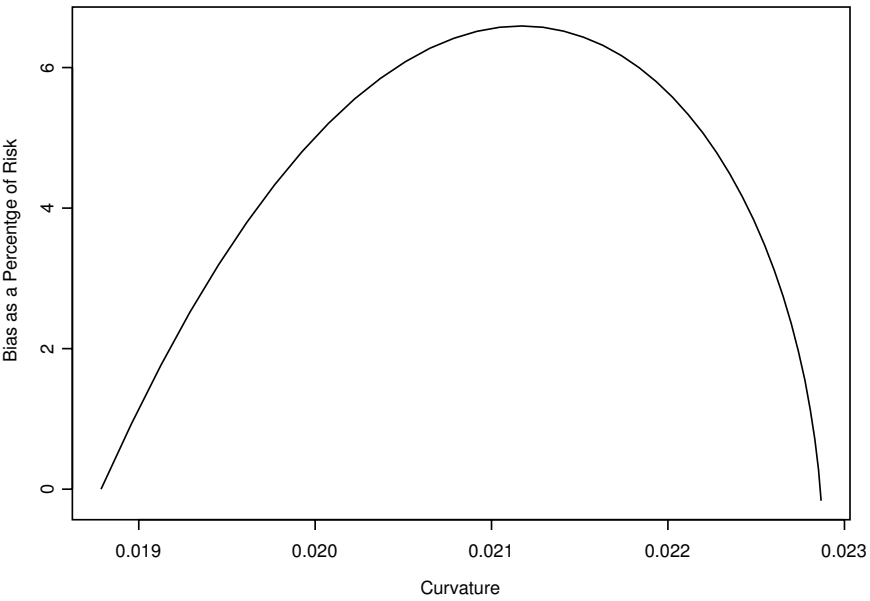


Fig. 1. Risk bias due to extrapolation and curvature.

even a small over- or under-estimation of risk may carry a high level of sensitivity in public health policy. The consequence of a linear dose-response assumption in policy-making remains the subject of further studies.

3. Clustered Multinomial Data: Developmental Toxicity

One of the most challenging and interesting aspects of reproductive and developmental toxicological data is the complex multivariate outcomes often encountered. This is the case because exposure to dangerous toxicants can affect many different stages in the reproductive process, including viability (sperm count, ovulation, etc.), fertilization and implantation. Once implantation has occurred, exposure can result in early pregnancy loss, malformations, lowered fetal weight or functional deficiency. Figure 2 shows some aspects of the developmental process, and illustrates how exposure may increase death rate, cause growth alterations such as lower fetal weight or malformations. In the figure, $\pi_1(d)$ denotes the risk of death including resorption and still birth, $\pi_2(d)$ the conditional risk of malformation among live births, all through maternal exposure at dose level d. We will focus on the multinomial outcomes of death and malformation only. A number of authors have discussed this issue in the context of dose response modeling.[11,22,23,37,44]

A second challenge is the clustering, i.e. the so-called litter effects or the tendency for littermates (offspring) of the same mother to respond similarly. In studies utilizing exposure to timed-pregnant animals over the period of organogenesis, the exposure effects on *in-utero* development are evaluated just before parturition would normally occur. Since the implants

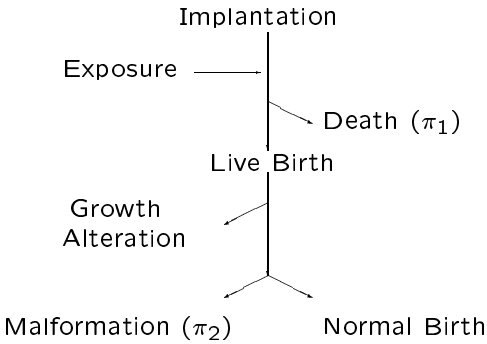


Fig. 2. Multivariate outcomes of developmental toxicity.

(offspring) of the same mother are genetically related and share a similar developmental environment, they are likely to respond to the toxic insult similarly.

To fit dose-response models for clustered multinomial data, we first adopt some notations. Consider an experiment in which pregnant dams are randomized to a control or one of D dose groups at exposure levels $d_0 = 0, d_1, \ldots, d_D$, respectively. Suppose the ith dose group has M_i dams, and the jth, $(j = 1, \ldots, M_i)$ dam has n_{ij} implants. Let $y_{ij} = (y_{1ij}, y_{2ij})^T$ denote the number of dead implants and malformed fetuses in the (ij)th dam.

3.1. *Generalized estimating equations*

Following Zhu *et al.*[44] and Krewski and Zhu,[23] we use Weibull models

$$\pi_1(d; \alpha) = 1 - \exp(-(\alpha_0 + \alpha_1 d^{\alpha_2}))$$

$$\pi_2(d; \beta) = 1 - \exp(-(\beta_0 + \beta_1 d^{\beta_2}))$$

to describe the incidences of death and malformation, respectively. With various values of the parameters $\theta = (\alpha^T, \beta^T)^T$ to be estimated from the data, the Weibull models are flexible in describing the various dose-response shapes observed in developmental toxicity studies, particularly a reversed L-shape or S-shape.[22] The models can mimic continuous threshold models without explicit use of a threshold.[38] The power parameters may be useful in other models such as the logistic and probit models.

Maximum likelihood estimation can be used to simultaneously fit dose-response models $\pi_1(d)$ and $\pi_2(d)$ in conjunction with a mixture of multinomial distributions. The infinite mixture of Dirichlet-multinomial distribution[9,44] and a finite mixture of multinomial[30] are perhaps the most common examples, because mixture is a simple mechanism to capture extra-multinomial variation due to clustering. However, computational complexity, uncertainty, and limitations about the distribution are disadvantages associated with ML estimation. Many favor alternative analyses based on quasi-likelihood, or more generally, generalized estimating equations (GEEs).[25] Ryan,[37] Zhu *et al.*[44] and Chen and Li[11] used variations of GEEs specifically in the modeling of death and malformation.

To estimate parameters θ using GEE, we need specification of only the mean and variance of the data. A general expression of the mean and variance functions for clustered multinomial data is

$$\mu_{ij} = E(Y_{ij}|n_{ij}) = n_{ij}(\pi_1(d_i), (1 - \pi_1(d_i))\pi_2(d_i))^T,$$

and

$$V_{ij} = (1 + (n_{ij} - 1)\rho_i) n_{ij} (\mathrm{diag}(\mu_{ij}) - \mu_{ij}\mu_{ij}^T) \,.$$

In the variance ρ_i is the coefficient of intra-cluster correlation, and $(n_{ij}-1)\rho_i$ is the variation component in excess of the multinomial variation. This variance function appears to cover all exchangeable multinomial data.[45] Now GEEs for θ take the following simple form:

$$\sum_{i,j} \frac{\partial(n_{ij}\mu_{ij})}{\partial\theta^T} V_{ij}^{-1} (y_{ij} - n_{ij}\mu_{ij}) = 0 \,. \tag{3}$$

To estimate the dispersion parameters ρ_i, a separate set of equations is required. The simplest example is the moment estimation given by

$$\sum_{i,j} \frac{\partial E(q(Y_{ij}))}{\partial\rho^T} W_{ij}^{-1} (q(y_{ij}) - Eq(Y_{ij})) = 0 \,, \tag{4}$$

where q can be a quadratic function of y_{ij} as well as θ, and W is chosen to approximate the variance of q. We obtain estimates for θ and ρ by iteratively solving the two sets of equations until convergence. An important addition in the GEE method is that the mean parameters and their variances will be estimated correctly even if the variance (V_{ij}) is incorrectly specified. This is achieved by the inclusion of an empirical variance "fix-up". There are still incentives to correctly specify the variance to improve statistical efficiency.[25]

One approach to improving the efficiency of estimation is to consider joint GEEs for θ and ρ. This can be done by forming an enlarged "data" vector $(y_{ij}^T, q_{ij}^T)^T$, and use its covariance matrix to construct GEEs in the form of Eq. (3). This approach is sometimes called GEE2. However, it requires the specification of the 3rd and 4th moments for the data Y_{ij} as well as iterative algorithms between estimates of θ and ρ because q_{ij} are likely dependent upon θ. Still other variations of GEEs are available. The extended GEEs[18] rely on the idea of constructing an extended quasi-likelihood by integrating the GEEs (3) and then differentiating the "likelihood" with respect to ρ to obtain the equations for ρ. If the extended quasi-likelihood is somewhat similar to the true likelihood, the second set of equations would be reasonably satisfactory.

In the current application we adopt the first approach, with $q(y_{ij}) = y_{ij}^2 - E(Y_{ij}^2)$ in Eq. (4). This approach is computationally simple, and numerically equivalent to the GEE2 approach. We further impose a distinct intra-litter correlation ρ_i for each dose group since it is common that ρ_i increases with dose level.[22]

A simple alternative approach is proposed by Krewski and Zhu[23] to avoid direct estimation of the intra-litter correlation ρ_i. The idea is to transform the data by dividing the death and malformation frequencies y_{ij} as well as the number of implants n_{ij} of each dam by the so-called design effects. The transformation effectively removes the over-dispersion component $(n_{ij} - 1)\rho_i$ from the data. As a result, we can use the mean and variance of a multinomial distribution to approximate that of the transformed data. The design effects associated with each dose group is in essence the "ratio" of the true variance to the variance of independent data. Denote the outcomes of malformation, death, and normal birth by an index $k = 1, 2, 3$. The design effect of each dose group can be estimated by

$$\hat{D}_i = \frac{1}{2}\left[\frac{\hat{v}_{i1}}{\hat{\mu}_{i1}} + \frac{\hat{v}_{i2}}{\hat{\mu}_{i2}} + \frac{\hat{v}_{i3}}{\hat{\mu}_{i3}}\right] \tag{5}$$

where $\hat{v}_{ik}$ $(k = 1, 2, 3)$ are the sample variances divided by the average number of implants and $\hat{\mu}_{ik}$ are the mean response rates in each dose group. An added advantage to this transformation procedure is that the usual Chi-squared goodness-of-fit test is applicable. Simulation studies reveals that this transformation procedure is valid even for moderate number of dams per dose group.[17]

3.2. *Illustration*

Data in Table 3 come from a study[35] that investigated the developmental effects of diethylene glycol dimenthyl ether (TGDM). The study included a vehicle control group and 3 dosed groups, each with from 20 to 30 pregnant rats. Measured outcomes included the number of implantation sites in each dam and the incidence of resorptions and/or fetal deaths. Fetuses surviving to sacrifice were weighed and evaluated for the presence of a variety of different types of malformation. It is clear from Table 2 that both the death and malformation incidences are elevated by increased exposure to TGDM.

The fitted death-malformation models are summarized in Table 4, where the transformation method is labeled by TR. While the incidence of malformation is strongly dose-related, the dose-response relationship of prenatal death is only marginally significant. Litter effects are apparent at the two highest dose levels, as indicated by the estimated intra-litter correlations as well as the design effects of 1.10, 0.69, 2.78, and 3.91 for the four dose groups respectively. The transformation method yielded a model comparable to that of GEEs, and a Pearson's goodness-of-fit statistic with $\chi^2 = 2.15$ and p-value $= 0.34$.

Table 3. Summary data of TGDM study.

Dose (g/kg)	*Dams*	*Total* Impl's	Death[a] No.	Death[a] (%)	Malf's[b] No.	Malf's[b] (%)
0.00	27	340	22	6.5	1	0.3
0.25	26	296	21	7.1	0	0.0
0.50	26	296	34	11.49	2	0.8
1.00	28	327	41	12.54	33	11.5

[a]Including dead or resorbed animals.

[b]Number of live animals exhibiting any malformation.

Table 4. Parameter estimates (S.E.)[a] of joint dose-response models for death and malformation in rats exposed to TGDM.

Model	Coefficient	Estimate	Estimates (TR)
Prenatal	Intercept (α_0)	0.0617	0.0642
Death		(0.0180)	(0.0154)
	Dose (α_1)	0.0956	0.0793
		(0.0563)	(0.0481)
	Power (α_2)	1.2650	1.2140
		(1.0584)	(1.0932)
Fetal	Intercept (β_0)	0.0006	0.0009
Malformation		(0.0009)	(0.0016)
	Dose (β_1)	0.1209	0.1222
		(0.0346)	(0.0356)
	Power (β_2)	4.8157	2.0224
		(2.0224)	(1.9803)
	Group 1 (ρ_1)	0.1546	
		(0.0688)	
Intralitter	Group 2 (ρ_2)	−0.0369	
		(0.0315)	
Correlation	Group 3 (ρ_3)	0.3288	
		(0.2829)	
	Group 4 (ρ_4)	0.2846	
		(0.1021)	

[a]S.E. = standard error.

We computed benchmark doses based on the risks of death (π_1), malformation (π_2), and the overall risk

$$\pi(d) = Pr(\text{death or malformation}|d) = 1 - (1 - \pi_1(d))(1 - \pi_2(d))\,.$$

We used both the multiplicative risk (1) and additive excess risk (2). The results are given in Table 5 along with their standard error (in parentheses)

Table 5. Estimates (S.E.) of BMD_{05} (mg/kg/d) based on the joint Weibull models for death and malformation.

	Prenatal Death	Malformation	Multivariate
Additive	0.643	0.837	0.568
	(0.343)	(0.072)	(0.212)
Multiplicative	0.611	0.837	0.548
	(0.333)	(0.072)	(0.216)

[a]S.E. = standard error.

derived by the delta method. The 95% lower confidence limit (LBD) of the BMD estimates can be approximated by $\text{LBD} = \text{BMD}_{05} - 1.645 \times \text{SE}$. The results in Table 5 confirm that a BMD based on the multiplicative risk is always smaller than that of the additive risk, although the difference becomes negligible when the spontaneous risk is small. Further, using the multiplicative risk formula, the BMD is always below the smallest BMD based separately on the univariate risk π_1 or π_2. The same needs not be true for the additive risk.[16]

4. Longitudinal Data: Neurobehavioral Toxicity Screening

In addition to its primary role in psychological functions, the nervous system controls most, if not all, other bodily processes. It is sensitive to perturbation from various sources and has limited ability to regenerate. There are evidences that even small anatomical, biochemical, or physiological insults to the nervous system may result in adverse effects on human health, transient or persistent.[41] Many chemicals in active commercial use may have, but are not tested for, neurotoxic potential. The US EPA has strongly recommended that neurotoxicity be used as an endpoint in regulating environmental toxicants. Neurotoxicity includes adverse effects in behavior, neurochemistry, neurophysiology, and neuropathology. It can also be evaluated based on neurological development and function in infants and children following prenatal and perinatal exposure. Our focus here is on neurobehavioral testing.

4.1. *Neurobehavioral screening test*

In a recent international study jointly sponsored by the International Program on Chemical Safety (IPCS) and US EPA,[32] the use of Functional Observational Battery (FOB)[29] was validated and tested for rapidly measuring neurobehavioral changes potentially caused by

chemical exposure. The FOB consists of a number of measurements that are grouped into several domains according to their neurological function: Autonomic, Neuromuscular, Activity, Sensorimotor, Excitability, and Physiology. Except for several measures in the domain of Neuromuscular as well as physiological measures, most measures are of observational nature, and criteria for adversities in individuals are not easy to develop. Thus, we assess the collective behavioral changes of the exposed group relative to that of a control group. If rare behaviors in the control become more common in the exposed group, there are indicators of adverse exposure effects. To unify all measures within a domain, every single measure, regardless of being binary, continuous or ordinal, is converted to a severity score. Under a 4-level Likert scale, "1" is most common or least severe in the control group, and "4" the least common or most severe. For instance, the measure of "ease of removal from cage" comes with six categories: "very easy", "easy", "moderately difficult", "rat flinches", "difficult", and "very difficult". If more than 50% of the control group are either "very easy" or "easy" to remove, then "very easy" and "easy" are converted to a "1"; Categories whose frequency is at least one rank away from the mode of the control distribution is assigned a "4". Further details on converting continuous measures can be found in MacDaniel and Moser.[29] Variation of this type of schemes can also be developed and tested. The average of individual severity scores within the same domain is a composite severity score for the domain, and is treated as a continuous measure in analysis.

Table 6 gives a summary of composite severity score of the "excitability" domain. The raw data are derived from an EPA study[31] in which 8 rats of the same strain and age, each from a different litter, were exposed at one of the four dose levels (150, 500, 1500, 5000 mg/kg) to tetrachloroethylenl(PER), a common chlorinated solvent. In addition, there was

Table 6. Mean (S.E.) excitability scores of rats exposed to PER.

Dose (mg/kg)	0	150	500	1500	5000
0 H	1.7938	2.0413	1.9188	1.7525	1.8350
	(0.0603)	(0.1070)	(0.2442)	(0.0545)	(0.0622)
4 H	1.6238	2.1650	1.9163	2.1650	2.5025
	(0.1414)	(0.1594)	(0.4352)	(0.3168)	(0.1273)
24 H	1.2064	1.5000	1.7088	1.6250	2.3300
	(0.0292)	(0.2876)	(0.6183)	(0.6187)	(0.0000)

a control group. FOB screening was conducted on each subject at just before dosing (time = 0), approximately 4, and 24 hours post exposure. There were four deaths in the 5000 mg/kg dose group by 24 hour. The composite severity score is the average of three individual measurements: handling reactivity, arousal, and ease of removal.

Repeated measurements are common in neurotoxicity studies in order to understand the often-transient effects of neurotoxicity. Traditionally,

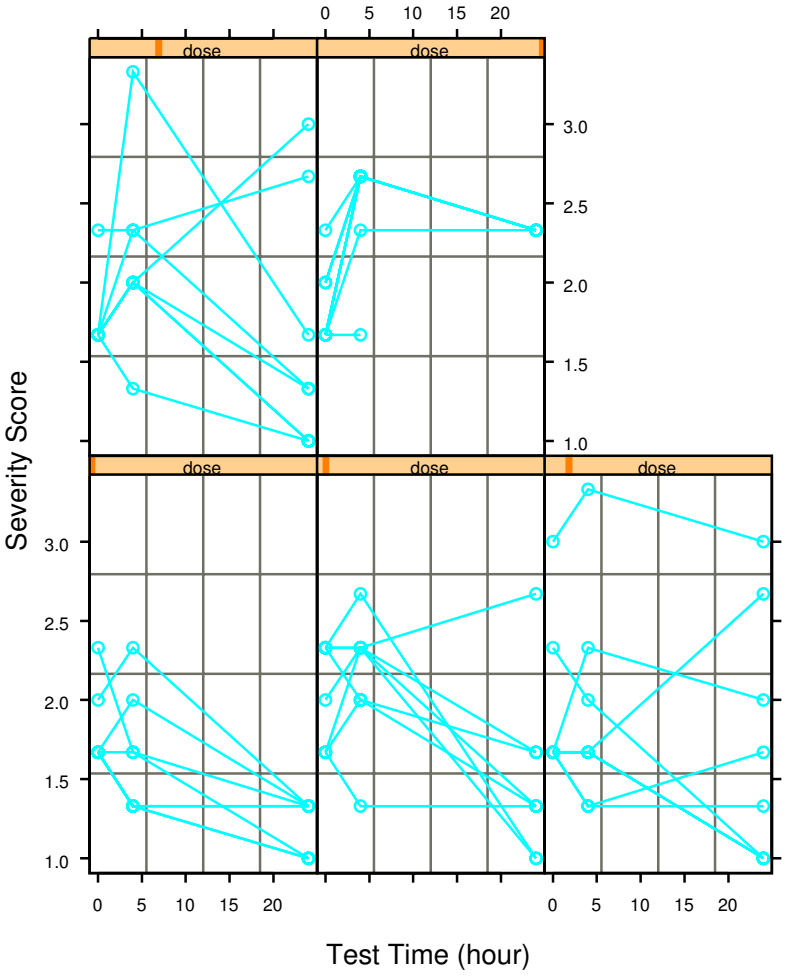


Fig. 3. Change of excitability in rats exposed to PER.

ANOVA with repeated measures is employed. This is also the case in the analysis of the composite severity scores of the FOB data.[31] However, ANOVA is designed for testing dose effects, not for dose-response modeling. Further, dose-response modeling is a prerequisite to the BMD approach. Data in Table 6 also showcase a typical situation in neurotoxicity study: small sample size per dose group, several repeated measures over time, and possible missing observations that often occur at higher dose levels or later times due to mortality. In the presence of missing data, ANOVA fails to utilize all available data because subjects with any missing data at some time points would have been removed to satisfy the requirement of balanced designs. This can be costly given an already-small sample per dose group in most neurotoxicity studies. Although linear or non-linear regression models may be used to model dose-response, they cannot differentiate distinct behavior trajectory among individual subjects. Between-subject variation often reflects important biological variation. Not only is it important to control to achieve better statistical power, it also sheds lights on identifying special risk groups. In Fig. 3 the excitability score is plotted against time for each subject, and the trajectories are grouped into separate panels by dose level. The bar on the top of each panel indicates an order of increasing dose, from left to right. The plot reveals some degree of between-subject

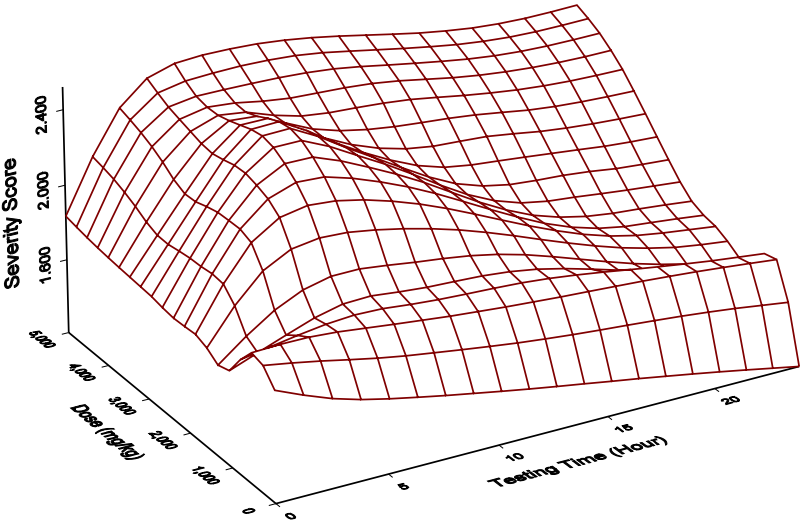


Fig. 4. Excitability of rats exposed to PER.

variation. To see the average trajectory of behavioral change, we have a 3-dimensional plot in Fig. 4 using a spline smoother. The control group's excitability score remained stable, only decreasing slightly over time, due perhaps to self-adjustment. As dose increased, however, there were slight increases in excitability, peaking at around 5 hour after exposure and then dropped somewhat by 24 hour. The trajectory of the 5000 mg/kg group was quite different: the excitability score increased over time, and did not drop. This indicates that there were sustained dose effects at excessively high dose levels. It was not clear though whether the effects could be permanent in nature. In summary, the fact that there were only a small number of subjects, potentially sizable between-subject variation, and missing data associated with some subjects points to the benefits of using mixed effects models for data analysis.

4.2. *Linear mixed effects models*

Examination of Fig. 4 reveals some interesting patterns of the dose-time-response of the excitability score. We found the following hybrid pharmocokinetic model

$$f(\phi(\text{dose}), t) = \frac{\phi_1(\text{dose}) + \phi_2(\text{dose})t}{1 + e^{\phi_3(\text{dose})t^{\phi_4(\text{dose})}}} \tag{6}$$

quite flexible in describing the various dose-response shapes observed in the FOB data. In this model, the dose effects are incorporated into the parameters ϕ_i as a function of dose. The simplest one is linear functions. Further, individual subject is allowed to have distinct behavior trajectory that is characterized by random coefficients associated with population parameters. After testing a number of options based on the actual data, we choose the following parametrization:

$$\begin{aligned}
\phi_1(d) &= \beta_{10} \\
\phi_2(d) &= \beta_{20} + (\beta_{21} + b_{21i})d \\
\phi_3(d) &= \beta_{31}d \\
\phi_4(d) &= \beta_{40}\,.
\end{aligned}$$

Note that ϕ_1 determines the initial (time = 0) excitability level that is not influenced by dose because the first test was done before dosing. Therefore we make $\phi_1(d) = \beta_{10}$ a constant. Although we can jitter the intercept β_{10}

by random effects to characterize unexplained variation in initial excitability among different subjects, the effects seem to be minimal in our actual data analysis. The time-slope ϕ_2 measures how excitability changes over time, either naturally (β_{20}), or under the influence of dosing ($\beta_{21} + b_{21i}$) with individual variation b_{21i}. The downturn time-slope β_{31} is simply to control for the non-monotone trend within some dose range. A constant term is unnecessary because it amounts to a scale in conjunction with the intercept term in ϕ_1. Finally, a power parameter β_{40} is employed to increase the flexibility of the dose-time-response shape. Although in principle every population parameter can be jittered with random effects, the use of random effects must be justified by actual data variation. Guidance for model selection as well as use of random effects can be found in Pinheiro and Bates.[34]

Before fitting the model (6), we briefly outline the methods of nonlinear mixed effects models. The notation here follows largely that of Chapter 7 of Pinheiro and Bates,[34] which emphasizes on the computational aspects and applications of mixed effects models. For theory and method, see for example Davidian and Glicknan.[14] Consider the following general model,

$$Y_{ij} = f(\phi_i, X_{ij}) + \epsilon_{ij}\,, \quad i = 1, \ldots, M;\ j = 1, \ldots, n_i \tag{7}$$

where $\epsilon_i = (\epsilon_{i1}, \epsilon_{ij}, \ldots, \epsilon_{in_i})^T$ is the error vector for the ith subject in the sample of M; X_{ij} is the covariate vector for the ith subject at the jth sampling time. The parameter vector $\phi_i = A_i\beta + B_i b_i$ has two components: the first component β is population parameters and the second component b_i ($q \times 1$ vector) represents random effects associated with the ith subject. The matrices A_i and B_i are chosen with appropriate dimensions to determine on an individual level how to incorporate the population parameters and random effects into the model. In the present case, the matrices would be the same for all individuals. More general formulations can be developed to allow for time-varying coefficients.[34] It is common to assume for continuous data that ϵ_i follow the $N_{n_i}(0, \sigma^2\Lambda_i(\theta_1))$ distribution, and are independent between different subjects. Various correlation options may be incorporated into $\Lambda_i(\theta_1)$ through the parameter θ_1. The random effects b_i are also conveniently assumed to follow the $N_q(0, \sigma^2\Omega(\theta_2))$ distribution. Although other distributions can be considered in principle, computer software is mostly limited to normal distributions.

To fit a non-linear mixed effects model, maximum likelihood estimation is often used via either the EM-algorithm[15] or other numerical algorithms

such as Newton-Raphson.[27] Here we use Newton-Raphson algorithm implemented in S-PLUS (Mathsoft, Seattle, Washington). Under the assumption of normality, the marginal distribution of $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ is

$$\begin{aligned} p(y_i|\beta, \sigma^2, \theta_1, \theta_2) &= \int p(y_i|b_i; \beta, \sigma^2, \theta_1) p(b_i; \theta_2) db_i \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n_i+q}{2}} |\Lambda_i|^{\frac{1}{2}} |\Omega|^{\frac{1}{2}}} \\ &\quad \times \int \exp\left\{ -\frac{(y_i - f(\phi_i))^T \Lambda_i^{-1} (y_i - f(\phi_i)) + b_i^T \Omega^{-1} b_i}{2\sigma^2} \right\} db_i, \end{aligned}$$

where $f(\phi_i) = (f(\phi_i, X_{i1})), \ldots, f(\phi_i, X_{in_i}))^T$. Using a first order Taylor series expansion about some initial estimates $\hat{\beta}^{(t)}$ and $\hat{b}_i^{(t)}$ to approximate the exponent of the integrand, we can obtain the following approximate marginal log likelihood function

$$\begin{aligned} l(\beta, \sigma^2, \theta_1, \theta_2; y_i) &= -\frac{n_i + q}{2} \log(\sigma^2) \\ &\quad - \frac{1}{2} \left\{ \log|\Sigma_i| + \frac{1}{\sigma^2} (w_i^{(t)} - Z_{1i}^{(t)} \beta)^T \Sigma_i^{-1} (w_i^{(t)} - Z_{1i}^{(t)} \beta) \right\} \end{aligned}$$

where $w_i^{(t)} = y_i - f(\hat{\beta}^{(t)}, \hat{b}_i^{(t)}) + Z_{1i}^{(t)} \hat{\beta}^{(t)} + Z_{2i}^{(t)} \hat{b}_i^{(t)}$, and

$$Z_{1i}^{(t)} = \frac{\partial f}{\partial \beta^T} \quad \text{and} \quad Z_{2i}^{(t)} = \frac{\partial f}{\partial b_i^T}$$

are the derivative matrices evaluated at $\hat{\beta}^{(t)}$ and $\hat{b}_i^{(t)}$, and

$$\Sigma_i(\theta_1, \theta_2) = \Lambda_i(\theta_1) + Z_{2i}^{(t)} \Omega(\theta_2) (Z_{2i}^{(t)})^T .$$

Maximum likelihood estimates $(\hat{\theta}_1^{(t)}, \hat{\theta}_2^{(t)}, \hat{\sigma}^{2(t)})$ are obtained from maximizing this approximate log-likelihood

$$\max_{\theta_1, \theta_2, \sigma^2} \sum_i^M l(\hat{\beta}^{(t)}, \sigma^2, \theta_1, \theta_2; y_i) . \tag{8}$$

The estimate of the regression parameters β is then updated by a least squares estimator

$$\hat{\beta}^{(t+1)} = \left\{ \sum_1^M (Z_{1i}^{(t)})^T (\Sigma_i^{(t)})^{-1} Z_{1i}^{(t)} \right\}^{-1} \left\{ \sum_1^M (Z_{1i}^{(t)})^T (\Sigma_i^{(t)})^{-1} w_i^{(t)} \right\}. \tag{9}$$

The estimate of the random effects b_i is updated also by a least squares estimator based on

$$\sum_{i}^{M} \Big\{ (y_i - f(\hat{\beta}^{(t+1)}, b_i))^T \Lambda_i^{-1}(\hat{\theta}_1^{(t)})(y_i - f(\hat{\beta}^{(t+1)}, b_i)) + b_i^T \Omega^{-1}(\hat{\theta}_2^{(t)}) b_i \Big\}.$$

Using a first order Taylor series approximation

$$f(\hat{\beta}^{(t+1)}, b_i, X_{ij}) = f(\hat{\beta}^{(t+1)}, \hat{b}_i^{(t)}, X_{ij}) + Z_{2i}^{(t)}(b_i - \hat{b}_i^{(t)}),$$

we have the update

$$\hat{b}_i^{(t+1)} = \{\Omega^{-1} + Z_{2i}^T \Lambda_i^{-1} Z_{2i}\}^{-1} Z_{2i}^T \Lambda_i^{-1} \Big\{ y_i - f(\hat{\beta}^{(t+1)}, \hat{b}_i^{(t)}) + Z_{2i} \hat{b}_i^{(t)} \Big\}, \tag{10}$$

with all terms evaluated at the most recent parameter estimates. We iteratively solve the three sets of Eqs. (8), (9) and (10) until convergence.

The maximization procedure outlined above relies heavily on the normality assumption for both the error terms and random effects. If the random effects follow a non-normal distribution, explicit marginal likelihood is often not available, and the EM-algorithm would be used to numerically approximate the marginal likelihood.

4.3. *Illustration*

The fitted mixed effects model (6) is summarized in Table 7, and the predicted dose-time-response surface is plotted in Fig. 5. The fitted model fits the data well. A comparison of Fig. 5 with Fig. 4 reveals that the model captures most of the trends observed in the raw data. The dose effects were

Table 7. Estimates of parameters in the excitability model.

Parameter	Value	S.E.[a]	DF	t-Value	p-value
β_{10}	1.8693	0.0763	72	24.51	< 0.0001
β_{20}	2.2611	0.1723	72	13.12	< 0.0001
β_{21}	-0.0003	0.0001	72	-4.29	0.0001
β_{31}	-0.0003	0.0001	72	-3.40	0.0011
β_{40}	1.1597	0.0320	72	36.30	< 0.0001
$\Omega(\theta_2)$	S.E.	7.72e-008			
$\Lambda(\theta_1)$	σ	0.4737			
	corr(i, j)[b]:	(1, 2): 0.63	(1, 3): 0.14	(2, 3): 0.30	

[a] S.E. = standard error.

[b] corr(j, k) = corr(ϵ_{ij}, ϵ_{ik}).

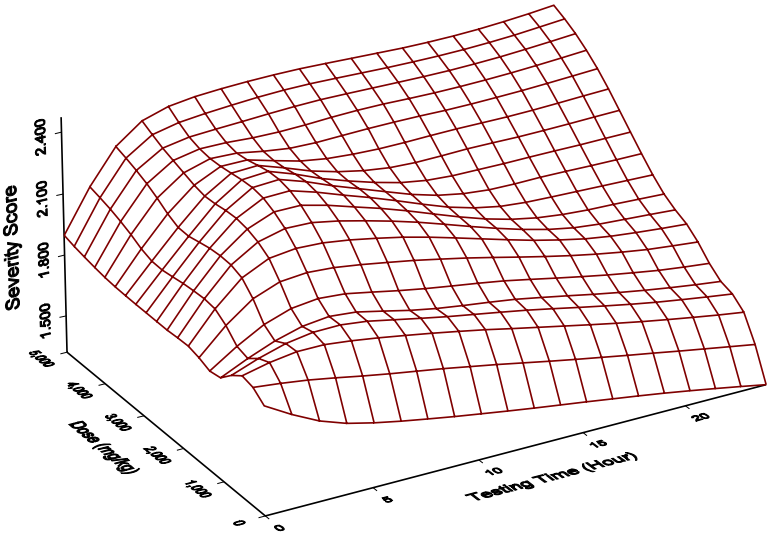


Fig. 5. Fitted model for the excitability score.

highly significant, and varied with testing time (Table 7) with a peak effect occurring at around 4.5 hour. However, it was difficult to accurately determine the true peak effect or peak effect time. The observed of peak effects are most likely taken to give the peak effect time. Several options were considered to include random effects in the model. Since using random effects for both the intercept β_{10} and the time slope β_{20}, or both β_{10} and the time-dose slope β_{21} led to an almost perfect correlation between the two random effects, both seemed an over-parameterization. Thus one random effect appeared sufficient for this dataset. Under several criteria, including likelihood ratio, information criterion and residual plots, the random effect associated with the time-dose slope β_{21} was found to be most effective. The standard error for b_{21i} was $7.72e^{-8}$ (Table 7), a small number compared with the estimate $\hat{\beta}_{21} = -3e^{-4}$, indicating a small between-subjects variation. We had also considered several correlation structures for the error terms, including that of complete independence, auto-regression with log 1, compound symmetric, and un-restricted. The unrestricted correlation matrix Λ (Table 7) seemed most flexible.

4.4. *Benchmark doses*

In computing BMDs, we must take into consideration the fact that risk depends not only on exposure level, but also on the time the screening test

took place. It is crucial that risk is assessed before, not after, the peak effect time. Mathematically, we can measure risk at any designated time, estimate BMDs as a function of time to create a timed-profile, and then search for the minimum BMD.

Since the excitability score is treated as a continuous measure, abnormality cannot be determined solely by a cutoff value, thus we adopt the population-level assessment approach by comparing the proportions of subjects whose score exceeds a given level between the exposed and controlled groups. This cutoff level can be chosen to include $\alpha \times 100\%$ of the subjects with the highest severity score in the control group. A significant increase in the proportion of subjects exceeding this level in the exposed group represents a risk because of the significant changes in the distribution of excitability.

Conceptually, let $\pi(0,t) = Pr(Y > c_t|d = 0, t) = \alpha$ be the proportion of subjects in the control group whose severity score exceeds c_t at testing time t, and $\pi(d,t) = Pr(Y > c_t|d,t)$ be that for the exposed group. Note again that π is not necessarily a risk, but merely a reference level for risk. We compute BMD corresponding to a $\gamma \times 100\%$ increase in multiplicative excess risk (1). Under a normal distribution for the severity score Y,

$$\pi(d,t) = 1 - \Phi\left(\frac{c_t - f(d,t)}{\sigma_0}\right),$$

where Φ is the standard normal cumulative distribution and with a constant standard deviation $\sigma_0 = 0.4737$ pooled across all dose groups (Table 7). Equation (1) simplifies to

$$\Phi\left(\frac{c_t - f(d,t)}{\sigma_0}\right) = \Phi\left(\frac{c_t - f(0,t)}{\sigma_0}\right)(1-\gamma) = (1-\alpha)(1-\gamma),$$

and $c_t = f(0,t) + z_{1-\alpha}\sigma_0$ with $z_{1-\alpha}$ being the $1-\alpha$ percentile of the standard normal distribution. The preceding equation further simplifies to

$$f(\mathrm{BMD}_t, t) = f(0,t) - \sigma_0(\Phi^{-1}((1-\alpha)(1-\gamma)) - z_{1-\alpha}). \tag{11}$$

Substituting for the parameters with their estimates and solving Eq. (11) lead to the estimate of BMD_t at chosen time t. In Fig. 6, we plotted the estimated BMDs over time to illustrate the influences of the reference and risk levels on BMDs with various choices of (α, γ): (0.05, 0.05), (0.05, 0.10), (0.10, 0.05) and (0.10, 0.10). We can see from Fig. 6 that the higher the reference level α is, the lower the risk detection limit is, hence the smaller the BMD would be. On the other hand, the lower the tolerance risk level γ is, the smaller the BMD would be because of increased sensitivity level

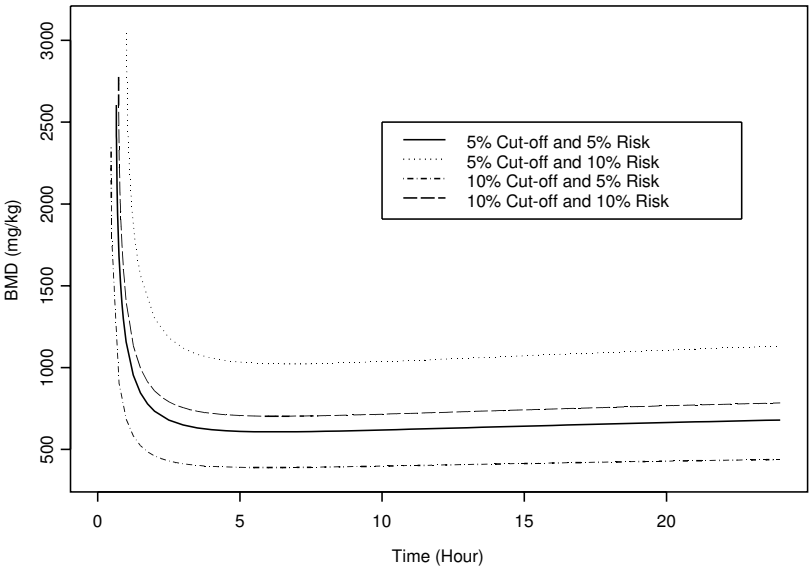


Fig. 6. Time-profiled benchmark doses.

to risk. A careful examination of the timed-profile of BMD suggests that, depending on the reference and risk levels, the minimum BMD is between 6.25 and 7 hour post exposure. This finding in turn suggests a sensitive time window for screening.

5. Discussion

We have illustrated in this article several dose-response modeling techniques with applications to health risk assessment, particularly BMD estimation. The examples used include binary, clustered multinomial and repeated measurements. Although the basic concepts of modeling and risk assessment are unambiguous, a number of technical issues warrant further discussion.

Defining adversity based on continuous measures is often arbitrary. We can circumvent this problem by assessing distributional changes to exposed populations relative to a control group. The baseline probability is only a reference, the changes to the baseline can be used as measures of risk. Bosch, Wypij and Ryan[5] proposed a non-parametric approach to establish the baseline reference for weight loss. Suppose Y_{ij} denotes the observed weight for the jth animal in the ith dose group ($i = 0, \ldots, D$; $j = 1, \ldots, n_i$),

with $d_0 = 0$ being the control. If there is no dose effects, $\pi(d_i) = Pr(Y_{0h} > Y_{ij}) = 1/2$ for all i, including in particular the case $i = 0$, i.e. $\pi(0) = 1/2$. We construct a "new" response vector of length mn_0, $m = \sum_{i=1}^{D} n_i$, representing the comparisons between each control subject and each exposed subject. That is, let

$$\mathbf{W} = (W_{0111}, W_{0112}, \ldots, W_{0hij}, \ldots, W_{0n_0Dn_D})', \quad W_{0hij} = I(Y_{0h} > Y_{ij}),$$

where the indicator function W_{0hij} takes the value 1 if $Y_{0h} > Y_{ij}$, and 0 otherwise. Since $E(W_{0hij}) = \pi(d_i)$, it is natural to estimate $\pi(d)$ from a binary regression model for W_{0hij}.

Developing generalized linear mixed effects models for ordinal data[19,20] would be useful for the analysis of individual measure of FOB data. There is one major technical challenge with regard to zero frequencies in categories of less common behavior particularly with a small sample size. Developing multivariate dose-response models is also interesting. For example, Catalano and Ryan[7] and Regan and Catalano[36] studied joint dose-response models with both binary (malformation) and continuous outcomes (body weight) in developmental toxicity studies, using a bivariate latent variable model. Attempts have also been made to further include death in the multivariate dose-response model.[8,10] Empirical evidences argue that multivariate models should result in improved precision for the purpose of risk assessment, as compared with a univariate approach.[22,37] This improved precision could arise in two ways: (1) a richer class of models and better fit, and (2) more efficient estimators. It would be very useful to explore good quality statistical techniques for evaluating goodness-of-fit of dose response models when they are fit using methods such as GEEs.

Better techniques for calculating lower confidence limits based on either likelihood or GEEs are needed. Given small to moderate sample size, the conventional methods[13,16] that use normal approximation to the estimated dose-response or BMD may not be accurate. Statistical calibration methods in conjunction with GEEs can be adopted and modified to estimate the benchmark dose. Nonparametric techniques such as bootstrap[43] should be further explored in this context to have robust estimate of confidence limits.

Acknowledgment

This work is in part supported by the US National Science Foundation grant DMS9978370.

References

1. Anger, W. K. (1984). Neurobehavioral testing of chemicals: Impact on recommended standards. *Neurobehavioral Toxicology Teratology* **6**: 147–153.
2. Anger, W. K. (1986). Workplace exposores. In *Neurobehavioral Toxicology*, ed. Z. A, Annau, John's Hopkins University Press, 331–347.
3. Barnes, D. G. and Dourson, M. (1988). Reference dose (RfD): Description and use in health risk assessments. *Regulatory Toxicology Pharmacology* **8**: 471–486.
4. Barnes, D. G. and Daston, G. P., Evans, J. S., Jarabek, A. M., Kavlock, R. J., Kimmel, C. A., Park, C. and Spitzer, W. L. (1995). Benchmark dose workshop: Criteria for use of a benchmark dose to estimate a reference dose. *Regulatory Toxicology Pharmacology* **21**: 296–306.
5. Bosch, R. J., Wypij, D. and Ryan, L. M. (1996). A semiparametric approach to risk assessment for quantitative outcomes. *Risk Analysis* **16**: 657–665.
6. Brown-Peterson, N., Krol, R., Zhu, Y. and Hawkins, W. (1999). Nitrosodiethylamine initiation of carcinogenesis in Japanese medaka (*Oryzias latipes*): Hepatocellular Proliferation, Toxicity, and Neoplastic Lesions Resulting from Short Term, Low Level Exposure. *Toxicological Sciences* **50**: 186–194.
7. Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of American Statistical Association* **87**: 651–658.
8. Catalano, P. J., Scharfstein, D. O., Ryan, L., Kimmel, C. and Kimmel, G. (1993). Statistical model for fetal death, fetal weight, and malformation in developmental toxicity studies. *Teratology* **47**: 281–290.
9. Chen J. J., Kodell, R. L., Howe, R. B. and Gaylor, D. W. (1991). Analysis of trinomial responses from reproductive and developmental toxicity experiments. *Biometrics* **47**: 1049–1058.
10. Chen, J. J. (1993). A malformation incidence dose-response model incorporating fetal weight and/or litter size as covariates. *Risk Analysis* **13**: 559–564.
11. Chen, J. J. and Li, L. A. (1994). Dose-response modeling of trinomial responses from developmental experiments. *Statistica Sinica* **4**: 265–274.
12. Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundamental Applied Toxicology* **4**: 854–871.
13. Crump, K. S. (1995). Calculation of benchmark doses from continuous data. *Risk Analysis* **15**: 79–89.
14. Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London.
15. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society* **B39**: 1–22.
16. Gaylor, D. W., Ryan, L., Krewski, D. and Zhu, Y. (1998). Procedures for calculating benchmark doses for health risk assessment. *Regulatory Toxicology Pharmacology* **28**: 150–164.
17. Fung, K. Y., Krewski, D., Rao, J. N. K. and Scott, J. A. (1994). Tests for trend in developmental toxicity experiments with correlated binary data. *Risk Analysis* **14**: 639–648.

18. Hall, D. B. and Severini, T. A. (1998). Extended generalized estimating equations for clustered data. *Journal of American Statistical Association* **93**: 1365–1375.
19. Hedeker D. and Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biomtetrics* **50**: 933–944.
20. Hedeker, D. and Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine* **49**: 157–176.
21. Kimmel, C. A. and Gaylor, D. W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicity. *Risk Analysis* **8**: 15–20.
22. Krewski, D. and Zhu, Y. (1994). Applications of multinomial dose-response models in developmental toxicity risk assessment. *Risk Analysis* **14**: 613–627.
23. Krewski, D. and Zhu, Y. (1995). A simple data transformation for estimating benchmark dose in developmental toxicity experiments. *Risk Analysis* **15**: 29–39.
24. Leisenring, W. and Ryan, L. (1992). Statistical properties of the NOAEL. *Regulatory Toxicology Pharmacology* **15**: 161–171.
25. Liang, K. Y and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
26. Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated-measures data (corr: 94v89 p1572), *Journal of American Statistical Association* **83**: 1014–1022.
27. Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed-effects models for repeated measures data, *Biometrics* **46**: 673–687.
28. McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.
29. MacDaniel and Moser (1993). Utility of a neurobehavioral screening battery for differentiating the effects of two pyrethroids, permethrin and cypermethrin. *Neurotoxicology Teratology* **15**: 71–83.
30. Morel, J. G. and Neerchal, N. K. (1993). A finite mixture distribution for modeling multinomial extra variation. *Biometrika* **80**: 363–371.
31. Moser, V. C., Cheek, B. M. and MacPhail, R. C. (1995). A multidisciplinary approach to toxicological screening III: Neurobehavioral toxicity. *Journal of Toxicology Environmental Health* **45**: 173–210.
32. Moser, V. C., Tilson, H. A., MacPhail, R. C., Becking, G. C., Cuomo, V., Frantik, E., Kulig, B. and Winneke, G. (1997). The IPCS collaborative study on neurobehavioral screening methods: II. protocol design and testing procedures. *NeuroToxicology* **18**: 929–938.
33. Office of Technology Assessment (OTA) (1990). Neurotoxicity: Identifying and controlling poisons of the nervous system. US Congress, Office of Technology Assessment. *OTA-BA-436*. US Government Printing Office, Washington, DC.
34. Pinheiro, J. C. and Bates, D. M. (2000) Mixed-effects models in S and S-plus. Springer-Verlag: New York.

35. Price, C. J., Kimmel, C. A., George, J. D. and Marr, M. C. (1987). The developmental toxicity of diethylene glycol dimenthyl ether in mice. *Fundamental Applied Toxicology* **81**: 113–127.
36. Regan, M. M. and Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Applications to developmental toxicology. *Biometrics* **55**: 760–768.
37. Ryan, L. M. (1992). Quantitative risk assessment for developmental toxicity. *Biometrics* **48**: 163–174.
38. Stiteler, W. M., Joly, D. A. and Printup, H. A. T. (1993). Monte Carlo investigation of issues relating to the benchmark dose method. Task No. 2-30, Environmental Criteria and Assessment Office. US Environmental Protection Agency, Cincinnati, Ohio.
39. US Environmental Protection Agency (1991). Guidelines for developmental toxicity risk assessment. *Federal Register* **56**: 63797–63826.
40. US Environmental Protection Agency (1995). The use of benchmark dose approach in health risk assessment. Office of Research and Development, Washington, DC. EPA/630/R-94/007.
41. US Environmental Protection Agency (1998). Guidelines for Neurotoxicity Risk Assessment *Federal Register* **63**: 26926–26954
42. Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**: 949–952.
43. Zeng, Q. and Davidian, M. (1997). Bootstrap-adjusted calibration confidence intervals for immunoassay. *Journal of American Statistical Association* **92**: 278–290.
44. Zhu, Y, Krewski, D and Ross, W. H. (1994). Multinomial models for developmental toxicity experiments. *Applied Statistics* **43**: 583–598.
45. Zhu, Y. and Fung, K. Y. (1996). Statistical methods in developmental toxicity risk assessment. In *Toxicology and Risk Assessment*, *Principles*, *Methods and Applications*, eds. Fan and Chang, 413–446.

About the Author

Zhu Yiliang is an Associate Professor at the College of Public Health, University of South Florida. He obtained his PhD in Statistics from University of Toronto, MS from Queen's University in Kingston, and BS from Shanghai University of Science and Technology. He was a visiting scientist at Environmental Health Center, Health Canada and US Environmental Protection Agency. He was also the principal Biostatistician at the headquarter of the Shriner's Hospital for Children. His current research focuses on statistical methodology and applications in environmental health risk assessment, including semiparametric methods for dose-response modeling of longitudinal data, effective designs for dose-response, benchmark

methods in environmental regulation, toxicokinetic-toxicodynamic models in toxicology, and software development. His research has been funded through grants from the US National Science Foundation and National Institute of Health. In addition to serving on many national and statewide committees, he has also done extensive research and consultation in clinical trials and management, health care outcome evaluation, epidemiology, and litigations.

CHAPTER 17

STATISTICAL MODELS AND METHODS IN INFECTIOUS DISEASES

HULIN WU

Frontier Science and Technology Research Foundation,
1244 Boylston Street, Suite 303, Chestnut Hill, Massachusetts 02467, USA
Tel: 617-632-5738; wu@sdac.harvard.edu

SHOUJUN ZHAO

Department of Neurology, University of California,
San Francisco, CA 94143, USA

1. Introduction

1.1. *Infectious diseases*

Infectious diseases are the illnesses caused by a specific infectious agent or its toxic products. Most of the agents are microorganisms, like bacteria, virus, parasites, etc. The transmission of the agent from an infected person, animal, or reservoir to a susceptible host, results in the infectious diseases of human, either directly or indirectly through an intermediate plant or animal host, vector, or the inanimate environment. For example, influenza, hepatitis, AIDS are caused by virus; dysentery, typhoid by bacteria; and schistosomiasis, filariasis by parasites.[1] Infectious diseases are also called communicable diseases.

Infectious disease is a great threat to human beings in the past centuries. In this new century, understanding and controlling the spread of infections is still vitally important to public health. The challenging problems in studies of infectious diseases include: (i) how to evaluate the epidemics of an infectious disease in a population; (ii) how to understand the pathogenesis of infections and transmissions; (iii) if intervention measures such as drugs and therapies, and prevention measures such as education program and vaccines are developed, how to evaluate their effectiveness. Mathematics and

statistics have played a central role in all these three aspects in the past decades. To accurately evaluate and project the epidemics of an infectious disease would help to determine the health care needs which is useful for the public health department or government to prepare and allocate the resources to fight the disease. It also signals a message on how serious a particular infectious disease is and draws attentions from the public on the dangerousness of the disease.

1.2. *Mathematical and statistical challenges from infectious diseases*

The application of mathematical methods to infectious diseases dates back to Daniel Bernoulli's paper in 1760[2] in which he used a mathematical model to evaluate the impact of smallpox on life expectancy. More analytical work has been done by Hamer[3] and Ross,[4] who tried to understand the mechanisms of disease transmission in early of last century. Kermack and McKendrick[5] studied the mass action principle and threshold theorem originally proposed by Hamer and Ross respectively. The well-known chain binomial models of disease spread may be traced back to En'ko,[6] and a stochastic counterpart of chain binomial models was introduced by Greenwood.[7]

Throughout the last century, theory and quantitative techniques have been developed to study both the dynamics of disease within individuals and the transmission of infections through populations. Mathematics and statistics have played an important role in the studies of infectious diseases. In particular, over the last two decades there has been a great deal of work on HIV/AIDS. Both mathematics and statistics have played and will continuously play an important roles in epidemic studies as well as intervention and prevention studies of infectious diseases. A recent brief review on these methods can be found in Farrington,[8] Heesterbeek and Roberts,[9] and Gani.[10]

1.3. *Outline*

Many mathematicians and statisticians have responded to the challenges from infectious diseases in the past two centuries, and continue to make the best efforts for meeting the new challenges in this new millennium.

In this chapter, first we will introduce the back-calculation method for a projection of epidemics, proposed recently for estimating epidemics of AIDS, followed by the introduction of models for natural history of

infectious diseases. Deterministic and stochastic models as well as their recent developments are presented in Sec. 2. In Sec. 3, we introduce viral dynamic models which are heavily studied for understanding pathogenesis of HIV, HBV, and HCV infection during the last decade. We briefly review the mathematical and statistical methods for evaluation of intervention and prevention measures in infectious diseases in Sec. 4. We conclude the chapter with a brief summary.

2. Epidemic Models

2.1. *Estimation and projection of epidemics — Back-calculation*

Estimation and projection of epidemics, such as disease incidence and prevalence, are very important for intervention and prevention of infectious diseases. It is also critical for a government to make decisions and to prepare public health needs. Back-calculation or back-projection method has been paid tremendous attention in estimating and projecting AIDS epidemics in the past 15 years. In this section, we briefly introduce the back-calculation method and its applications.

Back-calculation is a method for estimating past infection rates of an infectious disease by working backward from observed disease incidence using knowledge of the incubation period between infection and disease. Although it can be used to any infectious diseases in theory, it was first proposed to study AIDS epidemics by Brookmeyer and Gail.[11,12] It has been widely used in AIDS epidemics.

The basic idea is to use the convolution equation of the expected cumulative number of disease cases diagnosed by time t, $A(t)$, the infection rate $g(s)$ at time s, and the incubation period distribution $F(t)$, i.e.

$$A(t) = \int_0^t g(s) \cdot F(t-s) ds \,. \tag{1}$$

If the disease cases $A(t)$ are known (may be obtained from case reports) and the incubation period distribution $F(t)$ can be estimated from epidemiological studies, the infection rate $g(s)$ then can be estimated by deconvolution of Eq. (1). If the infection rate $g(s)$ and incubation period distribution $F(t)$ are known, the disease cases can be estimated or projected using the convolution Eq. (1).

We introduce the back-calculation using a discrete-time formulation since it is more realistic. We assume that we have n non-overlapping time

interval, (T_{j-1}, T_j), $j = 1, \ldots, n$; let Y_j be the number of disease cases diagnosed in the jth interval; denote f_{ij} as the probability of developing disease in time interval j given infection in interval i, or $f_{ij} = F(T_{j-i+1}) - F(T_{j-i})$ where $F(T_0) = 0$; let g_i denote the expected number of new infections in time interval i (infection rate). The discrete-time statistical convolution equation can be written as

$$\mathrm{E}(Y_j) = \sum_{i=1}^{j} g_i f_{ij}\,, \quad j = 1, \ldots, n\,. \tag{2}$$

Usually Y_j is assumed to follow a Poisson distribution. A Poisson regression analysis may be used to estimate parameters g_i while we regard f_{ij} as known covariates. The generalized linear model algorithms in the standard statistical packages such SAS or Splus can be used to fit the model. However, a difficulty with this model is that the number of parameters equal to the number of data points. This may result in unstable estimate of g_i. To resolve this problem, one may model $g(s)$ parametrically or nonparametrically. The parametric models include damped exponential model, log-logistic model, logistic (prevalence) model, and piecewise constant step function model. For nonparametric modeling methods, smoothing spline, kernel method, and series-based splines can be used. More details can be found in the book (Chapter 8) by Brookmeyer and Gail.[13]

If we model g_i as a parametric function, say, $g_i = g(i, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of parameters, the maximum likelihood method may be used to estimate the parameter vector $\boldsymbol{\beta}$ or the infection rate function $g_i = g(i, \boldsymbol{\beta})$. Assume that Y_j follows a nonhomogeneous Poisson process, the log-likelihood function of Y_j can be written as

$$L(\boldsymbol{Y}|\boldsymbol{\beta}) = \sum_{j=1}^{n} \left[Y_j \log \left(\sum_{i=1}^{j} g(i, \boldsymbol{\beta}) f_{ij} \right) - \sum_{i=1}^{j} g(i, \boldsymbol{\beta}) f_{ij} - \log Y_j! \right]. \tag{3}$$

The maximum likelihood estimate of $\boldsymbol{\beta}$ is obtained by maximizing $L(\boldsymbol{Y}|\boldsymbol{\beta})$ with respect of $\boldsymbol{\beta}$ using general numerical approaches such as the Newton-Raphson method or EM algorithm.[16] The variance of the estimate can be obtained using Fisher information or bootstrap method. As long as the infection rate function $g(i, \boldsymbol{\beta})$ is estimated, the number of infections and future disease cases can be estimated and projected.

The cumulative number of infections from time T_0 to time T_k can be estimated by

$$\hat{G}(T_k) = \int_{T_0}^{T_k} g(s, \hat{\boldsymbol{\beta}})ds \quad \text{(continuous time)} \tag{4}$$

$$= \sum_{i=0}^{k} g(i, \hat{\boldsymbol{\beta}}) \qquad \text{(discrete time)} \tag{5}$$

The variance of this estimate can be obtained by the delta method or by bootstrap method. Also note that the infection prevalence is defined to be the number of infected individuals who are alive. Thus, the estimate of the infection prevalence is $\hat{G}(T_k) - D(T_k)$, where $D(T_k)$ is the cumulative number of deaths during the same time interval.

The projection of disease incidence in a future time interval $[T_{l-1}, T_l)$ is obtained by projecting forward the number of individuals infected prior to the current time T_n, i.e.

$$\hat{A}(T_l) - \hat{A}(T_{l-1})$$

$$= \int_{T_0}^{T_n} g(s, \hat{\boldsymbol{\beta}})[F(T_l - s) - F(T_{l-1} - s)]ds \quad \text{(continuous time)} \tag{6}$$

$$= \sum_{i=0}^{n} g(i, \hat{\boldsymbol{\beta}}) f_{il} \qquad \text{(discrete time)}\,. \tag{7}$$

However, this estimate is a lower bound since it only considers the infected individuals prior to time T_n. To make an adjustment, the infections during time T_n and T_l need to be considered, that is, the following term needs to be added to the above projection,

$$\int_{T_n}^{T_l} g(s, \boldsymbol{\beta})[F(T_l - s) - F(T_{l-1} - s)]ds$$

in continuous time or

$$\sum_{i=n}^{l} g(i, \boldsymbol{\beta}) f_{il}$$

in discrete time. However, the future infection rate $g(s, \boldsymbol{\beta})$ or $g(i, \boldsymbol{\beta})$ is unknown. A guess or extrapolation of current infection rate is usually used.

A brief introduction on back-calculation can be found in Bacchetti[16] and a detailed description can be found in Brookmeyer and Gail.[13] Note that the above methods for projection of disease incidence or infection prevalence should be used with caution. There are many sources of uncertainty

in back-calculation methods. The first is the uncertainty in the incubation period distribution. The estimate of the incubation period distribution may be subject to errors and uncertainty of the designed epidemiological studies. The sensitivity analysis is usually used to evaluate these uncertainties. More details on the incubation period distribution can be found in the book (Chapter 4) by Brookmeyer and Gail.[13] The projection is also sensitive to the assumption of infection rate models, especially for the unknown future infection rates. Thus, the model of $g(s)$ needs to be chosen with care. Another problem is the reported disease incidence data. Different countries have different reporting systems for infectious diseases. Some of them may not be reliable. Reporting delay or underreporting occurs frequently. Some formal methods have been developed to account for the reporting uncertainty, see Harris[14] and Lawless and Sun.[15] Also note that the effect of immigration and emigration from one community (country) to another community (country) is not considered in above projection models. In summary, the back-calculation method only provides a rough (a lower bound) estimate or prediction for the disease incidence or infection prevalence.

2.2. *Model the natural history*

The natural history of a disease is the evolution of a disease in the absence of medical intervention. Today, however, most diseases are treated after they are diagnosed. The term "clinical course" is usually used to describe the natural history of a disease that has been affected by intervention. A broader definition of the term "natural history" may also include clinical course.

The endpoints of a natural history study may be dichotomous outcomes (such as death, relapse of a tumor or acquisition of AIDS following HIV infection, etc.), time-to-event (such as time to a clinical outcome occurs), or a repeated biomarker (such as CD4+ cell counts or HIV RNA copies in AIDS patients). To study the relationship between these endpoints and prognostic factors, standard statistical methods may be used. For example, logistic regression or tree-structured regression methods (see related chapter of this book) may be used to study dichotomous outcome endpoints. These methods are pretty standard, we omit the details here. To study the survival endpoints, a Kaplan-Meier curve or a product-limit estimator (life table) is widely used to describe the natural history. The popular proportional hazards model or Cox regression model can be used to study the relationship

between the survival endpoints and prognostic factors. Since neither the time of HIV infection nor the AIDS incidence can be observed exactly, the doubly censored or interval censored data need to be considered in this case. This problem motivated the development of new methods, see De Gruttola and Lagakos,[17] Kim *et al.*,[18], Jewell *et al.*,[19] Jewell[20] and Sun.[21] The detailed survival analysis methods can be found in related chapter of this book or other textbooks. For repeated measurement endpoints, statistical methods for longitudinal data have been developed in the past two decades. A good survey of these methods can be found in the book by Diggle *et al.*,[22] and others. Modeling biomarkers of HIV/AIDS such as CD4+ cell counts and HIV RNA copies (viral load) have been paid special attention in the last decade. Many new models and methods have been developed. In the following, we briefly introduce several new models and methods, but refer the readers to the original papers for details. Standard longitudinal data analysis methods can also be found in related chapter of this book.

In the early stage of HIV/AIDS research, CD4+ T cell count is the most important biomarker to study natural history of HIV infection and evaluate the treatment effects. Recently HIV RNA copies (viral load) became the new focus in HIV/AIDS research. But the methodology in modeling CD4+ T cell counts can be adopted to model viral load with minor modifications.

De Gruttola, Lange and Dafni *et al.*[23] proposed a linear mixed-effect model with errors-in-variables to model CD4+ T cell trajectory, i.e.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{a} + \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i\,, \quad i = 1, \ldots, n\,, \tag{8}$$

where design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are subject to measurement error since they depend on observed time measurements, $\boldsymbol{a}$ is population parameter, and $\boldsymbol{\beta}_i$ is subject-specific random effects with an i.i.d. normal distribution and is independent from $\boldsymbol{\varepsilon}_i$ which also follows an i.i.d. normal distribution. Taylor *et al.*[24] considered a linear mixed-effect model with within-subject covariance specified as an OU stochastic process, i.e.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{a} + \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{W}_i + \boldsymbol{\varepsilon}_i\,, \quad i = 1, \ldots, n\,, \tag{9}$$

where $\boldsymbol{W}_i$ is an OU process. They claimed that this model tracked CD4+ T cell data better compared to standard linear mixed-effect models.

To better track the nonlinearity of CD4+ T cells, some nonparametric and semiparametric models have been proposed. For example, Zeger and Diggle[25] introduced a semiparametric model,

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \mu(t_{ij}) + \varepsilon_i(t_{ij})\,, \tag{10}$$

where $\boldsymbol{x}_{ij}$ is a covariate matrix (prognostic factors) and $\mu(t_{ij})$ is an unknown smooth function of t. They proposed a back-fitting algorithm to fit the model, i.e. estimating $\boldsymbol{\beta}$ and fitting $\mu(t_{ij})$ (using kernel or other nonparametric regression methods) iteratively. See Zeger and Diggle[25] for details.

Nonparametric mixed-effects models have been proposed to model CD4+ T cell courses.[26,27] The basic idea is to decompose a population (cohort) of CD4+ T cell curves into two parts, a population effect and a subject-specific random effect, $y_i(t) = f(t) + h_i(t) + \varepsilon_i(t)$, where $f(t)$ and $h_i(t)$ denote the population curve and the subject-specific random effect curve respectively, and both of them are assumed to be smooth functions of t. Cubic B-spline method was proposed to fit $f(t)$ and $h_i(t)$. Let $B(t)$ be a vector of a cubic B-spline basis. Assume $f(t) = \alpha B(t)$ and $h_i(t) = \gamma_i B(t)$, then the CD4+ T cell model can be written as

$$y_i(t) = \alpha B(t) + \gamma_i B(t) + \varepsilon_i(t)\,. \tag{11}$$

This is a standard linear mixed-effects model by treating $B(t)$ as covariates, and α and γ_i as fixed and random effects respectively. The existing statistical packages such as SAS procedure MIXED or Splus lme function can be used to fit this model. Standard inference procedures for linear mixed-effects models are also available. Similarly, Wang and Taylor[28] also proposed a piecewise cubic polynomial model for CD4+ T cell changes and used the model to conduct inferences such as treatment comparisons.

Recently more flexible models such as functional linear models or varying-coefficient models[29–32] have been proposed. The model can be written as

$$Y_i(t_{ij}) = \boldsymbol{X}_i^T(t_{ij})\boldsymbol{\beta}(t_{ij}) + \boldsymbol{e}_i(t_{ij})\,, \tag{12}$$

where $\boldsymbol{\beta}(t_{ij})$ is a time-varying coefficient vector which is assumed to be a smooth function of t. Fan and Zhang[29,30] proposed a two-step procedure to fit the model. That is, for fixed time t, fit a standard linear regression model to obtain the raw estimates of $\boldsymbol{\beta}(t_{ij})$, and then smooth the raw estimates using one of the existing smoothing techniques. Hoover *et al.* and Wu *et al.*[31,32] proposed smoothing spline and local polynomials methods.

Hierarchical Bayes models have been introduced by Lange *et al.*[33] to model CD4+ T cell counts. This model is similar to a mixed-effects model, but in Bayes framework. De Gruttola and Tu[34] and Tsiatis *et al.*[35] also proposed a method for jointly modeling survival endpoints and longitudinal biomarkers. They modeled the longitudinal biomarkers (CD4+ T cell

counts) as a linear mixed-effect model and model survival data using a standard Cox model, and then construct a joint log-likelihood function of these two models. Thus, the likelihood-based method can be applied to the models.

Although above models are developed to model CD4+ T cell counts in AIDS research, the methodology is generally applicable to other similar repeatedly measured biomarker data for other infectious diseases. However, in most countries, especially for developed countries such as United States, patients with infectious diseases such as HIV/AIDS are mostly under active treatments. How to model the natural history or clinical course of infectious diseases under effective treatments is a great challenge, since the treatment may affect the changes of biomarkers and disease progression dramatically. Also note that there are many resources of biases and uncertainties in natural history studies. For examples, sampling or selection bias in study subject selection process, follow-up length bias due to study length limitations and long latent period of some infectious diseases such as AIDS, drop-out or missing data bias when the drop-out or missing pattern is not random. Another problem is that the time zero of a natural history may not be well-defined and exactly observed in a study, for instance, the exact time of HIV infection is difficult to obtain for some cohorts. See more discussions in Cnaan.[36] In summary, a careful design of a natural history study is necessary to eliminate or reduce these biases.

2.3. *Deterministic models for epidemic transmission*

A standard deterministic model for epidemic transmission is a compartmental model. For example, a general susceptible-infection-removal (SIR) compartment model can be written as

$$\dot{S} = \mu - \beta SI - d_S S\,, \tag{13}$$

$$\dot{I} = \beta SI - rI - d_I I\,, \tag{14}$$

where S and I represent the proportions of susceptible and infectious subjects in the population, and $\dot{S}$ and $\dot{I}$ denote their derivatives respectively. Parameter μ denotes the birth rate of susceptible subjects per time unit, β represents the infection rate when S and I are randomly interacted (mixed). Parameters d_S and d_I denote the death rates of susceptible and infectious subjects, r denotes the recovered (removal) rate of infectious subjects. The basic reproduction ratio is defined as

$$R_0 = \mu\beta/[d_S(r + d_I)]\,.$$

R_0 is an important summary measure of the infectiousness of a disease. This is the mean number of secondary cases generated by a single infective in a totally susceptible population. The higher the value of R_0, the more infectious the disease. If $R_0 \leq 1$, transmission of the infection cannot be sustained and will eventually die out.

If the infection has a latent stage with a proportion of E (during this stage, they are not infectious), a standard SEIR model is

$$\dot{S} = \mu - \beta SI - d_S S\,, \tag{15}$$

$$\dot{E} = \beta SI - \alpha E - d_E E\,, \tag{16}$$

$$\dot{I} = \alpha E - rI - d_I I\,, \tag{17}$$

where α is the transmission rate from latent to infectious, and d_E is the death rate of latently-infected subjects. For some infectious diseases, we may assume that $d_S = d_E = d_I = d$, but this may not be true in general. These compartment models are derived from a principal of mass action and homogeneously mixing pattern.

The age of subjects is another important factor in the epidemics of infectious diseases. The age-structured compartment model may be used to account for age effects. Here is a simple example,

$$\frac{\partial S}{\partial t} + \frac{\partial S}{\partial a} = \mu - \lambda S - d_S S\,, \tag{18}$$

$$\frac{\partial I}{\partial t} + \frac{\partial I}{\partial a} = \lambda S - rI - d_I I\,, \tag{19}$$

$$\frac{\partial R}{\partial t} + \frac{\partial R}{\partial a} = rI - d_R R\,, \tag{20}$$

where a is the age of subjects, R is the proportion of recovered (removal) of subjects in the population. Parameter λ is the so-called force of infection (age-specific hazard rate of infection) which is a function of time t, and can be defined by

$$\lambda(t) = \int_0^\infty k(a')I(t,a')da'\,,$$

where $k(a')$ is a kernel function. The partial differential equation system (18)–(20) can be solved numerically with appropriate initial conditions.

Note that the above models are very general for infectious diseases. But for a particular disease, these models may need to be modified to accommodate the special feature or characteristics of the disease. For example, HIV-infected patients cannot be cured or recovered from infection with

current treatments, instead they may progress to AIDS or death. Thus, in the differential equation of infectious population (I), r is the rate of progression to AIDS. An additional equation of AIDS cases, $\dot{A} = rI - d_A A$, may be added.

Hepatitis B virus (HBV) infection is another example of infectious diseases with its transmission to be characterized by a model of five compartments. A community population can be divided into five compartments: (1) susceptible $S(a,t)$; (2) latent period (the time interval from infection to development of infectiousness), $L(a,t)$; (3) temporary HBV carriers, $T(a,t)$; (4) chronic HBV carriers $C(a,t)$; and (5) immune $I(a,t)$.[37,38] Here "a" represents the age and "t" represents the length of follow-up. Among the five stages, compartments 3 and 4 are infectious. In this model, birth rate is considered as a constant; age specific death rates are collected from death notification systems. The immune status is assumed to be life-long and newborns are assumed susceptible. For simplicity of modeling, the rare intrauterine HBV infection,[39,40] the short period of newborn maternal antibody and the sex differences ignored. The model parameters are defined as the following: $\lambda(a,t)$ is the force of infection; α is the rate of transition from latent period to temporary HBV viremia; $\beta(a)$ is the risk of transient viremia progressing to chronic HBV carriage; ε is the rate of transition from temporary HBV viremia to immune per time unit; $\nu(a)$ is the rate of HBV clearance in chronic HBV carriers; $\tau(a)$ is the mortality rate of HBV related diseases; $\mu(a)$ is the age-specific mortality rate of non HBV related diseases; $V_c(a,t)$ is the effectiveness of hepatitis B vaccine immunization. Then the age-structured compartment model for HBV can be written as,

$$\frac{\partial S(a,t)}{\partial a} + \frac{\partial S(a,t)}{\partial t} = [\lambda(a,t) + V_c(a,t) + \mu(a)]S(a,t)\,, \tag{21}$$

$$\frac{\partial L(a,t)}{\partial a} + \frac{\partial L(a,t)}{\partial t} = \lambda(a,t)S(a,t) - [\alpha + \mu(a)]L(a,t)\,, \tag{22}$$

$$\frac{\partial T(a,t)}{\partial a} + \frac{\partial T(a,t)}{\partial t} = \alpha L(a,t) - [\beta(a) + \varepsilon + \mu(a)]T(a,t)\,, \tag{23}$$

$$\frac{\partial C(a,t)}{\partial a} + \frac{\partial C(a,t)}{\partial t} = \beta(a)T(a,t) - [\nu(a) + \tau(a) + \mu(a)]C(a,t)\,, \tag{24}$$

$$\begin{aligned}\frac{\partial I(a,t)}{\partial a} + \frac{\partial I(a,t)}{\partial t} &= V_c(a,t)S(a,t) + \varepsilon T(a,t) \\ &\quad + \nu(a)C(a,t) - \mu(a)I(a,t)\,.\end{aligned} \tag{25}$$

After all the parameters were estimated from the data of epidemiological studies,[41–43] the probabilities or variables, $S(a,t), L(a,t), T(a,t), C(a,t)$ and $I(a,t)$ at age a and time t in the model can be calculated by the integral of the partial differential equations. These estimates can describe the dynamics of HBV transmission in the population at the pre-vaccination period or predict the trend with different vaccination coverage $V_c(a,t)$ in the population. Detailed information about HBV modeling and parameter estimation can be found in references.[44,45]

The deterministic models for other diseases such as Malaria and Helminths can be found in Heesterbeek and Roberts.[9] More details on deterministic compartment models can be found in the books by Bailey,[46] Becker,[47] Anderson and May[48] and Daley and Gani.[49]

2.4. *Stochastic models for epidemic transmission*

2.4.1. *Branching processes*

In the cases in which it is reasonable to assume an unlimited pool of susceptibles, for instance during the initial stage of an epidemic, a branching process can be used to model the spread of infection. Let Y_0 denote an initial number of infectives at generation 0. These Y_0 individuals infect Y_1 individuals as the next generation. In turn, these Y_1 individuals infect Y_2 individuals as the third generation, and so on. Let Z denote the number of infections directly caused by one individual, which is a random variable with a mean of μ, variance σ^2 and a probability density function of $g(z)$. Thus, for each i, $Y_i = Z_1 + \cdots + Z_{Y_{i-1}}$, where Z_j are independent variables with density $g(z)$.

Harris[50] proposed a nonparametric maximum likelihood estimator for μ:

$$\hat{\mu} = \sum_{i=1}^{k} Y_i \Big/ \sum_{i=1}^{k} Y_{i-1}\,.$$

The properties of this estimator are discussed by Keiding.[51] Becker[52] suggested an alternative estimator:

$$\hat{\mu} = \begin{cases} (Y_k/Y_0)^{1/k} & \text{if } Y_k > 0\,, \\ 1 & \text{if } Y_k = 0\,. \end{cases}$$

Note that the expected number of infections caused by one individual, μ, plays the same role as the basic reproduction ratio (R_0) in deterministic models. It can be shown that if $\mu \leq 1$, the process will become extinct with

probability one.[50] Inferences for branching processes are usually conditional on extinction and non-extinction. Heyde[53] suggested a Bayesian approach which allows the extinction ($\mu \leq 1$) and non-extinction ($\mu > 1$) to be treated without distinction. Becker[47] gave several applications of branching processes to smallpox epidemics.

2.4.2. *Chain binomial models*

The branching process is unsuitable for the epidemics within a small communities such as households. In this context, chain binomial models are more appropriate. The chain binomial model, or refer to the Reed-Frost epidemic model was introduced by the biostatistician Lowell J. Reed and the epidemiologist Wade Hampton Frost around 1930, as a teaching tool at Johns Hopkins University. Although they did not publish their results formally, their model was introduced in later publications.[54,55]

Consider a fixed number of community (such as household, sexual partners, needle sharing group) with n individuals. At generation k there are X_k susceptibles exposed to Y_k infectives. The distribution of the number of infectives in the next generation, Y_{k+1}, conditional on X_k and Y_k, is binomial:

$$\Pr(Y_{k+1} = z | X_k = x, Y_k = y) = \frac{x!}{z!(x-z)!} p_k^z (1-p_k)^{x-z},$$

where p_k is the probability that a susceptible of generation k will acquire infection from one of the y_k infectives. The parameter p_k can be modeled under different assumptions. One assumption due to Reed and Frost[54] is that contacts with infectives occur independently, so that

$$p_k = 1 - (1-\pi)^{y_k},$$

where π is the probability of infection for the contact with the infectives. An alternative assumption, due to Greenwood,[7] is that the probability of infection does not depend on the number of infectives that the susceptible is exposed to, then

$$p_k = \begin{cases} \pi & \text{if } y_k > 1, \\ 0 & \text{otherwise}. \end{cases}$$

Under this assumption, it is usually called the Greenwood chain binomial model. More complicated models for p_k can be developed for complicated transmission mechanisms such as HIV infection. Some other extensions to the Reed-Frost model can be found in Longini and Koopman.[56]

Inference for chain binomial models is usually based on likelihood methods. See Bailey,[46] Becker,[47] Longini and Koopman,[56] Longini *et al.*[57] and Saunders.[58] A brief introduction can be found in Longini.[59] For an updated review, see Becker and Britton.[60]

2.4.3. *Stochastic compartment models*

The stochastic version of the SIR model (13) is useful to capture stochastic features of epidemics in a small population or in the early stage of the epidemics. Consider $S(t)$ and $I(t)$ as the number (rather than the proportion) of susceptibles and infectives respectively. Then the transition probabilities in a short time interval $(t, t+\delta t)$ are

$$\Pr[S(t+\delta t) = S(t) - 1; I(t+\delta t) = I(t) + 1] = \beta S(t) I(t) \delta t\,, \tag{26}$$

$$\Pr[S(t+\delta t) = S(t); I(t+\delta t) = I(t) - 1] = (r + d_I) I(t) \delta t\,, \tag{27}$$

$$\Pr[S(t+\delta t) = S(t) + 1; I(t+\delta t) = I(t)] = \mu \delta t\,, \tag{28}$$

$$\Pr[S(t+\delta t) = S(t) - 1; I(t+\delta t) = I(t)] = d_S S(t) \delta t\,. \tag{29}$$

The solution of this stochastic systems is not straightforward. Monte Carlo methods may be used to solve it.[61–63]

Tan and Hsu[64] proposed a stochastic SEIR model (including a latent stage of infection) for AIDS epidemics. Recently, Wu and Tan[65] suggested a multiple stage stochastic models (the chain multinomial model) for AIDS epidemics in a homosexual population. Here we briefly introduce this model.

Let $S(t), I_r(t)$, and $A(t)$ denote the numbers of susceptible people, people at rth infection stage $(r = 1, 2, \ldots, k)$ and people on set of AIDS at time t respectively. Then we are entertaining a $(k+2)$-dimensional discrete stochastic process $\boldsymbol{X}(t) = [S(t), I_1(t), I_2(t), \ldots, I_k(t), A(t)]^T$, where $[\cdot]^T$ denote the transpose of a vector or matrix. To formulate the dynamic model (the chain multinomial model) for this process, let $\alpha_S(t)$ denote the conditional probability of $S \to I_1$ given $\boldsymbol{X}(t)$ during $[t, t+1)$ and give the other notations of the transition probabilities and numbers of various transitions of the HIV epidemic in Table 1.

By using the chain multinomial model, we obtain the following stochastic difference equations:

$$S(t+1) = R_S(t) + S(t) - F_S(t) - D_S(t)\,, \tag{30}$$

Table 1. Notation for transitions of the HIV epidemic in homosexual populations during $[t, t+1)$.

Transition	Transition probability	Transition numbers
Immigration $\longrightarrow S$	$\mu_S(t)$	$R_S(t)$
Immigration $\longrightarrow I_r$	$\mu_r(t)$	$R_{I_r}(t)$
$S \longrightarrow I_1$	$\alpha_S(t)$	$F_S(t)$
$I_r \longrightarrow I_{r+1},\ r = 1, 2, \ldots, k-1$	$\alpha_r(t)$	$F_{I_r}(t)$
$I_1 \longrightarrow S$	$\beta_1(t) = 0$	$B_{I_1}(t) = 0$
$I_r \longrightarrow I_{r-1},\ r = 2, 3, \ldots, k$	$\beta_r(t)$	$B_{I_r}(t)$
$A \longrightarrow I_k$	$\beta_{k+1}(t) = 0$	$B_A(t) = 0$
$I_r \longrightarrow A,\ r = 1, 2, \ldots, k$	$\omega_r(t)$	$A_{I_r}(t)$
$S \longrightarrow$ Death	$d_S(t)$	$D_S(t)$
$I_r \longrightarrow$ Death	$d_r(t)$	$D_{I_r}(t)$
$A \longrightarrow$ Death	$d_A(t)$	$D_A(t)$

$$I_r(t+1) = R_{I_r}(t) + F_{I_{r-1}}(t) + B_{I_{r+1}}(t) + I_r(t) - \left[F_{I_r}(t) + B_{I_r} + A_{I_r}(t) + D_{I_r}(t)\right], \tag{31}$$

$$A(t+1) = \sum_{r=1}^{k} A_{I_r}(t) + A(t) - D_A(t), \tag{32}$$

where $r = 1, 2, \ldots, k$, and $F_{I_0}(t) = F_S(t)$, $F_{I_k}(t) = 0$. The distributional properties of the quantities in the equations are listed as follows:

- $R_S(t) \sim$ Binomial $[S(t), \mu_S(t)]$, independent of $F_S(t)$ and $D_S(t)$.
- $[F_S(t), D_S(t)]|\boldsymbol{X}(t) \sim$ Multinomial $[S(t); \alpha_S(t), d_S(t)]$.
- $R_{I_r}(t)|I_r(t) \sim$ Binomial $[I_r(t); \mu_r(t)]$, independent of $F_{I_r}(t), B_{I_r}(t), A_{I_r}(t)$ and $D_{I_r}(t)$.
- $[F_{I_1}(t), A_{I_1}(t), D_{I_1}(t)]|\boldsymbol{X}(t) \sim$ Multinomial $[I_1(t); \alpha_1(t), \omega_1(t), d_1(t)]$.
- $[F_{I_r}(t), B_{I_r}(t), A_{I_r}(t), D_{I_r}(t)]|\boldsymbol{X}(t) \sim$ Multinomial $[I_r(t); \alpha_r(t), \beta_r(t), \omega_r(t), d_r(t)]$, for $r = 2, \ldots, k-1$.
- $[B_{I_k}(t), A_{I_k}(t), D_{I_k}(t)]|\boldsymbol{X}(t) \sim$ Multinomial $[I_k(t); \beta_k(t), \omega_k(t), d_k(t)]$.
- $D_A(t)|A(t) \sim$ Binomial $[A(t), d_A(t)]$.

Equations (30)–(32) provide an avenue for computing the probability distributions of $\boldsymbol{X}(t)$. Although the exact probability distributions of $\boldsymbol{X}(t)$ are quite complicated, one may use these equations to derive equations for the means, the variances and higher cumulants of $\boldsymbol{X}(t)$ as well as other results. Wu and Tan[65] also proposed using a state-space model to

approximate the above stochastic model, and then Kalman filter can be used for estimation and projections. See Wu and Tan[65] for details.

3. Viral Dynamic Models

Recently a great attention has been paid for modeling interaction and dynamics of virus and immune systems at cellular level within a host. It brought up a breakthrough in studying pathogenesis of HIV, HBV and HCV infections. In this section, we briefly introduce the basic models and their extensions, and summarize the important results obtained by applying these models to clinical data. Some statistical methods for parameter estimation will be briefly introduced.

3.1. *HIV dynamics*

Modeling HIV dynamics within a host can be traced back to 1980s.[66–68] In the early stage of HIV modeling, the focus is to understand the mechanism and pathogenesis of HIV infection and antiviral drug action using computer simulations based on the developed models. When the simplified version of the complicated simulation models were successfully applied to the clinical data in the last several years,[69–72] it has led to a new understanding of the pathogenesis of HIV infection. Mathematical models and statistical methods played an important role in this breakthrough. Here we briefly introduce the models and the results.

In the seminar papers, Ho *et al.*[69] and Wei *et al.*[70] proposed simple compartment models (one or two compartments) for their clinical data of plasma HIV viral load (the number of RNA copies) in HIV-1-infected patients treated with potent antiviral agents. In Ho *et al.*,[69] a simple one-compartment model, $\frac{d}{dt}V = P - cV$, was proposed, where V denotes the concentration of virus (measured by the number of HIV RNA copies per ml plasma), P denotes the production rate of virus, and c denotes the clearance rate of virus. If we assume that the antiviral treatment is perfect, or $P = 0$ after initiation of a potent antiviral treatment, the solution to above ordinary differential equation is $V = V_0 e^{-ct}$, where V_0 is the initial viral concentration. When we have repeated measurements of V on individual patients, we can fit a nonlinear model, $Y(t) = V_0 e^{-ct} + \varepsilon$ or a linear model in a log scale, $Y(t) = \log(V_0) - ct + \varepsilon$, to obtain the parameter estimate of c. The mean life-span or half-life of HIV can be estimated by $1/c$ and $\ln 2/c$ respectively. Ho *et al.*[69] applied this simple method to 20 HIV-infected patients, and they obtained that the half-life of HIV (in fact, it is the

half-life of productively infected cells) is 2.1 ± 0.4 days with a range of 1.3 to 3.3 days. Wei *et al.*[70] obtained similar results. This estimated rapid turnover rate of HIV virus (or infected cells) has important implications for HIV therapy and pathogenesis. One of the implications is that the rapid turnover of HIV may generate viral diversity and increase the opportunities for viral escape from antiviral agents. This motivated the idea of the therapy with combination of several antiviral agents (or so-called "cocktail" therapy).

To refine the estimate of viral replication, Perelson *et al.*[71] considered a more complicated compartment model when HIV infected patients are treated with more potent protease inhibitor (PI) antiviral agents. The mechanism of the PI drug antiviral action is to block the replication (generation) of infectious virus. Under the assumption of perfect PI drug treatment, the model can be written as

$$\frac{dT^*}{dt} = kV_I T - \delta T^*\,,$$

$$\frac{dV_I}{dt} = -cV_I,$$

$$\frac{dV_{NI}}{dt} = N\delta T^* - cV_{NI}\,,$$

where T represents the concentration of uninfected CD4+ T cells; T^* denotes the concentration of productively infected T cells; V_I denotes the concentration of noninfectious virions; V_{NI} denotes the concentration of noninfectious virions; c denotes the clearance rate of virus; δ denotes the clearance rate of infected cells; k is the infection rate. A closed-form solution to above differential equations under the assumption of constant T (it is reasonable at initial stage of infection) can be obtained:

$$V(t) = V_I(t) + V_{NI}(t) = V_0 \exp(-ct) + \frac{cV_0}{c-\delta} \times \left\{ \frac{c}{c-\delta}[\exp(-\delta t) - \exp(-ct)] - \delta t \exp(-ct) \right\}.$$

When frequent repeated measurement data on $V(t)$ are available, a nonlinear model, $Y(t) = V(t) + \varepsilon$, can be fitted to obtain the estimates of important parameters such as clearance rate of virus (c) and infected cells (δ). Perelson *et al.*[71] applied this method to the data from 5 HIV infected individuals, and obtained that the refined estimate of half-life of free HIV is 0.24 ± 0.06 days (about 6 hours in average) which is much more rapid than

previous estimate in Ho *et al.*[69] and Wei *et al.*[70] The estimated half-life of infected cells is 1.55 ± 0.57 days.

Furthermore, Perelson *et al.*[72] developed a compartment model for the observed biphasic viral load data. They speculated that the first phase is due to viral replication from productively infected cells such as CD4+ T cells, and the second phase as latent or long-lived infected cells such as macrophages or dendritic cells. Based on clinical data, Perelson *et al.*[72] estimated that the half-life of short-lived productively infected cells is about 1.1 ± 0.4 days, for long-lived infected cells is 14.1 ± 7.5 days, and for latently infected cells is 8.5 ± 4.0 days. Using their model and the estimated results, they predicted that it might need 2.3–3.1 years to eliminate the HIV virus by the potent antiviral therapies, although later it was shown that this estimate was too optimistic.

Wu and Ding[73] recently proposed a unified approach for modeling observed HIV dynamic data. First Wu and Ding[73] proposed a comprehensive mathematical model for HIV dynamics considering all the potential cell and virus compartments: (1) uninfected target cells, such as T cells, macrophages, lymphoid mononuclear cells (MNCs), and tissue langerhans cells, which are possible targets of HIV-1 infection; (2) mysterious infected cells, cells other than T cells, such as tissue langerhans cells and microglial cells whose behavior is not completely known so far; (3) long-lived infected cells, such as macrophages, that are chronically infected and long-lived; (4) latently infected cells, infected cells that contain the provirus but are not producing virus immediately, and only start to produce virus when activated; (5) productively infected cells, infected cells which are actively producing virus; (6) infectious virus, virus that are functional and capable of infecting target cells; (7) noninfectious virus, virus that are dysfunctional and cannot infect target cells. We denote the concentration of the variety of these cells and virus by $T, T_m, T_s, T_l, T_p, V_I$, and V_{NI} respectively.

Without the intervention of antiviral treatment, the uninfected target cells may either decrease due to HIV infection or be in an equilibrium state due to the balancing between the regeneration and proliferation of uninfected target cells and HIV infection. Some uninfected target cells (T) are infected by infectious virus (V_I) and may become mysterious infected cells (T_m), long-lived infected cells (T_s), latently infected cells (T_l) or productively infected cells (T_p) with proportions of $\alpha_m k V_I, \alpha_s k V_I, \alpha_l k V_I$, and $\alpha_p k V_I$ respectively, where $\alpha_m + \alpha_s + \alpha_l + \alpha_p = 1$. The latently infected T cells may be stimulated to become productively infected cells with a rate of

δ_l. The infected cells, T_m, T_s and T_p, are killed by HIV at the rates of δ_m, δ_s and δ_p respectively after producing an average of N virions per cell during their lifetimes. The infected cells, T_m, T_s and T_l may also die at the rates of μ_m, μ_s and μ_l respectively without producing virus. We assume that the proportion of noninfectious virus produced by infected cells is η without the intervention of protease inhibitor (PI) antiviral drugs. The elimination rates for infectious virus and noninfectious virus are assumed to be the same, say c.

We assume that the antiviral therapy consists of one or more protease inhibitor (PI) drugs and reverse transcriptase inhibitor (RTI) drugs. We model the effect of RTI drugs by reducing the infection rate from k_0 to $(1-\gamma)k_0$, where $0 \leq \gamma \leq 1$. Parameter γ reflects the RTI drug efficacy. If $\gamma = 0$, the RTI drugs have no effect; if $\gamma = 1$, the RTI drugs are perfect and completely block HIV infection. The PI drugs are assumed to be so potent that the production of infectious virions is almost blocked except for a small fraction. To account for some compartments where the PI drugs cannot reach and some persistent virus that the PI drugs cannot completely block the production, we consider an additional virus production term with a constant (average) rate, P, in the model. If only a small fraction of persistent virus can escape from the attack of PI drugs, it may be considered as a Poisson process, and thus also be modeled by a constant production in a deterministic model. Thus after initiation of combination treatment of PI and RTI drugs, the HIV dynamic model can be written as,

$$\begin{aligned}
\frac{d}{dt}T_m &= (1-\gamma)\alpha_m k_0 T V_I - \delta_m T_m - \mu_m T_m\,,\\
\frac{d}{dt}T_s &= (1-\gamma)\alpha_s k_0 T V_I - \delta_s T_s - \mu_s T_s\,,\\
\frac{d}{dt}T_l &= (1-\gamma)\alpha_l k_0 T V_I - \delta_l T_l - \mu_l T_l\,,\\
\frac{d}{dt}T_p &= (1-\gamma)\alpha_p k_0 T V_I + \delta_l T_l - \delta_p T_p\,,\\
\frac{d}{dt}V_I &= (1-\eta)P - cV_I\,,\\
\frac{d}{dt}V_{NI} &= \eta P + N\delta_m T_m + N\delta_s T_s + N\delta_p T_p - cV_{NI}\,.
\end{aligned} \tag{33}$$

where $\alpha_m + \alpha_s + \alpha_l + \alpha_p = 1$. Under the assumption of constant T and perfect treatments, then the system of Eq. (33) can be solved analytically and the final solution for the total virus $V = V_I + V_{NI}$ has a form of (see

Appendix in Wu and Ding[73]),

$$V(t) = P_0 + P_1 e^{-\lambda_1 t} + P_2 e^{-\lambda_2 t} + P_3 e^{-\lambda_3 t} + P_4 e^{-\lambda_4 t} + (P_5 + P_6 t) e^{-\lambda_5 t} + P_7 e^{-\lambda_6 t} + P_8 e^{-\lambda_7 t}, \quad (34)$$

where $P_i, i = 0, \ldots, 8$ are functions of model parameters and $\lambda_1 = \delta_p, \lambda_2 = \delta_m + \mu_m, \lambda_3 = \delta_s + \mu_s, \lambda_4 = \delta_l + \mu_l, \lambda_5 = c, \lambda_6 = r$, and $\lambda_7 = c + r$. At time $t = 0$, $V(0) = \sum_{i \neq 6} P_i$. Parameter P_i represents the initial viral production rate, and Parameter λ_i represents the exponential decay rate of virus due to the corresponding compartment. This model is too complicated (too many parameters) to be used in practice. Wu and Ding[73] suggested to use simplified version of the model based on available data. For example, if only the biphasic data are available, a bi-exponential model, $V(t) = P_1 e^{-\delta_p t} + P_2 e^{-\lambda_l t}$ or a one-exponential plus a constant model, $V(t) = P_0 + P_1 e^{-\delta_p t}$ may be used.[73] To fit the model with sparse individual data, Wu *et al.*[74] and Wu and Ding[73] also proposed using nonlinear mixed-effect model approach.[75,76] The two-stage nonlinear mixed-effect model is briefly introduced as follows.

Stage 1. Intra-patient variation in viral load measurement:

$$y_{ij} = \log(V(t_{ij}, \boldsymbol{\beta}_i)) + e_{ij}, \quad \boldsymbol{e}_i | \boldsymbol{\beta}_i \sim (\mathbf{0}, \boldsymbol{R}_\mathbf{i}(\boldsymbol{\beta}_\mathbf{i}, \boldsymbol{\xi})), \quad (35)$$

where y_{ij} is the log-transform of the total viral load measurement for the ith patient and at the jth time point t_{ij}, $i = 1, \ldots, m$; $j = 1, \ldots, n_i$. The log-transformation of raw data is used to stabilize the variance (it is also more normally distributed). The function $V(t_{ij}, \boldsymbol{\beta}_i)$ is a nonlinear function of treatment time t which may be selected based on the available data and model assumptions. See Wu and Ding[73] for details.

Stage 2. Inter-patient variation:

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i. \quad (36)$$

Population parameters are $\boldsymbol{\beta}$, and random effects are $\boldsymbol{b}_i \sim (\mathbf{0}, \boldsymbol{D})$.

More detailed inferences regarding the nonlinear mixed-effect model can be found in the books by Davidian and Giltinan[75] and Vonesh and Chinchilli.[76]

A comparison for viral dynamic model-fitting procedures can be found in Ding and Wu.[77] Application to more adult HIV-1 infected patients and pediatric patients can be found in Wu *et al.*[78] and Luzuriaga *et al.*[79] Recently, Ding and Wu[80] and others have suggested using viral dynamics (decay rates) to evaluate the potency of antiviral therapies. Statistical

methods has been proposed to implement this idea by Ding and Wu.[81] Modeling drug resistance can be found in Nowak *et al.*[82] and others. For a good review of viral dynamic modeling and their extensions, see Perelson and Nelson[83] and Nowak and May.[84]

3.2. *Hepatitis virus dynamics*

Following the success of HIV dynamics modeling, similar studies have been done for hepatitis B and C virus (HBV and HCV). For example, Nowak *et al.*[85] proposed a simple compartment model for HBV dynamics. Let X, Y and V be uninfected cells, infected cells and free HBV virus respectively, then a mathematical model for HBV dynamics is

$$\dot{X} = \lambda - \beta XV - d_x X\,, \tag{37}$$

$$\dot{Y} = \beta XV - d_y Y\,, \tag{38}$$

$$\dot{V} = \alpha Y - d_v V\,, \tag{39}$$

where λ is the production rate of susceptible cells. Uninfected cells die at a rate of $d_x X$ and become infected at rate βXV. Infected cells are produced at rate βXV and die at rate $d_y Y$. Free virions are produced from infected cells at rate αY and are removed at rate $d_v V$. Nowak *et al.*[85] assumed that the potent treatment (the reverse transcriptase inhibitor, lamivudine) is perfect, i.e. $\alpha = \beta = 0$. Thus, $V(t) = V_0 \exp(-d_v t)$ and $Y(t) = Y_0 \exp(-d_y t)$. If the treatment is not perfect (more likely in reality), a model for free virion is $V(t) = V_0[1-r+r\exp(-d_v t)]$, where r is an efficacy parameter, which can be estimated from the viral load data. Nowak *et al.*[85] fitted a clinical data to these models, and found that the half-life of HBV free virions is about 1 day, the half-life of infected cells ranges from 10 to 100 days in different patients.

Many researchers have studied HCV dynamics using models similar to HBV under the antiviral treatment with Interferon-α.[86–90] The recent report from Neumann[90] showed that the half-life of HCV free virions was, on average, 2.7 hours, the half-life of infected cells was 1.7 to 70 days. All modeling techniques and statistical methods for HIV dynamics are applicable to both HBV and HCV with minor modifications.

4. Intervention and Prevention

Intervention and prevention measures are critical to stop the epidemic of infectious diseases. To evaluate the effectiveness of the intervention

and prevention methods, clinical trials are usually conducted. Since other chapters have addressed the general methods of clinical trials, here we only emphasize some special features of clinical trials for infectious diseases, in particular, AIDS clinical trials.

4.1. *Medical intervention*

The general design of clinical trials can be found in clinical trial textbooks.[91–94] One of the most important issues in clinical trials is the selection of an endpoint which can be used to measure the effectiveness of interventions such as medical treatments.

In general, an endpoint of a clinical trial should possess the following properties: (i) relevant to the treatment effectiveness and easy to interpret; (ii) clinically apparent and easy to diagnose (or measure); (iii) sensitive to treatment differences. An earlier discussion on the choice of an endpoint for AIDS clinical trials can be found in Amato and Lagakos.[95] In the early stage of AIDS clinical trials (before 1994), the time to progression to AIDS or survival (time to death) was used. Since many different types of events or symptoms were defined as AIDS, the endpoints of progression to AIDS are often referred to as "combined endpoints". After reviewing and comparing the clinical data, Neaton *et al.*[96] recommended that the survival, instead of combined endpoints, be a preferred primary endpoint of antiviral trials. However, due to long incubation period of AIDS, the trial requires a long-term follow-up if the survival was used as the endpoint. Recently, surrogate markers such as viral load (HIV RNA copies) or CD4+ cell counts have been proposed and used as endpoints after validation of these markers (it is out of scope of this chapter to discuss how to validate a surrogate marker, see Prentice[97] for details).

The commonly used primary endpoints in recent AIDS clinical trials are viral load based endpoints which include: (i) the magnitude of reduction in viral load (HIV-1 RNA level) from baseline to a prespecified primary follow-up time (e.g. Week 24 or Week 48); (ii) the proportion of patients having viral load below the limit of quantification of the assay being used at the primary follow-up time; (iii) the durability based on the time-to-virologic-failure (the time until plasma viral load becomes detectable again). Marschner *et al.*[98] studied the first endpoint, the magnitude of reduction in viral load, and proposed statistical methods to deal with censored (below detection limit) viral load measurements. They argued that the dichotomous endpoint (the proportion of patients having viral load below the limit

of detection) was more straightforward and less subject to bias than the analysis of the magnitude of viral load reduction. Standard methods of analysis for binary data would be appropriate. However, the dichotomous endpoint may lead to a lower power to detect the treatment difference than the endpoint of the actual magnitude of viral load change. By classifying virologic responses as either successes or failures, information is lost regarding the degree of virologic response. However, the endpoint based on the magnitude of viral load change involves complicated censored data problem due to the limit of detection of viral load assays, which may be subject to bias, see Marschner *et al.*[98] and Hughes.[99]

Gilbert *et al.*[100] studied the time-to-virologic-failure endpoints. They recommended the endpoint of time-to-virologic-failure from randomization due to its advantages in flexibility and sample size. They argued that the time-to-failure endpoint is generally more powerful than the binary endpoint, and it is flexible for evaluating covariate effects and for extending the study by prolonging the follow-up period. Also the interpretation of time-to-failure endpoint is more close to clinical practice than that of a binary endpoint, since physicians monitor viral load levels in patients over time for treatment managements.

Other endpoints such as the area under the curve (AUC) of viral load change, time-to-below-detection in viral load, and viral dynamic parameters (viral decay rates) were also suggested, but not widely used in large AIDS clinical trials. For the comparison of some of these endpoints, see Weinberg and Lagakos.[101] To evaluate the short-term potency of antiviral therapies using viral dynamic parameters, see Ding and Wu.[80,81]

Although the general clinical trial design methods can be used in most AIDS clinical studies, some new issues have arised from the complicated treatments for AIDS patients. See De Gruttola *et al.*[102] and Hughes[99] for some discussions on the design issues in AIDS clinical trials. Also note that the computer-assisted design techniques or clinical trial simulations (CTS)[103] may be useful for designing the complicated clinical trials.

Successful medical interventions will result in the change of epidemic patterns.[104–106] How to evaluate the epidemic trends under medical interventions is challenging. In this regards, computer simulations based on the epidemic models with considering treatment effects will be helpful.

Clinical studies on HBV and HCV are currently very active. HBV patients are treated with antiviral agents such as lamivudine and famciclovir,[85] and HCV patients are treated with interferon (IFN) and

ribavirin therapy.[90] New anti-HBV and anti-HCV agents are under development, some of them are already in the stage of clinical trials. The methods for studying anti-HIV medical interventions are generally applicable to HBV and HCV.

4.2. *Prevention*

Preventive interventions are widely used to stop or reduce the epidemic of infectious diseases. Prevention measures include "lifestyle" maneuvers such as the change of social behavior to reduce the exposure risk to infectives, and modifications of sexual behavior for sexually transmitted infectious diseases via public education and advertisement. Another effective prevention measure is to prevent the infectious diseases by vaccination. To evaluate the effectiveness of these prevention measures is very challenging in terms of designing and implementing a prevention study due to the high cost and long-term follow-up.

4.2.1. *Prevention trials*

A prevention trial is different from standard randomized clinical trials. The goal of prevention trials for infectious diseases is to evaluate the effectiveness of prevention measures to protect individuals from infections. If the infection rate in a community or population is low for a particular infectious disease, a prevention trial generally needs a large sample size with tens of thousands of subjects. If it takes time for the prevention measures to start to work or for an individual to acquire the infection via exposure to infectives, it may require long duration such as several years of follow-up to evaluate the effectiveness of the prevention programs. Thus, the prevention study is logistically challenging with high cost.

Traditionally an observation study, for example, a cohort or community study, is employed to evaluate the prevention programs (historical control may be used in this case). When prevention measures are taken in a cohort or a community, the infection rate may be evaluated within a prespecified time period, and then compare this rate with a historical infection rate in this cohort or community. This kind of observational studies are subject to several problems such as within-subject variations, measurement errors, confounding factors, and adherence to prevention measures. For example, to evaluate the promotion of condom use and education of safe sexual practices among homosexual men community to prevent HIV infection, not all subjects in the study adhere to the condom use during the study, and their

sexual behavior may change during the study period (within-subject variation). Also the epidemics of an infectious disease under the prevention may confound with other factors. The causal inferences for epidemiologic associations with corresponding preventive strategies are also subject to measurement errors.

Ideally it is most effective and informative to conduct a randomized and controlled prevention trial to evaluate prevention programs. This would avoid the difficulties in observational studies. The study subjects may be selected from a high risk community of an infectious disease to reduce the sample size and the cost in a randomized prevention trial. For example, homosexual men, IV drug users, or sexual workers are high risk communities that are targeted for prevention from HIV infection. The design, conduct, monitoring, and analysis for randomized prevention trials are similar to standard therapeutic clinical trials. However, in some cases, it may not be ethical and practical to conduct a randomized prevention trial. For example, the needle sharing is a confirmed cause of HIV infection among IV drug users and safe sex such as condom use may reduce the risk of sexually transmitted infectious diseases, education or advertisement to the public on these knowledge is a good prevention measure. However, it is not ethical and practical to randomize high-risk subjects into the arm without accessing the education or advertisement of needle sharing risk and safe sex. For more discussion on prevention studies, see Prentice[107] and Jacobs.[108]

4.2.2. *Vaccine studies*

Vaccine studies are designed to evaluate different effects of vaccination during different stages of vaccine development. The major purpose of vaccines is to protect the vaccinated person against infection or reduce the severity or risk of disease progression after being infected. Successful vaccination can reduce person-to-person transmission and change the pattern of epidemics of an infectious disease within a population by reducing the infectiousness of an vaccinated person or by preventing individuals from infection. Thus, vaccination is an important intervention tool to the epidemics of infectious diseases and has great contributions to the public health.

It is important to understand the biological background of a vaccine in order to evaluate its efficacy. A vaccine is usually composed of an antigen and an adjuvant. The antigen contains either a piece of or the whole infectious agent in question and is the component of the vaccine that induces

the immune response which is specific to the infectious agent. An adjuvant may increase the immunogenicity of the antigen. Thus, active immunization by vaccination does not prevent infection or disease, but the immune responses induced by vaccination interfere with infection or disease. For this reason, the efficacy of a vaccine also depends on the condition of the host's immune system. An important part of vaccine studies is evaluation of immunogenicity of the vaccine which is the ability of the vaccine to produce a measurable immune response in a host.

Three different types of population level effects of vaccination are identified. The indirect effects are the effects or benefits on those people not receiving the vaccine in the targeted population. The total effects in vaccinated individuals are the combination of the indirect effects with the individual-level effects of vaccination. Overall public health effect of the vaccination in the entire population of interest is a weighted average of the indirect effects on the unvaccinated people and the total effects on the vaccinated people. Vaccine studies can be used to evaluate the indirect, total, or overall effects of vaccination in a population. Note that safety is also an important aspect for evaluation of vaccines since vaccination can cause side effects due to the induction of the immune system.

A general definition of vaccine efficacy is the percentage reduction in the attack rate attributable to the vaccine, or

$$VE = \frac{p_u - p_v}{p_u} = 1 - \frac{p_v}{p_u} = 1 - \rho\,,$$

where p_u and p_v denote the risk of infection in unvaccinated and vaccinated individuals respectively, and $\rho = p_v/p_u$ is the relative risk of infection. Alternatively, the vaccine efficacy can be defined as the relative hazard of infection:

$$VE = 1 - \frac{\lambda_v}{\lambda_u}\,,$$

where $p_u = 1 - \exp(-\lambda_u t), p_v = 1 - \exp(-\lambda_v t)$. Vaccine efficacy can also be defined based on the cumulative incidence rates (attack rates) at the end of a study,

$$VE = 1 - \frac{C_v}{C_u}\,,$$

where $C_v q$ and C_u denote the cumulative incidence (infection) rates for vaccinated and unvaccinated individuals. To define the vaccine efficacy for infectiousness, we need to know the infection rate in exposures to vaccinated individuals (r_v) and unvaccinated individuals (r_u), then,

$$VE = 1 - \frac{r_v}{r_u}\,.$$

These definitions are used to evaluate vaccine efficacy in vaccine clinical trials and in the field. Vaccine efficacy can be estimated in a cohort study or a clinical trial involving n_v and n_u individuals in vaccinated and unvaccinated cohorts or groups respectively. Assume that r_v and r_u are the number of infection cases from vaccinated and unvaccinated cohorts, respectively, during a prespecified follow-up period, then the vaccine efficacy is estimated as

$$\widehat{VE} = 1 - \frac{r_v/n_v}{r_u/n_u}\,.$$

The vaccine efficacy can also be similarly estimated by a case-control study using the case-control study methodology. The screening method may also be used to estimate the vaccine efficacy in a population. Let θ denote the proportion of infection cases from vaccinated individuals, and let π as the proportion of the population vaccinated (known). The vaccine efficacy is estimated by

$$\widehat{VE} = 1 - \frac{\theta}{1-\theta}\frac{1-\pi}{\pi}\,.$$

In the screening method, the vaccination rate π is fixed and known, while θ is estimated. To investigate covariate effects, the generalized linear model such as logistic regression techniques can be used.

Vaccine development and studies can be divided into several phases. Phase 0 is the candidate vaccine development. In this early phase of vaccine development, the focus is the search for antigen candidates. A broad types of vaccine candidates may be investigated. Phase I is safety and immunogenicity testing in animals. In this phase, a study is designed to demonstrate the safety and immunogenicity of the vaccine candidate in animals. The question of whether the vaccine candidate is safe or effective in animals is a primary interest in the study. Usually the sample size of this kind of studies is small (only several animals involved). Thus the exact statistics such as Fisher's exact test are usually used for inferences Phase II is safety and immunogenicity testing in humans. Since infectious agents tend to be host-specific, and immune responses and the adverse reactions to a vaccine candidate may be different between animals and humans, safety and immunogenicity studies in humans are required before large-scale trials. Phase II trials may also try to determine dose levels and vaccination schedules. The sample size of Phase II vaccine trials can be very small or as large as several hundreds. Experimental challenges with infectious agents are usually not ethical in humans. The use of the immune response as a surrogate for protective immunity is still questionable since the correlation

between the measurable immune response and the actual protection against infection or disease by the vaccine cannot be confirmed without investigation. Thus, it is a very difficult decision to move the vaccine study from Phase II to a large scale Phase III field efficacy testing. In fact, a very small proportion of vaccine candidates move to Phase III studies, although sometimes a rare Phase IIb, a small field study, may be conducted.

The primary objective of Phase III field trials is to estimate the protective efficacy of vaccination, rather than to test whether there is an effect. Usually a randomized and double-blinded trial is ideal for a Phase III vaccine trial, but may be limited due to ethical or implemental problems. Since the number of infection cases depends on the exposure to infection and transmission rate (difficult to estimate), the sample size of Phase III trials is usually difficult to determine. A liberal strategy may be taken. Many vaccine trials have been inconclusive due to unpredicted transmission or exposure rate. If efficacy and safety of a vaccine is demonstrated in Phase III trials, the vaccine may be licensed by the responsible agency. However, the postlicensure Phase IV studies are still needed to evaluate: (i) protective efficacy under normal usage; (ii) safety under normal usage; (iii) duration of protection; and (iv) indirect and overall effects. Postlicensure studies are usually nonrandomized observational studies (subject to potential biases). Case-control studies are commonly used. Since Phase III efficacy trials are often too small to detect rare adverse events of vaccination, much larger postlicensure studies are very useful in this case. Thus, the design of a vaccine study depends on the scientific questions of interest and the phase of vaccine development. The corresponding statistical methods need to be selected for different studies.

The design of vaccine studies can be traced to early 19th century.[109] More detailed discussions can be found in Smith and Morrow[110] and Farrington and Miller.[111] Most materials of this section are taken from Halloran[112] and Farrington.[8]

4.2.3. *Mathematical modeling and simulations*

Prevention strategies or measures can be included in the deterministic or stochastic epidemic models of infectious diseases introduced in Sec. 2 of this chapter. Computer simulations may be used to evaluate or project how the pattern of epidemics will be changed by effective prevention strategies. Here we introduce an example of HBV infection with vaccine interventions.

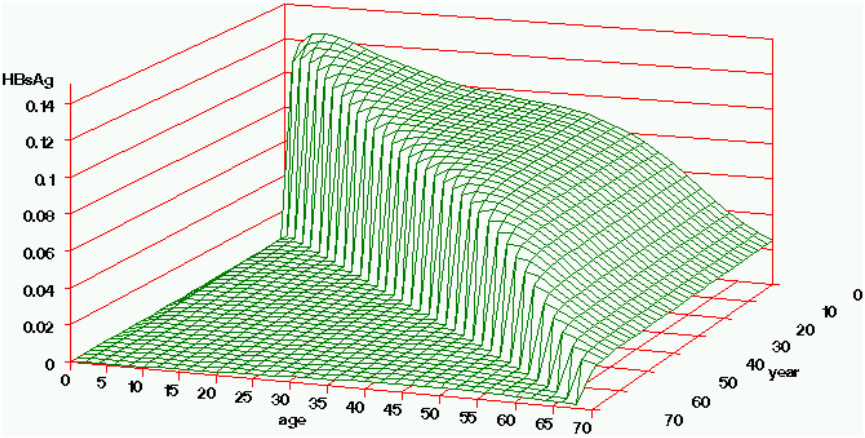


Fig. 1.

The HBV model (21–25) introduced in Sec. 2 can be used to simulate the hepatitis B transmission dynamics before and after vaccination, so we utilize the models to predict the long-term effectiveness of hepatitis B immunization, and to describe the transmission dynamics of HBV in the population. The HBV carriers in a vaccinated cohort will decrease sharply. If all newborn babies can be immunized, the proportion of HBV carriage for immunized children will decrease to a very low level ($< 1\%$). Following up the immunization program with 100% coverage, the transmission dynamics of HBV carriers can be described by the model (21–25). The majority of HBV carriers will shift gradually from children to the elderly. After the vaccination program has been implemented for 70 years or more, the average HBV carrier rate will decrease to a lower level (Fig. 1).

In order to evaluate the impact of the vaccination program on future incidence rate of hepatitis B, we can also define two incidence ratios: an acute hepatitis B and a chronic hepatitis B incidence ratio. Both of them can be calculated based on the dynamics of HBV carriers, that is, as a linear function of carrier proportion since vaccination implementation. The incidence ratio of acute hepatitis B, $R_a(a_1, a_2 : t)$ is the number of acute cases in the age range from a_1 to a_2 at time t divided by the corresponding number of acute cases at $t = 0$, the baseline before vaccination. Mathematically,

$$R_a(a_1, a_2 : t) = \int_{a_1}^{a_2} T(a, t)da \Big/ \int_{a_1}^{a_2} T(a, 0)da\,.$$

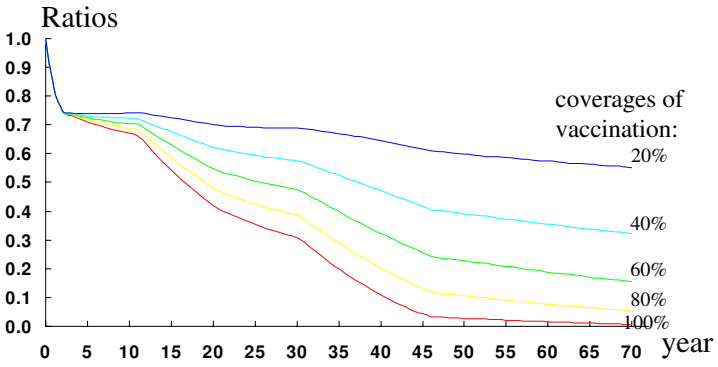


Fig. 2.

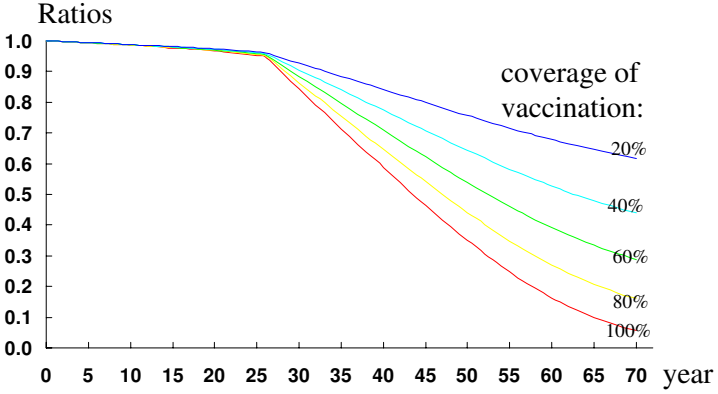


Fig. 3.

The range in the equation has been defined as $a_1 = 10$ and $a_2 = 45$, because the peak of the incidence curve for acute hepatitis B was observed in the age interval of 10 to 45 years old. The incidences in other age groups are at very low levels. The $R_a(a_1, a_2 : t)$ at time t with different vaccination coverage is shown in Fig. 2. It decreases steeply at the beginning of the hepatitis B vaccination program. The higher the vaccination coverage, the steeper the decrease of the ratio. The decrease slows down in a few years after the start of the vaccination program.

The incidence ratio of chronic hepatitis B, $R_c(a_1, a_2 : t)$ is defined as the number of chronic HBV carriers in the age range from a_1 to a_2, at time t divided by the corresponding number of chronic HBV carriers at $t = 0$,

the baseline before vaccination. Mathematically

$$R_c(a_1, a_2 : t) = \int_{a_1}^{a_2} C(a,t)da \Big/ \int_{a_1}^{a_2} C(a,0)da\,.$$

In the equation, $a_1 = 25$ and $a_2 = 70$. Most of chronic liver diseases were observed in adults of 25 years and older. The disease incidence in the younger age group was negligible. The $R_c(a_1, a_2 : t)$, at time t with different vaccination coverage, is shown in Fig. 3. It remains almost unchanged at the beginning of the vaccination program, and drops rapidly after 25 years of immunization. Again, the decrease in the ratio is closely related to the vaccination coverage.

5. Summary

Infectious diseases are dangerous and threat the public health as a whole. To evaluate and project the epidemics of an infectious disease is a great challenge to biomathematicians and statisticians. Enormous efforts have been made in the past century. In this chapter, we have briefly reviewed the epidemic models and methods, especially for HIV and hepatitis viruses, the two most active research areas. Mathematics and statistics have contributed to understanding the pathogenesis of infectious diseases, particularly to understanding the mechanisms of HIV, HBV, and HCV infection in recent years via viral dynamics modeling. We have introduced these models and statistical methods. Statistics always plays an important role in evaluating interventions and preventions of any diseases via clinical trials. Motivated by clinical studies, many advanced statistical methods have been developed. For infectious diseases, statistics even plays more critical role in evaluating medical interventions, prevention measures and vaccine efficacy. Apparently, the study of HIV/AIDS has spurred the statistical research in infectious diseases in the past two decades. Brief reviews on statistical issues in HIV research can be found in an special issue of Journal of Royal Statistical Society A (Vol. 161, Part 2, 1998). A good review on the references can be found in Foulkes.[113]

Acknowledgment

The first author was supported by NIAID/NIH grants No. R29 AI43220, RO1 AI45356 and U01 AI38855.

References

1. Last, J. M. (1988). *A Dictionary of Epidemiology*, Oxford University Press, New York, 27.

2. Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalite causee par la petite verole et des avantages de l'inoculation pour la prevenir. Memoires de Mathematiques et de Physique, in Histoire de l'Academie Royale des Sciences, Paris, 1–45.
3. Hamer, W. H. (1906). Epidemic disease in England: The evidence of variability and persistency. *Lancet* **ii**: 733–739.
4. Ross, R. (1911). *The Prevention of Malaria*, 2nd edn. John Murray, London.
5. Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society*, *Series* **A115**: 700–721.
6. En'ko, P. D. (1889). *The Epidemic Course of Some Infectious Diseases*. Vrach, St Petersburg, **10**: 1008–1010, 1039–1042, 1061–1063 (in Russian).
7. Greenwood, M. (1931). On the statistical measure of infectiousness. *Journal of Hygiene*, Cambridge, **31**: 336–351.
8. Farrington, C. P. (1998). Communicable diseases. In *Encyclopedia of Biostatistics*, Vol. 1, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 795–815.
9. Heesterbeek, J. A. P. and Roberts, M. G. (1998). Epidemic models, deterministic. In *Encyclopedia of Biostatistics*, Vol. 2, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 1332-1340.
10. Gani, J. (1998). Epidemic models, stochastic. In *Encyclopedia of Biostatistics*, Vol. 2, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 1345–1351.
11. Brookmeyer, R. and Gail, M. H. (1986). Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. *Lancet* **2**: 1320–1322.
12. Brookmeyer, R and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* **83**: 301–308.
13. Brookmeyer, R. and Gail, M. H. (1994). *Journal of Acquired Immune Deficiency Syndromes Epidemiology*: *A Quantitative Approach*. Oxford University Press, New York.
14. Harris, J. E. (1990). Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association* **85**: 915–924.
15. Lawless, J. and Sun, J. (1992). A comprehensive back-calculation framework for the estimation and prediction of AIDS cases. In *Journal of Acquired Immune Deficiency syndromes Epidemiology*: *Methodological Issues*, eds. N. Jewell, K. Dietz and C. Farewell, Birkhauser, Boston, 81–104.
16. Bacchetti, P. (1998). Back-calculation. In *Encyclopedia of Biostatistics*, Vol. 1, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 235–242.
17. De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45**: 1–11.
18. Kim, M. Y., De Gruttola, V. and Lagakos, S. W. (1993). Analysis of doubly censored data with covariates; with application to AIDS. *Biometrics* **49**: 13–22.

19. Jewell, N. P., Malani, H. M. and Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of the American Statistical Association* **89**: 7–18.
20. Jewell, N. P. (1994). Nonparametric estimation and doubly-censored data: General ideas and applications to AIDS. *Statistics in Medicine* **13**: 2081–2096.
21. Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics* **51**: 290–295.
22. Diggle, P. J, Liang, Ky. and Zeger, S. L. (1994). Analysis of Longitudinal Data, Oxford University Press, New York.
23. De Gruttola, V., Lange, N. and Dafni, U. (1991). Modeling the progression of HIV infection. *Journal of the American Statistical Association* **86**: 569–577.
24. Taylor, J. M. G., Cumberland, W. G. and Sy, J. P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**: 727–736.
25. Zeger, S. L. and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**: 689–699.
26. Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996). An analysis of paediatric CD4 counts for Acquired Immune Deficiency Syndrome using flexible random curves. *Applied Statistics* **45**: 151–163.
27. Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**: 253–259.
28. Wang, Y. and Taylor, J. M. G. (1995). Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine* **14**: 1205–1218.
29. Fan, J. and Zhang, J. T. (1998). Comments on "smoothing spline models for the analysis of nested and crossed samples of curves" (by B. Brumback and J. A. Rice). *Journal of the American Statistical Association* **93**: 980–983.
30. Fan, J. and Zhang, J. T. (1999). Two-step estimation of functional linear models with application to longitudinal data. *Journal of the Royal Statistical Society, Series* **B62**: 303–322.
31. Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**: 809–822.
32. Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association* **93**: 1388–1402.
33. Lange, N., Carlin, B. P. and Gelfand, A. E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). *Journal of the American Statistical Association* **87**: 615–632.

34. De Gruttola, V. and Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50**: 1003–1014.
35. Tsiatis, A. A., De Gruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**: 27–37.
36. Cnaan, A. (1998). Natural history study of prognosis. In *Encyclopedia of Biostatistics*, Vol. 4, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 2956–2962.
37. Xu, Z. Y. (1991). Impact and control of viral hepatitis in China. In *Viral Hepatitis and Liver Diseases*, eds. F. Blaine Hollinger, Williams and Wilkins, 700–706.
38. Szmuness, W. (1978). Sociodemograghic aspect of the epidemiology of HB. In *Viral Hepatitis.*, eds. G. N. Vyas, S. N. Cohen and R. Schmidt, Franklin Institute Press, Philadelphia, Pennsylvania, 296–320.
39. Tang, S. X. (1991). Study of the mechanisms and influential factors of intrauterine infection of hepatitis B virus. *Chinese Journal of Epidemiology* **12**(6): 325–328.
40. Wang, S. S. (1991). Transplacental transmission of hepatitis B virus. *Chinese Journal of Epidemiology* **12**(1): 33–35.
41. Zhao, S. J., Xu, Z. Y., Ma, J. C. *et al.* (1994). A follow up study of spontaneous clearance rates on hepatitis B surface agent persistent carriers. *Chinese Journal of Preventive Medicine* **29**(6): 378–379.
42. Zhao, S. J. and Xu, Z. Y. (1991). A study of hepatitis B infection force in China. *Chinese Journal of Epidemiology* **14**(suppl.): 70–74.
43. Zhao, S. J., Xu, Z. Y., Cao, H. L. *et al.* (1995). An estimating model of relationship between the infection age of hepatitis B virus and the age-specific chronic carrier rate. *Chinese Journal of Experimental and Clinical Virology* **9**(Suppl.): 101–104.
44. Zhao, S. J. and Xu, Z. Y. (1995). Mathematical simulation of hepatitis B transmission and application in immunization police. In *Proceeding of Epidemiology*, ed. X. W. Zhen, Beijing, **8**: 162–181.
45. Zhao, S., Xu, Z. Y. and Lu, Y. (2000). A mathematical model of hepatitis B virus transmission and its application for vaccination strategy in China. *International Journal of Epidemiology* **29**: 744–752.
46. Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*, 2nd edn., Griffin, London.
47. Becker, N. J. (1989). *Analysis of Infectious Disease Data*, Chapman and Hall, London.
48. Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford.
49. Daley, D. J. and Gani, J. (1999). *Epidemic Modelling*, Cambridge University Press, Cambridge.
50. Harris, T. E. (1948). Branching processes. *Annals of Mathematical Statistics* **19**: 474–494.

51. Keiding, N. (1975). Estimation theory for branching processes. *Bulletin of the International Statistical Institute* **46**(4): 12–19.
52. Becker, N. (1977). Estimation for discrete time branching processes with applications to epidemics. *Biometrics* **33**: 515–522.
53. Heyde, C. C. (1979). On assessing the potential severity of an outbreak of a rare infectious disease: A Bayesian approach. *Australian Journal of Statistics* **21**: 282–292.
54. Frost, W. H. (1976). Some conceptions of epidemics in general. *American Journal of Epidemiology* **103**: 141—151.
55. Fine, P. (1977). A commentary on the mechanical analogue to the Reed-Frost epidemic model. *American Journal of Epidemiology* **106**: 87–100.
56. Longini, I. and Koopman, J. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics* **38**: 115–126.
57. Longini, I., Koopman, J., Haber, M. and Otsonis, G. (1988). Statistical inference for infectious diseases: Risk-specified household and community transmission parameters. *American Journal of Epidemiology* **128**: 845–859.
58. Saunders, I. (1980). An approximate maximum likelihood estimator for chain binomial models. *Australian Journal of Statistics* **22**: 307–316.
59. Longini, I. (1998). Chain binomial model. In *Encyclopedia of Biostatistics*, Vol. 1, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 593–597.
60. Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society*, *Series* **B61**: 287–307.
61. Bartlett, M. S. (1957). Measles periodicity and community size. *Journal of the Royal Statistical Society*, *Series* **A120**: 48–70.
62. Bartlett, M. S. (1960). The critical community size for measles in the United States. *Journal of the Royal Statistical Society*, *Series* **A123**: 37–44.
63. Tan, W. Y. and Wu, H. (1998). Stochastic modeling the dynamics of CD4+ T cell infection by HIV with Monte Carlo studies, *Mathematical Biosciences* **147**: 173–205.
64. Tan, W. Y. and Hsu, H. (1989). Some stochastic models of AIDS spread. *Statistics in Medicine* **8**: 121–136.
65. Wu, H. and Tan, W. Y. (2000). Modeling HIV epidemic: A state-space approach. *Mathematical and Computer Modelling* **32**: 197–215.
66. Merrill, S. (1987). AIDS: Background and the dynamics of the decline of immunocompetence. In *Theoretical Immunology*, Part 2, ed. A. S. Perelson, Addison-Wesley, Redwood City, Calif.
67. Anderson, R. M. and May, R. M. (1989). Complex dynamical behavior in the interaction between HIV and the immune system. In *Cell to Cell Signaling*: *From Experiments to Theoretical Models*, ed. A. Goldbeter, Academic, New York.
68. Perelson, A. S. (1989). Modeling the interaction of the immune system with HIV. In *Mathematical and Statistical Approaches to Journal of Acquired Immune Deficincy Syndromes Epidemiology* (Lect. Notes Biomath., Vol. 83), ed. C. Castillo-Chavez, Springer-Verlag, New York.

69. Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**: 123–126.
70. Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. S. and Shaw, G. M. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**: 117–122.
71. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. and Ho, D. D. (1996). HIV-1 dynamics *in vivo*: Virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**: 1582–1586.
72. Perelson, A. S., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M. and Ho, D. D. (1997). Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387**: 188–191.
73. Wu, H. and Ding, A. (1999). Population HIV-1 dynamics *in vivo*: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**: 410–418.
74. Wu, H., Ding, A. and De Gruttola, V. (1998). Estimation of HIV dynamic parameters, *Statistics in Medicine* **17**: 2463–2485.
75. Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, New York.
76. Vonesh, E. F. and Chinchilli, V. M. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Marcel Dekker, New York.
77. Ding, A. A. and Wu, H. (2000). A comparison study of models and fitting procedures for biphasic viral dynamics in HIV-1 infected patients treated with antiviral therapies. *Biometrics* **56**: 16–23.
78. Wu, H., Kuritzkes, D. R., McClernon, D. R. *et al.* (1999). Characterization of viral dynamics in Human Immunodeficiency Virus Type 1-infected patients treated with combination antiretroviral therapy: Relationships to host factors, cellular restoration and virological endpoints. *Journal of Infectious Diseases* **179**(4): 799–807.
79. Luzuriaga, K., Wu, H., McManus, M. *et al.* (1999). Dynamics of HIV-1 replication in vertically-infected infants. *Journal of Virology* **73**: 362–367.
80. Ding, A. A. and Wu, H. (1999). Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics. *Mathematical Biosciences* **160**: 63–82.
81. Ding, A. A. and Wu, H. (2001). Assessing antiviral potency of anti-HIV therapies *in vivo* by comparing viral decay rates in viral dynamic models. *Biostatistics* **2**: 13–29.
82. Nowak, M. A., Bonhoeffer, S., Shaw, G. M. and May, R. M. (1997). Anti-viral drug treatment: Dynamics of resistance in free virus and infected cell populations. *Journal of Theoretical Biology* **184**: 203–217.
83. Perelson, A. S. and Nelson, P. W. (1999). Mathematical Analysis of HIV-1 dynamics *in vivo*. *SIAM Review* **41**(1): 3–44.
84. Nowak, M. A. and May, R. M. (2000). *Virus Dynamics*: *Mathematical Principles of Immunology and Virology*, Oxford University Press.

85. Nowak, M. A., Bonhoeffer, S., Hill, A. M. *et al.* (1996). Viral dynamics in hepatitis B virus infection. *Proceedings of National Academic Sciences* **93**: 4398–4402.
86. Fukumoto, T., Berg, T., Ku, Y. *et al.* (1996). Viral dynamics of hepatitis C early after orthotopic liver transplantation: Evidence for rapid turnover of serum virions. *Hepatology* **24**: 1351–1354.
87. Zeuzem, S., Schmidt, J. M., Lee, J. H. *et al.* (1996). Effect of interferon alfa on the dynamics of hepatitis C virus turn over *in vivo*. *Hepatology* **23**: 366–371.
88. Lam, N. P., Neumann, A. U., Gretch, D. R. *et al.* (1997). Dose-dependent acute clearance of hepatitis C genotype 1 virus with interferon alfa. *Hepatology* **26**: 226–231.
89. Yasui, K., Okanoue, T., Murakami, Y. *et al.* (1998). Dynamics of hepatitis C viremia following interferon-α administration. *The Journal of Infectious Diseases* **177**: 1475–1479.
90. Neumann, A. U., Lam, N. P., Dahari, H. *et al.* (1998). Hepatitis C viral dynamics *in vivo* and the antiviral efficacy of interferon-α therapy. *Science* **282**: 103–107.
91. Friedman, L. M., Furberg, C. D. and DeMets, D. L. (1981). *Fundamentals of Clinical Trials*, John Wright, PSG Inc.
92. Pocock, S. J. (1983). *Clinical Trials on Protocol Approach*, New York, Wiley.
93. Meinert, C. L. (1986). *Clinical Trials*, *Design*, *Conduct and Analysis*, Oxford, Oxford University Press.
94. Piantadosi, S. (1997). *Clinical Trials*, *A Methodologic Perspective*, New York, Wiley.
95. Amato, D. A. and Lagakos, S. W. (1990). Considerations in the selection of end points for AIDS clinical trials. *Journal of Acquired Immune Deficiency Syndromes* **3**(Suppl): S64–S68.
96. Neaton, J. D., Wentworth, D. N., Rhame, F. *et al.* (1994). Methods of studying interventions, Considerations in choice of a clinical endpoint for AIDS clinical trials. *Statistics in Medicine* **13**: 2107–2125.
97. Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* **8**: 431–440.
98. Marschner, I. C., Betensky, R. A., De Gruttola, V. *et al.* (1999). Clinical trials using HIV-1 RNA-based primary endpoints: Statistical analysis and potential biases. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **20**: 220–227.
99. Hughes, M. D. (2000). Analysis and design issues for studies using censored biomarker measurements, with an example of viral load measurements in HIV clinical trials. *Statistics in Medicine* **19**: 3171–3191.
100. Gilbert, P., Ribaudo, H. J., Greenberg, L. *et al.* (2000). Considerations in choosing a primary endpoint that measures durability of virologic suppression in an antiretroviral trial. *Journal of Acquired Immune Deficiency Syndromes* **14**: 1961–1972.
101. Weinberg, J. and Lagakos, S. W. (2001). Efficiency comparisons of rank and permutation tests based on summary statistics computed from repeated measures data. *Statistics in Medicine* **20**: 705–731.

102. De Gruttola, V., Hughes, M. D., Gilbert, P. and Phillips, A. (1998). Trial design in the era of highly effective antiviral drug combinations for HIV infection. *Journal of Acquired Immune Deficiency Syndromes* **12**: S149–S156.
103. Krall, R. L., Engleman, K. H., Ko, H. C. and Peck, C. C. (1998). Clinical trial modeling and simulation–work in progress. *Drug Information Journal* **32**: 971–976.
104. Mocroft, A., Vella, S., Benfield, T. L. *et al.* (1998). Changing patterns of mortality across Europe in patients infected with HIV-1. *Lancet* **352**: 1725–1730.
105. Palella, F. J., Jr, Delaney, K. M., Moorman, A. C. *et al.* (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine* **338**: 853–860.
106. Vittinghoff, E., Scheer, S., O'Malley, P. *et al.* (1999). Combination antiretroviral therapy and recent declines in AIDS incidence and mortality, *Journal of Infectious Diseases* **179**: 717–720.
107. Prentice, R. L. (1998). Prevention trials. In *Encyclopedia of Biostatistics*, Vol. 4, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 3494–3499.
108. Jacobs, Jr., D. R. (1998). Preventive medicine. In *Encyclopedia of Biostatistics*, Vol. 4, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 3500–3502.
109. Greenwood, M. and Yule, U. G. (1915). The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Proceedings of the Royal Society of Medicine* **8**: 113–194.
110. Smith, P. G. and Morrow, R. N. (1991). *Methods for Field Trials of Interventions Against Tropical Diseases*: *A Toolbox*, Oxford, Oxford University Press.
111. Farrington, P. and Miller, E. (1996). Clinical trials. In *Methods in Molecular Medicine*: *Vaccine Protocols*, eds. A. Robinson, G. Farrar and C. Wiblin, Totowa, Humana Press.
112. Halloran, M. E. (1998). Vaccine studies. In *Encyclopedia of Biostatistics*, Vol. 6, eds. P. Armitage and T. Colton, John Wiley and Sons, New York, 4687–4694.
113. Foulkes, M. A. (1998). Advances in HIV/AIDS statistical methodology over the past decade. *Statistics in Medicine* **17**: 1–25.

About ad Author

Hulin Wu is currently a senior statistician and project leader at Frontier Science and Technology Research Foundation. He also holds an adjunct faculty position at Biostatistics Department, Harvard University and Section Head at Center for Biostatistics in AIDS Research, Harvard School of Public Health. He obtained his BS and MS degrees in Automatical Control from the National University of Defense Technology of China, and PhD

in Statistics from Florida State University, USA. He has been a visiting assistant professor at University of Memphis, Tennessee, USA before he joined the current position. His current research interests include nonparametric regression for longitudinal data, linear and nonlinear mixed-effects models, state-space models and dynamic prediction, modeling HIV/AIDS epidemics and pathogenesis, AIDS clinical trials, and clinical trial simulations.

This page intentionally left blank

CHAPTER 18

SPECIAL MODELS FOR SAMPLING SURVEY

SUJUAN GAO

Division of Biostatistics, Indiana University School of Medicine, 1050 Wishard Blvd., RG 4101, Indianapolis, IN 46202-2872, USA
Tel: (317)274-0820; sgao@iupui.edu

1. Introduction

Sample surveys have been widely used in everyday life and scientific research, from our attitude towards a specific television program to major economic indices. Sample surveys are broadly classified into two types — descriptive and analytical. In a descriptive survey the objective is simply to obtain information about large groups. For example, the total number of men, women and children living in a certain geographic area may be the objective of a descriptive survey. In an analytical survey, comparisons are made between different subgroups of the population in order to ascertain whether true differences exist among them and to verify hypotheses about the reasons for these differences. Sample surveys in medical research fields are mostly taken for analytical purposes. Early well-known surveys include survey of the teeth of school children before and after fluoridation of water, of the death rates and causes of death of people who smoke different amount, and the huge study on the effectiveness of the Salk polio vaccine.

Recent surveys in the medical fields, still analytical in nature, can be further divided into two categories. The first category is large scale surveys with the intent to make inference on a large population. These surveys are designed to derive reliable estimates on various health, nutrition, medical expenditure, etc. on a national level. Examples of surveys of this nature are the National Health and Nutrition Examination Survey (NHAINES), Survey of Asset and Health Dynamics of the Oldest Old (AHEAD), Health and Retirement Survey (HRS), National Long Term Care Survey, to name

just a few. These surveys have elaborate sampling frames with oversampling of certain subgroups and extensive questionnaire containing a vast amount of information. They tend not to be disease specific, concerning instead about national trends on the general health of the entire population. The second category of medical surveys are community-based surveys, where a geographically defined catchment area is first defined and the survey is given to individuals selected from a sampling plan within this community only. Community-based surveys are usually focused on a specific disease, advantageous specifically on diseases that require extensive diagnosis, such as dementia and Alzheimer's disease. Community-based surveys offer the advantages of extensiveness on studying specific disease, but may suffer from limitations on the scope of populations they represent.

The basic steps and components of a survey are described in details in classical sample survey books such as Cochran (1977) and Kish (1965). We will review here some terminology central to the development in this chapter.

Population: The aggregate from which the sample is chosen. The population to be sampled (the *sampled population*) should coincide with the population about which information is desired (the *target population*). Sometimes, due to practical constraints, the sampled population is more restricted than the target population. It should be noted that the conclusion drawn from the samples under these situations should only apply to the sampled population. The extent to which the conclusion from the sampled population apply to the target population depends on many additional information.

Sampling plan: The rule by which the samples are selected.

Sampling units: The parts the population is divided into and selection rules are based upon. They must cover the entire population without any overlap. Sometimes the choice of sampling units is obvious, as in a surveys of hospitals where each hospital constitutes a sampling unit. In other situations, there may be many possible choices of sampling units. For example, sampling of individuals in a city can be done by selecting the whole family, or selecting individuals living in a whole city block. The decision on which sampling unit to use is often based on an array of factors such as logistic, economic and convenience.

Sampling frame: The list containing all sampling units so that random selection from the population is accomplished by random sampling from

the list. An ideal sampling frame should be complete containing every unit in the population and without any duplicates. The legitimacy of a survey often rests on how well the sampling frame is constructed to represent the population.

Simple random sampling: A method of selecting n units out of N such that every distinct sample has an equal chance of being drawn. Conventional statistical inference and software all apply to simple random sampling.

Stratified random sampling: In stratified random sampling, the population of N units is first divided into non-overlapping subpopulations of $N_1, \ldots, N_S$ units, respectively. These subpopulations are called *strata*. If a simple random sampling is taken within each stratum, the whole procedure is described as stratified random sampling. Stratification may produce a gain in precision in the estimates of characteristics of the whole population by creating relative homogeneity within each stratum.

Sample survey data differ from data collected from conventional observational studies in two fundamental ways. The first is that samples from surveys are usually selected with unequal probability, which if ignored may create a distorted picture of the target population. For example, if a health survey oversamples elderly individuals, the simple frequencies on diseases associated with increasing age are likely to overestimate disease rates in the population. The second difference between survey data and conventional studies is that survey data often present a natural clustering inherited from the sampling design. For example, if all family members in a selected household are interviewed in a survey, the outcomes on social-economical scales, behavior measures and attitudes are likely to be correlated.

A vast body of literature on methods of analyzing survey data exists, see, for example, Skinner, Holt and Smith[28] for review of earlier works in the area. The most used approach for analyzing survey data is the so-called pseudo-likelihood method where the score equation in a general likelihood framework is modified by including appropriate sampling weights. Variance estimation is achieved by taking the sample design into consideration and by using Taylor series linearization. Software packages implementing the pseudo-likelihood approach are also available, e.g. SUDAAN from the Research Triangle Institute[25] and WesVarPC from Westat.[22]

This chapter focuses on several special models used in analyzing survey data from epidemiological studies. First, we discuss the use of special models for estimating prevalence rate of a rare disease from community-based surveys, followed by a discussion about the use of random effect models

for small area estimation on both continuous and discrete outcomes. The final section is devoted to capture recapture models in epidemiology.

2. Models for the Estimation of Disease Prevalence

Disease prevalence is the percentage of individuals with a disease at the study time in a certain population. It describes the disease's effect on the population. Multi-phase samplings are often used in epidemiological studies where a disease is rare and diagnosis of the disease is expensive or difficult. The design has been used interchangeably with "multi-stage" sampling by medical researchers without distinction from multi-phase sampling. However, multi-stage sampling is a standard terminology from sampling theory which usually implies that different sampling units are used at various stages of sampling (for example, city blocks in the first stage, households in the second stage, and individuals in the third stage). We therefore prefer the term multi-phase sampling for the type of study where individuals are the sampling units in all sampling selections.

Two-phase sampling is by far the most often used design of all multi-phase studies. In the first phase of the study a large random sample from the targete population is screened with less intensive and expensive screening tests for the disease. Based on the results of the screening tests subjects are stratified and randomly selected within each stratum for extensive clinical evaluations at the second phase to determine disease status. The sampling plan are usually designed to identify as many diseased subjects as possible for risk factor studies and at the same time allow efficient estimation of disease prevalence for the population. The two-phase sampling design has been used to estimate the prevalence rates of Alzheimer's disease,[1] heart disease[4] and sexually transmitted disease.[31]

Data for our first example comes from the Indianapolis Study of Health and Aging, an on-going longitudinal study of dementia in the elderly African Americans age 65 and over living in Indianapolis, USA. Population-based two-phase surveys were conducted to estimate the prevalence rate of dementia in this population. At the first phase, 2212 individuals were randomly selected from the community and administered screening tests aimed at measuring their cognitive functions. Each individual received a cognitive score which ranged from 0 to 33. Based on the screening scores the subjects were grouped into three performance groups: good, intermediate and poor. The initial sampling plan was to invite 100% of the subjects from the poor performance group, 50% from the intermediate group, 5% from the

good performance group of which 75% should come from those older than 75 years of age. However, due to refusal, death, severe sickness and other reasons, the study had to sample more than the prespecified percentages in all groups except the poor performance group to achieve the targeted number of total clinical evaluations. The following table gives the number of demented subjects diagnosed from each of the sampling stratum by age group.

A weighting type estimator, also referred to as the direct standardization approach, is often used to estimate disease prevalence from a multi-phase sampling study. It assumes that subjects within each stratum are homogeneous and random sampling is used within each stratum. Suppose that in the first phase of the study N individual subjects are sampled by simple random sampling from the target population and information is collected from all N subjects on a set of characteristics X. X can be a vector containing several predictors, such as age, gender, screening scores, etc., that relate to the disease of interest. The N subjects are then divided into S strata, labeled as $I_1, \ldots, I_S$, based on the values of X. The total numbers of subjects in the respective strata are denoted by $N_1, \ldots, N_S$. In the second phase n_s subjects are sampled from the N_s subjects in the sth stratum using stratified random sampling. Disease status is ascertained on the selected n_s subjects only. Let y_{si} represents disease status on the ith subject from the sth stratum, with $y_{si} = 1$ denoting disease and $y_{si} = 0$ for non-disease. The probabilities for second phase sampling can be different for subjects from different strata. However, subjects from the same stratum are assumed to have equal probability of sampling.

The weighting type estimator of prevalence rate for a stratified random sampling is:

$$\hat{p}_{\text{wt}} = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \frac{N_s}{n_s} y_{si} = \sum_{s=1}^{S} \frac{N_s}{N} \hat{p}_s \,, \tag{1}$$

where $\hat{p}_s = \sum_{i=1}^{n_s} \frac{y_{si}}{n_s}$ is the average disease rate from the sth stratum. The variance of the estimator is estimated by

$$\text{var}(\hat{p}_{\text{wt}}) = \sum_{s=1}^{S} \hat{p}_s (1 - \hat{p}_s) \frac{N_s^2}{N^2 n_s} \,. \tag{2}$$

Problems in the weighting type estimation approach can occur in groups with no affected individuals or no unaffected individuals ($\hat{p}_s = 0$ or $\hat{p}_s = 1$), or groups where no individual is sampled ($n_s = 0$). Prevalence estimate can be very unstable with possible variance 0 or infinity.

An alternative method of estimating disease prevalence from two-phase surveys is the modeling type estimator. A model is assumed for the population where the finite population is sampled from and smoothed estimates from the model are used to estimate disease prevalence. The modeling type estimator for binary data was first proposed by Roberts *et al.*[26] and used by Beckett *et al.*[1] to estimate the prevalence of Alzheimer's disease from two-phase surveys. The modeling type estimator is preferred in situations where the disease is rare and estimates from strata containing few or zero events are desired.

Let X_{si} be a set of covariates collected at the first phase. Therefore, X_{si} is available for all N subjects. Let $\text{Prob}(y_{si} = 1) = p_{si}$. A logistic regression model is assumed for the disease model:

$$\log \frac{p_{si}}{1 - p_{si}} = X_{si}\beta \,, \tag{3}$$

where β is a $p \times 1$ vector of parameter. If β is known, then the average of the predicted probability of disease from the model is an unbiased estimator of disease prevalence. In practice, one has to estimate β from the sample. A psudo-maximum likelihood estimate $\hat{\beta}$ is obtained using data from the second phase and estimate of disease prevalence is then obtained using the average predicted probabilities of disease on every subject in the population:

$$\hat{p}_{\text{model}} = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{N_s} \frac{1}{1 + e^{-X_{si}\hat{\beta}}} \,. \tag{4}$$

The estimated variance of the prevalence estimate is

$$\text{var}(\hat{p}_{\text{model}}) = W'QXVX'Q'W \,, \tag{5}$$

where W is an $N \times 1$ vector with elements equal to $\frac{1}{N}$, Q is an $N \times N$ diagonal matrix with elements $\hat{p}_{si}(1 - \hat{p}_{si})$, X is the covariate matrix and V is the estimated variance covariance matrix of the logistic regression parameter β. Note that the above variance estimator assumes that X is fixed. Variance estimators accounting for the variability in X for survey data is given by Graubard and Korn.[9]

Prevalence estimates from the modeling approach can be more efficient than the weighting type estimate if the logistic model fits the data well. The weighting type estimates may mask trends in prevalence rates due to small sample sizes in some groups. Lastly, the weighting type estimator requires the assumption of missing completely at random while the modeling type estimator is unaffected under the missing at random mechanism (using

Table 1. Number of demented subjects diagnosed from the three sampling groups in each age group. Data from the Indianapolis Study of Health and Aging.

Age Group	Performance Group	Number in Population	Number Sampled	Number with Dementia
65–74	1	1133	27	0
	2	77	34	0
	3	97	63	10
75–84	1	523	58	1
	2	69	35	5
	3	112	75	31
85+	1	127	14	0
	2	21	10	0
	3	53	37	18

Table 2. Estimated age-specific prevalence rates of dementia using the weighting method and the modeling method. Data from Indianapolis Study of Health and Aging.

	Weighting		Modeling	
Age Group	Rate	Std Err	Rate	Std Err
65–74	1.18	0.34	1.83	0.37
75–84	9.26	1.66	6.73	0.85
85+	12.83	2.17	17.07	2.31

the definitions of Little and Rubin[21]), provided that covariates related to missing data are incorporated in the model.[7]

We return now to the example data in Table 1. Prevalence estimates for the three age groups are desired. Note in age groups 65–74 and 85+, the first two sampling strata produced zero disease case, the weighting type estimator is expected to underestimate the true prevalence in these two age groups. Prevalence estimates for the three age groups using both the weighting type approach and the modeling estimator along with standard error estimates are included in Table 2.

It can be seen from Table 2 that the modeling type estimator yields larger prevalence estimates for two age groups whiles the estimates for age group 75–84 is smaller than the weighting type estimates. A simulation study conducted by Beckett *et al.*[1] demonstrated that the modeling type estimator can increase the accuracy and efficiency of the rate estimates substantially if the model fits the data well.

In this section we are mainly concerned with estimating disease prevalence which is the mean of a binary variable. It should be noted that similar approaches exist in survey theory to estimate the means of continuous variables. For further details see the section on regression estimator in Cochran.[5]

3. Random Effect Models for Small Area Estimation

The terms "small area" were initially used to denote a small geographical area, such as a county, or a census division. They may also describe a "small domain" which is a small subpopulation such as a specific age-sex-race group of people within a large geographical area. Small area estimation arises usually from secondary analysis of large survey data, where the survey is designed to estimate the characteristics of a large domain. For example, in a national hospital cost survey, the sample selection is designed to estimate mean hospital cost with a desired precision at the national level. However, it may also be of interest to use the survey to derive hospital cost estimates by census region, or by state or county, possibly for the apportionment of government funds, and in regional and city planning. Direct estimates by region, state or county in this case based on only data from the small area are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area.

Suppose that the population is divided into k small areas, each contains N_i samples. n_i out of N_i units are sampled from the ith area. Y_{ij} denote the jth unit value in the ith small area. For convenience we let the first n_i units in Y_{ij} be sampled, and the remaining $N_i - n_i$ not sampled. In addition, we assume auxiliary information X_{ij} is available on *every* unit in the population. Alternative models can also be formed when X_{ij} is only available on the area-specific level. See Ghosh and Rao[8] for further details. The focus of inference is to estimate the small area mean $\bar{Y}_{i.}$.

A conceptualized example is described here without direct reference to a specific survey or real data set. The example is in the context of a study by Taylor *et al.*[29] published in the *New England Journal of Medicine*. We have modified the setting so that the sampling units are hospitals instead of patients to make the inference straightforward. Taylor *et al.*[29] pointed out that it is important to study hospital cost and outcome and to investigate any differences among various hospitals. Suppose a national survey on hospital cost is conducted. The primary interest of the survey is to estimate the average hospital cost for various primary diagnoses such as hip fracture, stroke, coronary heart disease or congestive heart failure on the national

level. Suppose we are also interested in using the survey data to provide average hospital cost estimates for the counties within a region. Suppose in addition, we have information on Medicare claim data for *all hospitals in the region* in the same time period of the survey. Medicare is a federal program that purchases inpatient services, primarily for persons 65 years and older, from various types of hospital. It is reasonable to assume that as the amount of Medicare claim increase, hospital cost is also likely to increase. For this hypothetical example, we will assume that there is no major systematic difference between the hospitals. In a real world situation, the type (teaching or non-teaching), size (numbers of bed) and location (urban or rural) of the hospital are all likely to affect hospital cost, as concluded in the article by Taylor *et al.*[29] A simulated data set for 114 hospitals in 16 small areas is generated using the equation:

$$y_{ij} = -2.0 + 0.2x_{ij} + \mu_i + e_{ij}\,,$$

where $\mu_i \sim N(0, 20)$, and $e_{ij} \sim N(0, 40)$. Thirty-eight hospitals are sampled using simple random sampling. The sampled data is presented in Table 3.

Table 3. Data from a simple random sample drawn from a synthetic population ($n = 38, N = 114$).

Area No.	N_i	n_i	x_{ij}	y_{ij}	Area No.	N_i	n_i	x_{ij}	y_{ij}
1	1	0	–	–	9	27	10	86.54	12.51
2	6	1	169.99	38.19				101.55	15.27
3	4	2	6.16	7.43	10	5	2	61.34	11.56
			30.09	10.42				102.77	17.62
4	1	0	–	–	11	12	4	61.49	7.02
5	8	2	94.60	20.52				99.77	27.09
			86.34	25.61				52.08	10.39
6	6	2	90.53	22.01				136.46	39.48
			86.34	25.61	12	7	3	92.28	9.53
7	6	2	95.10	22.52				93.73	9.43
			130.45	35.24				62.29	10.89
8	6	1	46.09	8.38	13	4	1	164.14	35.79
9	27	10	89.88	12.44	14	6	2	134.88	35.30
			145.68	26.23				164.94	40.81
			113.54	8.77	15	13	5	64.44	14.92
			114.54	17.98				63.51	9.07
			96.51	23.22				73.10	8.34
			84.58	9.41				109.66	10.35
			139.38	15.61				123.09	29.39
			117.24	34.42	16	2	1	186.16	28.08

For each county (small area), N_i represents the total number of hospitals in that county and n_i is the number of hospitals sampled in the survey. x_{ij} is the Medicare claim amount and y_{ij} is the true hospital cost.

A problem immediately seen with this data set, present also in many small area estimation situation, is that there are two counties without sampled hospital so that direct estimates from samples from those counties are not possible. Another problem is that some counties have a very small numbers of hospital sampled. Hence direct estimates for these counties may be unstable. We will describe three approaches commonly used for estimation from small area.

A synthetic estimator is similar in spirit to the ratio estimator in sample survey.[5] The estimator uses the percentage of $\frac{\bar{X}_i}{\bar{X}_\cdot}$ in the estimate $\hat{\bar{Y}}_\cdot$ of the overall mean as the estimate of the total in the ith area, where $\bar{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$, $\bar{X}_\cdot = \frac{1}{N}\sum_{i=1}^{k} X_i$, and $\hat{\bar{Y}}_\cdot = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$. The synthetic estimator of small area mean is given by:

$$\hat{\bar{Y}}_i(S) = \frac{\bar{X}_i}{\bar{X}_\cdot}\hat{\bar{Y}}\,, \tag{6}$$

The bias of the synthetic estimator is given by:

$$E(\hat{\bar{Y}}_i(S)) - \bar{Y}_i = \bar{X}_i\left(\frac{\bar{Y}_\cdot}{\bar{X}_\cdot} - \frac{\bar{Y}_i}{\bar{X}_i}\right).$$

It can be seen that the bias is not zero unless $\frac{\bar{Y}_\cdot}{\bar{X}_\cdot} = \frac{\bar{Y}_i}{\bar{X}_i}$. Under the assumption that the sample average $\bar{X}_i$ equals to the population average $\bar{X}_\cdot$, the synthetic estimator will only be unbiased if each small area mean $\bar{Y}_i$ equals to the overall mean $\bar{Y}_\cdot$. Such an assumption can be very strong and can produce biased estimates in situations when the assumption does not hold.

In an effort to reduce or balance the bias of the synthetic estimator a weighted estimator is proposed in the form of

$$\hat{\bar{Y}}_i(SD) = w_i\hat{\bar{Y}}_{1i} + (1-w_i)\hat{\bar{Y}}_{2i}\,, \tag{7}$$

where $\hat{\bar{Y}}_{1i}$ is the direct estimator from the selected samples, $\hat{\bar{Y}}_{2i}$ is an indirect estimator and w_i is a pre-determined weight ($0 \le w_i \le 1$). An optimal weight may be obtained to minimize the mean squared error of $\hat{\bar{Y}}_i(SD)$. See Ghosh and Rao[8] for further details on obtaining the optimal weight. In practice a sample size dependent weight is chosen as $w_i = \frac{n_i N}{n N_i}$, where N and N_i are the total number of units and number of units in each small area in the population, respectively. n and n_i are the total selected sample

size and the selected sample size in each small area, respectively. Therefore, the sample size dependent estimator can be written as:

$$\hat{\bar{Y}}_i(SD) = \begin{cases} \hat{\bar{Y}}_i\,, & \text{if } w_i \geq 1\,, \\ w_i\hat{\bar{Y}}_i + (1-w_i)\hat{\bar{Y}}_i(S)\,, & \text{if } w_i < 1\,, \end{cases} \tag{8}$$

where $\hat{\bar{Y}}_i(S)$ is the synthetic estimator.

In a random effect model approach the finite population containing N units is itself assumed to be random samples from an infinite population, the so-called superpopulation. The finite population is further assumed to have the following distribution:

$$y_{ij} = x_{ij}\beta + \nu_i + e_{ij}\,, \quad i = 1, \ldots, k\,, \quad j = 1, \ldots, N_i\,, \tag{9}$$

where x_{ij} is the value of the auxiliary variable, β is the parameter for the auxiliary effect. ν_i and e_{ij} are two independent random variables with

$$E(\nu_i) = 0\,, \quad V(\nu_i) = \sigma_\nu^2\,, \quad E(e_{ij}) = 0\,, \quad V(e_{ij}) = \sigma^2\,.$$

In addition, normality of the two random variables is assumed.

Using matrix notations, and an asterisk for the nonsampled elements, the above random effect model can be written as:

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{y}_i^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{X}_i^* \end{bmatrix} \beta + \nu_i \begin{bmatrix} \mathbf{1}_i \\ \mathbf{1}_i^* \end{bmatrix} + \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_i^* \end{bmatrix}, \tag{10}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ represents the sampled units, and $\mathbf{y}_i^* = (y_{in_i+1}, \ldots, y_{iN_i})'$ the non-sampled ones. Other vectors in the equation are similarly defined.

The difference between the random effect model for sampling survey and the random effect models in classical statistics textbook is demonstrated by Eq. (10). A component of the outcome variable is unobserved because it is not sampled. Therefore, there are two steps in estimating the means of small areas. The first step involves estimating the parameters in the model, i.e. β, σ_ν^2 and σ^2, using the sampled data only. The second step uses the parameter estimates to predict $\mathbf{y}_i^*$.

Three estimation approaches exist for parameter estimation from the random effect model, namely, the best linear unbiased predictor approach (BLUP), the empirical Bayes approach (EB) and the hierarchical Bayes approach (HB). We will focus on the discussion of the BLUP and be brief on the EB and HB approaches. Interested readers can find a rather thorough review on all three approaches from Ghosh and Rao.[8]

For the estimation of parameters from the random effect model using sampled data only, we start with a more general mixed effect model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \mathbf{e}\,, \tag{11}$$

where

$$E(\boldsymbol{\nu}) = 0\,, \quad V(\boldsymbol{\nu}) = \mathbf{G}\,, \quad E(\mathbf{e}) = 0\,, \quad V(\mathbf{e}) = \mathbf{R}\,,$$

and $\boldsymbol{\nu}$ and $\mathbf{e}$ are assumed to be independent of each other. Parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are obtained by solving the following equations simultanenously:

$$\begin{aligned} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\nu} &= \mathbf{X}'\mathbf{R}^{-1}\boldsymbol{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\boldsymbol{\nu} &= \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}\,, \end{aligned} \tag{12}$$

which can be expressed alternatively as:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \hat{\boldsymbol{\nu}} &= \mathbf{G}\mathbf{Z}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\,, \end{aligned} \tag{13}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

With the variance-covariance matrices $\mathbf{G}$ and $\mathbf{R}$ known, it is proved that $\hat{\beta}$ derived above is the best linear unbiased estimator and $\hat{\nu}$ is the best linear unbiased predictor in the measures of mean squared error.

In practice the variance-covariance matrices are unknown and have to be estimated from the data. Estimation of the variance-covariance matrix can be accomplished by using the restricted maximum likelihood (REML) approach proposed by Patterson and Thompson.[23] REML yields unbiased estimates of the variance covariance parameters for balanced designs. Technically, the optimality of the BLUP is lost when one uses estimated variance covariance matrices. However, such an approach coincides with the empirical Bayes approach with normal error assumption. Therefore, the empirical BLUP (EBLUP) and EB lead to identical estimates.

In the HB approach, a prior distribution on the model parameters is specified and the posterior distribution of the parameters of interest is then obtained. Inferences are based on the posterior distribution; in particular, a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior variance.

The EBLUP approach is implemented by various statistical software packages. For example, SAS Proc MIXED derives parameter estimates and prediction using the EBLUP method. To illustrate the various approaches

Table 4. Small area estimates by the synthetic estimator ($\hat{Y}_{\text{SYN}}$), the sample size dependent estimator ($\hat{Y}_{\text{SD}}$) and the empirical best linear unbiased predictor estimator ($\hat{Y}_{\text{EBLUP}}$).

Area No.	N_i	n_i	$\bar{X}_i$	$\bar{Y}_i$	$\hat{Y}_{\text{SYN}}$	$\hat{Y}_{\text{SD}}$	$\hat{Y}_{\text{EBLUP}}$
1	1	0	137.70	32.68	26.93	26.93	28.41
2	6	1	106.18	24.12	20.76	29.48	22.53
3	4	2	23.38	4.64	4.57	8.92	6.67
4	1	0	45.64	4.12	8.93	8.93	8.97
5	8	2	102.08	21.08	19.96	22.05	20.77
6	6	2	77.96	22.53	15.25	23.81	18.21
7	6	2	125.81	28.36	24.60	28.88	28.28
8	6	1	79.80	13.50	15.61	11.99	16.00
9	27	10	104.06	17.33	20.35	17.59	17.60
10	5	2	84.57	13.41	16.54	14.59	16.33
11	12	4	93.84	19.43	18.35	20.99	21.03
12	7	3	85.55	15.12	16.73	9.95	13.85
13	4	1	120.90	23.79	23.64	32.75	25.33
14	6	2	177.58	38.00	34.73	38.06	39.77
15	13	5	94.96	12.92	18.57	14.41	17.31
16	2	1	123.30	21.01	24.11	28.08	22.59
Average relative error %					21.31	25.49	19.59
Average squared error:					13.46	16.98	7.10

on small area estimation discussed in this section we use the synthetic population in Table 3. In Table 4 we compare the small area estimates of means derived by the synthetic estimator ($\hat{Y}_{\text{SYN}}$), the sample size dependent estimator ($\hat{Y}_{\text{SD}}$) and the empirical BLUP estimator ($\hat{Y}_{\text{EBLUP}}$) to the true small area mean ($\bar{Y}_i$). Note that the true means are available to us because the data is simulated. We also define two criteria for comparing the overall performances of the estimators across all small areas.

$$\text{Average relative error} = \frac{1}{16}\sum_{i=1}^{16}\frac{|\hat{Y}_i - \bar{Y}_i|}{\bar{Y}_i}\,,$$

$$\text{Average squared error} = \frac{1}{16}\sum_{i=1}^{16}(\hat{Y}_i - \bar{Y}_i)^2\,.$$

These two criteria provide measures on the overall bias and efficiency of the various estimators. The EBLUP estimator is shown to perform the best in having the smallest average relative error and the smallest average squared error. The sample size dependent estimator failed to improve on

the performance of the synthetic estimator, perhaps due to the use of a non-optimal weight.

So far we have discussed methods for small area estimation appropriate for a continuous outcome. Small area estimation for discrete outcomes such counts and proportions are often desired as well. Random-effect models can also be applied in these situations.

In the generalized linear model framework, the discrete outcome has its first two moments specified as:

$$E(y_{ij}|\nu_i) = h(x_{ij}\beta + \nu_i) = \mu_{ij}\,, \tag{14}$$

$$V(y_{ij}|\nu_i) = a_i\nu(\mu_{ij})\,, \tag{15}$$

where ν_i is assumed to be normally distributed with mean 0 and variance covariance matrix D.

There are two steps involved in small area estimation from discrete outcome, similar to the continuous variable case. In the first step we estimate the model parameters using the sampled data only. The second step involves prediction using the estimated model parameters and the auxiliary values for the non-sampled units in the population.

Several estimation approaches exist for parameter estimation from the generalized linear mixed model. The first approach is the full likelihood approach which requires the specification of the distributions of the random effects and often requires numerical integration of the likelihood function over the distribution of the random effect variables. To overcome these problem, Breslow and Clayton[3] proposed the penalized quasi-likelihood approach (PQL) where only the first two moments are specified and parameter estimation can be achieved using iterative weighted least square estimation. Raghunathan[24] proposed a quasi-empirical Bayes method for small area estimation on discreate outcomes.

We concentrate on a description of the PQL method, simply because it is implemented in some statistical software packages. For a detailed derivation of the PQL equations see Breslow and Clayton.[3] The PQL method requires iteratively solving the following equations:

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\,,$$

$$\hat{\nu} = \mathbf{D}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})\,, \tag{16}$$

where $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{D}$, and $\mathbf{W}$ is a diagonal matrix with the elements:

$$w_{ij} = \frac{\mathrm{var}(y_{ij}|\nu_i)}{g'(\mu_{ij})^2}$$

and $g(\mu_{ij}) = h^{-1}(x_{ij}\beta + \nu_i)$, and

$$Y_{ij} = x_{ij}\beta + \nu_i + (y_{ij} - h(x_{ij}\beta + \nu_i))g'(\mu_{ij})\,.$$

In the context of the previous example on a hospital survey, suppose we wish to estimate the average cancer-specific remission rate for each county. Such rates can be used in part to assess the quality of a hospital. We use the same setting as in the previous example and assume that the same 38 hospitals are randomly selected. A complete data set on all 114 hospitals are generated using a two-stage model:

$$E(y_{ij}) = m_{ij}\mu_{ij}\,, \quad \log\frac{\mu_{ij}}{1-\mu_{ij}} = 0.2x_{ij} + \nu_i\,,$$

where $\nu_i \sim N(0, 0.9)$.

The synthetic samples are presented in Table 5. We are interested in estimating the average cancer remission rate for each county. In Table 6 we present the estimates of proportions obtained using the PQL method as implemented in the SAS Glimmix macro.[20] These estimates are compared

Table 5. Data from a simple random sample drawn from a synthetic population ($n = 38, N = 114$).

Area No.	N_i	n_i	$x_{ij} \times 100$	m_{ij}	y_{ij}	Area No.	N_i	n_i	$x_{ij} \times 100$	m_{ij}	y_{ij}
1	1	0	–	–	–	9	27	10	86.54	64	19
2	6	1	169.99	37	1				101.55	9	1
3	4	2	6.16	24	9	10	5	2	61.34	126	26
			30.09	3	1				102.77	10	2
4	1	0	–	–	–	11	12	4	61.49	3	1
5	8	2	94.60	52	26				99.77	57	24
			129.93	47	16				52.08	17	9
6	6	2	90.53	24	15				136.46	5	3
			86.34	22	13	12	7	3	92.28	1067	455
7	6	2	95.10	9	4				93.73	246	114
			130.45	72	37				62.29	8	2
8	6	1	46.09	37	7	13	4	1	164.14	25	6
9	27	10	89.88	3	1	14	6	2	134.88	64	32
			145.68	80	22				164.94	9	4
			113.54	246	58	15	13	5	64.44	94	12
			114.54	8	1				63.51	40	5
			96.51	25	7				73.10	3	0
			84.58	340	73				109.66	57	1
			139.38	390	87				123.09	3	1
			117.24	24	7	16	2	1	186.16	45	19

Table 6. Small area estimates of proportions (in %) by raw proportion and by the PQL method. $\bar{Y}_i$ is the true population rate for each small area.

Area No.	N_i	n_i	$\bar{X}_i \times 100$	$\bar{Y}_i$	$\hat{Y}_{\text{raw}}$	$\hat{Y}_{\text{PQL}}$
1	1	0	137.70	30.00	33.55	35.43
2	6	1	106.18	44.30	51.35	48.88
3	4	2	23.38	37.66	37.04	36.70
4	1	0	45.64	40.38	33.55	35.28
5	8	2	102.08	36.93	42.42	41.97
6	6	2	77.96	68.47	60.87	57.74
7	6	2	125.81	47.39	50.62	49.50
8	6	1	79.80	18.16	18.92	21.80
9	27	10	104.06	23.15	23.21	23.28
10	5	2	84.57	21.70	20.59	21.41
11	12	4	93.84	45.78	45.12	44.41
12	7	3	85.55	42.98	43.22	43.17
13	4	1	120.90	20.60	24.00	26.54
14	6	2	177.58	43.70	49.32	48.26
15	13	5	94.96	11.45	9.64	10.85
16	2	1	123.30	42.29	42.22	41.26
Average relative error %					6.89	6.71
Average squared error:					0.0011	0.0013

to the direct estimates using data from each small area only. For counties without sampled hospitals, the overall mean is used as the estimate. The average relative error and the average squared error are defined in the same way as in the previous example and are included in the table. Although the PQL estimates show smaller overall bias than the direct estimates, it has a slightly larger average squared error than the direct estimates.

Raghunathan[24] demonstrated the quasi-empirical Bayes method on a data set to estimate county-specific mean number of hospital admission for cancer chemotherapy. A Poisson model for count data was assumed.

We want to conclude this section by pointing out that the random effect model approach we described here is a general approach in modeling sampling data in that each small area mean is itself assumed to be random variables following certain distributions. Parameter estimation is always a concern because the estimates directly contribute to the prediction of the non-sampled units. It may be noted that we might appear biased in choosing to present in more details the techniques that are implemented in computer software packages. The emphasis simply reflects the convenience in deriving estimates for our example data. It does not, however, reflect

the superity of performances of the estimation approaches we presented. In fact many approaches for the discrete outcomes have yet to be compared in a well designed simulation study. Therefore, the readers are adviced to keep an open mind on estimation techniques when applying random effect models themselves and when more results on comparing various estimation methods are available.

4. Capture Recapture Models

The use of capture recapture models in epidemiology is generally different from the sampling surveys we discussed so far in the previous two sections. Capture recapture setting usually works with several sampling frames, instead of just one in conventional surveys. Capture recapture model concentrates on matching individuals identified by different sources rather than sampling selection from one sampling frame. However, capture recapture model do share a common goal with some sample survey in that it also focus on the estimation of the size of a population.

As an alternative to the community survey we introduced in Sec. 2, capture recapture systems can be thought as multiple surveys on the same population trying to estimate the same quantity. This is especially useful when there does not exist one complete sampling frame from which a conventional sample survey can be established to reliably estimate population characteristics.

Capture recapture methods have a long history. They were first introduced in the study of fish and wildlife populations before being adopted for other populations. In animal studies, the animals being captured by the first attempt will be marked and returned to wildlife. This allows cross-classify the animals captured by various attempts. Hence the name capture recapture. Various authors have argued against the use of capture recapture model for human population on the bases that the various capture attempts of humans are usually not random.[2,11] In epidemiological studies hospital records, doctors' medical files, medical prescription list are examples of various sources to locate individuals with certain disease. Each of these sources is incomplete by their true nature, and the problem is to estimate those missing from all sources.

The simplest capture recapture model is the so called two-source model used to estimate the unknown size of the population. The first sample provides the individuals for *marking* and the second sample provide the recapture.

Table 7. Layout of a two source capture recapture setting.

		Source B	
		Yes	No
Source A	Yes	n_{11}	n_{10}
	No	n_{01}	n_{00}

We begin with two source model having sources A and B. Let n_{11}, n_{10}, n_{01} and n_{00} be the numbers of individuals captured by both sources (n_{11}), by source A only (n_{10}), by source B only (n_{01}), and by neither source (n_{00}). Note that n_{00} is unobservable. By estimating n_{00}, the number of cases missing by both sources, we provide an estimate of the number of cases in the population. The layout of a two source capture recapture setting is illustrated in Table 7.

Four assumptions are implicitly made on capture recapture analysis:

(1) There is no change to the population during the investigation. Such a population is usually called a *closed* population.
(2) Individuals can be matched from sources A to source B.
(3) In each source each individual has the same chance of being included in the sample.
(4) The two sources are independent meaning a "Yes" from source A does not affect the chance of a "Yes" from source B.

In epidemiological studies assumptions 1 and 2 can generally be assumed true. However, assumptions 3 and 4 present the biggest problem and has been the subject of debate since the application of capture recapture model in epidemiological fields. Human subjects are known to be heterogeneous with regard to being "caught" by a specific source. Methods to incorporate covariate in the method is becoming available. Tilling and Sterne[30] gave the latest development including a review of previous work. Assumption 4 is invariably false and is perceived as the biggest weakness on the use of capture recapture models in epidemiology. Humans are not fish where the chance of being recaptured is truly independent of whether they have been marked. For example, if certain doctors refer their patients to certain hospitals, then hospital records and doctors' records will not be two independent sources. Fienberg[6] approached the interdependence among sources of capture using log-linear model framework. We will focus our presentation using this approach.

Depends on the ways of parametrization, a 2×2 contingency table can be represented by the following log-linear models:

$$\begin{aligned}
\log E(n_{11}) &= \mu \\
\log E(n_{01}) &= \mu + \mu_A \\
\log E(n_{10}) &= \mu + \mu_B \\
\log E(n_{00}) &= \mu + \mu_A + \mu_B + \mu_{AB}\,.
\end{aligned} \tag{17}$$

The parameters on the right hand side of the above equations represent the logarithm of the number expected for each cell. Notice that there are four parameters in the log-linear models but only three known quantities to use for the estimation. One parameter is unestimatable. The customary solution is to assume $\mu_{AB} = 0$ which is equivalent of assumption 4. Hence n_{00} can be estimated by its expected value under the log-linear models:

$$\hat{n}_{00} = e^{\hat{\mu}+\hat{\mu}_A+\hat{\mu}_B} = e^{\log n_{11}+\log \frac{n_{10}}{n_{11}}+\log \frac{n_{01}}{n_{11}}} = \frac{n_{10}n_{01}}{n_{11}}\,. \tag{18}$$

The estimate of the total sample size is:

$$\hat{N} = n_{11} + n_{10} + n_{01} + \hat{n}_{00} = n_{11} + n_{10} + n_{01} + \frac{n_{10}n_{01}}{n_{11}}\,. \tag{19}$$

An example was taken from Bruno[2] and modified for presentation here. Four sources were used to identify known cases of diabetes among the residents in the area of Casale Nonferrato in Northern Italy. Data are presented in Table 8. Here, we illustrate the example with the first two sources only.

Source A: A list of all patients with a previous diagnosis of insulin-dependent diabetes mellitus (IDDM) or non-insulin dependent mellitus (NIDDM), via diabetes clinic and/or family physicians.

Source B: A list of all patients discharged with a primary or secondary diagnosis of diabetes in all public and private hospitals in the region.

Table 8. Capture recapture models with two sources.[2]

		Source B	
		Yes	No
Source A	Yes	377	1417
	No	115	n_{00}

Table 9. Parameter estimates from a two source log-linear model with the interaction term set to zero.

Parameter	Estimate	Standard Error
μ	5.8201	0.0545
μ_A	−1.0752	0.1082
μ_B	1.4362	0.0606

Using the log-linear approach described above, we fit a log-linear model with the PROC GENMOD procedure in SAS with the log-link function. Parameter estimates were displayed in Table 9.

The estimated number of cases missed by both cases is:

$$\hat{n}_{00} = e^{\hat{\mu}+\hat{\mu}_A+\hat{\mu}_B} = 483.54\,.$$

The total number of diabetes estimated from using both sources is 2353. Notice in this example, source A identified 1754 cases of diabetes and source B identified 452 cases, corresponding to 75% and 19% of the estimated total cases by using both sources, respectively. Both sources are seen to be relatively incomplete.

Note in the two source setting, dependency between the two sources cannot be estimated without additional information. If external data is used to estimate the dependency, μ_{AB} in the above log-linear models, estimation of n_{00} may be possible. If we define P(A), P(B) and P(A∩B) be the probability of captured by source A only, by source B only and by both sources, respectively, we can define the dependence between the two sources to be positive if P(A ∩ B) > P(A)∗P(B), negative if P(A ∩ B) < P(A)∗P(B). It has been shown that positive dependence of sources tends to underestimate the true population size and negative dependence tends to overestimate.

Extensions to modeling more than two sources are straightforward. In a capture recapture model with k sources, it is customary to set the highest interaction parameter of order k to be zero. We illustrate the cases of more than two sources with a three-sources analysis.

Table 10. The layout of a three source (A, B and C) capture recapture model.

		Source B			
		Yes		No	
		Source C		Source C	
		Yes	No	Yes	No
Source A	Yes	n_{111}	n_{110}	n_{101}	n_{100}
	No	n_{011}	n_{010}	n_{001}	n_{000}

The layout of a three-source capture recapture model is presented in Table 10.

We use the saturated model to construct the log-linear models for the three-source analysis.

$$\log E(n_{111}) = \mu$$

$$\log E(n_{110}) = \mu + \mu_C$$

$$\log E(n_{101}) = \mu + \mu_B$$

$$\log E(n_{011}) = \mu + \mu_A$$

$$\log E(n_{100}) = \mu + \mu_B + \mu_C + \mu_{BC}$$

$$\log E(n_{010}) = \mu + \mu_A + \mu_C + \mu_{AC}$$

$$\log E(n_{001}) = \mu + \mu_A + \mu_B + \mu_{AB}$$

$$\log E(n_{000}) = \mu + \mu_A + \mu_B + \mu_C + \mu_{AB} + \mu_{BC} + \mu_{AC} + \mu_{ABC}\,. \qquad (20)$$

Recall that there are 8 models, 8 parameters and 7 observations (counts). Therefore, one parameter is unestimable. An untestable assumption is made so that inference is possible: $\mu_{ABC} = 0$.

With more than two sources there is the possibility that a model with fewer parameters will fit the model equally well as the saturated model with 7 parameters (μ_{ABC} is set to 0). Statistically, the model with the fewest number of parameters fitting the data is to be chosen to represent the data. This leads us to the discussion of model selection criteria.

Three methods are commonly used for log-linear model selection: the G^2 deviance statistic which is a likelihood ratio statistic comparing a current model to the saturated model, Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

For the three source model, the G^2 statistic can be expressed as:

$$G^2 = -2 \sum_{i,j,k} n_{ijk} \log \frac{n_{ijk}}{E(n_{ijk})}\,,$$

where $E(n_{ijk})$ is the expected cell counts under the assumed alternative model other than the saturated one. If a model represents the data, then the difference in deviance between the considered model and the saturated model for which $G^2 = 0$ has an approximate χ^2 distribution.

The Akaike's Information Criterion is defined to be:

$$\text{AIC} = G^2 - 2\,(\text{d.f. of the model})\,.$$

The Bayesian Information Criterion is defined as:

$$\text{BIC} = G^2 - \frac{\log N}{2\pi}(\text{d.f. of the model}),$$

where N is the total number of observed cases. Both AIC and BIC take the number of parameters in the model into consideration. The BIC also considers sample size. Both criteria select the model with the lowest value on the respective criterion.

A three-source example data from LaPorte *et al.*[17] is modified here to illustrate the issues in model selection. Capture recapture method was used to identify the most accurate and efficient approaches to monitor adolescent injuries. For our example, we consider the issue of accuracy only. We take three sources from the article: 127 identified by student monthly recalls at either 1 month or 4 months, 58 by medical excuses and 33 by daily attendance records. Data is presented in Table 11.

A series of seven possible models were fit by using the SAS system for log-linear model. Table 12 includes the values of three model selection criteria: the G^2, AIC and BIC, and the estimated number of cases missed by all three sources and the estimated total number of injuries from all three sources.

Table 11. Injuries captured by three sources[17]: A: student recall at either 1 or 4 month; B: medical excuses; C: daily attendance records.

		Medical Excuses			
		Yes		No	
		Attendance Record		Attendance Record	
		Yes	No	Yes	No
Student Recall	Yes	16	39	13	69
	No	0	3	4	n_{000}

Table 12. The values of G^2, AIC and BIC for all seven models on the LaPorte data.[17]

Number	Model	G^2	d.f.	p-value	AIC	BIC	$\hat{n}_{000}$	$\hat{N}$
1	A, B, C	8.1222	3	0.0436	2.1222	2.8509	6	150
2	A, B, C, AB	4.8084	2	0.0903	0.8084	1.2942	15	159
3	A, B, C, AC	7.4284	2	0.0244	3.4284	3.9142	5	149
4	A, B, C, BC	6.3624	2	0.0415	2.3624	2.8482	7	151
5	A, B, C, AB, AC	3.3987	1	0.0652	1.3987	1.6416	*	*
6	A, B, C, AB, BC	1.9983	1	0.1575	−0.0017	0.2412	21	165
7	A, B, C, BC, AC	5.8258	1	0.0158	3.8258	4.0687	5	149

Using the G^2 statistic, three models (models 2, 5 and 6) are not rejected at the 0.05 significance level which means that these models are not significantly different from the saturated model. Model 5 failed to converge in parameter estimates. The likelihood based statistic would have favored model 2 because it is the simplest model not rejected by G^2. In fact, this model was chosen by LaPorte *et al.*[17] in the original paper. This model predicts 15 cases missed by all three sources and total number of injuries is estimated to be 159. However, both AIC and BIC identified model 6 as the optimal model. The number of cases missed by all sources is estimated to be 21 using model 6 and the total number of cases is estimated to be 165.

Notice that the G^2 method is based on large sample theory which assumes that each cell count is reasonably large. When there are small or zero cell counts as in this example, the validity of the test is questionable. In the above example, we would favor the use of model 6 over that of model 2 based on this observation.

Notice also that the conclusion on the validity of each source is very much dependent on which model one has chosen to represent the data. For example, LaPorte *et al.*[17] stated that student recall is the most accurate source of identifying injury with an estimated accuracy of 86% (137/159) using estimates derived from model 2. If the alternative model 6 is chosen, the accuracy rate for student recall will be estimated to be 83% (137/165), although it remains the most accurate of the three sources.

AIC and BIC base their decision on minimization. Therefore, uniqueness of the selected model is generally satisfied. Simulation studies have been conducted to compare AIC, BIC and several modified forms of the two criteria.[12] In general, the two criteria are quite comparable.

To conclude this chapter, we would like to reiterate the need for proper statistical methods in analyzing complex survey data. Many large national survey data are now accessible to the public for secondary data analysis providing medical researchers unique opportunities to study relationship and trend on the national level. However, great care must be exercised in analytical methods if one is to draw proper conclusion. The intention of this chapter is not on a exclusive coverage of general techniques on analyzing sampling data. Instead our focus of this chapter is on "special models" used for sampling data in the field of epidemiology. Readers are referred to Cochran[5] and Kish[15] for the background knowledge on sampling theory, to Skinner, Holt and Smith[28] for more theoretical development on statistical inference on sampling data. Examples of analysis of health survey data can

be found in Leclerc *et al.*,[19] Korn and Graubard,[16] Graubard and Korn[9] and LaVange *et al.*[18]

Acknowledgment

The research os supported in part by NIH grant R01 15813.

References

1. Beckett, L. A., Scherr, P. A. and Evans, D. A. (1992). Population prevalence estimates from complex samples. *Journal of Clinical Epidemiology* **45**: 393–402.
2. Brenner, H. (1994). Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* **6**: 42–48.
3. Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9–25.
4. Chambless, L. E. and Boyle, K. E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazard models. *Communications in Statistics-Theory and Methods* **14**: 1377–1392.
5. Cochran, W. G. (1977). *Sampling Techniques*, John Wiley and Sons, New York.
6. Fienberg, S. E. (1972). The multiple-recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**: 591–603.
7. Gao, S., Hui, S. L., Hall, K. S. and Hendrie, H. C. Estimating disease prevalence from two-phase surveys with nonresponse at the second phase. *Statistics in Medicine*, in press.
8. Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science* **9**: 55–93.
9. Graubard, B. I. and Korn, E. L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research* **5**: 263–281.
10. Graubard, B. I. and Korn, E. L. (1999). Predictive margins with survey data. *Biometrics* **55**: 652–659.
11. Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews* **17**: 243–264.
12. Hook, E. B. and Regal, R. R. (1997). Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology* **145**: 1138–1144.
13. Hurvich, C. M and Tsai, C. L. (1995). Model selection for extended Quasi-likelihood models in small samples. *Biometrics* **51**: 1077–1084.
14. International working group for disease monitoring and forecasting (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology* **142**: 1047–1058.
15. Kish, L. (1965). *Survey Sampling*, John Wiley and Sons, New York.

16. Korn, E. L. and Graubard, B. I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of Royal Statistical Society* **A158**: 263–295.
17. LaPorte, R. E., Dearwater, S. R., Chang, Y. F. *et al.* (1995). Efficiency and accuracy of disease monitoring systems: Application of capture-recapture methods to injury monitoring. *American Journal of Epidemiology* **142**: 1069–1077.
18. LaVange, L. M., Stearns, S. C., Lafata, J. E., Koch, G. G. and Shah, B. V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research* **5**: 311–29.
19. Leclerc, A., Luce, D., Lert, F., Chastang, J. F. and Logeay, P. (1988). Correspondence analysis and logistic modelling: Complementary use in the analysis of a health survey among nurses. *Statistics in Medicine* **7**: 983–995.
20. Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). *SAS System for Mixed Models*, SAS Institute Inc., Cary, NC.
21. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley and Sons, New York.
22. Morganstein, D. and Brick, J. M. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, Bureau of Census, 861–866.
23. Patterson, H. D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**: 545–554.
24. Raghunathan, T. E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association* **88**: 1444–1448.
25. Research Triangle Institute (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*, Research Triangle Park, NC.
26. Roberts, G., Rao, J. N. K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**: 1–12.
27. Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**: 15–51.
28. Skinner, C. J., Holt, D. and Smith, T. M. F. (1989). *Analysis of Complex Surveys*, John Wiley and Sons, New York.
29. Taylor, D. H., Whellan, D. J. and Sloan, F. A. (1999). Effects of admission to a teaching hospital on the cost and quality of care for medicare beneficiaries. *The New England Journal of Medicine* **340(4)**: 293–299.
30. Tilling, K. and Sterne, A. C. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology* **149**: 392–400.
31. Warszawski, J., Messiah, A., Lellouch, J., Meyer, L. and Deville, J. C. (1997). Estimating means and percentages in a complex sampling surveys: Application to a French national survey on sexual behaviour. *Statistics in Medicine* **16**: 397–423.

About the Author

Sujuan Gao is currently an assistant professor in the Division of Biostatistics, Indiana University School of Medicine, Indianapolis, USA. She received her BS in Mathematics (1985) from Beijing Normal University, Beijing, China, and a PhD in Statistics (1991) from the University of Southampton, UK. She joined the Division of Biostatistics at Indiana University School of Medicine as a post-doctoral fellow in 1994 and became a faculty member shortly after. Sujuan Gaos research interests include the analysis of longitudinal data with missing values, the analysis of data collected using complex sampling scheme, the development and application of statistical methods in medical and epidemiological studies.

CHAPTER 19

THE USE OF CAPTURE-RECAPTURE METHODOLOGY IN EPIDEMIOLOGICAL SURVEILLANCE

ANNE CHAO

Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043
Tel: 886-3-571-5131 ext 3161; chao@stat.nthu.edu.tw

HSIN-CHOU YANG

Institute of Biomedical Science, Academia Sinica, Taipei, Taiwan

PAUL S. F. YIP

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

1. Introduction

Accurate and timely estimates of disease occurrence over time or across geographic area play an important role in disease monitoring and health care planning. Traditional simple random sampling and other probabilistic sampling schemes are not easily applicable to such situations or are prohibitively expensive. Multiple surveillance systems are usually employed to ascertain cases using different resources or efforts. Although some studies manage to locate almost all patients, most epidemiological approaches merging different lists and eliminating duplicate cases are likely to significantly underestimate true occurrence rates.[23,25] That is, the final merged list misses those who are in the target population but were missed by all lists. This chapter discusses the use of capture-recapture models to estimate the number of missing cases under proper assumptions. We use three real data sets to illustrate the use of the capture-recapture methodology to correct for under-ascertainment of cases in epidemiological surveillance.

1.1. *Example* 1. *Data on hepatitis A Virus (HAV)*

Chao *et al.*[12] documented the details on a large outbreak of the HAV that occurred in and around a college in northern Taiwan from April to July 1995. Cases of students in that college were ascertained by three sources: (1) P-list: records based on a serum test taken by the Institute of Preventive Medicine, Department of Health of Taiwan; 135 cases were identified. (2) Q-list: local hospital records reported by the National Quarantine Service; 122 cases were found. (3) E-list: records collected by epidemiologists; of which there were 126 cases. Merging the three lists by eliminating duplicate records resulted in 271 ascertained cases.

The categorical data are shown in Table 1 where all ascertained cases are classified according to their presence/absence in the three lists. Presence or absence on any list is denoted by 1 and 0, respectively. For three lists, we can use a sequence of three numbers (each is either 0 or 1) to denote the record of each individual. For example, a record (001) describes an individual on the third list only and a record (011) describes an individual on the second and third lists but not on the first list. The three lists are displayed in an order of P, Q and E; this ordering is arbitrary and any legitimate inference procedure should be independent of the ordering of the lists. Those patients who were missed by all lists have the record (000). There are seven observable records and their counts over all ascertained cases are denoted as Z_{001}, Z_{010}, Z_{011}, Z_{100}, Z_{101}, Z_{110} and Z_{111}. From Table 1, there were 63 people listed in the E-list only, 55 people listed in the Q-list only, and 18 people listed in both lists Q and E but not in the P-list. Similarly, we can interpret the other records. In the P-list, there were 135 cases, which means $Z_{1++} = Z_{100} + Z_{101} + Z_{110} + Z_{111} = 135$.

Table 1. Categorical data on hepatitis A virus.

Hepatitis A list			
P	Q	E	Data
0	0	0	Z_{000} =??
0	0	1	$Z_{001} = 63$
0	1	0	$Z_{010} = 55$
0	1	1	$Z_{011} = 18$
1	0	0	$Z_{100} = 69$
1	0	1	$Z_{101} = 17$
1	1	0	$Z_{110} = 21$
1	1	1	$Z_{111} = 28$

Here, when we add over a sample, the subscript corresponding to that sample is replaced by a "+" sign. Similar relationship holds for the other two lists.

The number of different cases ascertained in at least one of the lists, 271 in this case, is the sum of all observable cell counts. Epidemiologists suspected that the observed number of cases considerably undercounted the true number of infected and an evaluation of the degree of undercount was needed.[12] The purpose was then to estimate the number of missed cases, Z_{000}, or equivalently, to estimate the number of total individuals who were infected in the outbreak. This data set was analyzed in Chao *et al.*[11] As opposed to many real data sets, this one has the advantage of a known true number of infected because a screen serological check for all students was conducted after the three surveys. In this chapter, we therefore select the HAV data set as an illustrative example to assess the relative merit of various estimation methods.

1.2. *Example* 2. *Data on Neurologic Illness (Stratified by Diagnostic Group)*

Bobo *et al.*[4] reported a comprehensive surveillance system for acute neurologic illness in children from August 1987 to July 1988 in two States of USA. Three surveillance strategies were employed: (1) Hospital surveillance system (H-list): Cases were identified based on hospitals discharge records. (2) Provider surveillance system (P-list): Cases were reported by pediatricians and neurologists. (3) Study staff surveillance system (S-list): Cases collected by the staff members by visiting all participating facilities and checking clinical records of potential cases. For this data set, relevant covariates (auxiliary or explanatory variables) include geographic location (Oregon or Washington), gender and primary diagnostic groups (encephalopathies, infantile spasms, afebrile seizure and complex febrile seizure). These four groups are referred to as stratum A, B, C and D for convenience.

Bobo *et al.*[4] found that substantial difference exists in case ascertainment rates by diagnostic groups. The post-stratified data by diagnostic groups are shown in Table 2. This covariate (group) is also referred to as a post-stratifying or stratifying variable. The data structure for each group is similar to that in Example 1. The collapsed data over the four groups are shown in the last column. In the original data, there were 626 ascertained cases. In Table 2 and our analysis in Sec. 5, we only consider 619 cases with

Table 2. Categorical data on neurologic illness.

List			Diagnostic Group (Stratum)				
H	P	S	A	B	C	D	Total
0	0	0	?	?	?	?	?
0	0	1	11	7	131	103	252
0	1	0	2	1	38	4	45
0	1	1	6	5	31	20	62
1	0	0	6	1	37	26	70
1	0	1	7	2	62	44	115
1	1	0	1	1	14	11	27
1	1	1	2	4	31	11	48

known diagnostic groups. There were 260, 182 and 477 cases, respectively, in H-list, P-list and S-list.

Despite the comprehensive surveillance systems, Bobo *et al.*[4] concluded that there were still some people who could not be identified. They performed capture-recapture adjustment for the data within each stratum and the collapsed data in order to obtain an accurate occurrence rate for various sub-populations defined by the available covariates. Their results showed that the ascertainment rate for the four groups were 82%, 94%, 69% and 91%, respectively. The rate was substantially low for the afebrile seizures.

1.3. *Example* 3. *Drug Data (Stratified by the Length of Time on Drug)*

Wittes[50] presented an ascertainment data set on patients receiving synthetic penicillin called methicillin. Cases were identified by the following four systems: (1) intravenous nurses (100 cases); (2) hospital floor nurses (21 cases); (3) hospitals pharmacists (156 cases) and (4) medication sheets (348 cases). We refer to these four lists as list 1, 2, 3 and 4, respectively. A total of 428 cases were found. Wittes[50] indicated that the length of time a patient was given the drug was related to his/her probability of being recorded. The original data consist of four strata for the time length (1–3 days, 4–6 days, 7–10 days and 11+ days). We combine the last two strata and the data are shown in Table 3. For each stratum, there are 15 observable presence/absence records and each can be expressed by a sequence of four numbers.

Wittes[50] found that an independent model (see Sec. 3 for explanation of the model) in each stratum fitted well and obtained an estimate of 544

Table 3. Categorical data on drug.

List				Usage on Drug (stratum)			
1	2	3	4	1–3 days	4–6 days	7+ days	Total
0	0	0	0	?	?	?	?
0	0	0	1	48	83	66	197
0	0	1	0	18	13	12	43
0	0	1	1	14	33	27	74
0	1	0	0	1	4	1	6
0	1	0	1	1	1	1	3
0	1	1	0	1	0	0	1
0	1	1	1	0	3	1	4
1	0	0	0	8	6	6	20
1	0	0	1	8	16	17	41
1	0	1	0	1	2	5	8
1	0	1	1	1	6	17	24
1	1	0	0	0	0	1	1
1	1	0	1	2	0	2	4
1	1	1	0	0	0	1	1
1	1	1	1	0	0	1	1

(s.e. 22.4) for the total number of patient receiving the drug. Dependence was suspected between lists 3 and 4 because the records from the pharmacy were duplicates of the medication sheets. To eliminate this possible dependence, the lists 3 and 4 were combined to form only one list. Then based on this combined list, list 1 and list 2, an estimate of 536 was obtained under independence. Both models provide evidences that a non-negligible number of patients were missed by all four identification sources.

One common goal for the above three examples is to find out under what assumptions or models we can estimate the number of missing cases and adjust for under-ascertainment considering the relevant covariate in the analysis. This has analogues in the biological sciences: Estimating the number of unseen animals in a closed population considering environmental factors or individual covariates. Here, a closed population means that there is no addition and loss so that the population size is a constant during the study period. The estimation of population size is a classical problem and has been extensively discussed in the literature.

Biologists have long realized that it is almost impossible to determine the size of a population by counting every animal. Most animals cannot be drawn like balls in an urn or numbers on a list, thus special types of sampling schemes have been developed. Capture-recapture sampling has

been widely used to adjust for undercount in the biological sciences. The recapture information (i.e. source-overlap information or source intersection) collected by marking or tagging can be used to estimate the number of missing under proper assumptions. Therefore, it is not necessary to count every animal in order to obtain an accurate estimate of population size.

In contrast, epidemiologists have attempted to enumerate all relevant cases to obtain the prevalence rates for various diseases. Cases in various lists are usually merged and any duplicate cases are eliminated. The overlap information is thus ignored. This typical approach assumes complete ascertainment and does not correct or adjust for under-ascertainment. As Hook and Regal[23,25] indicated, most prevalence surveys merging several records of lists are likely to miss some cases and thus be negatively biased. There is relatively little literature in the health sciences on the assessment of the completeness of these types of surveys or on the adjustment for under-ascertainment. Therefore, as commented by LaPorte *et al.*,[35] people know more about the number of animals than the count of diseases. In a similar way that ecologists and biologists count animals, we introduce with proper modification in this chapter the use of capture-recapture models to count human populations.

In Sec. 2, the background and motivation of the capture-recapture technique and its adaptation for use in human populations are reviewed. Section 3 summarizes the capture-recapture models when no covariates are available. Section 4 presents stratified capture-recapture models including covariates (or stratifying variable) effects. The analyses of the above three data sets are shown in Sec. 5. Concluding remarks are discussed in Sec. 6.

2. Background and Motivation of Capture-Recapture

Capture-recapture sampling was originally developed for estimating demographic parameters of animal populations. In a typical animal capture-recapture experiment, traps or nets are placed in the study area at several times, often called capture occasions (or trapping samples). At the first occasion, a number of animals are captured. A tag or mark with a unique number or record is attached to each captured animal. These animals are then released back into the population. At each subsequent occasion, animals that are first-captures are similarly marked and the tag numbers of re-captures are recorded. At the end of the experiment, a sequence of samples is obtained and the complete capture history for each captured animal is known.

Why is marking or tagging necessary in animal studies? Clearly, many animals look the same to humans and individuals cannot be identified by sight. Marking or tagging is used to distinguish captured individuals so that the recapture information (overlap information) by marking or tagging can be used to estimate the number of missing animals in the experiment. Marks include banding and tagging, paint brushing, toe clipping, ear clipping and in some species individual animals can be identified. Intuitively, if the proportion of newly captured animals in the later capture occasions is high, we know that the population size is much higher than the number of distinct captures. On the other hand, if there is a low proportion of newly captured animals, then we are likely to have caught most of the animals and the population size is close to the number of captured animals.

According to Seber,[44] the original idea of two-sample capture-recapture technique can be traced back to Laplace, who implemented it to estimate the population size of France. The interesting history of Laplaces survey conducted in 1802 was described in Cochran.[13]

The earliest applications to ecology include Petersens and Dahls work on fish populations and Lincolns use of band returns to estimate waterfowl in 1930. More sophisticated statistical theory and inference procedures have been proposed since Darrochs[17] paper, in which the mathematical framework of this topic was founded. Seber,[44–46], Pollock[39] and Schwarz and Seber[43] provided comprehensive reviews on the methodologies and applications.

The capture-recapture technique has been applied to human populations under the term "multiple-record system".[20,32,33,47,50,52] The special two-sample cases are often referred to as the "dual-system" or "dual-record system". For ascertainment data, if each list is regarded as a trapping sample and identification numbers, names and other characteristics are used as tags or marks, then this framework is similar to a capture-recapture setup for wildlife estimation. Comparisons of the applications to human and animal populations are listed in Table 4.

The earliest references to the application of the capture-recapture techniques to health science included the pioneering paper by Sekar and Deming[47] for two samples, Wittes and Sidel[52] for three samples, Wittes[50] for four samples, Wittes *et al.*[51] and Fienberg[20] for five samples. Hook and Regal[23] also suggested the use of capture-recapture models even for apparently exhaustive surveys. In recent years, there has been growing interest in the use of this technique in various disciplines. For example, another important application area is software reliability.[5] Hook and Regal,[25]

Table 4. Comparison of capture-recapture applied to human and animal populations.

Human Populations (Multiple-List System)	Animal Populations (Capture-Recapture Sampling)
Similarities:	
Lists, sources, records	Trapping samples or occasions
Identification numbers and/or names	Marks or tags
Ascertainment probability	Capture probability
Differences:	
Usually only 2 to 5 lists	Any t number of samples or occasions ($t \geq 2$)
No natural time ordering among lists	Natural time ordering in samples or occasions
Different ascertainment methods	Identical trapping methods
Some Shared Models Considered in this Chapter:	
Log-linear models[14,20,32,33]	
Sample coverage method[10,11]	
Logistic regression models[2,30,31,53]	

IWGDMF[32,33] and Chao[8] provided overviews of the applications of the capture-recapture models specifically to epidemiological data. However, some critical comments and concerns about the use of capture-recapture models in analyzing ascertainment data have been expressed by several authorsd.[15,19,34,38,42] For some of the concerns, Chao *et al.*[11] provided relevant discussion from a statistical point of view.

As shown in Table 4, there are some principal differences between wildlife and human applications. Researchers in wildlife and human populations have developed models and methodologies along separate lines. In Table 4, we list the approaches that are applicable to both populations. We will address those approaches in the next two sections after the introduction of notational conventions.

Throughout this paper, we use the following notational conventions:

- The true unknown population size (i.e. the number of individuals in the target population) is N and all individuals can be conceptually indexed by $1, 2, \ldots, N$.
- There are t lists (samples, records, or sources) and they are indexed by $1, 2, \ldots, t$.
- There are M identified individuals, i.e. M equals to the sum of all observable cell counts.
- Denote $Z_{s_1,s_2,\ldots,s_t}$ as the number of individuals with record $s_1, s_2, \ldots, s_t$, where $s_j = 0$ denotes absence in list j and $s_j = 1$ denotes presence in list j.

- Denote n_j, $j = 1, 2, \ldots, t$ as the number of individuals identified in the jth list.
- Denote P_{ij} as the capture or ascertainment probability of the ith individual in the jth list.

Basic assumptions are:

- All individuals act independently.
- Interpretation or definition for the characteristic of the target population should be consistent for all data sources.
- Closure assumption: The size of the population is approximately a constant during the study period.
- Ascertainable assumption: Any individual must have a positive probability to be ascertained by any source; any un-ascertainment is purely due to small chance rather than impossibility. Remark: When a random sample is feasible in a dual-system, some special types of structure zeros are permitted; see Sec. 6.1 of Chao it et al.[11]
- For all sources, identification marks are correctly recorded and matched.

Traditional statistical approach further assumes that the samples are independent. In animal studies, this traditional assumption is in terms of an even more restrictive "equal-catchability" assumption, i.e. in each fixed trapping sample all animals have the same capture probability. (Equal catchability assumption implies independence among samples but the reverse is not true; see Sec. 3.) Dependence or unequal catchabilities may be caused by the following two sources:

(1) Local dependence (also called list dependence) within each individual/stratum: Conditional on any individual, the presence/absence in one source has a direct causal effect on this individual's probability of inclusion in other sources. In animal populations, local dependence arises mainly from a behavioral response to capture due to identical trapping method. Animals may become trap-happy, and have an increased probability of subsequent capture, if baited traps are used whereas they may become trap-shy, and have a decreased probability of subsequent capture, if mist nets or ear clipping are used. Local dependence within each individual/stratum may also arises for human populations. For example, the probability of going to a hospital for treatment for any individual depends on his/her result on the serum test of the HAV, leading to dependence between the ascertainment of the serum sample and that of the hospital sample.

(2) Heterogeneity among individuals: Even if the two lists are independent within an individual/stratum, the ascertainment of the two sources may become dependent if the capture probabilities are heterogeneous among individuals/strata. Hook and Regal[24] presented an interesting epidemiological example. For many populations, capture or ascertainment probability may vary with age, gender, location, activity, diagnostic symptom, severity of illness or other individual characteristics. For example, in animal populations, some females tend to be less likely to be captured in all trapping occasions, leading to dependence among samples. In human populations, severe cases are more ascertainable in all lists than less severe cases, also leading to dependence.

These two types of dependencies are usually confounded and cannot be easily disentangled in a data analysis. Lack of independence leads to a bias (called "correlation bias") for the usual estimator which assumes independence. For example, in two-list cases, the widely used Petersen estimator underestimates the true size if both samples are positively dependent. Conversely, it overestimates for negatively dependent samples (see Sec. 3). A similar conclusion holds for a general number of samples.

When only two lists are available, the data are insufficient for estimating dependence unless additional covariates are available. All existing methods unavoidably encounter this problem and adopt the independence assumption. Therefore, when there are no available covariates, at least three lists are required to model dependence between samples.

3. Models Without Covariates

For closed populations, the most commonly used models were proposed by Pollock.[37,39,49] This class of models considers time (or occasional) effect, behavioral response to capture and heterogeneity among individuals. For models incorporating behavioral response, which induces local dependence among samples, the capture probability in a sample depends on whether the animal was captured in the "previous" samples. Hence, the ordering of the trapping samples is involved and estimators do depend on the ordering of samples. Since there is usually no ordering among human lists or the ordering may vary with individuals, such models with behavioral response are rarely adopted in modeling local dependence for humans.

Heterogeneous model which allows different capture probabilities among animals are potentially useful in health science. A commonly used estimator for such a model is the jackknife estimator proposed by Burnham and

Overton.[6] To assure the jackknife work well, the number of trapping samples should be at least five.[37] Similar condition is also required for other estimation procedure.[11] However, only two to five human lists are usually available. Therefore, most ecological models have limited use in epidemiological applications. Consequently, we only focus on another two approaches: log-linear models and sample coverage approach.

3.1. *Log-linear models*

The log-linear models[14,20] have been extensively used to handle dependence among samples. Part of this approach is discussed in Chapter 13. The methodology applied to human diseases was well covered in IWGDMF.[32,33] In this approach, various log-linear models are fitted to the observed cells. How well a model fits the data is assessed using the deviance statistic and a model is usually selected based on the Akaikes information criterion (AIC). The chosen model is then projected onto the unobserved cell by assuming that there is no highest order interaction. The two types of dependencies can be modeled by including some specific interactions or common interaction in the models.

We use three lists for illustration. The log-linear approach models the logarithm of the expected value of each observable category. Let $I(A)$ denote the usual indicator function, i.e. $I(A) = 1$ if A is true, 0 otherwise. The most general model is

$$\begin{aligned}\log E(Z_{s_1,s_2,s_3}) = {} & u + u_1 I(s_1 = 1) + u_2 I(s_2 = 1) + u_3 I(s_3 = 1) \\ & + u_{12} I(s_1 = s_2 = 1) + u_{13} I(s_1 = s_3 = 1) \\ & + u_{23} I(s_2 = s_3 = 1) + u_{123} I(s_1 = s_2 = s_3 = 1)\,. \qquad (1)\end{aligned}$$

That is, $\log E(Z_{000}) = u$, $\log E(Z_{100}) = u + u_1$, $\log E(Z_{010}) = u + u_2$, $\log E(Z_{001}) = u + u_3$, $\log E(Z_{110}) = u + u_1 + u_2 + u_{12}$, $\log E(Z_{101}) = u + u_1 + u_3 + u_{13}$, $\log E(Z_{011}) = u + u_2 + u_3 + u_{23}$ and $\log E(Z_{111}) = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$. This is a reparametrization of the eight expected values.

For three-list data, we have seven observed categories, whereas there are eight parameters. Therefore, a natural assumption is that there is no three-list interaction term, i.e. $u_{123} = 0$. Intuitively, this means the complete 2×2 table formed with respect to lists 2 and 3 for individuals in list 1 and the incomplete 2×2 table for individuals not in list 1 have the same odds ratio. The sample odds ratio for the former table is $Z_{111}Z_{100}/(Z_{110}Z_{101})$

whereas the odds ratio for the latter table is $Z_{011}Z_{000}/(Z_{010}Z_{001})$. The assumption of $u_{123} = 0$ allows the following extrapolation formula $\hat{Z}_{000} = \hat{Z}_{001}\hat{Z}_{010}\hat{Z}_{100}\hat{Z}_{111}/(\hat{Z}_{110}\hat{Z}_{011}\hat{Z}_{101})$, which expresses the estimated missing cases as a function of the fitted values of other categories.

The independent model includes only main effects as given by $\log E(Z_{s_1,s_2,s_3}) = u + u_1 I(s_1 = 1) + u_2 I(s_2 = 1) + u_3 I(s_3 = 1)$, which is denoted by model (1, 2, 3) as used in categorical data analysis. The interaction terms are used to model dependence. If local list dependence arises in samples 1 and 2, then the interaction term u_{12} is included, and the model is denoted as model (12, 3) or 12/3. If local dependence also appears in samples 1 and 3, then the two interactions u_{12} and u_{13} are needed. The model is denoted as model (12, 13) or 12/13 and similarly for models (13, 2), (23, 1), (13, 23) and others.

The log-linear model can also be motivated by the Rasch[41] model and its generalizations which incorporate heterogeneity among individuals. The Rasch model assumes $\text{logit}(P_{ij}) = \alpha_i + \tau_j$, where P_{ij} denotes the capture probability of the ith individual on the jth list, $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ represents heterogeneity effects among individuals and $\{\tau_1, \tau_2, \ldots, \tau_t\}$ denotes the list effects. Since only dependence due to heterogeneity is considered in the Rasch model, the capture probability for the ith individual in any category can be determined by the product of $\{P_{ij}, j = 1, 2, \ldots, t\}$. A generalized Rasch model allows the heterogeneity effects $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ to be different for two or more separate groups of samples.

It has been verified[18,21] that the Rasch (generalized Rasch) model is equivalent to a quasi-symmetric (partial quasi-symmetric) model with some moment constraints. Except for the constraints, a quasi-symmetric model for the three-list case with no second-order interaction, i.e. $u_{123} = 0$, is equivalent to the model with first-order interactions identical; this is denoted by $(12 = 13 = 23)$ or simply H1 (which is called the first-order heterogeneity by IWGDMF).[32,33] Only one degree of freedom is used to model heterogeneity. A partial quasi-symmetric model which assumes the heterogeneity effects are identical only for the first and second lists, is equivalent to the model with $u_{13} = u_{23}$. This model is denoted as $(13 = 23, 12)$. Similarly, we have models $(12 = 13, 23)$ and $(12 = 23, 13)$ corresponding to other two partial quasi-symmetric models. Therefore, the dependence due to heterogeneity can be modeled by either a quasi-symmetric or a partial quasi-symmetric model. When both types of dependencies occur, they are inevitably confounded in the interaction or common interaction terms and cannot be separated.

The log-linear model can be similarly formulated when there are more than three lists. The basic assumption for 4 lists is the third-order interaction vanishes (i.e. $u_{1234} = 0$); that is, the 3-list interaction for individuals in list 1 is the same as that for individuals not in list 1. Local list dependence can be modeled by including the first-order interaction term $(u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34})$ and/or the second-order interaction $(u_{123}, u_{134}, u_{124}, u_{234})$. The Rasch model is equivalent to a model with the first-order heterogeneity H1 (i.e. $12 = 13 = 14 = 23 = 24 = 34$) and the second-order heterogeneity H2 (i.e. $123 = 124 = 134 = 234$), and thus denoted by (H1, H2). Two degrees of freedom are used to model heterogeneity. If additional local dependencies also occur between lists 1 and 2, lists 1 and 3, and lists 2 and 4, then we add three more parameters u_{12}, u_{13} and u_{24} to the model and the resulting model is denoted as (12/13/24, H1, H2). Parallel formulations can be obtained for the general case of t lists; see Lloyd[36] for details. The reader is referred to Agresti,[1] Coull and Agresti[16] and Fienberg *et al.*[21] for other related and useful models.

3.2. *Sample coverage approach*

This approach was proposed by Chao and Tsay[10] for the three-list case. The extension to a general case is presented in Tsay and Chao.[48] Details and relevant software are reviewed in Chao *et al.*[11] The approach aims to provide a measure to quantify the overlap information and also to propose parameters to quantify source dependence.

Dependence is modeled by parameters called the "coefficient of covariation" (CCV). To better understand the CCV parameters, we discuss the dependence measure only for the heterogeneous case. Let the two sets of probabilities, $\{P_{ij}; i = 1, 2, \ldots, N\}$ and $\{P_{ik}; i = 1, 2, \ldots, N\}$, denote the capture probabilities for N individuals in samples j and k, respectively. The CCV of samples j and k for a fixed-effect approach is defined as

$$\gamma_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(P_{ij} - \mu_j)(P_{ik} - \mu_k)}{\mu_j \mu_k} = \frac{1}{N} \frac{\sum_{i=1}^{N} P_{ij} P_{ik}}{\mu_j \mu_k} = -1\,, \qquad (2)$$

where $\mu_j = \sum_{i=1}^{N} P_{ij}/N = E(n_j)/N$ denotes the average probability of being listed in the jth sample. The magnitude of γ_{ij} measures the degree of dependence between samples j and k. The two heterogeneous samples are independent if and only if $\gamma_{ij} = 0$, i.e. $N^{-1} \sum_{i=1}^{N} P_{ij} P_{ik} = \mu_j \mu_k$ which means that the covariance between the two sets of probabilities is zero. Thus if only one set of probabilities is homogeneous, then it suffices to assure independence provided no local dependence exists.

Two samples are positively (negatively) dependent if $\gamma_{jk} > 0$ ($\gamma_{jk} < 0$), which is equivalent to $N^{-1}\sum_{i=1}^{N} P_{ij}P_{ik} > \mu_j\mu_k$ ($N^{-1}\sum_{i=1}^{N} P_{ij}P_{ik} < \mu_j\mu_k$), i.e. the average probability of jointly being listed in the two samples is greater (less) than that in the independent case. The CCV can be similarly defined for more than two sets of heterogeneous probabilities.

When there are only two lists, say, lists 1 and 2, the relative bias of Petersens estimator (bias divided by the estimate) is approximately $-\gamma_{12}$.[10,47] This explains the direction of the correlation bias for Petersens estimator, as stated in Sec. 2. Thus, the value of CCV also quantifies the correlation bias. The CCV for the general cases with two types of dependencies has been developed,[10] but it will not be addressed here. We remark that all CCVs in the general cases measure the mixed overall effect of the two types of dependencies.

The sample coverage is used as a measure of overlap fraction of the available lists. While the mathematical formula for the sample coverage is complicated, its estimator is intuitively understandable. The estimated sample coverage can be written as[10]

$$\begin{aligned}\hat{C} &= 1 - \frac{1}{3}\left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3}\right) \\ &= \frac{1}{3}\left[\left(1 - \frac{Z_{100}}{n_1}\right) + \left(1 - \frac{Z_{010}}{n_2}\right) + \left(1 - \frac{Z_{001}}{n_3}\right)\right],\end{aligned}$$

which is the average (over three lists) of the fraction of cases found more than once. Note that Z_{100}, Z_{010} and Z_{001} are the numbers of individuals listed only in one sample. Hence, this estimator is the complement of the averaged fraction of singletons. Obviously, singletons cannot contain any overlapping information. Define

$$\begin{aligned}D &= \frac{1}{3}[(M - Z_{100}) + (M - Z_{010}) + (M - Z_{001})] \\ &= M - \frac{1}{3}(Z_{100} + Z_{010} + Z_{001})\,.\end{aligned}$$

Here $(Z_{100} + Z_{010} + Z_{001})/3$ represents the average of the non-overlapped cases and recall that M denotes the total number of identified cases. Thus, D can be interpreted as the average of the overlapped cases. The sample coverage estimation procedures for the three-list case are summarized in the following:

(1) When the three sources are independent, a simple population size estimator is derived as:

$$\hat{N}_0 = D/\hat{C}\,. \tag{3}$$

It can also be intuitively thought of as ratio of overlapped cases to overlap fraction.

(2) When dependence exists and the overlap information is large enough (how large it should be will be discussed further below), we take into account the dependence by adjusting the above simple estimator $\hat{N}_0$ based on a function of two-sample CCVs. The resulting estimator has the following explicit form:

$$\hat{N} = \left[\frac{Z_{+11} + Z_{1+1} + Z_{11+}}{3\hat{C}}\right] \div \left\{1 - \frac{1}{3\hat{C}}\left[\frac{(Z_{1+0}Z_{+10})Z_{11+}}{n_1 n_2} + \frac{(Z_{10+} + Z_{+01})Z_{1+1}}{n_1 n_3} + \frac{(z_{0+1} + Z_{01+})Z_{+11}}{n_2 n_3}\right]\right\}. \tag{4}$$

(3) For relatively low sample coverage data, we feel the data do not contain sufficient information to accurately estimate the population size. In this case, the following "one-step" estimator $\hat{N}_1$ is suggested: (The estimator is called "one-step" because it is obtained by one iterative step from the above-mentioned adjustment formula.)

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}}[(Z_{1+0} + Z_{+10})\hat{\gamma}_{12} + (Z_{10+} + Z_{+01})\hat{\gamma}_{13} + (Z_{01+} + Z_{0+1})\hat{\gamma}_{23}]\,, \tag{5}$$

where CCV estimates are

$$\hat{\gamma}_{12} = \hat{N}'\frac{Z_{11+}}{n_1 n_2} - 1\,, \quad \hat{\gamma}_{13} = \hat{N}'\frac{Z_{1+1}}{n_1 n_3} - 1\,, \quad \hat{\gamma}_{23} = \hat{N}'\frac{Z_{+11}}{n_2 n_3} - 1\,, \tag{6}$$

and

$$\hat{N}' = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}}\left[(Z_{1+0} + Z_{+10})\left(\frac{D}{\hat{C}}\cdot\frac{Z_{11+}}{n_1 n_2} - 1\right) + (Z_{10+} + Z_{+01})\left(\frac{D}{\hat{C}}\cdot\frac{Z_{1+1}}{n_1 n_3} - 1\right) + (Z_{01+} + Z_{0+1})\left(\frac{D}{\hat{C}}\cdot\frac{Z_{+11}}{n_2 n_3} - 1\right)\right].$$

This one-step estimator can be regarded as a lower (upper) bound for positively (negatively) dependent samples. Hook and Regal[27] noted that most data sets used in epidemiological applications tend to have a net positive dependence. Thus, the one-step estimator is often used as a lower bound.

The above three estimators $(\hat{N}_0, \hat{N}, \hat{N}_1)$ will be simply referred to as sample coverage estimators if there is no confusion with $\hat{C}$. A bootstrap resampling method[10] was proposed to obtain estimated standard errors for the above three estimators and to construct confidence intervals using a log-transformation.[7] A relatively low overlap fraction means that there are many singletons. In this case, the undercount cannot be measured accurately due to insufficient overlap. Consequently, a large standard error is usually associated with the estimator $\hat{N}$ in Eq. (4). How large should the overlap information be? Chao *et al.*[11] suggested that the estimated sample coverage should be at least 55%. A practical data-dependent guideline can be determined from the estimated bootstrap s.e. associated with the estimator $\hat{N}$. If the estimated bootstrap standard error becomes unacceptable (say, it exceeds one-third of the population size estimate), then only the lower or upper bound in Eq. (5) is recommended.

The estimation procedure for the general t-sample case[11] is parallel to that for the 3-sample case as discussed above.

4. Models With Covariates

In animal populations, individuals covariates include age, gender, body weight, wing length and others; environmental covariates include temperature, rainfall, number of traps and others. For human populations, relevant covariates are age, gender, race, geographic area, marital groups, diagnostic group, time of onset, severity of diseases and many other explanatory variables. The covariate variables are also classified as either discrete (categorical type) or continuous (numerical type).

As discussed earlier, traditional approach depends on a crucial assumption of "equal-catchability". Heterogeneity in capture probabilities induces dependence among samples, which causes correlation bias in the usual estimator. One approach to assessing heterogeneity is based on the assumption that heterogeneity can be largely explained by some relevant observable covariates. If covariate is discrete, Sekar and Deming[47] were the first to suggest post-stratification to reduce the bias due to heterogeneity. That is, if the population can be divided into several homogeneous

sub-populations defined by relevant covariates, then a stratified analysis can be performed. That is, log-linear model or any other proper model is fitted to the data for each stratum, then all estimates are combined to obtain a final estimate.[32,33,50]

Pollock *et al.*[40] were the first to use a logistic model to include continuous covariates in the analysis. In this approach, covariates are used to model heterogeneous capture probabilities by a logistic regression. They developed an estimation procedure based on the full likelihood. However, the covariates for the un-captured animals are not observable. Therefore, they had to make some assumptions about the covariates for the un-captured animals. Huggins[30,31] and Alho[2] avoided this difficulty by using a likelihood conditional on the captured animals so that the covariates of the un-captured are not needed. After the coefficients of the logistic regression are obtained, a Horvitz–Thompson type of estimator[29] is then employed to obtain an estimate of population size. Alho *et al.*[3] applied this logistic regression approach to the 1990 census and the Post-Enumeration Survey of the United States. Yip *et al.*[53] extended this logistic regression to allow random removals in the experiments.

We now apply the logistic regression model to the data sets in Tables 2 and 3, in which there is only one stratifying variable. A unified model proposed in Huggins[30,31] and Yip *et al.*[53] can incorporate effects for covariates, capture occasions and behavioral response. If the times for the individuals being recorded on the respective lists are known, then the behavioral response effect for humans could be explored. Since such information for both examples is not available, we only consider a model with stratum effect and occasional effects. Assume that there are k strata. We need to construct $k-1$ dummy indicators to specify the effect of each stratum. That is, for the ith individual define the dummy variable $W_{is} = I$ (the ith individual is in the sth stratum), $s = 1, 2, \ldots, k-1$, a logistic regression model can be expressed as

$$\begin{aligned} \operatorname{login}(P_{ij}) &= \log\left(\frac{P_{ij}}{1-P_{ij}}\right) \\ &= a + c_j + \beta_1 W_{i1} + \beta_2 W_{i2} + \cdots + \beta_{k-1} W_{i,k-1}\,. \end{aligned} \tag{7}$$

In this model, the parameters are

a: baseline intercept,
$(c_1, c_2, \ldots, c_{t-1})$: occasional or list effect, $(c_t = 0)$,
$(\beta_1, \beta_2, \ldots, \beta_{k-1})$: stratum effect, β_2 denotes the effect of the sth stratum,

$s = 1, 2, \ldots, k-1$. (The effect of the kth stratum is assumed to be 0.)

Under this model, the capture probability for each capture record and thus the likelihood can be formulated. Then the maximum likelihood estimates of the parameters a, $(c_1, c_2, \ldots, c_{t-1})$ and $(\beta_1, \beta_2, \ldots, \beta_{k-1})$ are searched by numerical iteration. The population size is estimated by the Horvitz–Thompson estimator. Huggins[30,31] derived the variance of the resulting estimator by an asymptotic variance formula.

Note that the logistic type model has an advantage that it can be used to assess the effect of any type of covariate (both discrete and continuous). However, it does not take account of any local dependence and heterogeneity is entirely explained by covariates. To incorporate possible local dependence, a multivariate logistic model[22] might be needed and will be a worthwhile future research topic.

5. Analysis of Three Examples

5.1. *Hepatitis A virus data (Three lists)*

Table 5 shows the results for analyzing the HAV data. The first part of the table presents Petersens estimate based on any pair of lists. Although Petersens estimator is valid only under the restrictive independence assumption, they are practically useful as a preliminary analysis. It has been suggested[21,23] that estimates based on any two lists can be used to detect possible dependence. A substantially higher (lower) estimate signifies possible negative (positive) dependence for those two samples. For the HAV data, Petersens estimates are in the range of 330 to 380. The narrow range of these estimates would not indicate the possible direction of dependence at this stage.

The second part of Table 5 includes the results for all possible log-linear models fitted to the three-list data. The corresponding deviances and estimates of the total number of infected are also shown. The independent model produces an estimate of 388, which is close to the results for any two samples. Except for the saturated model, all the log-linear models, which consider local independence only and do not take into account heterogeneity, i.e., models (PE, Q), (QE, P), (PQ, E), (PQ, QE), (PQ, PE), and (QE, PE), do not fit the data well. All other models, which take heterogeneity only into account (quasi-symmetric and partial quasi-symmetric models) fit well. Those adequate models produce approximately the same estimates of 1300 with an approximate estimated s.e. of 520. In

Table 5. Analysis results for the HAV data.

Model/Method	Deviance	d.f.	AIC	Estimate of total infected (s.e.)
Petersens estimates for pair of lists:				
(P, Q)	0	0		336 (29)
(P, E)	0	0		378 (36)
(Q, E)	0	0		334 (30)
Log-linear models for three lists:				
(P, Q, E) independent	24.36	3	69.8	388 (23)
(PE, Q)	24.25	2	71.6	393 (27)
(QE, P)	21.33	2	68.7	413 (31)
(PQ, E)	21.14	2	68.5	416 (33)
(PQ, QE)	13.20	1	62.6	527 (79)
(PQ, PE)	19.42	1	68.8	464 (61)
(QE, PE)	19.90	1	69.3	452 (54)
(PQ, QE, PE)	0	0	51.4	1313 (521)
Quasi-symmetric (PQ = QE = PE)	0.96*	2	48.4	1313 (521)
Partial quasi-symmetric (PQ = QE, PE)	0.03*	1	49.4	1309 (519)
Partial quasi-symmetric (PQ = PE, QE)	0.86*	1	50.3	1306 (517)
Partial quasi-symmetric (QE = PE, PQ)	0.55*	1	49.9	1325 (528)

Sample coverage approach

	D	$\hat{C}$	$\hat{\gamma}_{12}$	$\hat{\gamma}_{13}$	$\hat{\gamma}_{23}$	Estimate (s.e.)
$\hat{N}_0$: Eq. (3)	208.7	0.513	0.21	0.08	0.22	407 (28)
$\hat{N}$: Eq. (4)	208.7	0.513	1.89	1.57	1.91	971 (925)
$\hat{N}_1$: Eq. (5)	208.7	0.513	0.51	0.34	0.52	508 (40)

*Deviance is not significant at the 5% level, which means a proper fit.

terms of AIC, the quasi-symmetric model is selected, but the relatively large estimated s.e. shows that the data are actually insufficient to fit a heterogeneous model.

The third part of Table 5 contains the sample coverage approach. The sample coverage is estimated to be $\hat{C} = 51.3\%$, and the average of the overlapped cases is equal to $D = 208.67$. If we ignore the possible dependence between samples, an estimate based on (3) is $\hat{N}_0 = D/\hat{C} = 208.67/0.513 = 407$, which is slightly higher than the estimate of 388 based

on the independent log-linear model. The estimator given in Eq. (4) is $\hat{N} = 971$, but a large estimated bootstrap s.e. (925) renders the estimate useless. The estimated s.e. was calculated by using a bootstrap method based on 1000 replications. We feel these data with a relatively low sample coverage estimate of 51% do not contain enough information to correct for undercount. The proposed one-step estimator in Eq. (5) is $\hat{N}_1 = 508$ with an estimated s.e. of 40 using 1000 bootstrap replications. The same bootstrap replications produce a 95% confidence interval of (442, 600) using a log-transformation.[7] We remark that the estimated s.e. and confidence intervals might vary from trial to trial because replications vary in the bootstrap procedures.

It follows from Eq. (6) that the CCV measures depend on the value of N. The CCV estimates in Table 5 based on the three estimates of N show that any two samples are positively dependent. As a result, the estimate $\hat{N} = 508$ can only serve as a lower bound. Also, the estimates assuming independence based on two samples should have a negative bias. However, we cannot distinguish which type of dependence (local dependence or heterogeneity) is the main cause of the bias.

In December 1995, the National Quarantine Service of Taiwan conducted a screen serum test for the HAV antibody for all students of the college at which the outbreak of the HAV occurred.[12] After suitable adjustments, they have concluded that the final figure of the number infected was about 545. Thus this example presents a very valuable data set with the advantage of a known true parameter. Our estimator $\hat{N}_1$ does provide a satisfactory lower bound. This example shows the need for undercount correction and also the usefulness of the capture-recapture method in estimating the number of missing cases.

5.2. *Stratified neurologic illness data (Three lists)*

Various models have been fitted to the stratified neurologic illness data and the results are shown in Table 6. Except for the first stratum, the Petersen estimate based on the H-list and P-list is much lower than the other two Petersens estimates. Thus positive dependence exists between the two lists. This finding is further confirmed by the CCV estimates (not reported) in the sample coverage approach.

If dependence is ignored, the pooled stratified estimate gives an estimate of 762 (s.e. 21), which is identical to that obtained from an un-stratified analysis. Also, the sample coverage estimate under independence for

Table 6. Various estimates of the population size for neurologic illness data (Standard error is in the parenthesis).

	Diagnostic Group (Stratum)					
Model/Method	A	B	C	D	Stratified	Un-stratified
Petersens estimate:						
(H, P)	59 (16)	18 (3)	365 (34)	192 (24)	634 (45)	631 (46)
(H, S)	46 (7)	24 (3)	395 (19)	298 (21)	763 (29)	761 (29)
(P, S)	36 (5)	22 (2)	469 (34)	264 (23)	791 (41)	789 (41)
Log-linear models:						
(H, P, S) Independent	43 (5)	22 (2)	429 (16)	268 (13)	762 (21)	762 (21)
(HP, S)	42 (4)	23 (2)	438 (19)	275 (15)	778 (25)	778 (24)
(HP, HS)	39 (4)	22 (2)	505 (44)	240 (12)	806 (46)	865 (76)
Quasi-symmetric	40 (6)	26 (8)	543 (82)	273 (30)	882 (88)	802 (41)
(HS = PS, HP)	43 (13)	23 (5)	533 (88)	242 (15)	841 (90)	808 (67)
Sample coverage:						
$\hat{N}_0$: Eq. (3)	43 (5)	23 (2)	436 (17)	255 (10)	757 (21)	762 (21)
$\hat{N}$: Eq. (4)	42 (20)	24 (15)	524 (68)	218 (15)	808 (74)	812 (65)
$\hat{N}_1$: Eq. (5)	42 (7)	23 (3)	468 (29)	239 (16)	772 (34)	782 (32)
Logistic regression model:						
Eq. (7) with covariate and list effects, Horvitz–Thompson estimate: 765 (s.e. 22)						

un-stratified data ($\hat{N}_0$, in Eq. (3)) yields exactly the same result. An analogous estimate of 765 (s.e. 22) is also obtained by a logistic regression model incorporating both covariate (stratifying variable) and the list effects. However, the deviance statistic of the logistic regression model is 38.8 with 22 degrees of freedom (P-value = 0.015), indicating an inadequate fit.

Since significantly positive dependence exists for the H-list and P-list, Bobo *et al.*[4] suggested fitting a log-linear model with the interaction term HP, model (HP, S), to the total data, and obtained an estimate of 787. In Table 6, both stratified and un-stratified estimates for model (HP, S) are 778. Note that we have excluded seven patients whose diagnostic groups are unknown. This explains the slight difference between our result and that in Bobo *et al.*[4] Adding up these seven to our estimate, we then obtain a very close result.

We also list the results for model (HP, HS), quasi-symmetric and a partial-quasi-symmetric model in Table 6. For these three log-linear models, there are substantial differences between the stratified and un-stratified

results. In contrast, such differences for the three sample coverage estimators are limited. Recall that the purpose of stratification is mainly to reduce the dependence due to heterogeneity. The overall dependencies are considered and adjusted in the sample coverage estimators $\hat{N}$ and $\hat{N}_1$. Therefore, the closeness of the stratified and un-stratified results is expected. (As will be seen in Sec. 5.3, post-stratification is not warranted because insufficient overlap may arise in some strata, leading to an unstable stratified result.) For this data set, the estimator $\hat{N}$ that can take account of two types of dependencies in each stratum has acceptable precision. Both stratified and un-stratified estimates can be recommended since they result in similar variation using 1000 bootstrap replications. The latter estimate is 812 with an estimated bootstrap s.e. of 65, which yields a 95% confidence interval of (720, 988). The former yields a slightly lower estimate of 808 with a higher s.e. of 74, which implies a 95% confidence interval of (709, 1015).

5.3. *Stratified drug data (Four lists)*

Table 7 shows the analysis results for the drug data. Estimates based on various models are presented for each stratum and for the collapsed data. Except for the stratum of 4–6 days, the Petersen estimate based on the lists 1 and 2 are lower than the others and thus positive dependence is expected. The CCV estimate (unreported) for the total data reveals that positive dependence also exists between the lists 1 and 3. Wittes[50] suspected that positive dependence may arise between list-3 and list-4, but the CCV estimate only shows very weak dependence.

Wittes[50] fitted an independent model to the data in each stratum and obtained a pooled estimate of 544 (s.e. 22.4), which is slightly different from our result under independence in Table 7 probably due to numerical rounding errors. The un-stratified estimate under an independent model is 524 (s.e. 18). Wittes[50] thus concluded that failure to account for stratification in the analysis would have afforded the investigators a false sense of precision.

Table 7 also presents the results for some other selected log-linear models that fit well, i.e. models (1, 2, 34), (12, 13, 4), (H1, 12, 13, 4), (12, 13, 14), and the quasi-symmetric model. For the quasi-symmetric model, the iterations failed to converge for two strata. Therefore, the stratified result for the quasi-symmetric model is not obtainable. The iteration steps for the un-stratified estimate did converge, but the s.e. is extremely large. It implies

Table 7. Various estimates of the total number of patients for drug data (Standard error is in the parenthesis).

	Usage on Drugs (Stratum)				
Model/Method	1–3 days	4–6 days	7+ days	Stratified	Un-stratified
Petersens estimate:					
(1, 2)	50 (14)	278[#] (183)	80 (16)	408 (184)	300 (71)
(1, 3)	350 (112)	214 (49)	133 (15)	697 (123)	459 (54)
(1, 4)	135 (22)	194 (18)	178 (12)	507 (31)	497 (28)
(2, 3)	175 (49)	152 (42)	171 (47)	498 (80)	468 (112)
(2, 4)	123 (28)	284 (69)	211 (43)	618 (86)	609 (99)
(3, 4)	173 (27)	193 (12)	184 (11)	550 (32)	527 (25)
Log-linear models:					
(1, 2, 3, 4) Independent	160 (19)	203 (11)	179 (7)	542 (23)	524 (18)
(1, 2, 34)	149 (21)	215 (20)	175 (8)	539 (30)	521 (24)
(12, 13, 4)	156 (18)	201 (11)	184 (9)	541 (23)	531 (20)
(H1, 12, 13, 4)	181 (73)	242 (58)	185 (17)	608 (95)	586 (62)
(12, 13, 14)	165 (25)	204 (13)	187 (12)	556 (31)	547 (26)
Quasi-symmetric	(diverge)	(diverge)	353 (322)	(diverge)	1027 (815)
Sample coverage:					
$\hat{N}_0$: Eq. (3)	151 (22)	226 (23)	178 (10)	555 (33)	541 (26)
$\hat{N}$: Eq. (4)	170 (575)	286 (63)	185 (26)	641 (579)	635 (93)
$\hat{N}_1$: Eq. (5)	157 (32)	247 (28)	182 (17)	586 (46)	579 (44)
Logistic regression model:					
Eq. (7) with covariate and list effects, Horvitz–Thompson estimate: 539 (s.e. 21)					

[#]Petersens estimate does not exist due to no overlapped cases; Chapmans estimator is calculated instead (see Seber[44]; 59–60).

that a Rasch model which can reflect heterogeneity among individual cannot be adopted.

The sample coverage estimator under independence gives an estimate of 541 (s.e. 26), which is very close to the result obtained from a logistic analysis. The logistic regression model in Eq. (7) provides a proper fit to the data because the deviance is 39.81 with 39 degrees of freedom (P-value = 0.43). In the first stratum, the relatively large bootstrap s.e. of the estimator $\hat{N}$ indicates that data information cannot provide a reliable estimate and thus only a reasonable lower bound can be obtained. Consequently, the stratified estimate based on $\hat{N}$ is not recommended for use. For the collapsed data, the precision is acceptable, showing the pooled data are sufficient

to incorporate both types of dependencies. We obtain an estimate of 635 with an estimated s.e. of 93 based on 1000 bootstrap replications. A 95% confidence interval of the size can be constructed as (520, 895).

6. Remarks and Discussion

Capture-recapture models provide a potentially useful method for assessing the extent of incomplete ascertainment in epidemiological studies but there are assumptions and limitations to this approach. We have reviewed three methods (log-linear models, sample coverage approach and logistic regression analysis) and applied them to three data sets with/without covariates. The three data analyses have demonstrated the usefulness of the capture-recapture analysis.

Basic assumptions must be checked to validate the implementation of the capture-recapture method. Hook and Regal[26,28] presented 17 recommendations for the use of the capture-recapture method in epidemiology. We also urge the readers to check the assumptions listed in Sec. 2 before capture-recapture analysis.

We have shown that for some data sets (e.g. the HAV data and the first stratum of the drug data), insufficient overlap information usually results in an imprecise estimate. This implies that a serious limitation of the capture-recapture methods is that sufficiently high overlapping information is required to produce reliable population size estimates and to model dependence among samples. Coull and Agresti[16] also indicated that the likelihood functions under some models for sparse information might become flat and the resulting estimates are likely to become unstable. In such cases, we feel that a precise lower bound is of more practical use than an imprecise point estimate.

Almost all methods discussed in this chapter require extensive numerical iterations or calculations to obtain estimators and standard errors. Therefore, user-friendly software is essential for applications. We have developed a program CARE (for capture-recapture) containing two parts: CARE-1 and CARE-2. CARE-1 is an S-PLUS program for analyzing epidemiological data; CARE-2, written in C language, calculates various estimates for ecological models. All estimates and standard errors given in Tables 5–7 were obtained by using CARE-1. A tutorial article[11] demonstrated the use of CARE-1. The reader is referred to Chao and Huggins[9] for the use of CARE-2 if ecological models are needed. The program CARE is available and can be downloaded from the first authors website at http://chao.stat.nthu.edu.tw/.

Acknowledgments

Research was supported (for Chao) by the National Science Council of Taiwan under Contract/grant number NSC89-2118-M007-006 and NSC-90-2119-M-007-003.

References

1. Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50**: 494–500.
2. Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**: 623–635.
3. Alho, J. M., Murly, M. H., Wurdeman, K. and Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system. *Journal of the American Statistical Association* **88**: 1130–1136.
4. Bobo, J. K., Thapa, P. B., Anderson, J. R. and Gale, J. L. (1994). Acute encephalopathy and seizure rates in children under age two years in Oregon and Washington State. *American Journal of Epidemiology* **140**: 27–38.
5. Briand, L. C., El Eman, K., Freimut, B. G. and Laitenberger, O. (2000). A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Transactions on Software Engineering* **26**: 518–538.
6. Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**: 625–633.)(
7. Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**: 783–791.
8. Chao, A. (1998). Capture-recapture. In *Encyclopedia of Biostatistics*, eds. P. Armitage and T. Colton, Wiley, New York, 482–486.
9. Chao, A. and Huggins, R. M. (2003). Closed capture-recapture models. In *The Handbook of Capture-Recapture Methods*, eds. B. Manly, T. L. McDonald and S. C. Amstrup, Princeton University Press.
10. Chao, A. and Tsay, P. K. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association* **93**: 283–293.
11. Chao, A., Tsay, P. K., Lin, S. H., Shau, W. Y. and Chao, D. Y. (2001). Tutorial in Biostatistics: The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20**: 3123–3157.
12. Chao, D. Y., Shau, W. Y., Lu, C. W. K., Chen, K. T., Chu, C. L., Shu, H. M. and Horng, C. B. (1997). A large outbreak of hepatitis A in a college school in Taiwan: Associated with contaminated food and water dissemination. *Epidemiology Bulletin*, Department of Health, Executive Yuan, Taiwan Government.
13. Cochran, W. G. (1978). Laplaces ratio estimators. In *Contributions to Survey Sampling and Applied Statistics*, ed. A. David, Academic Press, New York, 3–10.

14. Cormack, R. M. (1989). Loglinear models for capture-recapture. *Biometrics* **45**: 395–413.
15. Cormack, R. M. (1999). Problems with using capture-recapture in epidemiology: An example of a measles epidemic. *Journal of Clinical Epidemiology* **52**: 909–914.
16. Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* **55**: 294–301.
17. Darroch, J. N. (1958). The multiple-recapture census I. Estimation of a closed population. *Biometrika* **45**: 343–359.
18. Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88**: 1137–1148.
19. Desenclos, J. C. and Hubert, B. (1994). Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* **23**: 1322–1323.
20. Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**: 591–603.
21. Fienberg, S. E., Johnson, M. S. and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of Royal Statistical Society* **A162**: 383–405.
22. Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of Royal Statistical Society* **B572**: 533–546.
23. Hook, E. B. and Regal, R. R. (1992). The value of capture-recapture methods even for apparently exhaustive surveys: The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *American Journal of Epidemiology* **135**: 1060–1067.
24. Hook, E. B. and Regal, R. R. (1993). Effects of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. *American Journal of Epidemiology* **137**: 1148–1166.
25. Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiological Reviews* **17**: 243–264.
26. Hook, E. B. and Regal, R. R. (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology* **52**: 917–926.
27. Hook, E. B. and Regal, R. R. (2000). Accuracy of alternative approaches to capture-recapture estimates of disease frequency: Internal validity analysis of data from five sources. *American Journal of Epidemiology* **152**: 771–779.
28. Hook, E. B. and Regal, R. R. (2000). On the need for a 16th and 17th recommendation for capture-recapture analysis. *Journal of Clinical Epidemiology* **53**: 1275–1277.
29. Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**: 663–685.

30. Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**: 133–140.
31. Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47**: 725–732.
32. International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995a). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology* **142**: 1047–1058.
33. International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995b). Capture-recapture and multiple-record systems estimation II: Application in human diseases. *American Journal of Epidemiology* **142**: 1059–1068.
34. Kiemeney, L. A. L. M., Schouten, L. J. and Straatman, H. (1994). Ascertainment corrected rates (Letter to Editor). *International Journal of Epidemiology* **23**: 203–204.
35. LaPorte, R. E., McCarty, D. J., Tull, E. S. and Tajima, N. (1992). Counting birds, bees and NCDs. *Lancet* **339**: 494.
36. Lloyd, C. J. (1999). *Statistical Analysis of Categorical Data*. Wiley, New York.
37. Otis, D. L., Burnham, K. P., White, G. C. and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62**: 1–135.
38. Papoz, L., Balkau, B. and Lellouch, J. (1996). Case counting in epidemiology: Limitation of methods based on multiple data sources. *International Journal of Epidemiology* **25**: 474–477.
39. Pollock, K. H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: Past, present, and future. *Journal of the American Statistical Association* **86**: 225–238.
40. Pollock, K. H., Hines, J. E. and Nichols, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics* **40**: 329–340.
41. Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, University of California Press, 321–333.
42. Schouten, L. J., Straatman, H., Kiemeney, L. A. L. M., Gimbrere, C. H. F. and Verbeek, A. L. M. (1994). The capture-recapture method for estimation of cancer registry completeness: A useful tool? *International Journal of Epidemiology* **23**: 1111–1116.
43. Schwarz, C. J. and Seber, G. A. F. (1999). Estimating animal abundance: Review III. *Statistical Science* **14**: 427–456.
44. Seber, G. A. F. (1982). *The Estimation of Animal Abundance*, 2nd edn., Griffin, London.
45. Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* **42**: 267–292.

46. Seber, G. A. F. (1992). A review of estimating animal abundance II. *International Statistical Review* **60**: 129–166.
47. Sekar, C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44**: 101–115.
48. Tsay, P. K. and Chao, A. (2001). Population size estimation for capture-recapture models with applications to epidemiological data. *Journal of Applied Statistics* **28**: 25–36.
49. White, G. C., Anderson, D. R., Burnham, K. P. and Otis, D. L. (1982). *Capture-Recapture and Removal Methods for Sampling Closed Populations*, Los Alamos National Lab, LA-8787-NERP, Los Alamos, New Mexico, USA.
50. Wittes, J. T. (1974). Applications of a multinomial capture-recapture method to epidemiological data. *Journal of the American Statistical Association* **69**: 93–97.
51. Wittes, J. T., Colton, T. and Sidel, V. W. (1974). Capture-recapture methods for assessing the completeness of cases ascertainment when using multiple information sources. *Journal of Chronic Diseases* **27**: 25–36.
52. Wittes, J. T. and Sidel, V. W. (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases* **21**: 287–301.
53. Yip, P. S. F., Wan, E. C. and Chan, K. S. (2001). A unified approach for estimating population size in capture-recapture studies with arbitrary removals. *Journal of Agricultural Biological and Environmental Statistics* **6**: 183–194.

About the Authors

Anne Chao received her PhD in Statistics (1977) from the University of Wisconsin. Since 1978, Professor Chao has been teaching at the National Tsing Hua University, Taiwan, and is a professor of the Institute of Statistics. Her main research areas are biostatistics, including estimation of class numbers, estimation of population size, sampling methods and multiple system estimation.

Hsin-Chou Yang received his PhD in Statistics (2002) from National Tsing Hua University under the supervision of Anne Chao. He is now working as a post-doctor at the Institute of Biomedical Science, Academia Sinica, Taiwan, in fulfillment of his compulsory military service.

Paul Yip is a Senior Lecturer in the Department of Statistics and Actuarial Science and Director of the Hong Kong Jockey Club Centre for Suicide

Research and Prevention, the University of Hong Kong. He received his PhD in Statistics from La Trobe University, Australia. His research areas cover a wide spectrum of methodological and applied topics, with a major interest in developing and applying statistical techniques to medical, social and biological sciences.

This page intentionally left blank

CHAPTER 20

STATISTICAL METHODS IN THE EFFECT EVALUATION OF MASS SCREENING FOR DISEASES

QING LIU

Department of Medical Statistics and Epidemiology,
School of Public Health, Sun Yat-Sen University,
Guangzhou, Guangdong, 510080, PR China
Tel: 86-20-87330664; qliu@gzsums.edu.cn

It is an important way in chronic disease control to detect disease earlier by mass screening. The purpose of screening is to detect disease in an early stage, in an expectation of the better treatment effect, the improvement of patients prognosis and the reduction of disability or death. Mass screening may achieve its objective of early detection in two approaches: one is to encourage patient to visit doctor when the early sign and symptom of disease appears; second is to supply a regular physical test and to detect disease in an asymptomatic stage.

1. Basic Concept of Mass Screening for Disease

American Commission on Chronic Illness gave a definition of screening in 1957[1]: "The presumptive identification of unrecognized disease or defect by the application of test, examine, or other procedures which can be applied rapidly to sort out apparently well persons who probably have a disease from those who probably do not. A screening test is not intention to be diagnostic. Persons with positive or suspicious finding must be referred to their physicians for diagnosis and necessary treatment." This definition emphasized on two points. Firstly, the potential disease states identified by mass screening include two subgroups: one is the state in the high risk of disease and another is the state of disease unrecognized by patients himself. Person in first group may not be ill but he may have disease in a high probability, such as the people with multiple intestinal polyps, oesophageal epithelial dysplasia and mataplasia. People in second group have suffered from the disease but have not recognized it, such as patients with small liver cancer found by α-fetoprotein (AFP). Secondly, screening procedure

itself does not diagnose illness. Those who test positive are sent on for further evaluation by a subsequent diagnostic test or procedure to determine whether they do have the disease.

In 1968, World Health Organization (WHO)[2] suggest some governing principles and pre-requisites of mass screening:

(1) Disease is a serous health problem. It means a high morbidity, mortality and social burden.
(2) Early detection of disease may improve the prognosis, reduce proportion of disability and death of patients. It means that there are the effective treatments of early stage disease.
(3) The natural history of disease is well known and the screening test is able to detect disease in preclinical phase.
(4) There is an effective screening test. It means that the test is sensitive, specific, high predictive value and safe.
(5) The screening program is acceptable to population, simple and inexpensive, high compliance rate and low complication and pain.
(6) There is proper procedure of further diagnosis and follow-up.

Simply, the success of a screening program depends on how many deaths saved. By the exact statistical statement, screening is effective if the mortality rate of the disease in screened population is lower than that of in unscreened population. It may exist by the difference of the mortality rates or by the relative rates. The methods of statistical analysis for screening data is similar to that for general epidemiological study. The life-year savings of a disease is also a common index for screening assessment and especially useful in the cost-effectiveness evaluation. Mass screening does not only reduce the mortality rate of disease but sometimes also reduces the incidence rate and the medication cost of disease. However, the quantitative assessment on this aspect of screening is much more difficult than death reduction.

The cost must be considered in the assessment of screening. The direct cost includes charges of screening test, further diagnostic procedure and follow-up for positive result. The indirect cost includes expense of time and work, management and organization of program, etc. The evaluation of the cost on the psychological and biological impact of screening, such as the anxiety for positive results, risk of complication, harm and pain brought by screening test is much more complicate but must be taken into consideration.

The assessment of screening is very complicate and difficult. The current method of assessment for screening still needs to be improved. In the

beginning, ones hope to compare the survival time of the cases detected by screening and the cases diagnosed in clinic. If the survival time of the cases detected by screening is longer than that of the cases diagnosed in the clinical, the screening is said to be effective. However, this comparison suffers may have many biases. Firstly, screening population is not a random sample of general population. They may be different in some important demographic characteristics from general population where the cases come, such as occupation, life status and education, etc. Secondly, longer survival time in the cases detected by screening may be caused by earlier diagnostic time rather than the prolonged life. That is called as lead time bias.[3] As shown in Fig. 1, we suppose the survival time of patients are not changed no matter it is detected by screening or diagnosed in clinic. The average longer survival time may be due to the screening advances the diagnostic time. The lead time in an uncontrolled clinical trial appears to increase survival time although the natural history of the disease and the time of death are unchanged, whereas, patients stay longer in disease phase and suffer more from pain and anxiety. Finally, the probability that a disease will be detected by screening is directly proportional to the length of its preclinical detectable phase, which is inversely related to its rate of disease progression. Individuals with rapidly progressive disease — those with short preclinical phases — are more likely to die than the average longer and are less likely to be identified by screening. Therefore, long survival time may not be the effect of screening nor the selective effects of screening procedure on cases. That is called as length bias.[4] The screening tends to detect disease subsets with long preclinical phase, less aggressive progression and perhaps better inherent prognosis.

As shown in Fig. 2, for Cases 1 and 2, the disease is less aggressively progressive, detected by screening and predictable a better prognosis. For Cases 3 and 4, disease progresses rapidly, missing the opportunity of screening detection and a poor prognosis. In the comparison of survival time of patients detected by screening and detected clinically, these biases

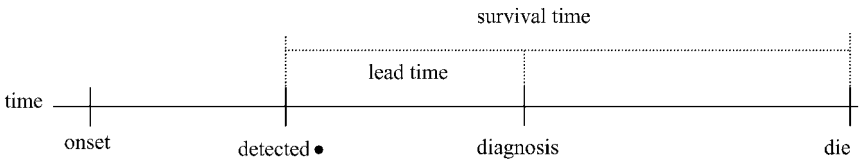


Fig. 1. Illustration of lead time bias.

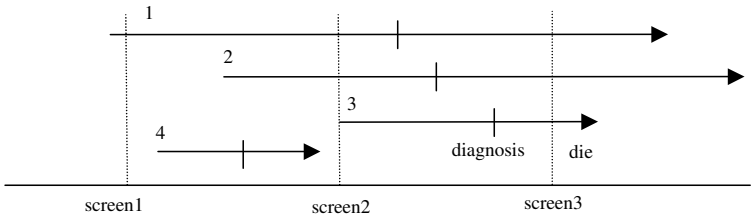


Fig. 2. Illustration of length bias.

must be adjusted. This adjustment is very difficult, depends on the fully perception of natural history of the disease and needs a very complex counting process.

Early diagnostic rate is one of indices for assessment in the pilot period of a screening program. However, this index only suggests that the screening may be effective but has not been proven. If there is no effective treatment for detected patients, early detection of disease will not improve the prognosis of patients and reduce the death or disability caused by disease. Only when is there the effective treatment for disease in early stage and not for later stage, the early detection of disease is of special meaning.

Obviously, the comparison of survival rates of patients detected by screening and detected in clinic is not an ideal method of screening assessment. Early diagnostic rate of disease is also not a good index and only an indirect index. An ideal assessment of screening is to compare the mortality rates of the disease and death rates in screening population and control population in a randomized controlled population-based design. The mortality rates of the disease and death rates are crucial in the assessment of screening effects. The comparison of mortality rates is not interfered by lead time bias and length bias. The results of the comparison reflect the true effectiveness of screening. For example, in the Health Insurance Plan (HIP)[5] in New York, the women aged 40 to 64 who participated of the plan are randomly separated into two groups, one group accept yearly physical check and mammography, another group receive a routine medical care service. Four repeat screening tests were given to first group in total. In the first 5 years, the mortality rates of breast cancer in screening group reduced about 40%. After 14 years follow-up, there is still a 20% of mortality reduction of breast cancer (Table 1). The results from a randomized control clinical trial of breast cancer screening in Sweden,[6] also proved that the screening of breast cancer may effectively reduces the deaths caused by the disease (Table 2).

Table 1. Cumulative deaths of breast cancer in screening population and control in HIP.

Years from first screens to diagnosis	Cases of breast cancer	Deaths from breast cancer according to follow-up years			
		5 years	7 years	10 years	14 years
5 years					
Screening group	306	39	71	95	118
Control group	300	63	106	133	153
Difference of rates (%)		38.1	33.0	28.6	22.9
7 years					
Screening group	425	39	81	123	165
Control group	443	63	124	174	212
Difference of rates (%)		38.1	34.7	29.3	22.2
10 years					
Screening group	600	39	81	146	218
Control group	604	63	124	192	262
Difference of rates (%)		38.1	34.7	24.0	16.8

Table 2. The mortality analysis of breast cancer between screening population and control.

Age	Groups	Deaths	Screening population	*RR* (95%CI)
40–49	Screen	28	19844	0.92(0.52–1.60)
	Control	24	15604	
50–59	Screen	45	23485	0.60(0.40–0.90)
	Control	54	16805	
60–69	Screen	52	23412	0.65(0.44–0.95)
	Control	58	16269	
70–74	Screen	35	10339	0.77(0.47–1.27)
	Control	31	7307	
Total	Screen	160	77080	0.69(0.55–0.88)*
	Control	171	55985	

*Adjusting for age.

It is shown in Tables 1 and 2 that the methods of statistical analysis are similar to those in the treatment of traditional epidemiological data. Mantel-Haenszel stratified analysis was used to estimate the relative risks and confidence interval of disease mortality. The methods of hypothesis testing are also same.

If a randomized control clinical trial is not feasible, a population-based cohort study is the next choice. For example, a study in England is to

compare the mortality rates of the disease between the screening area and the non-screening area. This kind of design requires a relative high participation rate in screening population. It means a good compliance. For comparison, the demographic characteristics of populations in different areas need to be adjusted.

The randomized control clinical trial or other observational design only evaluates a single screening scheme. The usage of them is limited because it cannot estimate the extra effects of a screening scheme applied in different population with different age distribution and different prevalence rates of disease, or the extra effects of different screening schemes, such as different frequency, different test. In practice, people cannot carry out a randomized control clinical trial for every screening scheme to evaluate its effects. In this situation, the mathematical model of natural history of disease based on the current data from a RCT or other observational studies may complement this limitation. The prevalence rate of disease on every screening and the incidence rates in the screening interval are estimated based on a stochastic model and the effects of different screening policies are evaluated.

The screening programs for a variety of diseases have been implemented in many countries in the world. The practice proves that the mortality rates of some diseases, such as breast cancer, cervical cancer, hypertension and diabetes may be reduced by screening. The question is which one in different screening test and different schemes detect disease earlier, with higher efficacy and higher cost-effectiveness. Two important parameters decide the effect of screening. One is the sensitivity of screening test. Another is the distribution of sojourn time in preclinical detectable phase (PCDP). A high sensitivity, or low false negative rate of screening test means a strong power to detect disease. A long sojourn time of preclinical detectable phase means more chance to be detected by screening and in an early stage of disease. A short sojourn time of preclincial detectable phase gives little chance to be detected by screening. That means that the proportion of the cases detected by screening is low, the effect of screening is poor and the screening may be not feasible. If the sojourn time of preclinical detectable phase is long, the interval between screens may be designed longer. When the distribution of preclinical detectable phase is known, the lead time bias may be estimated for the assessment of screening. Therefore, the core of analysis of screening data and optimization of screening schemes is to estimate these two parameters.

We assume that the disease progresses in the manner shown in Fig. 3. An individual enters the preclinical detectable phase of the disease, detectable

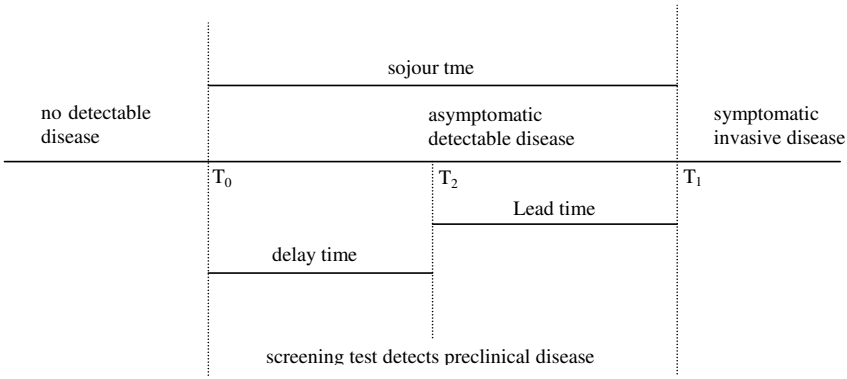


Fig. 3. Schema for the progression of a disease with the intervention of an early detection.

by the screening modality in question, at time T_0, and would begin to manifest symptoms, i.e. the disease would become clinically apparent, at time T_1, if no intervention were to take place. For this individual, the "sojourn time" is defined as $T_1 - T_0$. Suppose now that the individual is screened at time $T_2(T_0 < T_2 < T_1)$ and is diagnosed in the preclinical state. For this individual, the "lead time", the interval by which diagnosis is brought forward, is defined as $T_1 - T_2$. The probability that the screening test correctly identifies an individual as being in the preclinical detectable phase is termed the "sensitivity" of the test; the "false-negative rate" is one minus the sensitivity.

2. One-stage Models of the Natural History of Disease for Screening

The data from screening process consist of: (1) prevalent cases diagnosed in preclinical detectable state during the screening; (2) incidence cases diagnosed clinically in the interval of two screenings or both before and after screening. The purpose of natural history model of screening for a disease is to express these prevalence and incidence rates in terms of the false-negative rate and of the sojourn-time distribution. General model have been developed[7,8] to describe the effect of screening on the disease process in order to identify those parameters which determine the expected benefit. Estimation of parameters of interest is difficult with these general models since the number of unknowns is large. Here firstly the simplified model (NE Day and SD Walter)[9] is introduced.

In the absence of screening, the incidence of clinical disease at age t will be denoted by $I(t)$. For the screening modality in question, $f(y)$ will denote the probability density function of the length of the interval y during which the disease is preclinical but detectable, i.e. the sojourn time. For simplicity, we assume $f(y)$ to be independent of t. The function $J(t)$ will denote the incidence of the preclinical state, i.e. the rate at which individuals enter it. The "false-negative rate", i.e. the probability that an individual in the preclinical state is screened negative, will be denoted by β; thus $1-\beta$ is the "sensitivity". We assume that β is independent of both the lead time and the sojourn time.

The functions $I(t)$ and $J(t)$ are related through $f(y)$ by the equation

$$I(t) = \int_0^t J(s)f(t-s)ds\,. \tag{1}$$

Suppose that the population is screened at t_1. Then for $t > t_1$ the incidence is made up of two components, individuals with short sojourn time who entered the preclinical state after t_1, and individuals with a longer sojourn time falsely screened negative at t_1. Thus, after one screen, the incidence, $I_1(t)$ is given by

$$I_1(t) = \beta \int_0^{t_1} J(s)f(t-s)ds + \int_{t_1}^t J(s)f(t-s)ds\,. \tag{2}$$

Similarly, if screens occur at times $t_1, t_2, \ldots, t_n$, then the incidence $I_n(t)$ after the nth screen is given by

$$I_n(t) = \sum_{i=0}^{n} \beta^{n-i} \int_{t_1}^{t_{1+1}} J(s)f(t-s)ds\,, \tag{3}$$

where $t_0 = 0$ and $t_n + 1 = t$. We make an assumption that $J(t)$ is uniform for an individual over the duration of the study. For cancer, the screening interval usually is 1 or 2 years, in this interval the assumption of uniform incidence rate of preclinical state may hold approximately.

$$I_n(t) = J \sum_{i=0}^{n} \beta^{n-i} \int_{t-t_{i+1}}^{t-t_1} f(y)dy\,. \tag{4}$$

The prevalence, P_1 observed at a first screen at time t_1, is given by

$$P_1 = (1-\beta) \int_0^{t_1} J(s) \int_{t_1-s}^{\infty} f(y)dyds\,. \tag{5}$$

That is, for each time $s < t_1$, those individuals entering the preclinical state at time s will be prevalent cases at time t_1 if their sojourn time is greater

than $t_1 - s$. On reversing the order of integration and setting $J(s)$ constant, this expression becomes

$$P_1 = (1-\beta)J\int_0^{t_1} yf(y)dy + \int_{t_1}^{\infty} f(y)dy \tag{6}$$

If there are n previous screens, at times $t_1, \ldots, t_n$. Summation over the n intervals, gives the following expression for the total prevalence P_n at the nth screen at time t_n:

$$P_n = (1-\beta)J\sum_{i=1}^{n}\beta^{n-i}\int_{t_n-t_i}^{\infty}\min\{y-(t_n-t_i), t_i - t_{i-1}\}f(y)dy\,. \tag{7}$$

If the interval between screens is constant, i.e. if $t_{i+1} - t_i = \Delta$, $i = 1, 2, \ldots, n-1$. Above expressions may be further simplified. For incidence rates,

$$I_n(t) = J\sum_{i=0}^{n-1}\beta^{n-i}\int_{t-t_n+(n-i-1)\Delta}^{t-t_n+(n-i)\Delta} f(y)dy + J\beta^n\int_0^{t-t_n} f(y)dy\,. \tag{8}$$

For the prevalence rates,

$$P_n = (1-\beta)J\sum_{i=1}^{n}\beta^{n-i}\int_{(n-i)\Delta}^{\infty}\min\{y-(n-i)\Delta, \Delta\}f(y)dy\,. \tag{9}$$

We consider first the idealized situation where a total population is screened at regular intervals, each individual being screened with the same inter-screening interval Δ. The constant incidence rates are assumed known from a pre-existing disease registry. At the ith screen ($i = 1, \ldots, n$) one knows r_i, the number of cases of preclinical disease found, and n_i, the number screened; and between screen i and screen $i+1$ one knows the total c_i of cases diagnosed outside screening from a total of y_i person-years at risk. The probability q_i of a case developing between screen i and screen $i+1$ outside screening is given by

$$q_i = 1 - \exp\left\{-\int_{t_i}^{t_{i+1}} I_i(y)dy\right\}, \tag{10}$$

Which can be well approximated by

$$q_i = \int_{t_i}^{t_{i+1}} I_i(t)dt\,.$$

The cases at screen i can be taken to have a Poisson distribution with parameter n_iP_i and the cases emerging outside screening between screens i

and $i+1$ can be taken to have a Poisson distribution with parameter $y_i q_i$. Then the likelihood function is

$$L = \prod_{i=0}^{n-1} \frac{n_i P_i}{r_i!} e^{-n_i P_i} \prod_{i=1}^{n} \frac{y_i q_i}{c_i!} e^{-y_i q_i} \,. \tag{11}$$

Based on likelihood function, the maximum likelihood values of parameters may be estimated.

Three different forms for $f(y)$ may be considered: a step function with arbitrary probabilities defined over short time intervals, a lognormal and an exponential. Here, we only discuss the exponential distribution as an example for the application since it not only gives simpler expressions for the quantities of interest but also fits the data better.

2.1. *Example*

Breast cancer screening by the Health Insurance Plan of Greater New York (the HIP study).[5]

The data we have used are summarized in Tables 3 and 4. There were 4 screens at yearly interval and the cases arising between screens were identified. We use the data from the first 5 years of follow-up after the start of screening.

We assume that the distribution of sojourn time of preclinical phase is an exponential.

$$f(y) = \lambda \exp(-\lambda y) \,, \quad y \geq 0 \,.$$

With this assumption, for $t > t_n$, the incidence rate of screening interval is

$$I_n(t) = J - J \exp(-\lambda t) \{ \exp(\lambda t_n) - \sum_{i=1}^{n-1} \beta^{n-i} [\exp(\lambda t_{i+1}) - \exp(\lambda t_i)] - \beta^n \exp(\lambda t_i) \} \,. \tag{12}$$

Table 3. Prevalence rates of breast cancer in the first 5 years of the HIP study.

Years since start of study	No. of women screened	No of prevalent breast cancer cases		Prevalence (‰)
		Observed	Expected	
0	20166	55	59.5	2.73
1	15936	32	25.8	2.01
2	13679	17	20.4	1.24
3	11971	23	17.7	1.92

Table 4. Incidence rates of breast cancer in the first 5 years of the HIP study.

Years since start of study	No of previous negative screens	Women-months of follow-up	No of prevalent breast cancer cases		Annual incidence (‰)
			Observed	Expected	
0–1	1	240277	13	14.9	0.65
1–2	1	81337	7	9.0	1.03
2–3	1	38370	1	5.3	0.31
3–4	1	30942	3	4.7	1.16
4–5	1	26701	5	4.3	2.25
1–2	2	190474	8	9.6	0.50
2–3	2	52934	5	5.5	1.13
3–4	2	19036	2	2.6	1.26
4–5	2	12626	4	1.9	3.80
2–3	3	163642	10	8.1	0.73
3–4	3	45964	5	4.7	1.31
4–5	3	13151	2	1.8	1.82
3–4	4	145118	10	7.0	0.84
4–5	4	89371	10	9.2	1.34

The q_j are then given, for $j = 1, \ldots, n$ by the integrals of (12) from t_j to $t_j + 1$, so

$$q_j = J(t_{j+1} - t_j) - J\lambda^{-1}[\exp(-\lambda t_j) - \exp(\lambda t_{j+1})]$$
$$\times \left\{\exp(\lambda t_j) - \sum_{i=1}^{j-1} \beta^{j-i}[\exp(\lambda t_{i+1}) - \exp(\lambda t_i)] - \beta^j \exp(\lambda t_1)\right\}. \quad (13)$$

If the screens are equally spaced, q_j reduces to

$$q_j = J\Delta - J\lambda^{-1}[1 - \exp(-\lambda\Delta)]$$
$$\times \left\{1 - [\exp(\lambda\Delta) - 1]\sum_{i=1}^{j-1} \beta^i \exp(-i\lambda\Delta) + \beta^j \exp[-(j-1)\lambda\Delta]\right\}. \quad (14)$$

The expression for P_j, from (9) reduces to

$$P_j = (1-\beta)J\lambda^{-1}\sum_{i=1}^{j} \beta^{j-1}\{\exp[-\lambda(t_j - t_i)] - \exp[-\lambda(t_j - t_{i-1})]\}, \quad (15)$$

which for equal spaced intervals becomes

$$P_j = (1-\beta)J\lambda^{-1}[1 - \exp(-\lambda\Delta)]\sum_{i=0}^{j-1} \beta^i \exp(-i\lambda\Delta). \quad (16)$$

Since the example is a screening at an equal spaced interval, the expression (14) and (16) are used to develop the maximum likelihood function. Then it is relatively straightforward to compute the log likelihood as a function of λ and β. The results are: $\beta = 0.18$, $\lambda = 0.585$. According to the exponential distribution, the average sojourn time equals 1.71. Also shown in Table 4 are the expected numbers of cases based on the best-fitting exponential distribution, with an overall χ^2 goodness-of-fit test. The fit is clearly good.

3. Two-Stage Models of the Natural History of Disease for Screening

The models of natural history of the disease introduced before are based on a progressive disease model. The progressive disease model assumes individuals are in a healthy state until they enter the preclinical disease state and all individuals in this state eventually emerge with clinical symptoms if untreated. The key assumption of this model is that preclinical disease, if left untreated, would ultimately surface clinically. This assumption is true for a part of invasive diseases, such as breast cancer. When the mammography may detect the malignant tumor in breast, the tumor must progress until patient feels the symptom or sign and goes to visit physician. Similar cases are the chest radiography for lung cancer and α-fetoprotein (AFP) test for liver cancer. However, it is not always true for some of other cancers or diseases. For example, the Pap smear for screening of cervical cancer may detect the heavy epithelial dysplasia and mataplasia, and carcinoma *in situ*; the gastroscopy for screening of stomach cancer may detect the gastric mucous dysplasia and mataplasia; the enteroscopy for screening of colon cancer may detect multiple intestinal polyps; etc. These non-invasive lesions may progress to invasive disease but also may persist or revert to normal automatically. However, once a lesion becomes invasive, it almost never regresses without treatment and it is assumed all invasive lesions arise from a preinvasive lesion. According to this situation, R. Brookmeyer and NE Day[10] suggested a two-stage model for the analysis of cancer screening data. The two-stage model is illustrated schematically in Fig. 4.

We define the random variable X to be the duration of time a progressive lesion spends in the preclinical stage 1 and Y the duration of time a progressive lesion spends in the preclinical stage 2. The cumulative distribution function of the joint sojourn times (X, Y) is called $F(x, y)$ with density

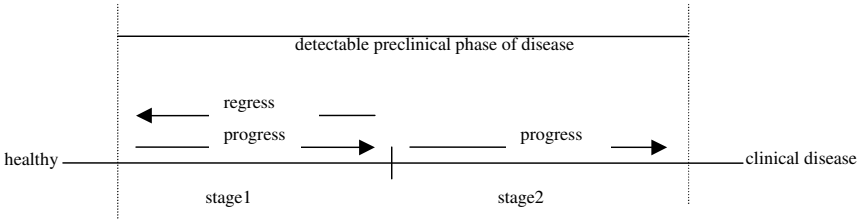


Fig. 4. Schematic illustration of two stage model for preclinical period of disease.

$f(x, y)$. Then the cumulative distribution function of the total preclinical duration $(X + Y)$ of progressive lesions is given by

$$F_T(t) = \int_0^t \int_0^s f(s - y, y) dy ds \,. \tag{17}$$

F_T is the distribution function of the total preclinical duration for progressive lesions only.

3.1. *The likelihood for the interval (clinical incident) cases*

Suppose the hazard function of clinical disease in the absence of screening is given by i and a steady state is assumed, before time t, the screening history of the jth individual in this risk set had a history of n_{ij} previous negative screens at time $H_{ij} = \{t_{ij1}, t_{ij2}, \ldots, t_{ijn_i}\}$. These times are given in reverse chronological order so that t_{ij1} refers to the time since the most recent screen and t_{ijn_i} refers to the time of first screens. By convention, $j = 1, \ldots, M_{1i}$ refers to the cases. Then the hazard of clinical disease at age t is given a screening history H_{ij} is approximately $I_\rho(H_{ij}; F_T, \beta)$, where $\rho(H_{ij}; F_T, \beta)$ is the probability that an individual who is destined to be clinically incident at age t in the absence of screening intervention, would have tested negative at prior times H_{ij} if the individual was in fact in the screening program. Then, the conditional likelihood of that the screening history H_{i0} corresponds to the interval case and the other screening histories among R_i screening subjects is

$$L_{1i} = \frac{\rho(H_{i0}; F_T, \beta)}{\sum_{j=1}^{R_i} \rho(H_{ij}; F_T, \beta)} \,. \tag{18}$$

It is shown that the constant I cancels out in (18). Suppose M_{1i} incident cases are clinically diagnosed at age a_i and there are an additional N_{1i} individuals at risk at a_i; that is, in addition to the cases there are N_{1i} individual in the cohort still at risk of being diagnosed with clinical disease

at age a_i. Suppose the jth individual in this risk set had a history of n_{ij} previous negative screens at time H_{ij}, where $R_{1i} = M_{1i} + N_{1i}$ is the size of the risk set. Then the partial likelihood contribution of incident cases at age a_i is

$$L_{1i} = \frac{\prod_{j=1}^{M_{1i}} \rho(H_{ij}; F_T, \beta)}{\sum_l \prod_{j \in sl} \rho(H_{ij}; F_T, \beta)}, \tag{19}$$

where S_l is the subsets consisting of M_{2i} individuals from the R_{2i} screened at age a_i,

$$l = 1, 2, \ldots, \begin{pmatrix} R_{1i} \\ M_{1i} \end{pmatrix}.$$

Suppose an interval case could have entered between the kth and $(k-1)$th most recent screen, and then the probability of falsely screened negative on all $k-1$ subsequent screens is

$$\rho(H_{ij}; F_T, \beta) = \sum_{k=1}^{n_{ij}+1} \beta^{k-1} [F_T(t_{ijk}) - F_T(t_{ijk-1})], \tag{20}$$

with the conventions $F_T(t_{ijn_{ij}+1}) = 1, \quad F_T(t_{ij0}) = 0.$

3.2. *Likelihood for screen-detected stage 2 prevalent cases*

The prevalence (probability) of stage 2 screen-detected disease at age t conditional on a screening history H_{ij} is

$$[(1-\beta) I \mu_2] \rho(H_{ij}; F_{B2}, \beta), \tag{21}$$

where $\mu_2 = \int y f(x, y) dx dy$ is the mean duration in stage 2. The first factor in brackets is the prevalence of screen-detected stage 2 disease unconditional on any screening history. The second factor $\rho(H_{ij}; F_{B2}, \beta)$ is the probability that an individual who is destined to be in stage 2 at age t in the absence of screening intervention, would have tested negative at prior times H_{ij} if the individual was in a screening program. F_{B2} is the backward recurrence distribution function. The backward recurrence time is the amount of time that a screen-detected lesion spent in the preclinical stage (stage 1 plus the time spent in stage 2 up to detection. The backward recurrence density is

$$f_{B2}(t) = \frac{\int_0^t \int_{t-x}^{\infty} f(x, y) dy dx}{\mu_2}. \tag{22}$$

This expression is derived by first noting the probability of being in stage 2 is $I_{\mu 2}$ and second, in order to have been in the preclinical phase for duration t and currently in stage 2 one must be in stage 1 for duration x and stage 2 for at least $t - x$, $0 < x < t$. It is defined $F_{B2}(t_{ijn_{ij}+1}) = 1$ and

$$F_{B2}(t_{ijo}) = 0\,.$$

$$L_{2i} = \frac{\rho(H_{i0}; F_{B2}, \beta)}{\sum_{j=1}^{R_i} \rho(H_{ij}; F_{B2}, \beta)}\,. \tag{23}$$

Similarly, suppose M_{2i} screen-detected stage 2 cases are detected at age a_i and an additional N_{2i} individuals also are screened at age a_i and are negative. Then the partial likelihood contribution of the screen-detected cases at a_i is

$$L_{1i} = \frac{\prod_{j=1}^{M_{1i}} \rho(H_{ij}; F_{B2}, \beta)}{\sum_l \prod_{j \in sl} \rho(H_{ij}; F_{B2}, \beta)}\,, \tag{24}$$

where meaning of s_l and l as same as before. If the screening times are randomly assigned, that is, the R_{2i} individuals who are screened at age a_i are a random sample of all individuals R_{1i} at risk at age a_i. Suppose the incident cases have C_l strata and prevalent cases screen-detected have C_2 strata, the partial likelihood is then the product of factors for the contributions from incident cases and contributions from screen-detected cases.

$$L = \prod_{i=1}^{c_1} L_{1i} \prod_{i=1}^{c_2} L_{2i}\,. \tag{25}$$

3.3. *The joint sojourn distribution of two stage model*

3.3.1. *The independent model*

The simplest model assumes that the sojourn times for the two stages X and Y are independent with cumulative distribution functions $F_1(x)$ and $F_2(y)$ and densities $f_1(x)$ and $f_2(y)$, respectively. The distribution function for the total sojourn time is

$$F_T(t) = \int_0^t f_1(x) F_2(1 - x) dx\,. \tag{26}$$

The backward recurrence cumulative distribution function is

$$F_{B2}(t) = \frac{1}{\mu_2} \int_0^t [1 - F_2(y)] F_1(t - y) dy\,. \tag{27}$$

For the two-stage independent model, with exponential sojourn distributions $[F_1(t) = 1 - e^{-\lambda_1 t}$, and $F_2(t) = 1 - e^{-\lambda_2 t}]$, both the total sojourn (F_T) and backward recurrence (F_{B2}) cumulative distribution functions are identical and given by

$$F_T(t) = F_{B2}(t) = \begin{cases} 1 + \dfrac{\lambda_2 e^{-\lambda_1 t} - \lambda_1 e^{-\lambda_2 t}}{\lambda_1 - \lambda_2}\,, & \lambda_1 \neq \lambda_2 \\ 1 - e^{-\lambda t}(1 + \lambda t)\,, & \lambda_1 = \lambda_2 = \lambda\,. \end{cases} \tag{28}$$

3.3.2. *Limiting dependent models*

For many diseases, the second stage (the preclinical invasive stage) is short relative to the first (the noninvasive stage). It is useful to consider the limiting behavior of $F_T(t)$ and $F_{B2}(t)$ as $\mu_2 \to 0$ with F_1 fixed. These limiting expressions could then be substituted into expression of likelihood. For example, consider the complete positive dependent exponential model, the relationship of X and Y may express as

$$\begin{aligned} Y &= \lambda_1 X/\lambda_2\,, \\ F_T(t) &= 1 - e^{-ut}\,, \\ F_{B2}(t) &= 1 - \left[e^{-ut} + \frac{\lambda_2}{\lambda_1}(e^{-ut} - e^{-\lambda_1 t})\right], \end{aligned} \tag{29}$$

where $u = \lambda_1\lambda_2/(\lambda_1 + \lambda_2)$. Under this model the limiting backward recurrence distribution is

$$\lim_{\mu_2 \to 0} F_{B2}(t) = 1 - e^{-\lambda_1 t}(1 + \lambda_1 t)\,. \tag{30}$$

For this limiting situation, $F_1(t) = 1 - e^{-\lambda_1 t}$ is substituted into expression (19) for total sojourn distribution while the cumulative distribution function in expression (30) is substituted into expression (24) for the backward recurrence distribution.

Brookmeyer and Day applied the two-stage model to the analysis of data from a case-control study. Data is from the case-control study of the Northeast Scotland Cervical Cancer Screening Program.[11] The program was started in 1960 when women were asked to come for an initial Pap smear. Records on all subsequent Pap tests were kept. The definition of a positive Pap test is given in MacGregor *et al.*[11] When a woman had a positive Pap test she was biopsied and/or treated. Thus, the natural history of the disease was interrupted at the time of the first positive Pap test. A case-control study was conducted and consisted of 85 women who

Table 5. Conditional likelihood analysis for independent (model 1) and limiting dependent (model 2) two-stage exponential models.

Maximum likelihood estimates	Model 1	Model 2
β	0.025	0.001
λ	0.003, 0.247*	0.013*
Proportion with < 5 years total sojourn	0.18	0.55
Proportion with < 10 years total sojourn	0.33	0.70
Maximum log-likelihood	−120.54	−119.35

*In time units of months.

were diagnosed with invasive squanmous carcinoma of the cervix between 1968 and 1982. Of these 85 cases, 35 were clinically incident (interval cases) and 50 were screen-detected with preclinical invasive disease (stage 2). Each interval cases was matched by year of birth to five controls who were healthy at the time of the cases diagnosis. Each stage 2 screen-detected case was matched by year of birth to a control who screened negative within 6 months of the date at which the case was screen-detected. The screening histories of all cases and controls were ascertained; these histories consisted of the number and timing of previous negative screens (prior to diagnosis date of the case).

The two-stage model was fitted to the data and results showed in Table 5.

The independent model gave two estimated values of λ, one is big and one is small. The author thought the development of preclincial invasive disease is very rapid so that he chose the big one as $\lambda 2$. The maximized log-likelihood was slightly higher with the dependent model and it suggest the limiting dependent model is better. Both the independent and limiting dependent analysis suggested a small false negative rate. However, there was some discrepancy in the estimates of the sojourn distribution. As expected, the value of λ is big in the limiting dependent model suggesting the shorter sojourn duration than the independent model.

4. Multiple Stages Markov model for the Natural History of Disease Screening

The two-stage model suggested by Brookmeyer and Day presented the concept of regression of disease development in preclinical phase. It describes the disease progression better and is of an important significance in evaluation and prediction of effect of screening for the disease in the

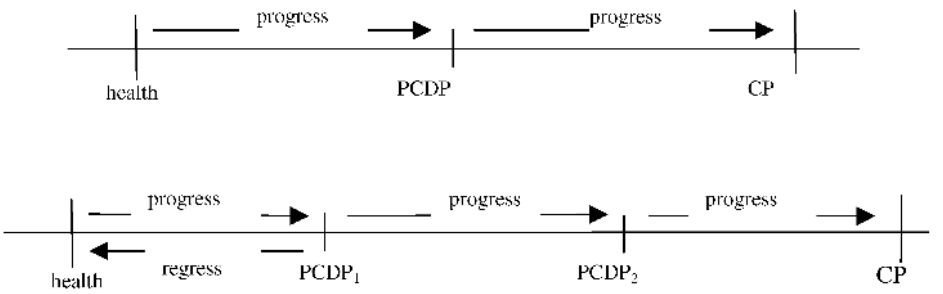


Fig. 5. Schematic illustration of state transition of disease screened.

different stages. But the two-stage model of Brookmeyer and Day does not fully describe the transition of disease in different stages. The parameter of the model may only use to estimate the total sojourn time in preclinical phase. The structure of model and the parameter estimation are relative complex. The history of a disease may look as a transition process of ones discrete healthy status. For example, in a certain period, the healthy status of an individual may transfer from healthy to potential illness, later may transfer further to clinical disease. The transition of status may be single direction or may also be double direction. Therefore, a stochastic process model is very convenient and reasonable to describe the transition of disease status and the sojourn time in each stage. If the future status only depends on the current status and is independent to all status before, that is called as the Markov property. We may use Markov process or Markov chain to describe the disease progression when it is of this property.

We assume that the single stage model and two-stage model of natural history of disease are illustrated in Fig. 5.

4.1. *Time homogeneous Markov chain model*

Duffy and Chen[12] suggested to describe the natural history of the disease for screening by the Markov process model. A Markov process with the following instantaneous transition matrix:

$$\begin{matrix} 0 \\ 1 \\ 2 \end{matrix} \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 \\ 0 & -\lambda_2 & \lambda_2 \\ 0 & 0 & 0 \end{bmatrix}.$$

Here 0 is the "no disease state", 1 is "preclinical but detectable disease" and 2 is "clinical disease". Implicit in this model is the assumption that diseases

are "born" into the preclinical state with an exponential distribution of time to birth with

$$P(\text{Time to birth} \leq t) = \int_0^t \lambda_1 e^{-\lambda_1 x} dx = 1 - e^{-\lambda_1 t}\,. \tag{31}$$

Time remaining in the preclinical phase conditional on being in the phase at time $t = 0$, is also assumed exponentially distributed with

$$\begin{aligned} P(\text{Time to transition to clinical state} \leq t) &= \int_0^t \lambda_2 e^{-\lambda_2 x} dx \\ &= 1 - e^{-\lambda_2 t}\,. \end{aligned} \tag{32}$$

Based on the solution of (dI-Q) to obtain the eigenvalues and eigenvectors, the transition probabilities for time t may be obtained:

$$P(t) = \begin{bmatrix} e^{-\lambda_1 t} & \dfrac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{(\lambda_1 - \lambda_2)} & 1 + \dfrac{\lambda_2 e^{-\lambda_1 t} - \lambda_1 e^{-\lambda_2 t}}{(\lambda_1 - \lambda_2)} \\ 0 & e^{-\lambda_2 t} & 1 - e^{-\lambda_2 t} \\ 0 & 0 & 1 \end{bmatrix}. \tag{33}$$

This can also readily obtained from the exponential distribution properties.

The transition probabilities in Eq. (33) are unconditional probabilities. There are two complications, however, which necessitate the replacement of some with conditional or compound probabilities. First, those found to be free of disease or to have preclinical disease at first screen are not from an entire cohort followed from birth; women with a previous and clinically confirmed disease were excluded from the trial. Thus the probabilities of being free of disease and of having preclinical disease at the first screen should be conditional on having no clinical disease between birth and first screen. Also, the time of entering the clincial phase is known exactly. Their probabilities should therefore be of becoming clinical at the time t_i rather than at some time between 0 and t_i. For one individual, suppose we know that the exact time the person becomees clinical is 5 years, for example. The probability of clinical disease at exactly five years is $P(\text{clinical at 5 years}) = P(\text{not clinical up to } 5 - \Delta t \text{ years}) \times P(\text{become clinical in the interval } (5 - \Delta t, 5))$. Since the model does not allow the possibility of instantaneous transition from no disease to clinical state, and since we wish to explicitly allow for the probability of both rapid and slow progression through preclinical phase, we use our limit of accuracy, in this case one month, and further approximate the correct probability as $P(\text{clinical at 5 years}) = P(\text{not clinical up to } 5 - \Delta t \text{ years})\ P(\text{become clinical in the interval } (5 - \Delta t, 5))$.

As 1 month = 0.08 years, approximately,

$$P = P_{00}(t_i - 0.08)P_{02}(0.08) + P_{01}(t_i - 0.08)P_{12}(0.08)\,.$$

Therefore, the probabilities of being free of disease at first and second screen, P_1 and P_2, were calculated as

$$P_1 = \frac{e^{\lambda_1 t}}{e^{-\lambda_1 t} + \frac{\lambda_1(e^{-\lambda_2 t} - e^{-\lambda_1 t})}{\lambda_1 - \lambda_2}} \tag{34}$$

$$P_2 = e^{-\lambda_1 t}\,. \tag{35}$$

The probabilities of having preclinical disease at first and second sceeen, P_3 and P_4 and the probability of clinical disease at time $t_i\,(I = 1, 2, \ldots, t)$ are

$$P_3 = 1 - P_1\,,$$

$$P_4 = \frac{\lambda_1(e^{\lambda_2 t} - e^{\lambda_1 t})}{\lambda_1 - \lambda_2} \tag{36}$$

$$P_{5i} = e^{-\lambda_1(t_i - 0.08)}\left(1 + \frac{\lambda_2 e^{-\lambda_1 0.08} - \lambda_1 e^{-\lambda_2 0.08}}{\lambda_1 - \lambda_2}\right) + \frac{\lambda_1(e^{-\lambda_2(t_i - 0.08)} - e^{-\lambda_1(t_i - 0.08)})(1 - e^{-\lambda_2 0.08})}{\lambda_1 - \lambda_2}\,. \tag{37}$$

The total likelihood function is

$$L = \prod_{l-1}^{n}(P_1^{1-\delta}P_3^{\delta})(P_2^{1-\delta}P_4^{\delta})P_{5i}\,. \tag{38}$$

δ is the index variable of screening results. If the result of screening test is negative, $\delta = 0$, otherwise, $\delta = 1$. The solution of the likelihood function is complicate, it must be iteratively maximized by a non-standard program. For simplicity, Duffy and Chen equate observed numbers of different types of observation to expected numbers and estimate the parameters by non-linear least squares. Thus, the least squares approximations to maximum likelihood estimates were obtained by a procedure similar to the method of moments. Then the non-linear procedure (NLIN) in SAS[13] may be used to estimate the parameters.

4.1.1. *Example*

A randomized trial was conducted in women aged 40–74 in two counties, Kopparberg and Ostergotland, in Sweden[14] to assess the effect on breast

Table 6. The estimated sojourn time in PCDP of breast cancer.

λ_1	λ_2	$SE(\lambda_2)$	Mean sojourn time	95%CI
0.0052	0.43	0.014	2.3	$2.1 \sim 2.5$

cancer mortality of screening by single-view mammography. The data from two screens were used in the example. The number of invited women at first and second screens was 5410 and 4823. Among those subjects, only 4383 and 3494, respectively, were actually screened. Of those who attended the second screen, 3347 had attended the first screen and 147 had not. Therefore following are the transition histories:

(1) 4383 women were screened at the first screen and 52 cancers were detected. There are 4331 (4383-52) women with transition history (72, 0-0), that is transition from no disease to no disease in 72 years (the average age at baseline was 72).
(2) The 52 cases detected at first screen have transition history (72, 0-1).
(3) 3494 women attended the second screen and 35 cancers were detected, all among the 3347 women who had attended the first screen. Thus 3312 (3347-35) have subsequent transition history (74.72, 0-0).
(4) The 35 cases detected at the second screen have subsequent history (2.75, 0-1).
(5) The 147 women who missed the first screen but attended the second screen have transition history (74.75, 0-0).
(6) There are 10 interval cancers between the first and second screens. Thus, of the above 4331, there are 10 with subsequent transition history (time to interval cancer, 0-2).
(7) There are 68 cases diagnosed clinically after the last screen, which is either the first screen (10 cases) or second (58 cases), depending on whether the subject attended the second screen. These have subsequent history (time to surface to clinical stage, 0-2).

The results of analysis were showed in Table 6.

4.1.2. *Estimation of sensitivity*

Day shows that under the constant incidence assumption, in a time interval T after a negative screen one would expect K new cases, where

$$K = J(1-s)\int_0^T (1 - F(t)dt) + J\int_0^T F(t)dt\,. \tag{39}$$

J is the annual (constant) incidence rate, S is the sensitivity and F is the comulative distribution function of the sojourn time. The first component in the formula for K is the number of cases missed at the screen and the second is the number of new cases "born" since the screen. This suggests as a formula for an estimate of sensitivity

$$\hat{S} \approx 1 - \frac{1 - \hat{K}/J}{1 - \frac{1}{T}\int_0^T f(t)dt}\,, \tag{40}$$

where K is the observed number of new cases. Using the proposed three-state Markov model, the corresponding expected number K of cases in time T after a negative screen at time t_i is

$$\begin{aligned} K = N(1-S)\int_0^{t_1} \lambda_1 e^{-\lambda_1 t} \int_{t_1-t}^{t_1+T-t} \lambda_2 e^{-\lambda_2 u} du dt \\ + N\int_{t_1}^{t_1+T} \lambda_1 e^{-\lambda_1 t} \int_0^{t_1+T-t} \lambda_2 e^{-\lambda_2 u} du dt\,, \end{aligned} \tag{41}$$

where N is the number screened at time t_1. After some integration and algebra, this given an estimate of sensitivity

$$\hat{S} = 1 - \frac{\hat{K}(\lambda_2 - \lambda_1)/N - a}{b}\,, \tag{42}$$

where

$$a = (\lambda_2 - \lambda_1)e^{-\lambda_1 t_1}(1 - e^{-\lambda_1 T}) - \lambda_1 e^{-\lambda_2 T} e^{-\lambda_1 t_1}(e^{-(\lambda_2 - \lambda_1)T} - 1)$$

and

$$b = \lambda_1(e^{-\lambda_1 t_1} - e^{-\lambda_2 t_1})(1 - e^{-\lambda_2 T})\,.$$

Thus, the same data as in the Markov model are used to estimate sensitivity, but in a second stage of estimation.

4.2. *Non homogeneous Markov model with covariables*

In the description of disease progression, the transition of disease states may be interfered by a lot of important factors. For example, some risk factors in living environment may decide the probability of transition from healthy state to ill state. A stochastic model of transition of disease states in the consideration of these factors will be benefit to identify the population with high risk of disease. This population is more appropriate to implement of screening program and is expected to get higher effectiveness. Another case

is that the clinical disease characters decide the probability of transition from preclinical state to clinical state. The inclusion of these variables may strengthen the identification power of stochastic model for different types of diseases and the precise of estimation for the sojourn time in preclinical detectable phase. Therefore, the purpose of developing a stochastic model describes the transition probability of healthy state of an individual not only by populational transition but also by the consideration of individual characters, such as age, gender, disease history, etc. To consider the individual variability, a regression combination of the variables of individual characters may be used to describe the transition probability and it is called as the non-homogeneous Markov model with covariables.

4.2.1. *Non-homogeneous time discrete Markov chain model*

J. Q. Fang and W. Q. Zhou[15,16] suggested a parameterized non-homogeneous Markov chain model in the analysis of screening data for disease. They assumed that the disease process is like in Fig. 5. The state space $S = \{0, 1, 2, 3\}$, The transition probability from state I to state j is defined as

$$P_{ij}(\tau, t) = P\{X(t) = j | X(\tau) = i\}\,. \tag{43}$$

In general screening practice, the screens were given in a fix interval and the time intervals may only differ in several days. This error may be neglected. Then the time t may be assumed as a fix unit, such as year or month. Among the Eq. (46), τ and t belong to screening time set $T \equiv \{ttt\}$. Based on the professional knowledge, it is assumed that the state of individual may just transfer one step in one interval. Suppose the one step transition matrix during the time from t to $t + I$ is

$$P(t) = \begin{bmatrix} p_{00}(t) & p_{01}(t) & 0 & 0 \\ p_{10}(t) & p_{11}(t) & p_{12}(t) & 0 \\ o & 0 & p_{22}(t) & p_{23}(t) \\ o & 0 & 0 & 1 \end{bmatrix} \quad i, j = 0, 1, 2, 3\,. \tag{44}$$

Among expression (44)

$$\begin{aligned} &P_{01}(t) = \alpha_{01} \cdot /\grave{\mathrm{E}}(t)\,, \quad P_{00}(t) = 1 - P_{01}(t)\,, \\ &P_{10}(t) = \alpha_{10} \cdot (1 - \theta(t))\,, \quad P_{12}(t) = \alpha_{12} \cdot /\grave{\mathrm{E}}(t)\,, \\ &P_{22}(t) = 1 - P_{10}(t) - P_{12}(t)\,, \end{aligned} \tag{45}$$

$$P_{23}(t) = \alpha_{23} \cdot /\grave{\mathrm{E}}(t)\,, \quad P_{22}(t) = 1 - P_{23}(t)\,,$$
$$\theta(t) = 1 - \exp(-\beta' Z(t))\,, \quad \beta = (\beta_1, \beta_2, \ldots, \beta_p)'\,. \tag{46}$$

Here the proportional factor α and vector β are the estimation parameters. The transition matrix of m steps during the time from t to $t+m$ is

$$A_m(t) = \prod_{\substack{k=0 \\ t+k \in T}}^{m-1} P(t+k)\,. \tag{47}$$

Suppose there is only one step discriminant error in the state 0, 1, 2, 3, $s \in S$, $P\{s+i|s\} = 1-\gamma$, here γ is the false negative rate. Therefore, the discriminant vectors are

$$B(0) = (1-\gamma, \gamma, 0, 0)'\,, \quad B(1) = (0, 1-\gamma, \gamma, 0)'$$
$$B(2) = (0, 0, 1-\gamma, \gamma)'\,, \quad B(3) = (0, 0, 0, 1-\gamma)'\,. \tag{48}$$

The maximum likelihood function is

$$L = \prod_{k=1}^{N} \prod_{j=1}^{q_k - 1} B'(s_{kj}) A m_{kj}(t_{kj}) B(s_{k,j+1})\,. \tag{49}$$

The hypothesis test of parameters may use the likelihood ratio test. The statistic is

$$G = 2(\ln L - \ln L)\,. \tag{50}$$

When the sample size is big enough and H_0 is true, the statistic G follows the chi-square distribution and the degree of freedom is the number of estimating parameters.

4.2.2. *Non-homogeneous time continuous Markov process model*

J. Q. Fang and J. H. Mao[15,16] suggested a time continuous Markov process model. They assumed that the transition power from state i to state j is

$$\lambda_{ij}(t) \cdot dt = P\{X(t+dt) = j | X(t) = i\} = P_{ij}(t, t+dt)\,,$$
$$t \in [0, \infty)\,, \quad i, j \in S\,. \tag{51}$$

They also assumed that the transition power of two stage model for disease screening was related with p covariables. The model is

$$\begin{cases} \lambda_{01}(t) = A_0 + A_1 Z_1(t) + \cdots + A_p Z_p(t)\,, \\ \lambda_{10}(t) = B_0 + B_1 Z_1(t) + \cdots + B_p Z_p(t)\,, \\ \lambda_{12}(t) = C_0 + C_1 Z_1(t) + \cdots + C_p Z_p(t)\,, \\ \lambda_{23}(t) = D_0 + D_1 Z_1(t) + \cdots + D_p Z_p(t)\,. \end{cases} \tag{52}$$

Among the expression (52) A_i, C_i, D_i, $(I = 1, 2, \ldots, p)$ are the estimating parameters. The one step transition probability and stay probability are

$$P_{00}(\tau,t) = \sum_{i\neq j}^{1,2} \frac{\rho_i + \lambda_{10} + \lambda_{12}}{\rho_i - \rho_j} e^{\rho_i(t-\tau)}\,, \quad P_{11}(\tau,t) = \sum_{i\neq j}^{1,2} \frac{\rho_i + \lambda_{01}}{\rho_i - \rho_j} e^{\rho_i(t-\tau)}\,,$$

$$P_{01}(\tau,t) = \sum_{i\neq j}^{1,2} \frac{\lambda_{01}}{\rho_i - \rho_j} e^{\rho_i(t-\tau)}\,, \quad P_{10}(\tau,t) = \sum_{t\neq j}^{1,2} \frac{\lambda_{10}}{\rho_i - \rho_j} e^{\rho_i(t-\tau)}\,,$$

$$P_{12}(\tau,t) = \sum_{i\neq j}^{1,2} \frac{\rho_i + \lambda_{01}}{\rho_i - \rho_j} \cdot \frac{\lambda_{12}}{\rho_i + \lambda_{23}} (e^{\rho_i(t-\tau)} - e^{-\lambda_{23}(t-\tau)})\,,$$

$$P_{23}(\tau,t) = 1 - e^{-\lambda_{23}(t-\tau)}\,, \quad P_{22}(\tau,t) = e^{-\lambda_{23}(t-\tau)}\,. \tag{53}$$

Among them, $\lambda_{ij} \equiv \lambda_{ij}(t)$, $i \to j \in S$ and

$$\rho_1 = \frac{\lambda_{01} - \lambda_{10} - \lambda_{12} + \sqrt{(\lambda_{10} + \lambda_{12} - \lambda_{01})^2 + 4\lambda_{01}\lambda_{10}}}{2}\,,$$

$$\rho_2 = \frac{\lambda_{01} - \lambda_{10} - \lambda_{12} + \sqrt{(\lambda_{10} + \lambda_{12} - \lambda_{01})^2 + 4\lambda_{01}\lambda_{10}}}{2}\,.$$

The multiple step transition probability is

$$P_{02}(\tau,t) = \int_\tau^t P_{01}(\tau,\xi) \cdot P_{12}(\xi,t) d\xi\,,$$

$$P_{03}(\tau,t) = \int_\tau^t P_{02}(\tau,\xi) \cdot P_{23}(\xi,t) d\xi\,,$$

$$P_{13}(\tau,t) = \int_\tau^t P_{12}(\tau,\xi) \cdot P_{23}(\xi,t) d\xi\,. \tag{54}$$

During the screening process, the total sample of screening population is N and individual i is observed staying in state $s \in S$ and with covariables $Z_j(t)$ at screening time t_i. Therefore, the likelihood model for parameter

estimation is

$$L = \prod_{i=1}^{N} \prod_{k=1}^{m_i - 1} P_{s(t_{ik})s(t_{ik+1})}(t_{ik}, t_{ik+1}) \,. \tag{55}$$

The estimation of parameters may use the maximum likelihood method and the hypothesis testing may use likelihood ratio test.

4.2.3. *Example*

The stochastic model of natural history of disease for nasopharyngeal carcinoma (NPC) screening is used as the example to introduce the development of non-homogeneous Markov model with covariables. The natural history of NPC is showed in Fig. 6

Three states were assumed in the progress of NPC: health, PCDP of NPC and clinical phase of NPC. When the individual transfers from health to PCDP of NPC, gender, age, antibody level and variability characters of Epstein Barr virus (EBV) are the covariables deciding the transition power. According to the natural history of NPC, the transition probability matrix of Markov chain is

$$P(i,j) = \begin{bmatrix} p_{11} & P_{12} & 0 \\ 0 & p_{22} & p_{23} \\ 0 & 0 & 1 \end{bmatrix} \quad i,j = 0,1,2,3 \,,$$

where 1 is state of health, 2 is state of PCDP of NPC and 3 is state of clinical phase of NPC.

Since there is no reverse transition between the states, the $p_{21} = 0$, $p_{32} = 0$ and $p_{32} = 0$. This assumption is reasonable for the progressive diseases, such as malignant tumor. The state 3 is the absorbable state. It is also assumed that the transition between states is no jump. That means

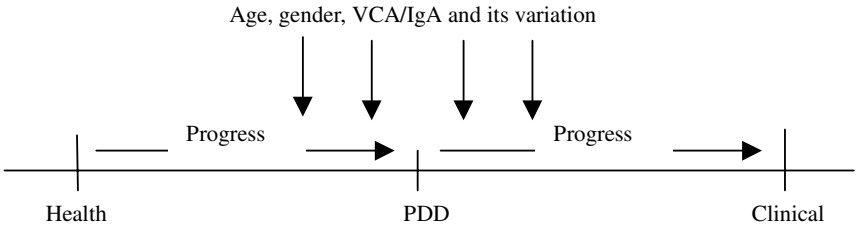


Fig. 6. Natural history model of NPC.

$p_{13} = 0$. Suppose θ is the parameter of transition intension. The transition probabilities are

$$P_{12}(t) = \alpha_{12}\theta(t)\,, \quad P_{11}(t) = 1 - P_{12}(t)\,,$$

$$P_{23}(t) = \alpha_{23}\theta(t)\,, \quad P_{22}(t) = 1 - P_{23}(t)\,, \tag{56}$$

$$\theta(t) = 1 - e^{-\beta X(t)}\,, \tag{57}$$

$$\beta = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\} \quad 0 \le \beta \le \infty\,,$$

$$X(t) = \{x_1, x_2(t), x_3(t), x_4(t)\}\,.$$

Among them, X_1 is gender (1 = woman, 2 = man) and does not change with time. X_2 is age at the time of screen. X_3 is the titer of antibody of EBV(VCA/IgA), 1 = negative, 2 = 1:5 1:10, 3 = 1:20 1:40, 4 = 1:80+. X_4 is the characteristics of variation of VCA/IgA, 1 = negative, 2 = low level of positive antibody, 3 = persistent high level, 4 = increasing level, 5 = both positive of VCA/IgA and EA/IgA. The age and level of antibody of EBV are covariables with time.

It is assumed that the maximum times of transition during a fixed screening interval is m. An individual stay in state i at the time t and transfer to state j during a fixed interval (for example 1 year) m. The matrix of m step transition probability is

$$A_m(t) = \prod_{\substack{k=0 \\ t+k\in T}}^{m-1} P(t+k)\,. \tag{58}$$

In order to estimate the false negative rate, it is assumed that the one step missing discriminant may happen only and suppose $s \in S = 1, 2, 3$. $P(\text{state} = s+1 | \text{diagnosis} = s) = \gamma \cdot \gamma$ is the false negative rate and the diagnostic vector is

$$B(1) = (1-\gamma, \gamma, 0)\,,$$

$$B(2) = (0, 1-\gamma, \gamma)\,,$$

$$B(3) = (0, 0, 1)\,.$$

Supposed N individuals are screened, individual $k = 1, 2, \ldots, N$ participates q_k screens and then the likelihood function is

$$L = \prod_{k=1}^{N} \prod_{j=1}^{q_k - 1} B'(s_{kj}) A m_{kj}(t_{kj}) B(s_{k,j+1})\,. \tag{59}$$

Table 7. Estimated values of parameters in natural history model of NPC and hypothesis test.

Parameters	Values	Likelihood test	*P value*
α_{12}	0.001908		
α_{23}	0.2051		
β_0	−0.4163		
β_1 (gender)	0.1651	10.7338	< 0.00005
β_2 (age)	0.00002217	0.0194	> 0.9
β_3 (VCA/IgA)	0.1553	34.2804	< 0.0001
β_4 (variation of VCA/IgA)	0.1879	25.1728	< 0.0001
γ (false negative rate)	0.0002		
M (Maxi. transitions)	2		

A mass screening for nasopharyngeal carcinoma was carried out in Guangzhou,[17] where 2970 cases with positive results of test for VCA/IgA and 3 cases of NPC were found. All the cases with positive VCA/IgA and 214 controls with negative VCA/IgA were followed up and 35 NPC were found during a 7-year period. And 2988 cases with positive VCA/IgA and 34 cases of NPC were found in first screen in Zhongshan. All cases with positive VCA/IgA and 2068 controls with negative VCA/IgA were followed up and 40 cases of NPC were found during a 7-year period. Also, 1297 cases with positive VCA/IgA and 13 cases of NPC were found in first screen in Sihui; 19 cases of NPC were found during a 7-year follow-up for the cases with positive VCA/IgA.

A Markov model with time dependent covariables was developed and the results showed in Table 7.

5. The Simulation and Optimization of Screening Policy

Based on the parameters of natural history of disease and screening implement, the disease process and the effects of screening intervention may be simulated and the cost-effectiveness may be evaluated. There are two purposes of analysis and assessment of screening data. One is to estimate the parameters of screening, including the attendance rate, cost, the characteristics of screening test (sensitivity and specificity) natural history of disease (the sojourn time in preclinical detectable phase PCDP), the impact of screening on the mortality and prevalence of disease, etc. Based on these parameters, one may make a conclusion if the mass screening is efficacy. Another thing is to choose an optimized screening policy. It means the choice of population in eligible age group, frequency of screening and the

interval between subsequent screening test, the combination of screening test and sequential diagnostic procedures. An optimized screening policy is expected to obtain the maximum health effects in the limited resource. The choice of a screening policy should preferably be based on the balance between expected health effects and costs. The development of natural history model of disease serves for first purpose and the simulation of disease process and screening serves for second one. The fundament of simulation for disease process is technique of Monte Carlo. Based on the known and hypothetic parameters, a screening process for disease in a large population is simulated by computer and the simulated effects of screening policy are evaluated. A variety of screening policies are simulated repeatedly and the optimized policy of screening are identified for the consideration of policy decision-maker.

Knox[18] first used the macro-simulation method to evaluate the health effects of screening for cervical cancer in England. He assumed that the duration of the interval between the point at which the disease first becomes detectable and the point at which it becomes incurable is a constant (A). The duration of the interval between incurability and death is B. The sensitivity of screening test is S. The interval between subsequent screens is I. The disease incidence rates in different age groups are P. Now a mass screening program was carried out starting in age B in a 100,000 population. In the population, the disease onsets in a speed of P. It is assumed that the disease is curable if it is detected by screening and the life may be saved. The disease is incurable if it is diagnosed in the hospital and the life lost is the mortality rate (D) of disease. If the main concern were to save life year rather than lives, the life year lost ($Y = De_x^0$ taken from the current life-table) and an another compromise is the weighted index ($\mathrm{Id} = D\sqrt{e_x}Y^{-1}$), that is long survivals are not weighted in proportion to their length. The health effects of different screening policy are demonstrated in Fig. 7. The dash line is the original Id caused by the disease. Each test, after an interval of B, produces a deep cut in the mortality, proportional in depth to the sensitivity of the test. The cut persists for a period A, and ends. Closely set tests involve some waste because of overlap, but later tests cut by the same proportion into the cases missed by earlier tests. The interval of first screen and second screen is wider than A so that the Id returns to original line when the health effects of early detection of screening on the disease mortality disappear. The interval of second screen and third one is shorter than A, so that Id goes down again before it return to original value. The sensitivity of screening test decides the depth of cut.

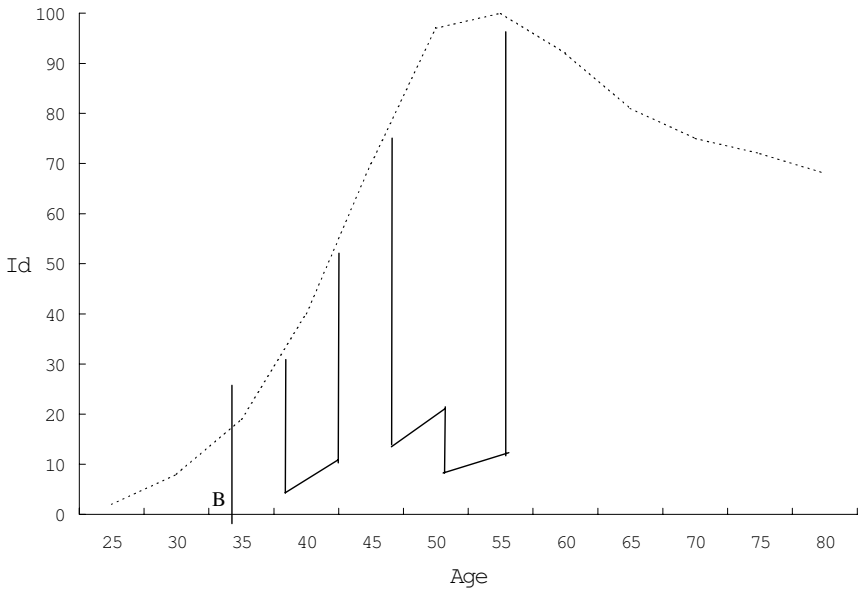


Fig. 7. Illustration of simulated effects of screening for disease.

A high sensitivity means the cut is close to the horizontal axis and a good health effect of screening. However, the expense of high sensitivity is a low specificity and a high false positive rate when the power of screening test is fixed. That means a increase of amount of following diagnostic work and the total cost of screening.

Knox simulated the cost-effectiveness of mass screening for cervical cancer in England by a simplified macro-simulation model. It is showed in Fig. 7 that the maximum health effects may obtain when screening is started just before the steep increase of disease incidence rate and more frequently screening in the age with high mortality rate of disease. It can be seen by simulation that a very wide range of results can be obtained from different deployments of the same resource, the range itself depending upon the natural history. For example, Knox assumed that the natural history distribution centered upon a mean interval is 6-year for cervical cancer, a 5-year spacing of tests beginning at age 35 gives something like 30 times the benefit of a one-year spacing beginning at age 20 and ending at age 29. The health effects of different screening policies can be roughly compared by macro-simulation model with relative simple calculation and the optimized scheme can be suggested. However, the parameters in macro-simulation are only assumed as the constants and the average disease process and

the health effects of screening in an whole population are simulated the variability of individuals is ignored. It is known that the sojourn time of preclinical detectable phase is a variable with a certain distribution. The sensitivity and specificity of screening test depends on the individual characteristics of disease. And the disease process may change in different individuals. Therefore the macro-simulation can not consider the variations between individuals and evaluate the cost-effectiveness of different screening policies precisely.

Habbema[19] developed a micro-simulation model of screening process by the assistance of computer. The disease process and the impact of screening intervention of every individual in 100,000 population were simulated by the method of Monte Carlo. This simulation model divided into two parts: the disease part and the screening part. The disease part generates a large number of life histories. Together, the life histories constitute the target population that will be screened in the screening part. The stochastic model underlying the simulation of the population is specified by the input of the program. The input related to the population (e.g. the life table), the epidemiology of the disease (e.g. age-specific incidence) and the disease process. Important aspects of the disease process include disease states into which preclinical and clinical disease is subdivided, the duration of preclinical disease, the probability that preclinical disease will regress spontaneously, etc. The output of disease part consists of the simulated life histories. All types of epidemiological data are computed from the aggregation of life histories: incidence of clinical disease, the prevalence of the disease states, the mortality, and survival figures. The input of the screening part consists of assumptions on the screening process (properties of the screening test, prognosis after early detection) and of a specification of the screening policy. The output of screening part consists of the simulated screening results (e.g. the number of cases detected, number of cases missed, mortality among screen-detected cases) and of the simulated effects of screening (e.g. the number of lives/life years saved, and the number of unnecessarily treated persons). Habbema applied this model to the evaluation of screening for breast cancer and colorectal cancer.

Here the structure of micro-simulation model is introduced by an example of screening for nasopharyngeal carcinoma.[20] The basic structure of model for disease process and screening process of NPC is illustrated in Fig. 8.

The main biomarker of NPC risk is the antibody level of Epstein Barr virus (VCA/IgA). In order to simplify, the positive rate of VCA/IgA is assumed as a constant, e.g. instantaneous transition rate λ. It is supposed

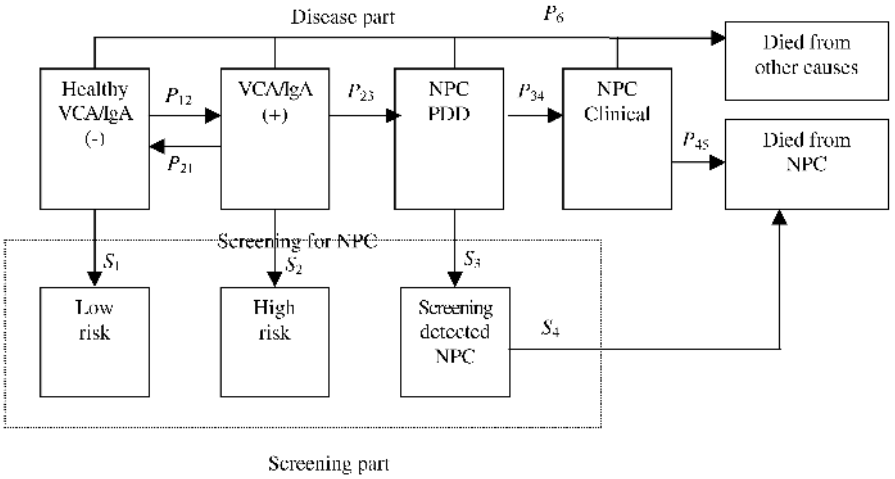


Fig. 8. Structure of the disease model for NPC and stages used in model.

that the distribution of λ depended upon the time is an exponential distribution and then the accumulated distribution probability at certain time can be estimated, e.g. transition probability (P_{12}). It is assumed that a part of population wit positive VCA/IgA may become negative and the transition probability is P_{21}. It is reasonable that the progression of nasopharyngeal carcinoma is irreversible. Therefore, the preclinical cases should progress to clinical without the intervention of medication. It means the incidence of preclinical NPC is same as the incidence of clinical NPC. The transition probabilities from normal population with negative VCA/IgA to preclinical NPC are the incidence rates (I_{0i}) of NPC in different age groups with negative VCA/IgA. The incidence rates among the population with positive VCA/IgA are $I_{1i} = I_{0i} \times RR_{EB}$ (the relative risk of positive VCA/IgA). These are the instantaneous transition probabilities. The transition probabilities (P_{13} and P_{23}) can be estimated if the distribution functions are assumed. If there is no intervention of screening, the preclinical NPC cases enter the clinical phase according to the distribution probability (P_{34}) of sojourn time in preclinical detectable phase. And then these cases will die according to survival rate (P_{45}) of clinical cases of NPC. All individuals may die from other cause in every stages depended upon the age-specific mortality rates (P_6). Finally, the death age of every individual are simulated. That is the disease part of model without the intervention of screening. The output of disease part includes the simulated life history of

every individual and other disease index such as incidence rate, prevalence rates in different disease stages, mortality rate, etc. The simulated results are compared to the actual figures. Some adjustment of parameters should be done to make simulation model close to realistic population.

In the screening part, the simulated population goes through the screening intervention according to the different policies of screening. For example, the population ise divided into the low and the high risk groups according the level of VCA/IgA and are screened in different frequencies. The early detected cases will be checked by further diagnostic procedure. Therefore, the simulated population will enter these three states according to the sensitivity and specificity of screening tests. The transition probability (S_4) from screening detected NPC to death is estimated depended upon the survival rate of preclinical NPC cases. In theory, $S_4 < P_{45}$. It means that the early detection may save the life of NPC cases. The simulated results of screening part are also the death age (d) of every individual. The total effect of screening is the life year saving ($Y = \sum(D-d)$). The total cost (C) of screening can be estimated according to the simulated screening process including physical check, test of VCA/IgA, further diagnostic procedure and the organization of screening program. The cost-effectiveness index of screening is the average cost per life year saving $= C/Y$. The lower the cost-effectiveness index, the more the total life year saving, the more the health effect of screening is.

The simulated results of different screening schemes for NPC are listed in Table 8. It is showed that the cost-effectiveness is the best when the positive value of VCA/IgA sets on 1:20. A lower value increases the false

Table 8. Simulated effects of screening in different policies.

VCA/IgA		VCA/IgA $\geq$ 1:5		VCA/IgA $\geq$ 1:20		VCA/IgA $\geq$ 1:80	
(−)	(+)	Life year saving	Cost per life year	Life year saving	Cost per life year	Life year saving	Cost per life year
1	1	5848	7210.30	6212	6787.53	5620	7503.04
3	1	5716	3243.40	5227	3268.65	4603	3617.20
3	2	4606	3788.55	4562	3697.57	4547	3653.35
3	3	4025	3738.33	4171	3607.44	3696	4070.09
5	1	4921	2605.28	4542	2523.95	3759	2967.23
5	2	4376	2772.54	4051	2809.80	3340	3336.29
5	3	4100	2871.37	3758	3005.80	3208	3466.84
5	4	3804	3053.20	3697	3046.84	3343	3326.62
5	5	3591	2679.13	3282	2930.68	3239	2969.72

positive rate and the cost for further diagnosis. A higher value increases the missing rate and the cost-effectiveness index since the less life saving obtains. The cost is the lowest and the total life year saving is median when the interval between screens is every year for population with positive VCA/IgA and every 5 years for population with negative VCA/IgA. That may be considered as the optimized screening scheme for NPC.

From the results of simulation, it is showed that the cost-effectiveness index of the poorest scheme is four times higher than that of the best one. It cannot be overemphasized the importance of simulation study for the assessment of screening.

References

1. United States Commission on Chronic Illness (1957). *Chronic illness in the United States*, Vol. 1, Cambridge, MA:harvard University Press, MA, 267.
2. Wilson, J. and Jungner, G. (1968). Principles and practices of screening for disease. *Public Health Paper 34*, World Health Organization, Geneva, 26–39.
3. Shapiro, S., Goldberg, J. D. and Hutchinson, G. B. (1974). Lead-time in breast cancer screening detection and implication for periodicity of screening. *The American Journal of Epidemiology* **100**: 357–360.
4. Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56**: 601–614.
5. Shapiro, S., Venet, W., Strax, P. *et al.* (1982). Ten- to fourteen-year effect of screening on breast cancer mortality. *Journal of National Cancer Institute* **69**: 249–255.
6. Tabar, L., Fagerberg, G., Duffy, S. W. *et al.* (1989). The Swedish two county trial of mamographic screening for breast cancer: Recent results and calculation of benefit. *Journal of Epidemiology Community Health* **43**: 107–114.
7. Albert, A., Gertman, P. M. and Liu, S. (1978). Screening for the early dectection of cancer: II. The impact of screening on the natural history of the disease. *Mathematical Biosciences* **40**: 61–109.
8. Eddy, D. (1980). *Screening for Cancer*: *Theory*, *Analysis and Design*, Englewood Cliffs, Prentice-Hall, New Jersey.
9. Day, N. E. and Walter, S. D. (1984). Simplified models of screening for chronic disease: Estimation procedures from mass screening programmes. *Biometrics* **40**: 1–14.
10. Brookmeyer, R. and Day, N. E. (1987). Two-stage models for the analysis of cacner screening data. *Biometrics* **43**: 657–669.
11. MacGregor, J. E., Moss, S., Parkin, D. M. *et al.* (1985). A case-control study of cervical cancer screening in Northeaset Scotland. *British Medical Journal* **290**: 1543–1546.
12. Duffy, S. W., Chen, H. H., Tabar, L. *et al.* (1995). Estimation of mean sojourn time in breast cancer screening using a markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine* **14**: 1531–1543.

13. SAS Institute Inc. (1989). SAS/STAT Users Guide, Version6, Volume 2, Cary NC: SAS Institute Inc.
14. Tabar, L., Fagerberg, G., Duffy, S. W. *et al.* (1992). Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiologic Clinics of North America* **30**: 187–210.
15. Fang, J. Q., Wu, C. B., Mao, J. H. *et al.* (1995). Two stage model of tumor sojourn time(I)-non-homogeneous Markov Model. *Applied Probability and Statistics* **11**(2): 205–212.
16. Fang, J. Q. and Wu, C. B. (1995). Two stage model of tumor sojourn time(II) — Counting process and Bootstrap method. *Applied Probability and Statistics* **11**(2): 213–222.
17. Liu, Q., Fang, J. Q., Hu, M. X. *et al.* (1997). Stochastic model of natural history of nasopharyngeal carcinoma. *Chinese Journal of Health Statistics* **14**(4): 12.
18. Knox, E. G. (1976). Ages and frequences for cervical cancer screening. *British Journal of Cancer* **34**: 444–452.
19. Habbema, J. D. F., Van Oortmarssen, G. J., Lubbe, J. T. N. *et al.* (1984). The MISCAN simulation program for the evaluation of screening for disease. *Computation Methods and Programs Biomedicine* **20**: 79–93.
20. Liu, Q. (1995). Simulation study of effect evaluation of screening for tumor. *Chinese Journal of Tumor* **17**: 182.

About the Author

Qing Liu got his BS in Public Health (1982) from Guangdong Pharmacological and Medical College. MS (1985) and PhD (1996) in Medical Statistics from Sun Yat-Sen University of Medical Sciences. Dr. Liu has worked in Sun Yat-Sen University as assistant professor, associate professor and professor since 1985. Attending to an international seminar of clinical epidemiology held in West Chinese Medical University in 1985 and being a fellow in International Agency of Research on Cancer during 1990–1991. He has been teaching medical statistics, epidemiology and clinical epidemiology for years. His main research interests are the epidemiology of chronic diseases and related statistical methods, the model of natural history of diseases and evaluation of cancer screening, genetic epidemiology of cancer.

This page intentionally left blank

CHAPTER 21

CAUSAL INFERENCE

ZHI GENG

Department of Probability and Statistics, School of Mathematical Sciences, Peking University, Beijing 100871, PR China
Tel: 86-10-62751837; zgeng@math.pku.edu.cn

1. Introduction

More than 2000 years ago, Aristotle pointed out that the real scientific knowledge was about causation. Since ancient times, exploration of causation has been the ultimate destination of almost all the scientific studies including philosophy, social sciences and medical sciences. Causation and association are two different and important concepts. Although causation is more important than association in many scientific studies, most of the statistical methods at present can only be suitable for associational studies. Even if there is not causation between two factors, there may be spurious association; on the other hand, causation may also appear spurious independence. A lot of examples may explain spurious association. For example, the watch times of Mr. John Doe 1 and Mr. John Doe 2 have very strong association, but changing the watch time of Mr. John Doe 1 would not affect that of Mr. John Doe 2. Freedman[7] gave an example that the reading ability of primary school students was related to sizes of their shoes, but there was not causation between them apparently because changing the shoes sizes would not change their reading ability. A few of statisticians and medical researchers wondered whether causation should present association. In fact, there are many examples of spurious independence. We may imagine that exercising Taijiquan (shadow boxing) can build up health, that is, doing Taijiquan has causal effects on health. However, people who exercise Taijiquan may show little difference in health from, or even worse than, those who do not exercise Taijiquan. That may be because

people exercising Taijiquan were all in poor health and would be worse if they had not exercised Tajiquan, and thus spurious independence appears. Another example is that the life of uranium miners was as long as, or even longer than, that of others who were not exposed to uranium mines. It cannot be explain that exposure to uranium mines would not affect one's life. Perhaps uranium miners were selected from the stronger. If they had not been exposed to uranium mines, they would have longer life. This is called healthy worker effect.

In the history of statistical science, research on causal inference has not received deserved emphasis and development. The early statistical theories and methods of causal inference include contingency tables, path analysis and structural equation models. Although association and causation are the well-known different concepts, association obtained by statistical inference is not unusually misused to explain the relation between cause and effect. In application studies, one often overlooks assumptions on causal mechanisms and interprets parameters of association as those of causation. Rubin[35] proposed a causal effect model, which was similar to the counterfactual philosophy of Lewis[21] and was called the counterfactual model. Pearl[26] proposed concepts of causal diagram and external intervention and established the method of diagram for causal inference that combined knowledge about causal mechanisms with data from observational studies. These causal models need various assumptions that cannot be tested by using data from observational studies. The statistical theories and methods are lack in causal inference and identification of confounders. Causal inference is a complex problem involving statistics, philosophy and related application areas and has been discussed by many authors in recent years.[7,14,33]

2. Experimental Studies and Observational Studies

The difference between observation and experiments is that information from observation seems to emerge by itself while that from experiments is knowledge on the truth.[1] Observation is a method for collecting facts while experiment is a means for obtaining knowledge.[1] The information from an experiment has an essential distinction from that from observation. Experiments try to explore information about causation, and observational studies can get information only about association. Only under some conditions or assumptions can they are equivalent. Holland[17] pointed out that it was impossible to make causal inference based on observational studies without untestable assumptions. From philosophic views of the

Popperian, an affirmation is not scientific if it is not empirically testable. Randomized experiment is the best scientific method to assess causal effects. However, randomized experiments or even experimental methods are prohibited in many studies, and only observational studies can be conducted. A well-known example is the epidemiological study on cancer and smoking. When randomized experiments cannot be carried out, a case-reference study attempts to seek a control group that is comparable to the treatment group. Grace *et al.*[13] gave the results of 51 studies on the portacaval shunt, in which some are randomized experimental studies, some are controlled studies and some are uncontrolled studies, as shown in Table 1, where 24 of 32 uncontrolled studies and 10 of 15 nonrandomized controlled studies were markedly enthusiastic about the shunt, while all of 4 randomized controlled studies showed that the surgery does not have effect. This shows that different study methods may lead to completely different conclusions.[8]

In 1948 and 1949, Doll and Hill carried out a case-control study on lung cancer and smoking in 20 hospitals in London and found significant association between smoking and lung cancer (see Table 2). Fisher pointed out in 1957 that association from case-control studies could not be explained simply as causal relation between smoking and lung cancer and this association might be explained by other two alternative theories: (1) Cancer causes smoking; (2) Genes cause both cancer and smoking. He illustrated the association between genes and smoking habit from data of twins. Up to the present, epidemiologists have shown by various methods and data that smoking is a risk factor of lung cancer: (1) Dose response relation: The

Table 1. 51 studies on the portacaval shunt.[8,13]

Design	Degree of Enthusiasm		
	Marked	Moderate	None
No Controls	24	7	1
Controls, but not Randomized	10	3	2
Randomized Controlled	0	1	3

Table 2. The case-control study on lung cancer and smoking.

	Male		Female		
	Smoking	No smoking	Smoking	No smoking	Total
Lung Cancer	647	29	2	41	719
Controlled	622	27	28	32	709
	$P = 0.64 \times 10^{-6}$		$P = 0.025$		

Table 3. Water source and death rate from cholera.[8]

	No. of Houses	Deaths from Cholera	Rate per 10,000
SV Company	40,046	1,263	315
L Company	26,107	98	37

more one smokes, the bigger risk is; (2) The risk increases as the years of smoking increases; (3) The risk decreases as the years of giving up smoking increases.

Freedman[7] reviewed Snows study on cholera. In 1855, Snow, a physician in London, found that cholera was a kind of infectious disease through drinking water. He developed a series of arguments to support his germ theory about cholera. For example, cholera spread along the tracks of human commerce. When a ship stopped at a port where cholera was prevailing, sailors who contacted local residents would contract the disease. These could show that cholera was an infectious disease and could not be explained by miasma or bad air. In August and September 1854, cholera broke out in London and the patients gathered mainly near water pumps of Broad Street in Soho District. A number of groups in the district, fortunately, escaped from cholera. One was a brewery, where the workers preferred ale to water and it had a private pump. Another was a poor-house, which owned a water pump. Snow could explain that most of patients in other areas also drank water from the pump in Broad Street. For example, a female patient in Hampstead had lived in Soho before and she liked the taste of the water of Broad Street and routinely draw water from there. There were several water supply companies in London in the 19th century. There was no difference between the customers of the Southwark and Vauxhall (SV) Company and the Lambeth (L) Company, which imitated an experiment of nature. The water source of the SV Company contained polluted water while the L Company drew relatively pure water. It was found that the mortality rate from the SV Company was about 9 times the death rate for the L Company, as shown in Table 3. Therefore, it was concluded that the water source was the cause of cholera. Snows study shows that non-experimental study can also be applied for causal inference.

3. Simpsons Paradox and Standardization

We first illustrate the paradox proposed by Simpson[36] in 1951, which is usually called Simpsons Paradox although many statisticians had noted it

Table 4. A numerical example of smoking and lung cancer.

	Case	Control	Total
Smoking	80	120	200
No Smoking	100	100	200
	$RR = 0.80$, $OR = 0.67$, $RD = -0.10$		

Table 5. Stratification with sex.

	Male		Female	
	Case	Control	Case	Control
Smoking	35	15	45	105
No Smoking	90	60	10	40
	$RR_1 = 1.17$		$RR_2 = 1.50$	
	$OR_1 = 1.56$		$OR_2 = 1.71$	
	$RD_1 = 0.10$		$RD_2 = 0.10$	

before 1951. Suppose that we obtain observed data on smoking and lung cancer given in Table 4, which may be regarded as the distribution from the population. Thus the relative risk is $RR = (80/200)/(100/200) = 0.80$, the odds ratio is $OR = (80 \times 100)/(100 \times 120) = 0.67$ and the risk difference is $RD = (80/200) - (100/200) = -0.10$. From these measures, we can see that the prevalent rate of cancer in the smoking population is lower than that in the no smoking population, and thus smoking seems to be a preventive factor.

Suppose that sex is also included in the observation, and data stratified by sex are shown in Table 5. For male, the relative risk is $RR_1 = 1.17$, the odds ratio $OR_1 = 1.56$ and the risk difference $RD_1 = 0.10$. For female, the relative risk is $RR_2 = 1.50$, the odds ratio $OR_2 = 1.71$ and the risk difference $RD_2 = 0.10$. Smoking appears to be a risk factor. It follows from Tables 4 and 5 that smoking is harmful to both male and female, but is good for human. This phenomenon is called Simpsons Paradox.

Suppose furthermore that the observation includes age, classified as two levels: Age under 40 and above 40. The $2 \times 2 \times 2 \times 2$ contingency table is shown in Table 6, where smoking seems to be again a preventive factor across the strata of sex and age. This numerical example illustrates that covariates should carefully considered in survey design and data analysis; otherwise, a spurious association may be obtained.

Table 6. Stratification with sex and age.

Age	≤ 40				> 40			
Sex	Male		Female		Male		Female	
Control	Case	Control	Case	Control	Case	Control	Case	Control
Smoking	5	5	40	50	30	10	5	55
No smoking	60	55	5	5	30	5	5	35
	$RR_{11} = 0.96$		$RR_{12} = 0.89$		$RR_{21} = 0.88$		$RR_{22} = 0.67$	
	$OR_{11} = 0.92$		$OR_{12} = 0.80$		$OR_{21} = 0.50$		$OR_{22} = 0.64$	
	$RD_{11} = -0.02$		$DR_{12} = -0.06$		$RD_{21} = -0.11$		$RD_{22} = -0.04$	

Table 7. Admission of graduate students at the University of California, Berkeley.

	Admission	Rejection	Sum	Percentage of Admission
Men	3738	4704	8442	44
Women	1494	2827	4321	35

Table 8. Stratification with major.

Major	A	B	C	D	E	F
No. of Male Applicants	825	560	325	417	191	373
Percentage of admission	62	63	37	33	28	6
No. of Female Applicants	108	25	593	375	393	341
Percentage of Admission	82	68	34	35	24	7
Sum of Applicants	933	585	918	792	584	714

Bickel *et al.*[2] reanalyzed the data of the observation study on sex discrimination in admissions of graduate students at the University of California, Berkeley. The data on sex and admission are shown in Table 7. The admission rate for the men was about 44% and that for the women was about 35%, which appeared the existence of sex discrimination.

By looking at each major separately, however, it was found that there did not exist sex discrimination in any major and, even if there existed, it discriminated against the men. Data for the six largest majors are shown in Table 8, where the candidates in the six majors covered over one-third of all candidates in over 100 majors of the university. It can be seen that the reason why the admission rate for the women was lower than that for the men was that most of the women applied to the hard majors with lower admission rates. In order to eliminate the influence of different distributions

of the men and women in application to majors, we may select, for example, the sums of each majors applicants in the last row of Table 8 as the standard distribution of majors, and then we adjust the distributions of the men and women in majors. The standardized admission rate for the men after adjustment is

$$\frac{0.62 \times 933 + 0.63 \times 585 + 0.37 \times 918 + 0.33 \times 792 + 0.28 \times 584 + 0.06 \times 714}{4526} = 0.39\,,$$

and the standardized admission rate for the women after adjustment is

$$\frac{0.82 \times 933 + 0.68 \times 585 + 0.34 \times 918 + 0.35 \times 792 + 0.24 \times 584 + 0.07 \times 714}{4526} = 0.43\,,$$

It can be seen that the standardized admission rate for the women is slightly higher than that for the men after eliminating the influence of different distributions of majors.

4. Counterfactual Model for Causal Inference

Rubin[35] proposed the counterfactual model for causal inference, which can be traced back to Neyman.[23] The basic idea is the potential outcomes. If we could observe the responses of a unit under both exposed and unexposed to the treatment, we would assess the causal effect of the treatment on the unit by the difference between the two responses. In epidemiological and medical studies, however, every unit can be under only one exposure status, exposed or unexposed. As the motto of Heraclitus: "You cant step into the same river twice," only one response can be observed and the other cannot. Some untestable assumptions are necessary for causal inference from observational studies. Thus, there are arguments about the counterfactual model,[5] but this property reflects a strength of the counterfactual model.[16] The counterfactual model is widely applied to causal inference, which gives the most precise definition and description of causal effects.

4.1. *Causal effects*

Definitions of causal effects at three different levels are introduced in this section.[18,35] Let E be a binary variable denoting treatment (or exposure) $E = e$, denotes treated (or exposed) and $E = \bar{e}$ denotes untreated (or unexposed). Let $D_e(u)$ denote the response of unit u under treated $E = e$, and $D_{\bar{e}}(u)$ denote the response under untreated $E = \bar{e}$.

Definition 4.1. The individual causal effect (ICE) of the unit u. For a particular unit u and an interval of time from the exposure time t_1 to the

response time t_2, the causal effect of the e versus $\bar{e}$ treatment on the unit u is defined as:

$$ICE(u) = D_e(u) - D_{\bar{e}}(u)\,.$$

Since any unit u can be under only one exposure status, it is impossible to observe the both D_e and $D_{\bar{e}}$ on the same unit, and thus it is impossible to observe $ICE(u)$. So this model is called the counterfactual model or the potential model. If there exist two units u_1 and u_2 so that $D_e(u_1) = D_e(u_2)$ and $D_{\bar{e}}(u_1) = D_{\bar{e}}(u_2)$, and if the units u_1 and u_2 receive different treatments, then the individual causal effect $ICE(u_1)$ of the unit u_1 can be observed.

Definition 4.2. The average causal effect (ACE) over U. Suppose that there are N units in the population U. The average causal effect (ACE) over U is defined as the mean of individual causal effects of all units:

$$ACE = \frac{1}{N}\sum_{u\in U} ICE(u) = E(D_e - D_{\bar{e}})\,.$$

The average causal effect denotes the difference between the response average if all units had been exposed and that if all units had not been exposed. Since every unit can be under only one exposure status, ACE is also a potential quantity.

We define a covariate as a variable unaffected by treatment. For example, a variable that occur before treatment is a covariate because it is not affected by the treatment.

Definition 4.3. Let X be a discrete covariate. The population U is stratified into K subpopulations by the covariate $X = 1, \ldots, K$. The average causal effect of the subpopulation $X = k$ is defined as:

$$ACE_k = E(D_e - D_{\bar{e}}|X = k) = P(D_e = 1|X = k) - P(D_{\bar{e}} = 1|X = k)\,.$$

ACE_k denotes the difference between the average response if all units in the subpopulation $X = k$ had been exposed and that if all of the units had not been exposed.

4.2. *Randomized experiments*

Randomized experiment is the most powerful method to assess average causal effects. Randomization can balance the joint distribution of all known confounders and all unknown confounders in the treated group and that in

the control group. That is, randomization can eliminate not only the known confounders but also the unknown confounders.

In a randomized experiment, treatment E is independent of any other covariates, and particularly, E is independent of both D_e and $D_{\bar{e}}$, denoted by $E \perp (D_e, D_{\bar{e}})$. Thus, we have $E(D_i|E = i) = E(D_i)$ for $i = e$ or $\bar{e}$, where $E(D_i|E = i)$ denotes the expectation of the response D_i in the group $E = i$, which can be estimated by using the mean of responses of all units in the group $E = i$. If units are randomly sampled from the population, $E(D_i|E = i)$ is asymptotically equal to the average causal effect when all units in the whole population receive the treatment $E = i$. For a randomized experiment, we have

$$E(D_e|E = e) - E(D_{\bar{e}}|E = \bar{e}) = E(D_e) - E(D_{\bar{e}}) = E(D_e - D_{\bar{e}}) = ACE\,.$$

For a randomized experiment, we can obtain the following unbiased estimate of ACE from sample means:

$$\begin{aligned}\widehat{ACE} &= \text{the average of all responses in the treated group}\\ &\quad - \text{the average of all responses in the control group}\\ &= \frac{1}{\text{the number of units in the treated group}}\\ &\quad \times \sum_{u \text{ in the treated group}} D_e(u)\\ &\quad - \frac{1}{\text{the number of units in the control group}}\\ &\quad \times \sum_{u \text{ in the control group}} D_{\bar{e}}(u)\,.\end{aligned}$$

4.3. *Ignorability and the propensity score*

Rosenbaum and Rubin[34] proposed the following assumption, called strong ignorability. This assumption is widely applied in observational studies.

Definition 4.4. We say that a treatment assignment E is strongly ignorable given a covariates X if

$$E \perp (D_e, D_{\bar{e}})|X \quad \text{and} \quad 0 < P(E = e|X = x) < 1\,.$$

If a treatment assignment is strongly ignorable, then we have, for $i = e$ and $i = \bar{e}$,

$$E(D_i|E = i, X = x) = E(D_i|X = x)\,.$$

Therefore, the average causal effect in subpopulation $X = x$ is

$$\begin{aligned} ACE_x &= E(D_e|X = x) - E(D_{\bar{e}}|X = x) \\ &= E(D_e|E = e, X = x) - E(D_{\bar{e}}|E = \bar{e}, X = x)\,, \end{aligned}$$

where $E(D_e|E = e, X = x)$ and $E(D_{\bar{e}}|E = \bar{e}, X = x)$ can be estimated by using observed data.

Theorem 4.1. *Under the assumption of strong ignorability, the average causal effect* (ACE) *in the whole population equals the expectation of the average causal effects in subpopulations. When* X *is discrete, it can be denoted as*:

$$\begin{aligned} ACE = E(ACE_k) &= \sum_k ACE_k P(X = k) \\ &= \sum_k \{[P(D_e = 1|E = e, X = k) \\ &\quad - P(D_{\bar{e}} = 1|E = \bar{e}, X = k)]P(X = k)\}\,. \end{aligned}$$

Under the assumption of strong ignorability, the average causal effect ACE *is identifiable and can be estimated from observed data.*

In an observational study, if we can observe enough covariates X so that the assumption of strong ignorability X holds, then we first stratify the population by X, next compute the average causal effects in subpopulations and finally obtain the average causal effect ACE with the above weighted average. The assumption of strong ignorability is one of the important assumptions for causal inference in observational studies. Note that this assumption cannot be tested empirically. Thus we must depend on knowledge and experience of experts in related disciplines for judgment, or we develop experiments (for example, stratified randomization) to ensure the assumption of strong ignorability. Stone[39] discussed various assumptions needed for causal inference.

When X is a continuous covariate or a discrete covariate with a lot of levels, even if stratification could ensure the strong ignorability, it makes each stratum sparseness and thus makes statistical inference inefficient. One solution to this problem is to use the propensity score to stratify the population as coarsely as possible so that units in each stratum can be as more as possible.

Definition 4.5. Let X be a continuous or discrete covariate. The propensity score is defined as the conditional probability:

$$f(X) = P(E = 1|X)\,,$$

where $f(X)$ is a function of X.

Rosenbaum and Rubin[34] presented the following result.

Theorem 4.2. *If treatment assignment E is strongly ignorable given covariates X (i.e. $E \perp (D_e, D_{\bar{e}}|X$ and $0 < P(E = e|X = x) < 1$), then treatment assignment E is also strongly ignorable given the propensity score $f(X)$ (i.e. $E \perp (D_e, D_{\bar{e}})|f(X)$ and $0 < P(E = e|f(x)) < 1$).*

Since the propensity score $f(X)$ is a function of X, stratifying the population by the propensity score $f(X)$ is coarser than that by X. When X is a continuous variable, we may apply a logistic regression model to the propensity score.

5. Confounding and Confounders

Confounding is one of the most important concepts in observational studies. Greenland *et al.*[16] provided an overview on confounding. In epidemiological studies, selection of a control (unexposed) group should ensure comparability of the exposed group with the control group. The difference between the response distributions in comparison groups is called confounding. When there exists confounding, methods of control and adjustment for confounders are usually used to reducing and remove confounding. There is, however, inconsistency on definitions of confounding biases and confounders in the epidemiological literature. There are two main approaches for assessing confounding and a confounder:

(1) Collapsibility-based criterion: A covariate is a confounder if a measure of association across strata of the covariate is not equal to the marginal measure.[3,9]
(2) Comparability-based criterion: A covariate is a confounder if adjusting for the covariate reduces confounding.[22]

The formalized definitions of confounding and a confounder and the relation between these two criteria have been discsussed by Geng *et al.*[11,12]

5.1. *Collapsibility-based criterion*

We say that a covariate is collapsible (or ignorable) if the measure of association of interest remains unchanged after ignoring the covariate.

When a covariate is not collapsible, ignoring it will confound causal effects and give biased inference. Therefore, the covariate is called a confounder. Suppose that there exists a sufficient set of covariates such that the measure of association is constant across the strata stratified by all covariates in the set and the constant measure also equals to the marginal measure. Then, ignoring these covariates in the set does not produce confounding, and thus they are not confounders. For discussions on collapsibility, see Kleinbaum *et al.*,[19] Whittemore[42] and Geng.[9]

Consider two binary random variables A and B which take value 0 or 1, and a K-valued random variable C, which may consists of many covariates. Let $P_{ijk} = P(A = i, B = j, C = k)$ denote the probability for the cell (i, j, k) in a $2 \times 2 \times K$ contingency table. For the stratum $C = k$, the relative risk is defined as:

$$RR_k = \frac{P(A = 1|B = 1, C = k)}{P(A = 1|B = 0, C = k)} = \frac{p_{1|1k}}{p_{1|0k}}.$$

The risk difference is defined as:

$$RD_k = P(A = 1|B = 1, C = k) - P(A = 1|B = 0, C = k) = p_{1|1k} - p_{1|0k}.$$

The odds ratio is defined as:

$$OR_k = \frac{P(A = 1, B = 1|C = k)P(A = 0, B = 0|C = k)}{P(A = 1, B = 0|C = k)P(A = 0, B = 1|C = k)} = \frac{p_{11|k}p_{00|k}}{p_{10|k}p_{01|k}}.$$

If a measure of association (e.g. the relative risk) is constant across the strata, for example $RR_1 = \cdots = RR_k = CRR$, then we say that the relative risk is homogenous, and CRR is called the common relative risk. The common relative difference CRD and the common odds ratio COR may be defined similarly. If the common measure of association equals the marginal measure of association in the 2×2 marginal table obtained by pooling the K strata, then the measure is said to be simply collapsible, or simply called collapsible. For example, let $OR_+ = p_{11+}p_{00+}/(p_{11+}p_{00+})$ denote the odds ratio in the 2×2 marginal table obtained by pooling the K tables, where $p_{ij+} = \sum_k p_{ijk}$. If $OR_1 = \cdots = OR_K = COR = OR_+$, then the odds ratio is collapsible. When a measure is collapsibility, the phenomenon of Simpsons paradox on that measure can be avoided and statistical inference on the measure can be done in the marginal table.

If the common measure of association equals the measure of association in the partially marginal table obtained by pooling any several tables, then the measure is said to be strongly collapsible. Let ω be a subset of $\{1, \ldots, K\}$ and pool the strata in ω to obtain a 2×2 partially marginal

table, where the probability is sum of probabilities in corresponding cells of these strata:

$$p_{ij\omega} = \sum_{k\in\omega} p_{ijk}\,.$$

In survey design and data analysis, we may consider based on collapsibility, which covariates should be included and which may be ignored, and we may determine if the data from different survey studies may be combined for analysis. Collapsibility also provides a method for categorizing covariates. Assume that the probability $p_{ijk} > 0$ for every cell (i, j, k). We shall show conditions of simple collapsibility and strong collapsibility for measures of associations.

5.1.1. *Conditions of collapsibility for relative risk*

Geng[9] discussed the collapsibility of relative risks. If relative risks in all the strata are the same (i.e. $RR_1 = \cdots = RR_K$), we say the relative risks are homogenous. The marginal relative risk obtained by pooling all the strata of C is defined as:

$$RR_+ = \frac{P(A=1|B=1)}{P(A=1|B=0)} = \frac{p_{1|1+}}{p_{1|0+}}\,.$$

If the relative risk is homogenous across the strata, and also equals the marginal relative risk (i.e. $RR_1 = \cdots = RR_K = RR_+$), then we say that the relative risks are simply collapsible. Let $\omega \subseteq \{1, \ldots, K\}$ be a subset of levels of C. We define the relative risk of the partially marginal table obtained by pooling all the strata in ω as:

$$RR_\omega = \frac{P(A=1|B=1, C\in\omega)}{P(A=1|B=0, C\in\omega)} = \frac{p_{1|1\omega}}{p_{1|0\omega}}\,.$$

If the relative risk keeps constant for any partially marginal table obtained by pooling several arbitrary strata (i.e. $RR_\omega = RR_+$ for any set $\omega \subseteq \{1, \ldots, K\}$), then we say that the relative risks are strongly collapsible.

Theorem 5.1. *The relative risks are simply collapsible if one of the following conditions holds*:

(1) *A is conditionally independent of C given B (written as $A \perp C|B$);*
(2) *B is marginally independent of C (written as $B \perp C$), and the relative risks are homogenous (i.e. $RR_1 = \cdots = RR_K$); and*
(3) *B is marginally independent of C (written as $B \perp C$), and B is conditionally independent of C given $A = 1$ (written as $B \perp C|A = 1$).*

It can be proven that the above condition (2) is equivalent to (3).

Theorem 5.2. *A necessary and sufficient condition for the relative risks to be strongly collapsible (i.e. $RR_\omega = RR_+$ for any $\omega \subseteq \{1, \ldots, K\}$) is*:

(1) *A is conditionally independent of C given B (written as $A \perp C|B$); or*
(2) *B is marginally independent of C (written as $B \perp C$), and the relative risks are homogenous (i.e. $RR_1 = \cdots = RR_K$).*

We may explain the theorem as follows: If one of the conditions (1) and (2) holds, then the relative risks are strongly collapsible; on the other hand, if the relative risks are strongly collapsible, then one of the conditions (1) and (2) must hold.

5.1.2. *Conditions of collapsibility for risk differences*

Similar to definitions of collapsibility for relative risks, we may define the homogenous, simple collapsibility and strong collapsibility for risk differences. The risk differences are homogenous if $RD_1 = \cdots = RD_K$. The risk difference in the marginal table obtained by pooling all the strata of C is defined as:

$$RD_+ = P(A = 1|B = 1) - P(A = 1|B = 0) = p_{1|1+} - p_{1|0+}\,.$$

We say that the risk differences are simply collapsible if $RD_1 = \cdots = RD_K = RD_+$. Let $\omega \subseteq \{1, \ldots, K\}$. We define the risk difference of the partially marginal table obtained by pooling all the strata in as:

$$\begin{aligned} RD_\omega &= P(A = 1|B = 1, C \in \omega) - P(A = 1|B = 0, C \in \omega) \\ &= p_{1|1\omega} - p_{1|0\omega}\,. \end{aligned}$$

If $RD_\omega = RD_+$ for any set $\omega \subseteq \{1, \ldots, K\}$, then the risk differences are strongly collapsible.

Theorem 5.3. *The risk differences are simply collapsible if one of the following conditions holds*:

(1) *A is conditionally independent of C given B (written as $A \perp C|B$);*
(2) *B is marginally independent of C (written as $B \perp C$), and the risk differences are homogenous (i.e. $RD_1 = \cdots = RD_K$).*

Theorem 5.4. *A necessary and sufficient condition for the risk differences to be strongly collapsible (i.e. $RD_\omega = RD_+$ for any set $\omega \subseteq \{1, \ldots, K\}$) is*:

(1) *A is conditionally independent of C given B (written as* $A \perp C|B$*), or*
(2) *B is marginally independent of C (written as* $B \perp C$*), and the risk differences are homogenous (i.e.* $RD_1 = \cdots = RD_K$*).*

5.1.3. *Conditions of collapsibility for odds ratios*

The odds ratios are homogenous if $OR_1 = \cdots = OR_K$. The odds ratio of the marginal table obtained by pooling all the strata in C is defined as:

$$OR_+ = \frac{P(A=1,B=1)P(A=0,B=0)}{P(A=1,B=0)P(A=0,B=1)} = \frac{p_{11+}p_{00+}}{p_{10+}p_{01+}}.$$

We say that the odds ratios are simply collapsible if $OR_1 = \cdots = OR_K = OR_+$. Let $\omega \subseteq \{1,\ldots,K\}$. We define the odds ratio of the partially marginal table obtained by pooling all the strata in ω as:

$$OR_\omega = \frac{P(A=1,B=1|\omega)P(A=0,B=0|\omega)}{P(A=1,B=0|\omega)P(A=0,B=1|\omega)} = \frac{p_{11|\omega}p_{00|\omega}}{p_{10|\omega}p_{01|\omega}}.$$

If $OR_\omega = OR_+$ for any set $\omega \subseteq \{1,\ldots K\}$, then the odds ratios are strongly collapsible.

Theorem 5.5. *The odds ratios are simply collapsible if one of the following conditions holds*:

(1) *A is conditionally independent of C given B (written as* $A \perp C|B$*);*
(2) *B is conditionally independent of C given A (written as* $B \perp C|A$*).*

Theorem 5.6. *A necessary and sufficient condition for the odds ratios to be strongly collapsible (i.e.* $OR_\omega = OR_+$ *for any set* $\omega \subseteq \{1,\ldots,K\}$*) is*:

(1) *A is conditionally independent of C given B (written as* $A \perp C$*); or*
(2) *B is conditionally independent of C given A (written as* $B \perp C|A$*).*

If C contains sufficient covariates such that relative risks, risk differences or odds ratios are measures of causal effects in each stratum $C = k$, then ignoring C will not affect the value of the association measure and C is not a confounder if the measure of association is collapsible over C. We usually say that covariate C is a confounder when a measure of association is not collapsible over C. Note that we may probably get different conclusions when we use different measures of association. For example, risk differences are collapsible while relative risks may be not. Therefore, whether or not a covariate is a confounder may depend on what measure of association is used.

5.2. *Comparability-based criterion*

If the distribution of response (e.g. the probability of diseased) in the unexposed population (e.g. the population of non-smokers) would be the same as that in the exposed population (e.g. the population of smokers) had all individuals in the exposed population not been exposed (i.e. non-smoking), and if the distribution of response in the exposed population would be the same as that in the unexposed population had all individuals been exposed, then the exposed population is said to be exchangeable with the unexposed population. In this case, the difference between the response distribution in the exposed population and that in the unexposed population equals the population-averaged causal effect. Therefore, we can estimate the population-average causal effect with observed data from the exposed and unexposed populations. If the exposed population is not exchangeable with the unexposed population, but the exposed subpopulation is exchangeable with the unexposed subpopulation defined by a covariate, then the covariate is called as a confounder, i.e. confounding can be eliminated by stratifying the population with this covariate. In this case, the subpopulation-average causal effects may be identified and the population-average causal effect can be obtained by adjusting the subpopulation-average causal effects with the distribution of the confounder as weight. The comparability-based criterion for confounders was presented by Miettinen and Cook[22] (written as M and C hereafter), and it was discussed further by many authors.[11,12,15,43]

It has been discussed above that the collapsibility-based criterion for confounders depends on the selected measure of association. It is possible that some measures are collapsible while others are not, and thus assessment of a confounder depends on which measure is selected. The collapsibility-based criterion is widely used in practical data analysis because this criterion could be tested simply. However, some epidemiologists think that the basic criteria for confounders should not depend on selection of measures. M and C[22] used many examples to induce the following comparability-based criteria for confounders, which does not depend on selection of measures. A confounder C must satisfy the following two conditions:

(1) C must be predictive of risk in the unexposed population, and
(2) C must have different distributions between the exposed and unexposed populations.

Now, we use the counterfactual model to describe the comparability-based criterion for detecting confounders. Epidemiological studies usually

focus on causal effects in the exposed population rather than those in the whole population. The average causal effect in the exposed population is defined as:

$$\begin{aligned} ACE_e &= E[D_e(u) - D_{\bar{e}}(u)|E = e] \\ &= P(D_e = 1|E = e) - P(D_{\bar{e}} = 1|E = e)\,, \end{aligned}$$

where $P(D_{\bar{e}} = 1|E = e)$ denotes the hypothetical probability of disease if the individuals in the exposed population had not been exposed. In epidemiological studies, one usually selects an unexposed reference population of $E = \bar{e}$ and uses the probability of disease $P(D_{\bar{e}} = 1|E = \bar{e})$ in the unexposed population to estimate the hypothetical probability $P(D_{\bar{e}} = 1|E = e)$. Therefore, if the probability $P(D_{\bar{e}} = 1|E = \bar{e})$ in the unexposed population equals the hypothetical probability $P(D_{\bar{e}} = 1|E = e)$, then there does not exist confounding, and we say that the exposed population is comparable with the unexposed population.

Rosenbaum and Rubin[34] proposed two important assumptions needed for causal inference in observational studies, one of which is strong ignorability: $(D_e, D_{\bar{e}}) \perp E|C$ and the other is week ignorability: $D_e \perp E|C$ and $D_{\bar{e}} \perp E|C$. Wickramaratne and Holford[43] gave a weaker assumption for causal inference in epidemiological studies: $D_{\bar{e}} \perp E|C$, i.e. there is no confounding in any subpopulation: for every k,

$$P(D_{\bar{e}} = 1|E = e, C = k) = P(D_{\bar{e}} = 1|E = \bar{e}, C = k)\,.$$

All of these assumptions are untestable with observed data. Wickramaratne and Holford[43] proved the comparability-based criteria of M and C under their assumption, see the following theorem.

Theorem 5.7. *Assume that $D_{\bar{e}} \perp E|C$. If one of the following conditions holds*:

(1) $D_{\bar{e}} \perp C|E = \bar{e}$, *or*
(2) $E \perp C$,

then the exposed population is comparable with the unexposed population, i.e.

$$P(D_{\bar{e}} = 1|E = e) = P(D_{\bar{e}} = 1|E = \bar{e})\,.$$

Condition (1) means that the covariate C is not predictive to $D_{\bar{e}}$ in the unexposed population, and condition (2) means that the covariate C has the same distribution in both exposed and unexposed populations. This just

verifies M and Cs criteria. This theorem for judging confounding depends on categorization of a covariate C. For example, one may get different conclusions when a covariate, say age C, is categorized by every 10 years and every 20 years. Geng *et al.*[12] proposed a concept of uniform non-confounding. If for any set $\omega \subseteq \{1, \ldots, K\}$,

$$P(D_{\bar{e}} = 1|E = e, C \in \omega) = P(D_{\bar{e}} = 1|E = \bar{e}, C \in \omega)\,.$$

i.e. there is no confounding in any coarse subpopulation, then we call it uniform non-confounding. This means that no confounding occurs no matter how to categorize the covariate C. For example, when C denotes age, the exposed population always is comparable with the unexposed population no matter how to stratify the population by age, every 10 years or every 20 years.

Theorem 5.8. *Assume that $D_{\bar{e}} \perp E|C$. A necessary and sufficient condition for uniform non-confounding is*:

(1) $D_{\bar{e}} \perp C|E = \bar{e}$, or
(2) $E \perp C$.

This theorem shows that M and Cs criteria is a necessary and sufficient condition for uniform non-confounding. But both Theorems 5.7 and 5.8 need the untestable assumption of ignorability $D_{\bar{e}} \perp E|C$.

5.3. *Formal definitions and criteria for confounders*

M and C[22] derived inductively the criteria for confounders without the assumption of ignorability. Greenland and Robins[15] exemplified that the comparability-based criteria of M and C is not a sufficient condition for a confounder, but only a necessary condition. Thus, M and Cs criteria cannot be used as a definition of confounders. The collapsibility-based criteria neither are sufficient for confounders, which also depend on which measure of association is used and how to categorize a covariate. The common qualitative definition of a confounder is that it is a risk covariate for disease, controlling for which can reduce bias for estimating causal effects.[15,19] In this section, we discuss formal definitions of confounders without the assumption of ignorability.[11] Let C be a discrete covariate, which is not affected by the exposure. The following common standarization in epidemiology is used to estimate the hypothetical probability $P_\Delta(D_{\bar{e}} = 1|E = e)$:

$$P_\Delta(D_{\bar{e}} = 1|E = e) = \sum_{k=1}^{K} P(D_{\bar{e}} = 1|E = \bar{e}, C = k)P(C = k|E = e)\,.$$

Definition 5.1. C is said to be a confounder[11] if

$$|P(D_{\bar{e}} = 1|E = e) - P_{\Delta}(D_{\bar{e}} = 1|E = e)| < |P(D_{\bar{e}} = 1|E = e) - P(D_{\bar{e}} = 1|E = \bar{e})| .$$

The above definition of a confounder means that the standardized probability $P_{\Delta}(D_{\bar{e}} = 1|E = e)$ obtained by adjusting for the confounder C is closer to the hypothetical probability $P(D_{\bar{e}} = 1|E = e)$ than is the crude probability $P(D_{\bar{e}} = 1|E = \bar{e})$. Thus, we can adjust for the confounder C to reduce confounding bias. The inequality in Definition 5.1, however, is untestable with observed data since the hypothetical probability usually is unknown. Thus, we need to introduce the concept of an irrelevant factor for empirically detecting a confounder.

Definition 5.2. C is said to be an irrelevant factor if

$$P_{\Delta}(D_{\bar{e}} = 1|E = e) = P(D_{\bar{e}} = 1|E = \bar{e}) .$$

From the definition, we can see that an irrelevant factor is not a confounder. Adjustment for an irrelevant factor cannot reduce confounding bias. According to the above definition of a confounder, however, it is possible that C_1 is a confounder, C_2 is a confounder but $\{C_1, C_2\}$ as a composite covariate is not a confounder. To avoid this counter-intuitive property, we present the concept of an occasional confounder. The above definitions also depend on the choice of categorization for the covariate C under consideration. For example, age may be a confounder or irrelevant factor when it is categorized by every 10 years of age, but it may not be a confounder when categorized by every 20 years. We consider that the definition of a confounder should not depend on the categorization of a covariate and we give the following definition of an occasional confounder.

Definition 5.3. If there exists a partition p of the categories of C (i.e. $p = \{\omega_1, \ldots, \omega_M\}$, where $\omega_k \neq \phi$, $\omega_i \cap \omega_j = \phi$, $\cup_k \omega_k = \{1, \ldots, K\}$, $M \leq K$) such that

$$P(D_{\bar{e}} = 1|E = e) - P_p(D_{\bar{e}} = 1|E = e)| < |P(D_{\bar{e}} = 1|E = e) - P(D_{\bar{e}} = 1|E = \bar{e})| ,$$

then C is said to be an occasional confounder, where $P_p(D_{\bar{e}} = 1|E = e)$ is the standardized probability based on the partition p:

$$P_p(D_{\bar{e}} = 1|E = e) = \sum_{m=1}^{M} [P(D_{\bar{e}} = 1|E = \bar{e}, C \in \omega_m)P(C \in \omega_m|E = e)] .$$

Similarly, the inequality is also untestable, and the concept of a uniform irrelevant factor is introduced below.

Definition 5.4. C is said to be a uniform irrelevant factor if, for any possible partition p,

$$P_p(D_{\bar{e}} = 1|E = e) = P(D_{\bar{e}} = 1|E = \bar{e}) .$$

From definitions, we can see that any confounder is also an occasional confounder, but the reverse is not true. Definition 5.3 implies that C is an occasional confounder if there exists a partition of C such that adjustment for C with respect to the partition p can reduce confounding bias, but such a partition cannot be recognized from observed data. If C is not an occasional confounder, then confounding bias cannot be reduced by controlling for C, no matter what categorization is chosen for C, or no matter how the subpopulations are pooled together. If C is an occasional confounder, then any covariate set containing C must also be an occasional confounder. Definition 5.4 implies that C is a uniform irrelevant factor if there does not exist any partition of C such that adjustment for it can reduce confounding bias.

It can be shown from the formal definition of a confounder that the collapsibility-based and comparability-based criteria for assessing a confounder are not contradictory, but mutually complementary. Combination of these two criteria may eliminate more non-confounders from the set of potential confounders.

Three necessary conditions for a covariate C to be a confounder are as follows:

(A1) $C \not\perp E$ and $D_{\bar{e}} \not\perp C|E = \bar{e}$;
(A2) the risk difference is not collapsible over C; and
(A3) the relative risk is not collapsible over C.

The condition (A1) is the comparability-based criteria of M and C, and (A2) and (A3) are the collapsibility-based criteria. If C is a confounder, then it must satisfy all of these three conditions (A1), (A2) and (A3).

Three necessary conditions for a covariate C to be an occasional confounder are:

(B1) $C \not\perp E$ and $D_{\bar{e}} \not\perp C|E = \bar{e}$;
(B2) the risk difference is not strongly collapsible over C; and
(B3) the relative risk is not strongly collapsible over C.

If C is an occasional confounder, then it must satisfy all of the three conditions (B1), (B2) and (B3). It can be shown that condition (B1) implies both conditions (B2) and (B3). Therefore, condition (B1) is the essential necessary one for an occasional confounder. Condition (B1) is just the comparability-based criterion of M and C, while (B2) and (B3) are the strong collapsibility criteria, but not the simple collapsibility criteria.

5.4. *Numerical examples of confounders*

As mentioned above, the comparability-based criteria of M and C is only a necessary condition and cannot be used as a definition, and Greenland and Robins[15] illustrated this. The collapsibility-based criteria are not sufficient either. In the section, we use the individual effect model presented by Greenland and Robins[15] to illustrate necessity of these conditions and why the criteria cannot define confounders.

Suppose that the response of every individual is determinate under any exposure status. Then all individuals in the population may be classified into the following four types:

Type 1. No effect (individual "doomed"): $D_e = D_{\bar{e}} = 1$.

Type 2. Exposure causative (individual susceptible): $D_e = 1, D_{\bar{e}} = 0$.

Type 3. Exposure preventive (individual susceptible): $D_e = 0, D_{\bar{e}} = 1$.

Type 4. No effect (individual immune to disease): $D_e = D_{\bar{e}} = 0$.

Example 5.1. Suppose there is no exposure effect, i.e. there are only individuals of Types 1 and 4, and that the joint distribution of response, exposure and a covariate is given in Table 9. We cannot know if an individual is Types 1, 2, 3 or 4, but only know if he developed the disease. It can be seen from Table 9 that neither RD nor RR are collapsible, and thus we cannot assessing from the collapsibility-based criteria if C is a confounder. On the other hand, it follows from Table 9 that $C \perp E$, and thus we can judge from the comparability-based criteria that C is not a confounder. Without stratifying and adjusting for C, we can obtain directly from the crude marginal table that and $RD = 0$ and $RR = 1$, which correctly evaluate no causal effect.

Example 5.2. Suppose that exposure has no causal effect, i.e. there are only Types 1 and 4 individuals, and that the joint distribution is shown in Table 10. It can be seen from Table 10 that both RD and RR are collapsible,

Table 9. Hypothetical joint distribution: Non-confounder C that is not collapsible.

Type	$C=1$		$C=2$		Crude	$(C \in \{1,2\})$
	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$
1("doomed")	15	5	40	50	55	55
4("immune")	85	95	60	50	145	145
Total	100	100	100	100	200	200
RD	0.10		−0.10		0.00	
RR	3.00		0.80		1.00	

Table 10. Hypothetical joint distribution: Non-confounder C that is collapsible.

Type	$C=1$		$C=2$		$C=3$		Crude	$(C \in \{1,2,3\})$
	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$
1("doomed")	95	95	5	5	75	25	175	125
4("immune")	5	5	95	95	75	25	175	125
Total	100	100	100	100	150	50	350	250
RD	0.00		0.00		0.00		0.00	
RR	1.00		1.00		1.00		1.00	

and thus we can judge from the collapsibility-based criteria that C is not a confounder. On the other hand, it follows from Table 10 that $C \not\perp E$ and $D_{\bar{e}} \not\perp C|E=\bar{e}$, and thus we cannot decide from the comparability-based criteria if C is a confounder. Without stratifying and adjustment for C, we can obtain directly from the crude marginal table that $RD=0$ and $RR=1$, which correctly assess no causal effect.

Example 5.3. Suppose exposure has no causal effect, i.e. there are only Types 1 and 4 individuals, and that the joint distribution is given in Table 11. It can be seen from Table 11 that RD is collapsible, but RR is not. Therefore, we can judge from the collapsibility-based criteria that C is not a confounder. On the other hand, it follows from Table 11 that $C \not\perp E$ and $D_{\bar{e}} \not\perp C|E=\bar{e}$, and thus we cannot decide from the comparability-based criteria if C is a confounder. Stratifying and adjusting for C, we can obtain that

$$P_{\Delta}(D_{\bar{e}}=1|E=e)=\frac{66}{100}\cdot\frac{100}{700}+\frac{184}{400}\cdot\frac{100}{700}+\frac{46}{100}\cdot\frac{400}{700}+\frac{26}{100}\cdot\frac{100}{700}=0.46\,,$$

which equals $P_{\Delta}(D_{\bar{e}}=1|E=\bar{e})=\frac{322}{700}=0.46$ without stratifying or adjusting for C. It shows that adjustment for C neither decreases nor increases confounding biases, and C is an irrelevant factor.

Table 11. Hypothetical joint distribution: Non-confounder C that is simply collapsible.

Type	$C=1$		$C=2$		$C=3$		$C=4$		Crude ($C \in \Delta$)	
	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$
1("doomed")	60	66	40	184	160	46	20	26	280	322
1("immune")	40	34	60	216	240	54	80	74	420	378
Total	100	100	100	400	400	100	100	100	700	700
RD	-0.06		-0.06		-0.06		-0.06		-0.06	
RR	0.91		0.87		0.87		0.77		0.87	

Table 12. $p = \{[1,2],[3,4]\}$, an occasional confounder C.

Type	$C \in [1,2]$		$C \in [3,4]$	
	$E=e$	$E=\bar{e}$	$E=e$	$E=\bar{e}$
1("doomed")	100	250	180	72
4("immune")	100	250	320	128
Total	200	500	500	200
RD	0.00		0.00	
RR	1.00		1.00	

Let $p = \{\{1,2\},\{3,4\}\}$. Table 12 represents the distribution of the coarse subpopulations obtained by pooling levels 1 and 2 and pooling levels 3 and 4 based on the partition p. It can be seen that the risk difference RD is not strongly collapsible. The hypothetical probability $P(D_{\bar{e}} = 1|E = e)$ is usually unknown, so we cannot decide if C is an occasional confounder. Under the assumption of no causal effect, we have

$$P(D_{\bar{e}} = 1|E = e) = P(D_{\bar{e}} = 1|E = e) = \frac{280}{700} = 0.40\,.$$

We can compute the standardized probability $P_p(D_{\bar{e}} = 1|E = e)$ based on the partition p as follows:

$$P_p(D_{\bar{e}} = 1|E = e) = \frac{250}{500} \cdot \frac{72}{200} + \frac{72}{200} \cdot \frac{500}{700} = \frac{280}{700} = 0.40\,.$$

Since $|P(D_{\bar{e}} = 1|E = e) - P_p(D_{\bar{e}} = 1|E = e)| = 0$ is less than $|P(D_{\bar{e}} = 1|E = e) - P(D_{\bar{e}} = 1|E = \bar{e})| = 0.06$, C is an occasional confounder. Confounding bias can be completely eliminated by controlling for C with respect to the partition p. Furthermore, we can obtain the adjusted risk difference $|P(D_e = 1|E = e) - P_p(D_{\bar{e}} = 1|E = e)| = 0$ and the adjusted relative risk $|P(D_e = 1|E = e)/P_p(D_{\bar{e}} = 1|E = e)| = 1$, which correctly assess no causal effect. However, this example by no means suggests that

we should try to merge the levels of a covariate to correct confounding since it is impossible in practice to determinate an occasional confounder based on observed data.

6. Causal Diagrams

Graphical models have been widely applied for large complex systems.[4,6,20,42] Sprites, Glymour and Scheines[38] and Pearl[25–38] proposed causal diagrams for causal inferences for observational studies. They supposed first a completely constructed causal diagram. Pearl[26] showed several of sufficient conditions for no confounding, and he gave several rules for identifying causal effects. Greenland *et al.* applied causal diagrams to epidemiological research and provided a criterion for detecting multiple confounders, which is more efficient than the traditional criterion. In many practical cases, however, it is difficult to construct such a complete causal diagram. Geng and Li[10] discussed conditions of non-confounding without a completely constructed causal diagram.

6.1. *Definitions and notations*

Let Γ denote a directed acyclic graph, $X = \{X_1, \ldots, X_n\}$ be a set of nodes, where a node represent a discrete random variable. X_j is called a parent node of X_i, or X_i is a son of X_j if there is a directed arrow from X_j to X_i. Let pa_i denote the set of parents of X_i. In a diagram Γ, a path between X_i and X_j is a succession of arcs from X_i to X_j, regardless of their directions. All nodes with an arrow pointing path X_i from are called to be descendants of X_i.

Definition 6.1. Let C be a set of nodes in a diagram Γ. A path between X_i and X_j is said to be connected if every node on the path satisfies the following two conditions:

(1) if it has converging arrows along the path, then either it or one of its descendants is in C; and
(2) if it does not have converging arrows along the path, then it is not in C.

Otherwise, the path is said to be blocked by C.

Definition 6.2. Let A, B and C be three disjoint subsets of nodes in a diagram Γ. C is said to d-separate A from B, denoted $d(A, B, C)$, if and only if C blocks every path from any node in A to any node in B.

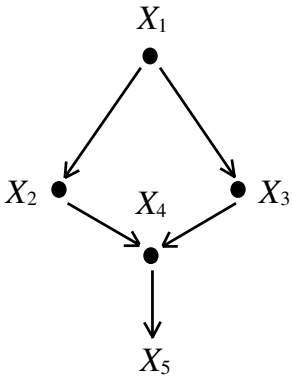


Fig. 1. A directed acyclic graph.

There is a one-to-one correspondence between d-separation $d(A, B, C)$ and conditional independency $A \perp B|C$ under the stability condition of distribution.[25] Thus, conditional independence between variables can be read off from a diagram Γ by using the d-separation criterion.

Example 6.1. In Fig. 1, $C = \{X_1\}$ blocks the path $X_2 \leftarrow X_1 \rightarrow X_3$; neither X_4 nor X_5 are in C, and thus C blocks the path $X_2 \rightarrow X_4 \leftarrow X_3$ too. Therefore, $C = \{X_1\}$ d-separates $A = \{X_2\}$ from $B = \{X_3\}$, and thus $X_2 \perp X_3|X_1$.

If we let $C' = \{X_1, X_4\}$, then X_4 is in C' and has converging arrows along the path $X_2 \rightarrow X_4 \leftarrow X_3$. So C' does not d-separate $A = \{X_2\}$ from $B = \{X_3\}$, and $X_2 \not\perp X_3|(X_1, X_4)$.

A graph model defined by Γ represents that $X_1, \ldots, X_n$ have a joint probability distribution as follows:

$$P(x_1, \ldots x_n) = \prod_{i=1}^{n} P(x_i|pa_i)\,,$$

where $P(x_i|pa_i)$ stands for the conditional probability of $X_i = x_i$ given $Pa_i = pa_i$. We assume that $P(x_1, \ldots, x_n) > 0$ for every $x_1, \ldots, x_n$.

Example 6.2. The directed graph in Fig. 1 describes the following joint probability distribution:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4)\,.$$

Assume that a causal system can be represented by a directed acyclic graph, where an arrow means the direction from cause to effect. The causal

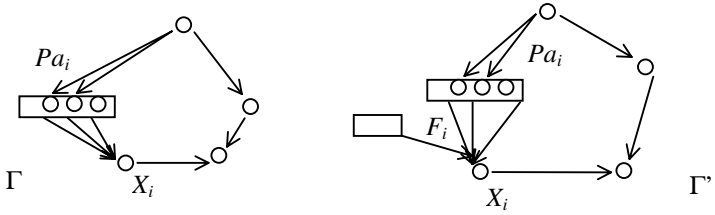


Fig. 2. External intervention F_i.

effect of X_i on the whole system can be evaluated by an external intervention to X_i. Consider Fig. 2. An arrow $F_i \to X_i$ in Γ represents the external intervention to X_i. F_i is a new variable with value, $set(x_i)$ or $idle$, where $set(x_i)$ stands for that F_i forces X_i to take on some value x_i and $idle$ stands for no external intervention influences to X_i. Γ' represents the diagram after the external intervention.

In the diagram Γ' after the external intervention, the set of parent nodes is augmented to $Pa'_i = Pa_i \cup \{f_i\}$. The conditional distribution of X_i given Pa'_i is:

$$P'(x_i|pa_i) = \begin{cases} p(x_i|pa_i)\,, & \text{if } F_i = idle\,; \\ 0\,, & \text{if } F_i = set(x'_i) \text{ and } x_i \neq x'_i\,; \\ 1\,, & \text{if } F_i = set(x'_i) \text{ and } x_i = x'_i\,. \end{cases}$$

The joint distribution after the intervention $set(x'_i)$ is given by

$$P_{x'_i}(x_i, \ldots, x_n) = P'(x_1, \ldots, x_n | F_i = set(x'_i))\,.$$

$P_{x'_i}(x_i, \ldots, x_n)$ represents the joint distribution of $X_1, \ldots, X_n$ if X_i of all units in the target population had been forced to x'_i.

The transformation formula between the pre-intervention and the post-intervention distributions is given by

$$P_{x'_i}(x_i, \ldots, x_n) = \begin{cases} \dfrac{P_{x'_i}(x_i, \ldots, x_n)}{P(x_i|pa_i)}\,, & \text{if } x_i = x'_i\,; \\ 0\,, & \text{if } x_i \neq x'_i\,. \end{cases}$$

Let X_j and X_i denote the response and treatment variables respectively, and Ω_j and Ω_i their corresponding domains. The causal effect of X_i on X_j is defined as $P_{x'_i}(x_j)$, which means the post-intervention distribution of X_j under the intervention to X_i.

Given a causal diagram, we say that there is no confounding for the causal effect of X_i on X_j if $P_{x'_i}(x_j)$, the post-intervention distribution of

X_j under the intervention $X_i = x_i$ in the whole target population, is equal to $P(x_j|x_i')$, the distribution of X_j in the subpopulation $X_i = x_i'$ without external intervention. The formal definition is given below.

Definition 6.3. No confounding. There is no confounding for the causal effect of X_i on X_j if for every $x_j \in \Omega_j$ and $x_i' \in \Omega_i$,

$$P_{x_i'}(x_j) = P(x_j|x_i') \,.$$

No confounding implies that the causal effect $P_{x_i'}(x_j)$ can be evaluated with the observed association measure $P(x_j|x_i')$. Therefore, we may define the confounding bias as the difference between $P_{x_i'}(x_j)$ and $P(x_j|x_i')$. If there is confounding in the population, but there is no confounding in all the subpopulations stratified by some other variables, we call these variables as confounders. Let $C = \{X_{t_1}, \ldots, X_{t_k}\}$ be a set of variables, Ω_{t_k} be the domain of X_{t_k}, and $P_{x_i'}(x_j|x_{t_1}, \ldots, x_{t_k}) = \frac{P_{x_i'}(x_j, x_{t_1}, \ldots, x_{t_k})}{P_{x_i'}(x_{t_1}, \ldots, x_{t_k})}$ denotes the conditional distribution under the external intervention $X_i = x_i$. It means that the condition is taken after the intervention.

Given a causal diagram together with the corresponding joint distribution, we can infer the post-intervention distribution from the transformation formula (1). Thus, we can estimate the causal effects of interventions. The joint distribution, however, is usually unknown when some variables are unobservable. Let $X = X_O \cup X_U$, where X_O denotes the set of observed variables and X_U denotes the unobserved variables. The basic problem of causal inference is to identify the causal effect of X_i on X_j, $P_{x_i'}(x_j)$, from the distribution of observed variables, $P(X_O)$. The causal effect of X_i on X_j is said to be identifiable if $P_{x_i'}(x_j)$ can be represented by $P(X_O)$. Pearl[26] proposed a set of inference rules for identifying causal effects. No confounding implies that the causal effect $P_{x_i'}(x_j)$ is identifiable.

The following properties can be obtained from the transformation defined in formula (1):

(1) An intervention $set(x_i)$ can affect only the descendants of X_i in Γ.
(2) $P_{x_i'}(S|pa_i) = P(S|x_i', pa_i)$ holds for any set S of variables.
(3) $X_j \perp pa_i|X_i$ is a sufficient condition for no confounding.

Property (1) implies that $P_{x_i'}(S) = P(S)$ if S is not the descendants of X_i. Property (2) implies no confounding, which means that the causal effect of X_i on any set S of variables is equal to the conditional distribution when the parents of X_i are given. Spiegelhatter *et al.*[37] gave an example to show that the conditional independency in property (3) is not necessary for

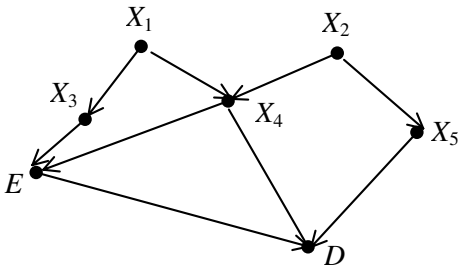


Fig. 3. An example of the back-door criterion.

no confounding. Geng and Li[10] proved that the conditional independency (3) is a necessary and sufficient condition for uniform non-confounding.

6.2. *The back-door and front-door criteria*

In this section, we shall introduce two criteria, the back-door criterion and the front-door criterion, proposed by Pearl.[26]

Definition 6.4. Back-door path. A path between X_i and X_j with an arrow into X_i is said to be a back-door path between X_i and X_j.

Definition 6.5. Back-door criterion. A set S of variables is said to satisfy the back-door criterion if: (i) No node in S is a descendant of X_i, and (ii) S blocks every back-door path between X_i and X_j.

Example 6.3. Consider the causal effect of E on D in Fig. 3. The sets $S_1 = \{X_3, X_4\}$ and $S_2 = \{X_4, X_5\}$ meet the back-door criterion, but $S_3 = \{X_4\}$ does not meet it because S_3 does not block the back-door path between E and $D(E \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \leftarrow X_2 \rightarrow X_5 \rightarrow D)$.

Theorem 6.1. *Two conditions of the back-door criterion are equivalent to*

$$S \perp F_i \quad \text{and} \quad X_j \perp F_i | (X_i, S)\,.$$

Theorem 6.2. *If a set S of variables satisfies the back-door criterion, we have*

$$P_{x_i'}(x_j) = \sum_s P(x_j | s, x_i') P(s)\,,$$

for every $x_j \in \omega_j$ and $x' \in \omega_i$.

It follows from Theorem 6.2 that the causal effect of X_i on X_j, $P_{x_i'}(x_j)$, can be estimated from the joint distribution of observed variables if we find a subset of variables $S \subseteq X_0$ satisfying the back-door criterion.

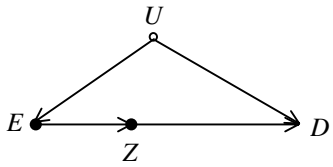


Fig. 4. A diagram for the front-door criterion.

We cannot derive the causal effect $P_{x_j'}(x_j)$ from Theorem 6.2 if the set S of observed variables does not satisfy the back-door conditions. Consider the causal diagram in Fig. 4, where U denotes an unobserved variable, and E, Z and D represent three observed variables respectively. Z is on a directed path from E to D and blocks all directed paths from E to D (there is only one directed path from E to D in Fig. 4). Following traditional epidemiological theories, we cannot control or adjust for intermediate variables on the causal path to evaluate the causal effect of E on D. Now we introduce the front-door criterion proposed by Pearl[26] through the causal diagram in Fig. 4.

It follows from Fig. 4 that the joint probability distribution of (E, D, Z, U) is given by

$$P(e, d, z, u) = P(d|z, u)P(z|e)P(e|u)P(u)\,,$$

and from the back-door criterion,

$$P_{e'}(d) = P(d|set(e')) = \sum_u P(d|e', u)P(u)\,,$$

which is not directly identifiable because U is unobserved. We can obtain two conditional independencies from Fig. 4: (1) $D \perp E|(Z, U)$ and (2) $Z \perp U|E$, and we can show

$$P_{e'}(d) = \sum_z \left[P(z|e') \sum_e P(d|e, z)P(e) \right].$$

Because the probabilities on the right side of Eq. (2) are all conditional distributions, the causal effect $P_{e'}(d)$ is identifiable when D, Z and E are observed. This result is summarized in the following theorem, which is called the front-door criterion.[26]

Theorem 6.3. *Suppose that a set Z of variables satisfies the following conditions relative to an ordered pair of variables (E, D) : (1) Z intercepts all directed paths from E to D. (2) There is no unblocked back-door path*

between E and Z. (3) *Every back-door path between Z and D is blocked by E. Then the causal effect of E on D is identifiable and is given by Eq.* (2).

Example 6.4. Pearl[26] presented an example to illustrate the front-door criterion. For the observational research of smoking and lung cancer, Fisher once proposed a conjecture: There exists an unknown carcinogenic genotype related to smoking. If this is true, the genotype is an unobservable confounder and the confounding biases could not be eliminated using traditional standardization methods of epidemiology. If we observe an intermediate variable, tar deposit, on the causal path from smoking to lung cancer, and the unobserved genotype does not have a causal path to the tar deposit, then we can obtain a causal diagram as shown in Fig. 4, where E, Z, U, and D represent, respectively, smoking, tar deposits, the carcinogenic genotype and lung cancer. It follows from Theorem 6.3 that the causal effect of smoking on lung cancer can be identified by the distribution of observed variables.

6.3. *Criterion for multiple confounders*

If there is an association between exposure E and response D after removing all causal effects of E on any factor, then the association cannot be a causal effect of E on D and thus there must be confounders. We now introduce the algorithm presented by Greenland *et al.*[14] for checking confounders based on the back-door criterion. Given a set of covariates, $S = \{S_1, \ldots, S_n\}$, which does not contain descendants of E and D, we perform the following steps:

(1) Delete all arrows emanating from E (i.e. remove all causal effects of E).
(2) For every node S_i, draw undirected edges to connect every pair of nodes that share a common child which is either in S or has a descendant in S.
(3) If the nodes in S block all paths from E to D in the new graph, then the set S is sufficient for controlling confounding bias. Otherwise, S is not sufficient.

Example 6.5. Consider the causal diagram in Fig. 5. Delete the arrow $E \to D$, and thus E has no effect on D. It follows from the diagram that there is still an association between E and D. Therefore, there exists confounding bias. When controlling for C, the unique back-door path from E to D is blocked by C, and thus the confounding is completely eliminated. So, C is a confounder.

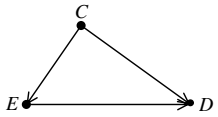


Fig. 5. C is a confounder.

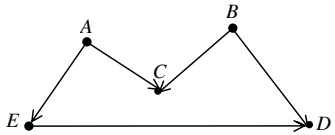


Fig. 6. C is not a confounder.

Example 6.6. Consider the causal diagram in Fig. 6. Delete the arrow $E \to D$, thus E has no effect on D. The node C blocks the unique back-door path $E \leftarrow A \to C \leftarrow B \to D$. It follows from the new diagram obtained by deleting $E \to D$ that E is independent of D and the association between E and D in the original diagram is attributed to causation. Neither A, B nor C is a confounder. However, if controlling for C, then A and B must be connected by an undirected edge in the step 2, and thus the path $E \leftarrow A - B \to D$ is not blocked. In the diagram obtained by deleting $E \to D$, E is not independent of D conditional on C, and the association between E and D is not attributed to causation, that is, confounding bias is introduced by controlling for C.

7. Causal Inference for Longitudinal Studies

Longitudinal study is one of the important methods in epidemiological and medical research. Causal inference and criteria for confounders in longitudinal studies have been discussed.[30–32] In a longitudinal study, the same unit is repeatedly observed during the follow-up period. Longitudinal observation includes treatment variables, covariates, intermediate variables and response variables of interest in every time points, and it can provide more information on causation.

In this section, we introduce notations and conceptions of causal inference in longitudinal studies proposed by Robins *et al.*[32] Let $m = 1, 2, \ldots, M$ denote the time points at which data are observed; let $Y_{m,i}$ and $A_{m,i}$ denote the response and exposure of unit i at time m, respectively. Let X_i denote a vector of time-independent covariates of unit i, and $L_{m,i}$ denote a vector

of time-independent covariates of unit i at time m. Suppose that $Y_{m,i}$ and $L_{m,i}$ occur before $A_{m,i}$ at the each time m.

Based on the counterfactual model, let $Y_{m,i}^{(0)}$ denote the response of unit i at m if the unit were not unexposed at any time during the follow-up (i.e. continuous unexposed). Similarly, let $Y_{m,i}^{(1)}$ denote the response of the unit i at m if the unit had been exposed continuously during the follow-up (i.e. continuous exposed). If a unit i changes its exposure status during the follow-up, its response at m is neither $Y_{m,i}^{(0)}$ nor $Y_{m,i}^{(1)}$. Therefore, neither $Y_{m,i}^{(0)}$ nor $Y_{m,i}^{(1)}$ can be observed for such units. Let $\bar{Z}_m = (Z_1, \ldots, Z_m)$ for $m \geq 1$ and $\bar{Z}_m = 0$ for $m \leq 0$. Assume that N series of data, $(\bar{Y}_{M,i}^{(0)}, \bar{Y}_{m,i}^{(1)}, \bar{Y}_{M,i}, \bar{A}_{M,i}, \bar{L}_{M,i}, X_i)$ for $i = 1, \ldots, N$, are independently and identically distributed. For the subpopulation $X = x$ at the time m, define expectation of responses for continuous unexposed and that for continuous exposed respectively as

$$\theta_m^{(0)}(x) = E[Y_m^{(0)}|X = x] \quad \text{and} \quad \theta_m^{(1)}(x) = E[Y_m^{(1)}|X = x]\,.$$

The average causal effect of continuous exposed versus continuous unexposed is defined as

$$\delta_m(x) = \theta_m^{(1)}(x) - \theta_m^{(0)}(x)\,.$$

The average causal effect cannot be identified without any assumptions. An assumption was proposed[32] for sufficiently controlling confounding at every time k, which is similar to Rosenbaum and Rubins[34] the strongly ignorable assumption for cross-section studies:

$$\{(Y_m^{(0)}, T_m^{(1)}); m = k+1, k+2, \ldots, M\} \perp A_k|(\bar{A}_{k-1}, \bar{Y}_k, \bar{L}_k, X) \qquad (3)$$

for all $k \geq 1$. This assumption means that the strongly ignorability holds for all k, and it cannot be tested empirically.

If there are no time-dependent confounders, that is, the set $\bar{L}_k$ in model (3) is empty, then conditional probabilities of potential responses can be identified as following:

$$\begin{aligned} &P(Y_m^{(j)} = 1|\bar{Y}_{m=1}^{(j)} = \bar{y}_{m-1}, X = x) \\ &\qquad = P(Y_m = 1|\bar{Y}_{m-1} = \bar{y}_{m-1}, \bar{A}_{m-1} = j^{[m-1]}, X = x)\,, \end{aligned}$$

where $j^{[m-1]}$ is a vector with $m-1$ elements all of which are equal to j (i.e. $j^{[m-1]} = (j, \ldots, j)$). Substituting the expression into the following g-algorithm formula presented by Robins,[28,29] the causal parameter $\theta_m^{(j)}(x)$

can be identified by:

$$\begin{aligned}\theta_m^{(j)}(x) &= P(Y_m^{(j)} = 1|X = x)\\ &= \sum_{\bar{y}_{m-1}} [P(Y_m^{(j)} = 1|\bar{Y}_{m=1}^{(j)} = \bar{y}_{m-1}, X = x)\\ &\quad \times \prod_{k=1}^{m-1} P(Y_k^{(j)} = y_k|\bar{Y}_{k-1}^{(j)} = \bar{y}_{k-1}, X = x)] .\end{aligned}$$

Similarly, if there are time-dependent confounders, that is, the set $\bar{L}_k$ in model (3) is not empty, then the causal parameter $\theta_m^{(j)}(x)$ can be identified by the following expression:

$$\begin{aligned}\theta_m^{(j)}(x) &= P(Y_m^{(j)} = 1|X = x)\\ &= \sum_{\bar{y}_{m-1}, \bar{l}_{m-1}} \Bigg\{ P(Y_m = 1|\bar{Y}_{m-1} = \bar{y}_{m-1}, \bar{L}_{m-1} = \bar{l}_{m-1} ,\\ &\quad \bar{A}_{m-1} = j^{[m-1]}, X = x) \prod_{k=1}^{m-1} [P(Y_k = y_k|\bar{Y}_{k-1} = \bar{y_{k-1}},\\ &\quad \bar{L}_{k-1} = \bar{l}_{k-1}, \bar{A}_{k-1} = j^{[k-1]}, X = x)\\ &\quad \times P(L_k = l_k|\bar{Y}_k = \bar{y}_k, \bar{L}_{k-1} = \bar{l}_{k-1}, \bar{A}_{k-1} = j^{[k-1]}, X = x) \Bigg\},\end{aligned}$$

where l_k and $\bar{l}_{k-1}$ consist of part elements of $\bar{l}_{m-1}$. From the causal parameter $\theta_m^{(j)}(x)$, we can identify the average causal effect $\delta_m(x)$ of continuous exposed versus continuous unexposed for the subpopulation $X = x$ at any time m.

References

1. Bernard, C. (1920). *Introduction to Research on Experiment Medicine* (Chinese version).
2. Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* **187**: 398–404.
3. Boivin, J. F. and Wacholder, S. (1985). Conditions for confounding of the risk ratio and of the odds ratio. *American Journal of Epidemiology* **121**: 152–158.
4. Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies*: *Models, Analysis and Interpretation*, Chapman and Hall, London.
5. Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association* **95**: 407–448.

6. Edwards, D. (1995). *Introduction to Graphical Modelling*, Springer, New York.
7. Freedman, D. (1999). Association to causation: Some remarks on the history of statistics. *Statistical Science* **14**: 243–258.
8. Freedman, D., Pisani, R., Purves, R. and Adhikari, A. (1991). *Statistics*, 2nd edn., W. W. Norton and Company, Inc., New York.
9. Geng, Z. (1992). Collapsibility of relative risks in contingency tables with a response variable. *Journal of the Royal Statistical Society* **B54**: 585–593.
10. Geng, Z. and Li, G. W. (2002). Conditions for non-confounding and collapsibility without knowledge of completely constructed causal diagrams. *Scandinavian Journal of Statistics* **29**: 169–181.
11. Geng, Z., Guo, J. and Fung, W. (2002). Criteria for confounders in epidemiological studies. *Journal of the Royal Staistical Society* **B64**: 3–15.
12. Geng, Z., Guo, J., Lau, T. and Fung, W. (2001). Confounding, homogeneity and collapsibility for causal effects in epidemiologic studies. *Statistica Sinica* **11**: 63–75.
13. Grace, N. D., Muench, H. and Chalmers, T. C. (1966). The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology* **50**: 684–691.
14. Greenland, S., Pearl, J. and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**: 37–48.
15. Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiologic confounding. *International Journal of Epidemiology* **15**: 413–419.
16. Greenland, S., Robins, J. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**: 29–46.
17. Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**: 945–970.
18. Holland, P. W. and Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Review* **12**: 203–231.
19. Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H. (1982). *Epidemiologic Research*: *Principles and Quantitative Methods*. Van Nostrand Reinhold, New York.
20. Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.
21. Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge.
22. Miettinen, O. S. and Cook, E. F. (1981). Confounding: Essence and detection. *American Journal of Epidemiology* **114**: 593–603.
23. Neyman, J. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society* **2**(Suppl.): 107–180.
24. Neufeld, E. (1995). Simpson's paradox in artificial intelligence and in real life. *Computational Intelligence* **11**: 1–10.
25. Pearl, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge University Press, Cambridge.
26. Pearl, J. (1995a). Causal diagrams for empirical research (with discussion). *Biometrika* **83**: 669–710.
27. Pearl, J. (1995b). From Bayesian networks to causal networks. ed. A. Gammerman, *Probabilistic Reasoning and Bayesian Belief Networks*, Henley-on-Thames, 1–31.

28. Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effects. *Mathematical Modelling* **7**: 1393–1512.
29. Robins, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease* **40**(Supp.): 139s–161s.
30. Robins, J. M. (1989). The control of confounding by intermediate variables. *Statistics in Medicine* **8**: 679–701.
31. Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane, Springer-Verlag, New York, 69–117.
32. Robins, J. M., Greenland, S. and Hu, F. C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association* **94**: 687–712.
33. Rosenbaum, P. R. (1999). Choice as an alternaive to control in observational studies. *Statistical Science* **14**: 259–304.
34. Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55.
35. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701.
36. Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society* **B13**: 238–241.
37. Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. C. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science* **8**: 219–283.
38. Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction and Search*, Springer, New York.
39. Stone, R. (1993). The assumptions on which causal inference rest. *Journal of the Royal Statistical Society* **B55**: 455–66.
40. Wagner, C. H. (1982). Simpson's paradox in real life. *American Statistics* **36**: 46–48.
41. Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.
42. Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society* **B40**: 328–340.
43. Wickramaratne, P. J. and Holford, T. R. (1987). Confounding in epidemiologic studies: The adequacy of the control groups as a measure of confounding. *Biometrics* **43**: 751-765.

About the Author

Zhi Geng is Professor of School of Mathematical Sciences, Peking University. He got his PhD (1989) from Kyushu University, Japan. He has been an ISI selected member since 1996. He is associated editors of *Computation Statistics and Data Analysis* and *Statistica Sinica*. His research interests include categorical data, incomplete data, causal inference, graphical models, uncertainty inference and probabilistic expert systems, and biostatistics. These researches were supported by NSFC, EYNSFC of NSF and TCTP of SECC.

Section 4
Advanced Statistical Theory and Methods

This page intentionally left blank

CHAPTER 22

SURVIVAL ANALYSIS

D. Y. LIN

Department of Biostatistics, University of North Carolina,
3101E McGavran-Greenberg Hall, CB#7420,
Chapel Hill, NC 27599-7420, USA
Tel: 919-843-5134; lin@bios.unc.edu

1. Introduction

The primary response variables in many medical studies pertain to the time to occurrence of a clinically important event, such as death, development or progression of a disease, or occurrence of a clinically significant morbid event such as a serious infection, stroke or major organ failure. A complexity that frequently arises in studies having time-to-event outcome measures is that a substantial fraction of the study subjects have not experienced the event of interest at the time of data analysis. These subjects who provide this incomplete information are referred to as being censored, or more precisely right-censored, since it is only known that the true time-to-event for that subject exceeds the duration of follow-up.

The complexities created by the presence of censored observations have led to the development of a special field of statistical methodology. Because the analysis of clinical trials data with time-to-death outcomes provided the original motivation for this development, the field has become known as survival analysis. This article provides an overview of the key ideas and methods in survival analysis. After introducing some basic terminology, we will discuss the estimation of the survival distribution, the comparison of two survival distributions as well as regression models. We will focus on non- and semi-parametric methods, which do not impose parametric assumptions on the survival distribution.

2. Basic Concepts

Let T denote the true time-to-event or failure time for a study subject in a medical study. Primary interest usually lies in estimation and testing regarding the distribution of T. This distribution can be characterized by the survival function $S(t) \equiv \Pr(T > t)$. Because of censoring, it is more convenient to deal with the hazard function, which is the instantaneous probability of dying at time t given that the subject is alive just prior to t. If T is continuous with density function f, then the hazard function is defined by

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \Pr(t \le T < t + \Delta t | T \ge t)/\Delta t = f(t)/S(t)\,.$$

The function $\Lambda(t) \equiv \int_0^t \lambda(u)du$ is called the cumulative hazard function for T, and it is easily shown that $S(t) = e^{-\Lambda(t)}$ for a continuous survival time T.

Let U denote the censoring time, that is, the time beyond which the study subject cannot be observed. Then (T, U) are referred to as latent data, while the observed data are denoted by (X, δ), where $X = \min(T, U)$, $\delta = I(T \le U)$, and $I(\cdot)$ is the indicator function. The study subjects having $\delta = 0$ are referred to as having censored observations.

While the distribution function $S(t)$ can be consistently estimated when data are uncensored, neither $\lambda(t)$ nor $S(t)$ is identifiable or consistently estimable if one only observes (X, δ).[18,78] Observing (X, δ) rather than T for all subjects only allows one to consistently estimate $S^{\#}(t) \equiv \exp\{-\int_0^t \lambda^{\#}(u)du\}$ for all t such that $\Pr(X > t) > 0$, where

$$\lambda^{\#}(t) = \lim_{\Delta t \downarrow 0} \Pr(t \le T < t + \Delta t | T \ge t, U \ge t)/\Delta t\,. \tag{1}$$

We refer to $\lambda^{\#}(t)$ as the *crude* hazard and $\lambda(t)$ as the *net* hazard.[14] In most survival analysis applications, a key assumption is made regarding the equality of the crude hazard (that is estimable) and the net hazard (that is of interest), i.e.

$$\lambda^{\#}(t) = \lambda(t) \text{ for all } t \text{ such that } \Pr(X > t) > 0\,. \tag{2}$$

A sufficient condition for the validity of assumption (2) is the independence of T and U.

3. Estimation of the Survival Distribution

A fundamental problem in survival analysis is the estimation of the hazard function $\lambda(t)$ and the survival function $S(t)$. Several parametric models are

available, and the maximum likelihood approach can be used for estimation of the parameters under the assumption of independent censoring. For example, assuming a constant hazard function, i.e. $\lambda(t) = \lambda$ for all $t > 0$, one obtains the exponential distribution, where the maximum likelihood estimator for λ is the number of observed events divided by the summation of duration of follow-up over all subjects. Kalbfleisch and Prentice[37] and Lawless[42] provided detailed discussion of parametric methods.

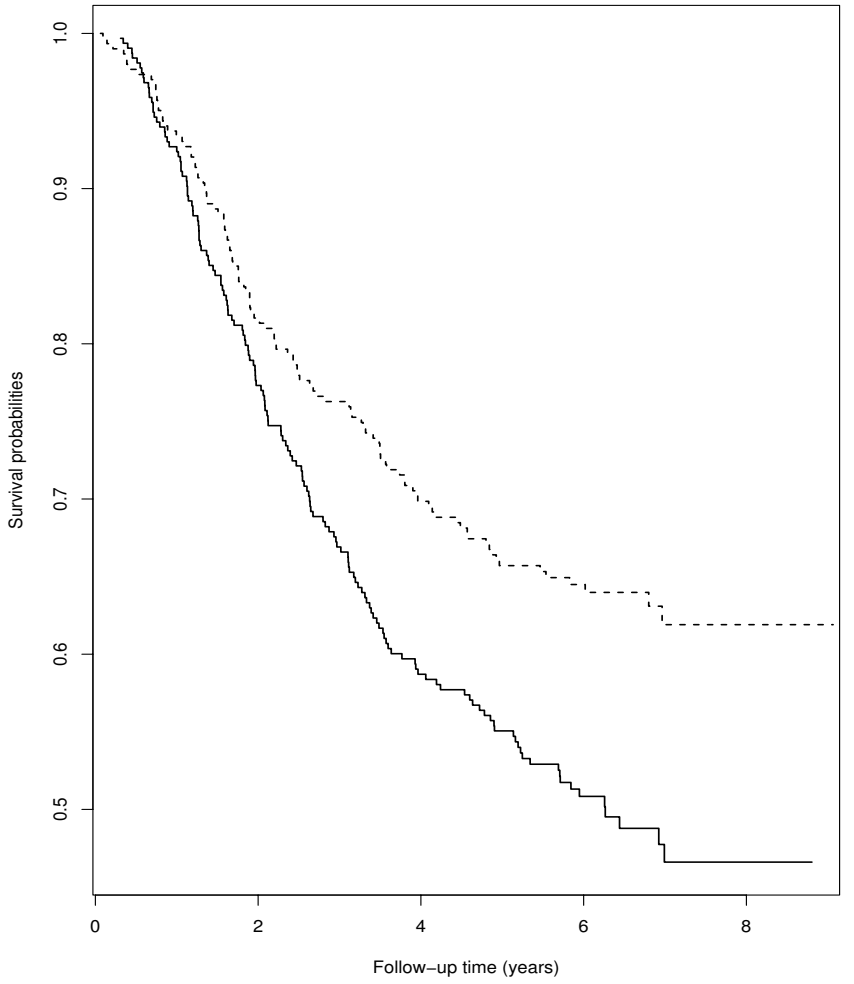


Fig. 1. Kaplan–Meier estimates for the survival probabilities in the colon cancer study: ————: observation groups, - - - - - - - - : therapy groups.

It is more desirable to estimate the failure time distribution without parametric modelling. Nelson[60] introduced a nonparametric estimator of the cumulative hazard function $\Lambda(t)$. This estimator is given by a step function, with steps occurring at times of observed events and having size D/Y, where D events occur among Y subjects at risk. Recognizing the relationship between $S(t)$ and $\Lambda(t)$ through the differential equation, $-\{dS(t)/dt\}/S(t-) = \lambda(t)$, one motivates the relationship,

$$-\{\Delta\hat{S}(t)\}/\hat{S}(t-) = \Delta\hat{\Lambda}(t)\,,$$

where one estimates $\Lambda(t)$ using Nelson's estimator and then recursively solves for the estimator of $S(t)$. The resulting estimator is that proposed by Kaplan and Meier.[38] It is a step function, with value reduced by the multiplicative factor $\{1-(D/Y)\}$ at times of observed events. The asymptotic properties of the Kaplan–Meier estimator have been studied by Breslow and Crowley,[9] Gill[30] and Ying[84] among others.

Figure 1 displays the Kaplan–Meier estimates for the survival probabilities in a randomized clinial trial, which was designed to assess whether a new therapy, levamisole plus fluorouracil, prolongs the survival time for patients with Duke's Stage C colon cancer.[47] There are 315 and 304 patients in the observation and therapy groups. By the end of the study, 155 patients in the observation group and 108 patients in the therapy group had died. The Kaplan–Meier estimates given in Fig. 1 show the average survival experiences of the patients in the two groups over the entire follow-up period. Clearly, the patients on the therapy tend to have higher survival probabilities than the patients in the observation group.

4. Counting Process Theory

It is difficult to study the properties of the statistics used in survival analysis, such as the Nelson and Kaplan–Meier estimators, by using standard statistical techniques because they are not sums of independent random variables. Aalen[1] introduced an elegant martingale-based approach to survival analysis, where statistical methods can be cast within a unifying counting process framework. This approach uses an integral representation for censored data statistics that provides a simple unified form for estimators, test statistics and regression methods. These martingale methods allow one to obtain simple expressions for moments of complicated statistics and asymptotic distributions for test statistics and estimators, and to examine the operating characteristics of censored data regression methods.

Detailed presentation of this approach has been provided in textbooks by Fleming and Harrington[27] and Andersen *et al.*[4]

In the counting process approach for analyzing data on time-to-a-single-event, the data for the ith subject, (X_i, δ_i), is represented as $\{N_i(t), Y_i(t)\}$ $(t > 0)$, where

$$N_i(t) = I(X_i \leq t, \delta_i = 1)\,, \quad \text{and} \quad Y_i(t) = I(X_i \geq t)\,. \tag{3}$$

The right-continuous process $N(t)$ is referred to as the counting process, since it essentially counts the number of events observed up to and including time t, while the left-continuous process $Y(t)$ is referred to as the at-risk process, indicating whether the subject is at risk at time t.

A simple yet important illustration of the counting process approach is provided by examining the properties of the Nelson estimator $\hat{\Lambda}(t)$ of $\Lambda(t)$. The hazard integrated over the region in which one has data is

$$\Lambda^*(t) \equiv \int_0^t I\{\bar{Y}(u) > 0\}\lambda(u)du\,,$$

where $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$, and n is the sample size. One can write

$$\hat{\Lambda}(t) - \Lambda^*(t) = \sum_{i=1}^n \int_0^t H_i(u)dM_i(u)\,, \tag{4}$$

where $H_i(t) \equiv I\{\bar{Y}(t) > 0\}/\bar{Y}(t)$ is a left-continuous process, and

$$M_i(t) \equiv N_i(t) - \int_0^t Y_i(u)\lambda(u)du \tag{5}$$

is the subject-specific martingale. The martingale M_i in Eq. (5) represents the difference over the interval $(0, t]$ between the observed number and the model-predicted number of events for the ith subject. The left-continuity of the process H_i and the martingale property for M_i render the entire expression in Eq. (4) to be a martingale transform. This structure directly yields moments and large-sample properties. For example, since the martingale M_i has expectation zero, it follows that the Nelson estimator $\hat{\Lambda}(t)$ has expectation $\int_0^t \Pr\{\bar{Y}(u) > 0\}\lambda(u)du$. This martingale-based approach enables an elegant development of the small- and large-sample properties of the Nelson and Kaplan–Meier estimators, as shown by Gill.[29]

5. Two-Sample Statistics

The primary objective of many clinical trials is to provide a reliable comparison of the efficacy and safety of two treatments, where efficacy often

is assessed in terms of a time-to-event outcome measure. Similarly, many epidemiologic studies are concerned with the comparisons of exposed and unexposed groups in the time to disease occurrence. A variety of parametric and nonparametric two-sample statistics have been proposed to compare two survival distributions based on censored data. Parametric methods are described by Kalbfleisch and Prentice[37] and Lawless.[42]

The most popular nonparametric two-sample statistic is the so-called logrank statistic, which was originally proposed by Mantel.[55] The subjects at risk at the time of an event are classified into a 2×2 table, according to event status (yes versus no) and group membership. The numerator of the logrank statistic is obtained by computing the observed and the expected (conditioning on the margins of the 2×2 table) events in the first group, and by summing the differences of these over all distinct event times. Within each 2×2 table, the variance of the number of events in the first group is obtained using the hypergeometric distribution. These are then summed over all distinct event times to provide the variance estimator for the logrank statistic. For the colon cancer study mentioned in Sec. 3, the observed chi-squared value of the logrank statistic is 11.2, providing strong evidence for the benefit of the therapy.

The logrank statistic has been extended to a broad class of weighted logrank statistics. Any member of this class can be written as a weighted sum of the "observed minus expected" events in Mantel's 2×2 tables. These statistics can be formulated as in Eq. (4). Using this structure, Gill[29] derived small- and large-sample properties for statistics of this wide class. He developed criteria for consistency of these tests against ordered hazards and stochastic ordering alternatives. He also provided asymptotic distribution results, not only under the null hypothesis of equality of survival distributions, but also under contiguous alternatives, allowing him to provide a characterization of the alternatives against which the tests are efficient. Among these results was a proof that the logrank statistic provides an efficient test under proportional hazards alternatives.

6. Regression Models

In medical studies designed to assess the effect of a treatment or exposure on a time-to-event outcome, it is important to be able to explore or adjust for the effect of an array of other covariates that may be associated with that outcome. Hence, the information collected on each study subject (X, δ) is expanded to be $(X, \delta, \mathbf{Z})$, where $\mathbf{Z}$ represents a p-vector of covariates.

The covariates can be treatment indicators or exposure levels; demographic variables, such as age, gender or race; laboratory measurements, such as levels of bilirubin, blood pressure or viral load; histologic assessments based on biopsy; or other descriptive measurements such as time from diagnosis of disease, type of disease, prior therapeutic exposures, or functional status of the subject. In regression models, these covariates can take a variety of functional forms, being dichotomous, ordered or continuous. The continuous variables may be transformations of original measures, such as the logarithm of bilirubin.

The linear regression model for survival time data takes the form

$$\log T = \boldsymbol{\beta}'\mathbf{Z} + \epsilon\,, \tag{6}$$

where $\boldsymbol{\beta}$ is a set of unknown regression parameters, and ϵ is an error variable independent of $\mathbf{Z}$. The logarithmic transformation is employed because T is positive; other appropriate transformations of T may also be selected. Exponentiation of Eq. (6) yields $T = e^{\boldsymbol{\beta}'\mathbf{Z}}T_0$, where $T_0 = e^{\epsilon}$. This expression shows that the role of $\mathbf{Z}$ is to accelerate (or decelerate) the time to failure. Thus, Eq. (6) is referred to as the accelerated failure time model.

Because of censoring, it is more convenient to model the survival data through the hazard function. Let $\lambda(t|\mathbf{Z})$ denote the hazard function associated with $\mathbf{Z}$, i.e.

$$\lambda(t|\mathbf{Z}) = \lim_{\Delta t \downarrow 0} \Pr(t \le T < t + \Delta t | T \ge t, \mathbf{Z})/\Delta t\,.$$

The proportional hazards model specifies that

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}}\,, \tag{7}$$

where $\lambda_0(t)$ is the so-called baseline hazard function, i.e. the hazard function under $\mathbf{Z} = \mathbf{0}$, and $\boldsymbol{\beta}$ is a set of unknown regression parameters. Under this model, the covariates have multiplicative effects on the hazard function, and the regression parameters are interpreted as the logarithms of the hazard ratios or relative risks.

Equation (6) can be rewritten as

$$\lambda(t|\mathbf{Z}) = \lambda_0(te^{-\boldsymbol{\beta}'\mathbf{Z}})e^{-\boldsymbol{\beta}'\mathbf{Z}}\,, \tag{8}$$

where $\lambda_0(t)$ is the hazard function of T_0. A comparison of Eq. (8) with Eq. (7) reveals that the only overlap in the accelerated failure time and proportional hazards models arises when $\lambda_0(t)$ is Weibull.[37]

In the regression setting, the independent censoring assumption given by Eq. (2) is extended so that, conditional on $\mathbf{Z}$, the crude and net hazard

functions are equal. Survival models, such as Eqs. (6) and (6.2), are referred to as parametric models if the distributional form of the failure time, i.e. $\lambda_0(t)$, is specified, and as semiparametric models otherwise. Analysis of parametric survival models has been discussed by Kalbfleisch and Prentice,[37] Lawless,[42] Cox and Oakes,[21] and Andersen *et al.*[4] Due to the complex nature of human diseases, it is difficult to specify the parametric form. Thus, semiparametric models are preferable to parametric models in most medical applications.

7. Cox Proportional Hazards Model

Cox[19,20] introduced an ingenious semiparametric approach to inference based on the proportional hazards model. These methodologic results are among the developments in the field of survival analysis that have had the most profound impact on medical research.

By fitting the proportional hazards model in Eq. (7) with an unspecified baseline hazard function $\lambda_0(t)$, Cox obtained a robust approach for studying the influence of covariates on outcome. However, with an infinite-dimensional nuisance function $\lambda_0(t)$, modifications to the classical likelihood approach would be needed. Thus, Cox[20] introduced the partial likelihood, which is based on the data that does not carry information about $\lambda_0(t)$. Specifically, one discards the times of observed events, and the number of events at those times. Assuming that censoring is independent and is uninformative for $\boldsymbol{\beta}$ (see Definition 4.3.1 of Fleming and Harrington[27]), one also discards the censoring times and the identity of subjects associated with the censored times. The partial likelihood is then based on, for all event times, the identity of the subject(s) failing at each event time, given the number failing and the identity of the subjects at risk at that time. It takes the form

$$L(\boldsymbol{\beta}) = \prod_{i \in \mathcal{D}} \frac{e^{\boldsymbol{\beta}' \mathbf{Z}_{(i)}}}{\sum_{j \in \mathcal{R}_i} e^{\boldsymbol{\beta}' \mathbf{Z}_j}}, \tag{9}$$

where $\mathcal{D}$ is the set of indices of observed event times, $\mathbf{Z}_{(i)}$ is the covariate vector for the subject failing at the ith observed event time T_i^0, and $\mathcal{R}_i$ is the set of subjects at risk at T_i^0. The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$. Given $\hat{\boldsymbol{\beta}}$, the cumulative baseline hazard function $\Lambda_0(t) \equiv \int_0^t \lambda_0(u)du$ is estimated by the Breslow estimator[8]:

$$\hat{\Lambda}_0(t) \equiv \sum_{i \in \mathcal{D}; T_i^0 \le t} \frac{1}{\sum_{j \in \mathcal{R}_i} e^{\hat{\boldsymbol{\beta}}' \mathbf{Z}_j}}. \tag{10}$$

Table 1. Regression analysis of the Mayo primary biliary cirrhosis data.

	Proportional Hazards			Accelerated Failure Time		
Parameter	Est	SE	95% conf int	Est	SE	95% conf int
Age	0.039	0.008	(0.024, 0.054)	−0.026	0.004	(−0.035, −0.018)
log(Albumin)	−2.533	0.648	(−3.803, −1.263)	1.656	0.368	(0.934, 2.378)
log(Bilirubin)	0.871	0.083	(0.709, 1.033)	−0.585	0.046	(−0.674, −0.496)
Oedema	0.859	0.271	(0.328, 1.391)	−0.734	0.178	(−1.083, −0.385)
log(Protime)	2.380	0.767	(0.877, 3.883)	−1.944	0.462	(−2.850, −1.038)

Cox[19,20] conjectured that $L(\boldsymbol{\beta})$ shares the asymptotic properties of a full likelihood. This conjecture was confirmed by a number of authors. The first published proof was provided by Tsiatis.[78] Andersen and Gill[5] provided an elegant asymptotic theory for $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_0(t)$ by observing that the partial likelihood score function can be formulated as a martingale transform, of the form given in Eq. (4).

The left panel of Table 1 summarizes the results of the Cox regression analysis for the Mayo primary biliary cirrhosis data.[27] The database contains 418 patients who were referred to the Mayo Clinic. As of the date of data listings, 161 patients had died. The Cox regression analysis not only quantifies the effects of the five covariates on the risk of death but also allows one to estimate the survival probabilities for patients associated with specific covariate values.[48]

8. Multiplicative Intensity Model

In many medical studies, the outcome of primary interest extends beyond the time of the first event to exploration of the rate of recurrent events over time. These recurrent events, for example, may be repeated otitis media infections in an infant, or repeated hospitalizations in an adult with a serious disease. To analyze such data, Aalen[2] introduced the multiplicative intensity model as a generalization of the proportional hazards model. In this model, the subject-specific martingale is

$$M(t) = N(t) - \int_0^t Y(u)\lambda_0(u)e^{\boldsymbol{\beta}'\mathbf{Z}(u)}du\,, \tag{11}$$

where N and Y are of more general forms than given in Eq. (3). Specifically, the counting process $N(t)$ still reflects the number of events that have occurred by time t, but now has range over all non-negative integers. The at-risk process $Y(t)$ can be any left-continuous process indicating, by 1

versus 0, whether or not the subject is at risk at time t. In addition, the covariate vector is allowed to be a stochastic process.

In the semiparametric setting where $\lambda_0(t)$ in Eq. (11) is unspecified, one can use the partial likelihood principle to make inference about $\boldsymbol{\beta}$ and the Breslow estimator to estimate $\Lambda_0(t)$, although now the set $\mathcal{D}$ in Eqs. (9) and (10) may involve multiple event times from the same subject. The corresponding large-sample theory was again provided by Andersen and Gill.[5]

9. Regression Model Diagnostics

Extensive development of residuals has provided a wide variety of model diagnostics that are useful for the Cox proportional hazards model as well as for the broader multiplicative intensity model. For simplicity, we consider the subject-specific martingale in Eq. (5) for the special case of the Cox model given by Eq. (7). The corresponding martingale residual is

$$\hat{M}_i(t) \equiv N_i(t) - \hat{\Lambda}_0(t \wedge X_i) e^{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i} ,$$

where $a \wedge b = \min(a, b)$. This residual, introduced by Barlow and Prentice[6] and explored by Therneau *et al.*[76] can be interpreted as the "observed" minus "estimated model predicted" events for subject i over the interval $(0, t]$. As $t \to \infty$, the martingale residual reduces to

$$\hat{M}_i \equiv \delta_i - \hat{\Lambda}_0(X_i) e^{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i} .$$

These residuals, symmetrized using the deviance transformation[56] can be used to detect outliers. The partial residuals, defined by

$$\hat{M}_i / \{\hat{\Lambda}_0(X_i) e^{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i}\} + \hat{\beta}_j Z_{ij} , \quad i = 1, \ldots, n ; \; j = 1, \ldots, p ,$$

where Z_{ij} and $\hat{\beta}_j$ are the jth components of $\mathbf{Z}_i$ and $\hat{\boldsymbol{\beta}}$, can be used to suggest the proper functional form for covariates in the model.

A class of martingale-transform residuals can be obtained by replacing $M_i(u)$ with $\hat{M}_i(u)$ for each i in Eq. (4). Important members of this class are the p "score residuals" for each subject. These residuals are defined by

$$L_{ij} \equiv \int_0^\infty H_{ij}(t) d\hat{M}_i(t) , \quad i = 1, \ldots, n ; \; j = 1, \ldots, p ,$$

where $H_{ij}(t)$ is chosen such that $\sum_i L_{ij}$ reduces to the jth component of the partial likelihood score statistic. These p score residuals can be used to assess the influence of each subject on the parameter estimates $\hat{\beta}_j$ $(j = 1, \ldots, p)$. They are also related to a class of residuals, proposed by

Schoenfeld,[74] that are useful for detecting departures from the proportional hazards assumption.

Lin *et al.*[52] studied the cumulative sums of martingale-based residuals over covariates or event times. The distributions of these stochastic processes under the assumed model can be approximated by zero-mean Gaussian processes. Each observed process can then be compared, both graphically and numerically, with a number of realizations from the approximate null distribution by computer simulation. These comparisons enable one to determine objectively whether a seemingly abnormal residual pattern reflects model misspecification or natural random variation. This methodology can be used to assess the functional forms of covariates, the proportional hazards assumption, as well as the overall fit of the model.

10. Alternatives to the Cox Model

Despite the great popularity and versatility of the Cox regression model, there are reasons to explore alternative models. First, the proportional hazards assumption may not be satisfied in some applications. Second, alternative models characterize different aspects of the associations between covariates and survival time. In this section, we describe briefly some alternative semiparametric models.

In contrast to the proportional hazards model, the additive hazards model specifies that covariates have additive rather than multiplicative effects on the hazard function, i.e.

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}'\mathbf{Z}(t)\,. \tag{12}$$

This model was discussed by Cox and Oakes,[21] Thomas[77] and Breslow and Day.[10] Using the counting-process martingale approach, Lin and Ying[54] obtained closed-form estimators for the regression parameters $\boldsymbol{\beta}$ and the cumulative baseline hazard function $\Lambda_0(t)$.

Semiparametric transformation models take the form

$$h(T) = \boldsymbol{\beta}'\mathbf{Z} + \epsilon\,, \tag{13}$$

where ϵ is a random error with a given distribution function F, and h is a completely unspecified function. If F is the extreme value distribution, then Eq. (13) is the proportional hazards model. If F is the standard logistic function, then Eq. (13) is the proportional odds model, under which the hazard ratio approaches unity as time increases. This class of models was studied by Clayton and Cuzick[17] and Dabrowska and Doksum,[24]

and the proportional odds model was studied by Pettitt,[68] Bennet[7] and Murphy *et al.*[59] A significant breakthrough was made by Cheng *et al.*[13] who provided simple and relatively efficient estimators of $\boldsymbol{\beta}$ for all members of Eq. (13).

The semiparametric accelerated failure time model takes the same form as Eq. (13), but with h specified, usually as $h(T) = \log T$, and ϵ unspecified. Various methods of estimation for this model were proposed around 1980. Specifically, Koul *et al.*[39] suggested to include in the least-squares estimator only the uncensored survival times, but to weigh them by the inversed probabilities of being uncensored. The resulting estimator is highly inefficient, especially in the presence of heavy censoring. However, the underlying idea of weighting uncensored observations by their inversed probabilities of being uncensored, to be referred to as the inverse probability of censoring weighting (IPCW) technique, turns out to be extremely useful in many other contexts. In fact, the Cheng *et al.*[13] estimators were based on this idea. A more efficient modification of the least-squares estimator was provided by Buckley and James,[11] which replaces the conditional expectations for the censored survival times by their estimates based on the Kaplan–Meier estimator of the residual lifetime distribution and which involves an iterative estimation scheme analogous to the EM algorithm.[25] Prentice,[69] on the other hand, showed how to adapt the rank estimation method for noncensored data to the censored data setting. The asymptotic properties of the Buckley–James and rank estimators were established in the early 1990's by Tsiatis,[80] Ritov,[72] Wei *et al.*[83] Lai and Ying[40,41] and Ying.[85]

Despite the theoretical advances, semiparametric methods for the accelerated failure time model have not been widely used in medical applications due to the lack of simple and reliable numerical algorithms. Recently, Jin *et al.*[36] provided a practical method for implementing the rank estimators. Using their method, we obtain the results for the accelerated failure time regression of the primary biliary cirrhosis data shown in the right panel of Table 1. These results are based on the log-rank estimating function. Although the conclusions are not qualitatively different, the analysis under the accelerated failure time model provides an alterative and more direct interpretation of the effects of the covariates on the survival time, as compared to the Cox regression analysis.

11. Multivariate Failure Time Data

Under the multiplicative intensity model described in Sec. 8, the risk of a recurrent event for a subject is unaffected by earlier events that occurred to

the subject unless time-dependent covariates that capture such dependence are included explicitly in the model. In medical applications, the dependence structures are complex and the forms of time-dependent covariates are unknown. Furthermore, the inclusion of such time-dependent covariates which are part of the response results in biased estimation of the overall treatment effect in a randomized clinical trial. Thus, it would be desirable to model the marginal distribution of the recurrent event times while leaving the dependence structures unspecified.

It is particularly appealing to consider the cumulative mean function $\mu(t) \equiv E\{N^*(t)\}$, where $N^*(t)$ is the number of events that the subject has actually experienced by time t (in the absence of censoring). This function was first considered by Nelson[61] and further studied by Lawless and Nadeau.[43] A number of authors[44,51,65] studied the following regression models for the cumulative mean function

$$E\{N^*(t)|\mathbf{Z}\} = \mu_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}}, \tag{14}$$

where $\mu_0(t)$ is an arbitrary baseline mean function, and $\boldsymbol{\beta}$ is a set of regression parameters. If $N^*(t)$ is a (non-homogeneous) Poisson process, then Eq. (14) is equivalent to the intensity model determined by Eq. (11). Although in general N^* is not a Poisson process, the maximum partial likelihood estimator for $\boldsymbol{\beta}$ of Eq. (11) remains consistent and asymptotically normal under Eq. (14). The covariance matrix, however, can no longer be estimated by the inversed information matrix. A sandwich variance estimator has to be used instead.

In the one-sample case, $\mu(t)$ can be consistently estimated by the Nelson estimator. Under model (14), the baseline mean function $\mu_0(t)$ can be consistently estimated by the Breslow estimator, and the covariate-specific mean function can be estimated in a similar fashion.[51] It is particularly informative to display the estimated mean functions for different treatment arms and for specific covariate patterns.

In some medical studies, each subject can potentially experience more than one type of event. Examples include the developments of physical symptoms or diseases in several organ systems (e.g. stroke and cancer) or in several members of the same organ system (e.g. eyes or teeth). Models such as Eqs. (11) and (14) are not applicable since the multiple events on the same subject are of different natures and in fact may not even be ordered.

It is convenient to formulate the marginal distributions of the multiple event times through the proportional hazards models while leaving the dependence structures completely unspecified. Let K denote the number

of potential events per subject. The hazard function for the kth event of the ith subject is postulated to take the form

$$\lambda(t|\mathbf{Z}_{ki}) = \lambda_{k0}(t)e^{\boldsymbol{\beta}'\mathbf{Z}_{ki}(t)}, \quad k = 1, \ldots, K; \quad i = 1, \ldots, n, \tag{15}$$

where $\mathbf{Z}_{ki}$ is the covariate vector for the ith subject with respect to the kth event, λ_{k0} $(k = 1, \ldots, K)$ are arbitrary baseline hazard functions, and $\boldsymbol{\beta}$ is a set of regression parameters. In some applications (e.g. an ophthalmologic study involving the left and right eyes), it is natural to impose the restriction that $\lambda_{10} = \cdots = \lambda_{K0}$ wheareas in others (e.g. the setting of multiple diseases) it is necessary to allow the λ_{k0}'s to be different.

If the event times were independent, then the partial likelihood could be easily constructed for $\boldsymbol{\beta}$ of model (15). The resulting estimator turns out to be consistent and asymptotically normal even if the event times are correlated. However, a sandwich variance estimator is again needed to account for the intra-class dependence. This approach was pioneered by Wei *et al.*[82] and further developed by Lee *et al.*,[45] Liang *et al.*[46] and Cai and Prentice[12] among others.

The marginal approach discussed above treats the dependence of related event times as a nuisance. An alternative approach is to explicitly formulate the nature of dependence by the so-called frailty. The term frailty was first introduced by Vaupel *et al.*[81] to illustrate the consequences of a lifetime being generated from several sources of variation. The use of frailty in bivariate survival time data was considered by Clayton.[15] Frailty models were studied extensively in the 1980s by Clayton and Cuzick,[16] Hougaard[34] and Oakes[63] among others. The frailty-model analog of Eq. (15) specifies that the hazard function for the kth event of the ith subject, given the frailty ν_i, takes the form

$$\lambda_{ki}(t|\mathbf{Z}_{ki}; \nu_i) = \nu_i \lambda_{k0}(t)e^{\boldsymbol{\beta}'\mathbf{Z}_{ki}(t)}, \tag{16}$$

where ν_i $(i = 1, \ldots, n)$ are independent random variables. Conditional on ν_i, the event times on the ith subject are assumed to be independent.

The parameter vector $\boldsymbol{\beta}$ has a population-average interpretation under model (15) and a subject-specific interpretation under model (16). Models (15) and (16) cannot hold simultaneously unless ν is a positive-stable variable. It is very challenging, both theoretically and computationally, to deal with frailty models such as (16). Major progress was made in the 1990s. In the special case of gamma frailty models, maximum likelihood estimation via the EM algorithm was studied by Nielsen *et al.*,[62] Murphy,[57,58] Andersen *et al.*[3] and Parner[64] among others.

Nonparametric estimation for the multivariate survival function is a fundamental problem in the analysis of multivariate failure time data. Using the IPCW technique, Lin and Ying[53] developed a simple estimator for the special case where there is a common censoring time for all event times of the same subject. Estimation in the general setting has been studied by Dabrowska,[23] Prentice and Cai,[70] and van der Laan[81] among others.

The occurrence of one event (e.g. death) may preclude the development of another (e.g. relapse of cancer). In some applications, such as cause-specific mortality studies, the subject can only experience one of several potential events. This type of data is referred to as competing risks. The simplest solution to this problem is to censor the event time of interest at the time of the competing events, and then apply the standard survival analysis methods such as the logrank test and Cox regression. The results pertain to the so-called cause-specific hazard function, which is given by Eq. (1) with U representing the time to the competing events.

An important limitation of the cause-specific hazard function is that the associated $S^{\#}(t)$ is not a survival function unless the cause of interest is independent of other risks and the other risks could be eliminated without altering the distribution of the cause of interest. Thus, in general the Kaplan–Meier estimator does not pertain to the survival function or disease incidence. Special methods have been developed to estimate disease incidence functions.[26,31,66]

12. Concluding Remarks

We have reviewed many areas of survival analysis in the previous sections. All these methods require the assumption of independent censoring. As discussed in Sec. 2, the survival distribution is not identifiable in the presence of dependent censoring. If one is willing to model certain aspects of the dependent censoring mechanism, then it is possible make inference about the survival distribution under dependent censoring.[50,73]

When the event of interest is asymptomatic, as is the case with cancer progression or HIV infection, the event time cannot be measured exactly, but is rather known to lie in an interval determined by two successive examinations. The resulting data are said to be interval censored. Non- and semi-parametric analysis of such data has been studied by Groeneboom and Wellner,[32] Huang,[35] Lin *et al.*[49] and Rabinowitz *et al.*[71] among others.

The applications of survival analysis methods to medical studies have been greatly facilitated by the developments of software packages. Standard

methods such as the Kaplan–Meier estimator, weighted logrank tests, Cox regression and parametric regression with (univariate) right-censored data are now available in virtually all software packages. The multiplicative intensity model and the sandwich variance estimators for models (14) and (15) have been implemented in major packages, such as SAS, S-Plus and STATA. However, most of the newer methods, such as those for the semiparametric analysis of models (6), (12) and (13), and those mentioned in this section, are not available in software packages.

Further developments are anticipated in many areas of survival analysis. For example, the Cheng *et al.*[13] estimators for Eq. (13) require modelling the censoring distribution, and it would be worthwhile to explore estimation procedures which do not involve such modelling. In the area of multivariate failure time data, efficient estimators for model (15) have yet to be identified, and further theoretical and numerical advances are warranted for model (16). Considerable activities are also expected in the areas of dependent censoring, interval censored data, causal inference, and joint modelling of longitudinal and failure time data. Finally, further expansion of software is anticipated.

References

1. Aalen, O. O. (1975). *Statistical Inference for a Family of Counting Processes.* PhD. Dissertation, University of California, Berkeley.
2. Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* **6**: 701–726.
3. Andersen, P. K., Klein, J. P., Knudsen, K. M. and Palacios, R. T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics* **53**: 1475–1484.
4. Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, London.
5. Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**: 1100–1120.
6. Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* **75**: 65–74.
7. Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**: 273–277.
8. Breslow, N. E. (1972). Contribution to the discussion on the paper by D. R. Cox cited below. *Journal of the Royal Statistical Society* **B34**: 216–217.
9. Breslow, N. E. and Crowley, J. J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics* **2**: 437–453.
10. Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. The Design and Analysis of Cohort Studies* **2**, IARC, Lyon.

11. Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**: 429–436.
12. Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**: 151–164.
13. Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**: 835–845.
14. Chiang, C. L. (1968). *Introduction to Stochastic Processes in Biostatistics*. John Wiley, New York.
15. Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**: 141–151.
16. Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society* **A148**: 82–117.
17. Clayton, D. G. and Cuzick, J. (1986). The semiparametrics Pareto model for regression analysis of survival times. In *Proceedings of the International Statistical Institute*, International Statistical Institute, Amsterdam, 23.3-1-18.
18. Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society* **B21**: 411–421.
19. Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **B34**: 187–220.
20. Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**: 269–276.
21. Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.
22. Cuzick, J. (1985). Asymptotic properties of censored linear rank tests. *The Annals of Statistics* **13**: 133–141.
23. Dabrowska, D. M. (1988). Kaplan–Meier estimate on the plane. *The Annals of Statistics* **16**: 1475–1489.
24. Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**: 1-23.
25. Dempster, A. P., Laird, N. M. and Rubin, D. (1977). Maximum likelihood estimation for incomplete data via the EM algorithms (with discussion). *Journal of the Royal Statistical Society* **B39**: 1–38.
26. Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**: 496–509.
27. Fleming, T. R. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.
28. Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52**: 203–223.
29. Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.
30. Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *The Annals of Statistics* **11**: 49–58.

31. Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics* **16**: 1141–1154.
32. Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.
33. Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**: 553–566.
34. Hougaard, P. (1987). Modelling multivariate survival. *Scandinavian Journal of Statistics* **14**: 291–304.
35. Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24**: 540–568.
36. Jin, Z. Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the semiparametric accelerated failure time model. *Biometrika*, in press.
37. Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.
38. Kaplan, E. L. and Meier, P. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association* **53**: 457–481.
39. Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *The Annals of Statistics* **9**: 1276–1288.
40. Lai, T. L. and Ying, Z. (1991a). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* **19**: 531–556.
41. Lai, T. L. and Ying, Z. (1991b). Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *The Annals of Statistics* **19**: 1370–1402.
42. Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.
43. Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**: 158–168.
44. Lawless, J. F., Nadeau, C. and Cook, R. J. (1997). Analysis of mean and rate functions for recurrent events. In *Proceedings of the First Seattle Symposium in Biostatistics*: *Survival Analysis*, eds. D. Y. Lin and T. R. Fleming, Springer-Verlag, New York, 37–49.
45. Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis*: *State of the Art*; eds. J. P. Klein and P. K. Goel, Kluwer Academic Publishers, Dordrecht, 237–247.
46. Liang, K. Y., Self, S. G. and Chang, Y. C. (1993). Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society* **B55**: 441–453.
47. Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine* **13**: 2233–2247.
48. Lin, D. Y., Fleming, T. R. and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**: 73–81.
49. Lin, D. Y., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**: 289–298.
50. Lin, D. Y., Robins, J. M. and Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**: 381–393.

51. Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society* **B62**: 711–730.
52. Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**: 557–572.
53. Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* **80**: 573–581.
54. Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**: 61–71.
55. Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Report* **50**: 163–170.
56. McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.
57. Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics* **22**: 712–731.
58. Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics* **23**: 182–198.
59. Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92**: 968–976.
60. Nelson, W. B. (1969). Hazard plotting for incomplete failure data. *Journal of Qual Technology* **1**: 27–52.
61. Nelson, W. B. (1988). Graphical analysis of system repair data. *Journal of Quality Technology* **20**: 24–35.
62. Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**: 25–43.
63. Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**: 487–493.
64. Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26**: 183–214.
65. Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**: 811–820.
66. Pepe, M. S. and Mori, M. (1993). Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* **12**: 737–751.
67. Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society* **A135**: 185–206.
68. Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society* **B44**: 234–243.
69. Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**: 167–179.

70. Prentice, R. L. and Cai, J. (1992). Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* **79**: 495–512.
71. Rabinowitz, D., Tsiatis, A. and Aragon, J. (1995). Regression with interval censored data. *Biometrika* **82**: 501–513.
72. Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics* **18**: 303–328.
73. Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*, eds. N. P. Jewell, K. Dietz and V. T. Farewell, Birkhäuser, Boston, 297–331.
74. Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**: 239–241.
75. Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**: 156–160.
76. Therneau, T., Grambsch, P. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**: 147–160.
77. Thomas, D. C. (1986). Use of auxiliary information in fitting nonproportional hazards models. In *Modern Statistical Methods in Chronic Disease Epidemiology*, eds. S. H. Moolgavkar and R. L. Prentice, Wiley, New York, 197–210.
78. Tsiatis, A. A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Science USA* **72**: 20–22.
79. Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics* **9**: 93–108.
80. Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18**: 354–372.
81. van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *the Annals of Statistics* **24**: 596–627.
82. Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**: 439–454.
83. Wei, L. J., Lin D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**: 1065–1073.
84. Wei, L. J., Ying, Z. and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**: 845–851.
85. Ying, Z. (1989). A note on the asymptotic properties of the product-limit estimator on the whole line. *Statisitics and Probability Letters* **7**: 311–314.
86. Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics* **21**: 76–99.

About the Author

D. Y. Lin is the Dennis Gillings Distinguished Professor of Biostatistics at the Univerisity of North Carolina at Chapel Hill. Professor Lin has

made significant contributions to medical statistics, especially in survival analysis and clinical trials. As a result, he was honored with the Mortimer Spiegelman Award from the American Public Health Association in 1999. Professor Lin is a Fellow of the American Statistical Association, and a Fellow of the Institute of Mathematical Statistics. He currently serves as an associate editor for *Biometrika* and *Statistica Sinica*.

This page intentionally left blank

CHAPTER 23

REGRESSION MODELS FOR THE ANALYSIS OF LONGITUDINAL DATA

COLIN O. WU

Office of Biostatistics Research, DECA, National Heart, Lung and Blood Institute, II Rockledge Center, Rm 8218, 6701 Rockledge Drive, MSC 7938, Bethesda, MD 20892, USA
Tel: 301-435-0440; wuc@nhlbi.nih.gov

KAI F. YU

Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, 6100 Executive Blvd. Room 7B05, National Institutes of Health, Bethesda, MD 20892-7510, USA

1. Introduction

1.1. *Structures of longitudinal data*

In biomedical and epidemiological studies, interests are often focused on evaluating the effects of treatments, dosage, risk factors or other covariates of interest on the outcomes of interest, such as disease progression and change of health status of a population, over time. Because the changes of outcome and covariates within each subject usually provide important information of scientific relevance, longitudinal samples that contain repeated measurements of the chosen subjects over time are often more preferable than the classical cross-sectional samples. In fact, by combining the characteristics of random sampling and time series observations, the usefulness of longitudinal samples goes far beyond biomedicine and epidemiology, and their trace is often found in economics, psychology, sociology and many other fields of natural and social sciences.

For a typical framework of longitudinal data, we define t to be a real-valued variable of time, $Y(t)$ a real-valued outcome and $X(t) =$

$(X^{(0)}(t), \ldots, X^{(k)}(t))^T$, $k \geq 1$, a R^{k+1}-valued covariate vector at time t. Depending on the choice of origin, the time variable t is not necessarily non-negative. As part of the general methodology, interest of statistical analysis with regression models is often focused on modeling and determining the effects of $(t, X(t))$ on the population mean of $Y(t)$. For n randomly selected subjects, each repeatedly measured over time, the longitudinal sample of $(Y(t), t, X(t))$ is denoted by $\{(Y_{ij}, t_{ij}, X_{ij}) : i = 1, \ldots, n, j = 1, \ldots, n_i\}$, where t_{ij} is the jth measurement time of the ith subject, Y_{ij} and $X_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(k)})^T$ are the observed outcome and covariate vector, respectively, of the ith subject at t_{ij} and n_i is the ith subject's number of repeated measurements. The total number of measurements is $N = \sum_{i=1}^{n} n_i$. In contrast to the classical independent identically distributed (i.i.d.) samples, the measurements within each subject are possibly correlated, although the inter-subject measurements are independent. A longitudinal sample is said to have a balanced design if all the subjects have their measurements made at a common set of time points, i.e. $n_i = m$ for some $m \geq 1$ and all $i = 1, \ldots, n$ and $t_{1j} = \cdots = t_{nj}$ for all $j = 1, \ldots, m$. An unbalanced design arises if the design time t_{ij} are different per subject. In practice, unbalanced designs may be caused by the presence of missing values in an otherwise balanced design or by the random variations of the time design points.

In general, two routes could be used to obtain biomedical observations: clinical trials and observational cohort studies. The main difference between a clinical trial and an observational cohort study is at their designs. In a clinical trial, the investigator has the power to determine, at least partially, the selection of the participants and the design of the trial, such as the treatment offered, the length of the trial and the time and methods of the measurement process. An observational cohort studies, on the other hand, is more complicated, because the risk factors, the treatments and the measurement process now depend on the participants of the study and are usually not controlled by the investigator. Consequently, it is possible to observe balanced longitudinal data from clinical trials. But, for various reasons that are out of the investigators' control, most observational cohort studies have unbalanced longitudinal designs.

1.2. *Examples of longitudinal studies*

The following two epidemiological examples illustrate some typical features of longitudinal samples. Although these examples share some similarities, such as both of them are cohort studies with unbalanced designs, they differ in the numbers of repeated measurements and the design of their covariates.

1.2.1. *Alabama small-for-gestational-age (ASGA) study*

This is a prospective study of risk factors and intrauterine growth retardation involving 1475 women who had their fetal anthropometry measurements made by ultrasound repeatedly during pregnancy. All the women were scheduled to have their measurements made at approximately 17, 25, 31 and 36 weeks of gestation. Their actual visits, however, numbered between 1 to 7 times per person, did not follow this schedule and were scattered throughout 12 to 43 weeks of gestation. This results in an unbalanced design in the sense that not all the subjects are measured at the same design points. Associated covariates that may affect fetal development include maternal behavioral factors, such as cigarette smoking, alcohol consumption and drug abuse, and maternal anthropometric measurements, such as pre-pregnancy weight, height and body mass index, and placental development measured by placental thickness at different stages of gestation. Some of these covariates, such as the maternal anthropometric measurements are time-dependent. But, the others, such as maternal behavioral factors, may be either time-dependent or time-invariant, depending on how these variables are defined. The outcome variables of fetal development, which, of course, are all time-dependent, include the fetal abdominal circumference, biparietal diameter, femur length and other ultrasound measurements. Figure 1 shows the observed fetal abdominal circumferences (in cm) at their corresponding gestational age in weeks. To see the trend of the individual repeated measurements, the line segments indicate the measurement sequences for a number of randomly selected subjects. Heuristically, we observe a linear upward trend on the growth of fetal size. But, there has been no prior study which justifies the goodness of a linear growth model for this population or any other statistical model on the covariate effects on fetal growth. A statistical analysis should then focus on two objectives: establishing an appropriate statistical model so that the effects of these covariates on the outcome of interest can be clearly interpreted; developing estimation and inference procedures to adequately quantify the covariate effects based on the chosen statistical models.

1.2.2. *HIV/CD4 depletion data*

The data set is from Multicenter AIDS Cohort Study (MACS) 1984–1991, which includes 400 homosexual men who were infected by the human immunodeficiency virus (HIV) between 1984 and 1991. Because CD4 cells (T-helper lymphocytes) are vital for immune function, an important component of the study is to evaluate the effects of risk factors, such as

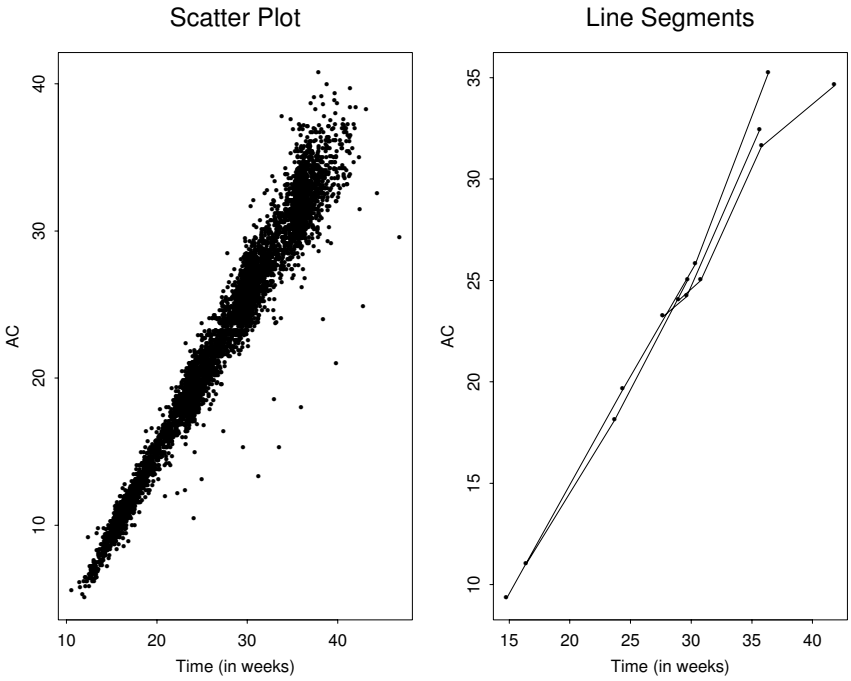


Fig. 1. Relationship between fetal abdominal circumference (in cm) and fetal gestation (in weeks). Left Panel: individual measurement. Right Panel: sequences of measurements for some randomly selected subjects.

cigarette smoking and drug use, and health status, such as CD4 cell levels before the infection, on the post-infection depletion of CD4 percent (CD4 percent of lymphocyte cells). Although all the individuals were scheduled to have their measurements made at semi-annual visits, the study has an unbalanced design because the subjects' actual visiting times did not exactly follow the schedule and the HIV infections happened randomly during the study. The numbers of repeated measurements range from 1 to 14 with a median of approximately 6. Compared with the previous example of ASGA Study, this data set has a smaller number of subjects and a wider range of repeated measurements. The covariates of interest in these data also include both time-dependent and time-invariant variables. However, as will be seen later in this chapter, these covariates have some important differences, hence, should be treated differently, from those considered in the ASGA Study. Further details of the design and medical importance of the MACS data can be found in Kaslow *et al.*[23]

1.3. *Overview of regression models*

1.3.1. *Main objectives*

Generally speaking, a proper longitudinal analysis should achieve at least three objectives:

- The model under consideration must give an adequate description of the scientific relevance of the data and be sufficiently simple and flexible to be practically implemented. In biomedical and epidemiological studies, we would prefer a model that gives a clear and meaningful biological interpretation and also has a simple mathematical structure.
- The methodology must contain proper model diagnostic tools to evaluate the validity of a statistical model for a given data set. Two important diagnostic methods are confidence regions and tests of statistical hypotheses.
- The methodology must have appropriate theoretical and practical properties, and can adequately handle the possible intra-subject correlations of the data. In practice, the intra-subject correlations are often completely unknown and difficult to be adequately estimated, so that it is generally preferred to use estimation and inference procedures that do not depend on modeling the specific correlation structures.

1.3.2. *Parametric models*

Naturally, the most commonly used modeling approach is to use parametric regression, such as the random and mixed effects linear models, the generalized linear models and nonlinear models. The simplest case of these models is the marginal linear model of the form

$$Y_{ij} = \sum_{l=0}^{k} \beta_l X_{ij}^{(l)} + \epsilon_i(t_{ij}) \,, \tag{1}$$

where $\beta_0, \ldots, \beta_k$ are constant linear coefficients describing the effects of the corresponding covariates, $\epsilon_i(t)$ are realizations of a mean zero stochastic process $\epsilon(t)$ at t and X_{ij} and $\epsilon_i(t_{ij})$ are independent. Similar to all regression models where a constant intercept term is desired, we set $X^{(0)} \equiv 1$, which produces a baseline coefficient β_0, representing the mean value of $Y(t)$ when all the covariates $X^{(l)}(t)$ are set to zero. A popular special case of the error process is to take $\epsilon(t)$ to be a mean zero Gaussian stationary process. Although (1) appears to be overly simplified for many practical

situations, its generalizations lead to many useful models which form the bulk of longitudinal analysis.

Estimation and inference methods based on parametric models, including the weighted least squares, the quasi-likelihoods and the generalized estimating equations, have been extensively investigated in the literature. The details can be found in some references.[5,7,8,20,21,24,25,28,34,35,37,43] The main advantage of parametric models is that they generally have simple and intuitive interpretations. User friendly computer programs are already available in most popular statistical software packages, such as SAS and S-Plus. However, these models suffer the potential shortfall of model misspecification, which may lead to erroneous conclusions. At least in exploratory studies, it is necessary to relax some of the parametric restrictions.

1.3.3. *Semiparametric models*

A useful semiparametric model, investigated by Zeger and Diggle[42] and Moyeed and Diggle,[26] is the partially linear model of the form

$$Y_{ij} = \beta_0(t_{ij}) + \sum_{l=1}^{k} \beta_l X_{ij}^{(l)} + \epsilon_i(t_{ij}) \,, \tag{2}$$

where $\beta_0(t)$ is an unknown smooth function of t, β_l are unknown constants and $\epsilon_i(t)$ and X_{ij} are defined in model (1). This model is more general than the marginal linear model (1), because $\beta_0(t)$ is allowed to change with t, rather than setting to be a constant over time. On the other hand, by including the linear terms of $X_{ij}^{(l)}$, Eq. (2) is also more general than the nonparametric regression studied,[2,14,15,30] which involves only (t_{ij}, Y_{ij}). However, because model (2) describes the effects of $X_{ij}^{(l)}$ on Y_{ij} through constant linear coefficients, this model is still, to some degree, based on mathematical convenience rather than scientific relevance. For example, there is no reason to expect that the influences of maternal risk factors on fetal development in the ASGA Study (Sec. 1.2.1) or the effects of cigarette smoking and pre-infection CD4 level on the post-infection CD4 cell percent in the HIV/CD4 Depletion Data (Sec. 1.2.2) are linear and constant throughout the study period. Thus, further generalization of model (2) is needed in many situations.

1.3.4. *Nonparametric models*

Although it is possible in principle to model (Y_{ij}, t_{ij}, X_{ij}) through a completely nonparametric high dimensional function, such approach is often

impractical due to the well-known problem of "curse of dimensionality". Moreover, numerical results obtained from high dimensional nonparametric fittings are often difficult to interpret. These problems motivate the consideration of nonparametric models that have certain meaningful structures. An important class of structural nonparametric regression models is the varying-coefficient models of the form

$$Y_{ij} = X_{ij}^T \beta(t_{ij}) + \epsilon_i(t_{ij}) \,, \tag{3}$$

where $\beta(t) = (\beta_0(t), \ldots, \beta_k(t))^T$ is a $(k+1)$-vector of smooth functions of t and $\epsilon_i(t)$ and X_{ij} are defined as in Eq. (1). Because Eq. (3) gives a linear model between $Y(t)$ and $X(t)$ at each fixed t, the linear coefficients $\beta_l(t)$, $l = 0, \ldots, k$, can be interpreted the same way as in Eq. (1). Taking $X_{ij}^{(0)} \equiv 1$, $\beta_0(t)$ represents the intercept at time t. On the other hand, because all the linear coefficients may change with t, we may obtain different linear models at different time points. Model (3) is a special case of the general varying-coefficient models discussed by Hastie and Tibshirani.[17]

Methods of estimation and inferences based on this class of models have been subjected to intense investigation recently in the literature. A number of different smoothing methods for the estimation of $\beta(t)$ have been proposed. These include the ordinary least squares local polynomials, the penalized least squares, the two-step and componentwise methods and the basis approximation approaches. Targeted to specific types of longitudinal designs, each of these methods has its own advantages and disadvantages in practice. We will present in Secs. 4 and 5 an overview of the above estimation and inference methods and demonstrate in Sec. 6 the applications of these methods.

2. Linear mixed effects models

2.1. *Models for covariate effects and correlations*

Statistical models for longitudinal observations generally serve two purposes: (i) describing the effects of the treatments and other factors on the mean response profile; (ii) describing the differences in response profiles between individuals. A model serving the first purpose is generally classified as a marginal model or a population average model. A model serving the second purpose is a random effects model or a subject specified model.[43] A mixed effects model then combines both the marginal and random effects. In particular, a linear mixed effects model is obtained when the marginal and random effects are additive and follow a linear relationship.

It is convenient to describe the model through a matrix representation. Let $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ be the $[n_i \times 1]$ vector of the response for the ith subject, $t_i = (t_{i1}, \ldots, t_{in_i})^T$ be the subject's time design points and X_i be the corresponding $[n_i \times (k+1)]$ covariate matrix whose jth row, for $j = 1, \ldots, n_i$, is $(1, X_{ij}^{(1)}, \ldots, X_{ij}^{(k)})$. Assuming that the error term $\epsilon_i(t)$ of Eq. (1) is a mean zero Gaussian process with covariate matrix $V_i(t_i)$, the responses Y_i are then independent Gaussian random vectors such that

$$Y_i \sim \mathbf{N}(X_i\beta, V_i(t_i)), \tag{4}$$

where $\beta = (\beta_0, \ldots, \beta_k)^T$ with β_j being defined in Eq. (1) and $\mathbf{N}(\mathbf{a}, \mathbf{b})$ denotes a multivariate normal distribution with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{b}$. Note that, because model (4) represents the conditional mean of Y_i at X_i through $X_i\beta$, it is a marginal model.

The covariance structures of model (4) are usually influenced by three factors: random effect, serial correlation and measurement error. The random effects characterize the stochastic variations between subjects within the population. In particular, we may view that, when the covariates affect the response linearly, some of the linear coefficients may vary from subject to subject. The serial correlations are the results of time-varying associations between different measurements of the same subject. Such correlations are typically positive in biomedical studies, and become weaker as the time interval between the measurements increases. Finally, the measurement errors, which are normally assumed to be independent both between and within the subjects, are induced by the measurement process or random variations within the subjects.

Suppose that, for each subject i, there is a $[r \times 1]$ vector of explanatory variables U_{ij} measured at time t_{ij}, which may or may not overlap with the original covariate vector X_{ij}. Using the additive decomposition of random effects, serial correlations and measurement errors, $\epsilon_i(t_{ij})$ can be expressed as

$$\epsilon_i(t_{ij}) = U_{ij}^T b_i + W_i(t_{ij}) + Z_{ij}, \tag{5}$$

where b_i is the $[r \times 1]$ random vector with multivariate normal distribution $\mathbf{N}(\mathbf{0}, D)$, D is a $[r \times r]$ covariance matrix with (p,q)th element $d_{pq} = d_{qp}$, $W_i(t_{ij})$ for $i = 1, \ldots, n$ are independent copies of a mean zero Gaussian process whose covariance at time points t_{ij_1} and t_{ij_2} is $\rho_W(t_{ij_1}, t_{ij_2})$, and Z_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ are independent identically distributed random variables with $N(0, \tau^2)$ distribution. Writing $\delta_i(t_{ij}) = W_i(t_{ij}) + Z_{ij}$, $\delta_i = (\delta_i(t_{i1}), \ldots, \delta_i(t_{in_i}))^T$ and U_i to be the $[n_i \times r]$ matrix whose jth row

is U_{ij}^T, (4) and (5) reduce to the linear mixed effects model of Laird and Ware[24]

$$Y_i = X_i\beta + U_i b_i + \delta_i\,. \tag{6}$$

The marginal effect β represents the influence of X_i on the population average of the response profile, while b_i describes the variation of the ith subject from the population conditioning on the given explanatory variable U_i. Thus, conditioning on X_i and U_i, model (6) implies that Y_i for $i = 1, \ldots, n$ are independent Gaussian vectors such that

$$Y_i \sim \mathbf{N}(X_i\beta, U_i D U_i^T + P_i + \tau^2 I_i)\,, \tag{7}$$

where P_i is the $[n_i \times n_i]$ covariance matrix whose (j_1, j_2)th element is $\rho_W(t_{ij_1}, t_{ij_2})$ and I_i is the $[n_i \times n_i]$ identity matrix.

A number of special cases can be derived for the variance-covariance structure of model (5). The classical linear models for the independent cross-sectional data (or the independent identically distributed data) is a special case of model (7) where $\epsilon_i(t_{ij})$ are only affected by the measurement errors Z_{ij}. When neither random effects nor measurement errors are present, the error term is of pure serial correlation $\epsilon_i(t_{ij}) = W_i(t_{ij})$. Moreover, if $W_i(t_{ij})$ are from a mean zero stationary Gaussian process, the covariance of $\epsilon_i(t_{ij_1})$ and $\epsilon_i(t_{ij_2})$, hence, Y_{ij_1} and Y_{ij_2}, can be specified by

$$\mathrm{Cov}(\epsilon_i(t_{ij_1}), \epsilon_i(t_{ij_2})) = \sigma^2\rho(|t_{ij_1} - t_{ij_2}|)\,, \tag{8}$$

where σ is a positive constant and $\rho(\cdot)$ is a continuous function. Useful choices of $\rho(\cdot)$ include the exponential correlation $\rho(s) = \exp(-as)$ for some constant $a > 0$ and the Gaussian correlation $\rho(s) = \exp(-as^2)$, among others. When $\epsilon_i(t_{ij})$ are affected by a mean zero stationary Gaussian process and a mean zero Gaussian white noise (measurement error), the variance of Y_{ij} is $\sigma^2\rho(0) + \tau^2$, while the covariance of Y_{ij_1} and Y_{ij_2}, for $j_1 \neq j_2$, is $\sigma^2\rho(|t_{ij_1} - t_{ij_2}|)$, for some $\sigma > 0$, $\tau > 0$ and continuous correlation function $\rho(\cdot)$. When serial correlations are not present, the intra-subject correlations are only induced by the random effects, so that P_i is not present in model (7).

2.2. *Likelihood based estimation and inferences*

2.2.1. *Conditional maximum likelihood estimation*

Suppose that the variance-covariance matrix $V_i(t_i)$ of model (5) is determined by a R^q-valued parameter vector α. Denote $V_i(t_i; \alpha)$ to be the

variance-covariance matrix parametrized by α. The log-likelihood function for model (4) is

$$L(\beta,\alpha) = c + \sum_{i=1}^{n}\Big\{ -\frac{1}{2}\log|V_i(t_i;\alpha)| - \frac{1}{2}(Y_i - X_i\beta)^T V_i^{-1}(t_i;\alpha)(Y_i - X_i\beta)\Big\}, \tag{9}$$

where $c = \sum_{i=1}^{n}[(-n_i/2)\log(2\pi)]$. For a given α, model (9) can be maximized by

$$\hat{\beta}(\alpha) = \left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)X_i)\right]^{-1}\left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)Y_i)\right]. \tag{10}$$

It is easy to verify that, under model (4), $\hat{\beta}(\alpha)$ is an unbiased estimator of β. Direct calculation also shows that the covariance matrix of $\hat{\beta}(\alpha)$ is

$$\begin{aligned}\mathrm{Cov}[\hat{\beta}(\alpha)] &= \left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)X_i)\right]^{-1} \\ &\quad\times\left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)\ \mathrm{Cov}(Y_i)V_i^{-1}(t_i;\alpha)X_i)\right] \\ &\quad\times\left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)X_i)\right]^{-1} \\ &= \left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)X_i)\right]^{-1}. \end{aligned}\tag{11}$$

It is interesting to note that the second equality sign of model (11) does not hold when the structure of the variance-covariance matrix is not correctly specified. Further derivation using Eqs. (4)–(11) shows that $\hat{\beta}(\alpha)$ has a multivariate normal distribution,

$$\hat{\beta}(\alpha) \sim \mathbf{N}\left\{\beta, \left[\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\alpha)X_i)\right]^{-1}\right\}. \tag{12}$$

When α is known, this result can be used to develop inference procedures, such as confidence regions and test statistics, for β.

2.2.2. *Maximum likelihood estimation*

When α is unknown, as in most practical situations, a consistent estimate of α has to be used. An intuitive approach is to estimate β and α by maximizing (9) with respect to β and α simultaneously. Maximum likelihood estimators (MLE) of this type can be computed by substituting (10) into (9) and then maximizing (9) with respect to α. Denote the resulting ML estimators by $\hat{\beta}_{ML}$ and $\hat{\alpha}_{ML}$. The asymptotic distributions of $(\hat{\beta}_{ML}, \hat{\alpha}_{ML})$ can be developed using the standard approaches in large sample theory.

Although $(\hat{\beta}_{ML}, \hat{\alpha}_{ML})$ has some justifiable statistical properties, as for most likelihood-based methods, it may not be desirable in practice. To see why an alternative estimation method might be warranted in some situations, we consider the simple linear regression with independent errors and $n_1 = \cdots = n_n = m$,

$$Y_i \sim \mathbf{N}(X_i\beta, \sigma^2 I_m)\,, \tag{13}$$

where I_m is the $[m \times m]$ identity matrix. The parameters involved in the model are β and σ. Let $\hat{\beta}_{ML}$ and $\hat{\sigma}_{ML}$ be the MLEs of β and σ, respectively, and RSS be the residual sum of squares defined by

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - X_i\hat{\beta}_{ML})^T (Y_i - X_i\hat{\beta}_{ML})\,.$$

The MLE of σ^2 is $\hat{\sigma}^2_{ML} = \mathrm{RSS}/(nm)$. However, it is well-known that, for any finite n and m, $\hat{\sigma}^2_{ML}$ is a biased estimator of σ^2. On the other hand, a slightly modified estimator $\hat{\sigma}^2_{REML} = \mathrm{RSS}/[nm - (k+1)]$ is unbiased for σ^2. Here, $\hat{\sigma}^2_{REML}$ is the restricted maximum likelihood (REML) estimator for the model (13).

2.2.3. *Restricted maximum likelihood estimation*

This class of estimators was introduced by Patterson and Thompson[29] for the purpose of estimating variance components in the linear models. The main idea is to consider a linear transformation of the original response variable so that the distribution of the transformed variable does not depend on β. Let $\mathbf{Y} = (Y_1^T, \ldots, Y_n^T)^T$, $\mathbf{X} = (X_1^T, \ldots, X_n^T)^T$ and $\mathbf{V}$ be the block-diagonal matrix with $V_i(t_i)$ on the ith main diagonal and zeros elsewhere. Then, with $\mathbf{V}$ parameterized by α, model (4) is equivalent to

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\beta, \mathbf{V}(\alpha))\,. \tag{14}$$

The REML estimator of α, the variance component of (14), is obtained by maximizing the likelihood function of $\mathbf{Y}^* = A^T\mathbf{Y}$, where A is a $[N \times$

$(N-k-1)]$, $N=\sum_{i=1}^{n} n_i$, full rank matrix such that $A^T\mathbf{X}=\mathbf{0}$. A specific construction of A can be found in Diggle, Liang and Zeger.[8] It follows from (14) that $\mathbf{Y}^*$ has a mean zero multivariate Gaussian distribution with covariance matrix $A^T\mathbf{V}(\alpha)A$. Harville[16] showed that the likelihood function of $\mathbf{Y}^*$ is proportional to

$$L^*(\alpha)=\left|\sum_{i=1}^{n} X_i^T X_i\right|^{1/2}\left|\sum_{i=1}^{n} X_i^T V_i^{-1}(t_i;\alpha)X_i\right|^{-1/2}\left\{\prod_{i=1}^{n}|V_i(t_i;\alpha)|^{-1/2}\right\}$$
$$\times\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_i-X_i\hat{\beta}(\alpha))^T V_i^{-1}(t_i;\alpha)(Y_i-X_i\hat{\beta}(\alpha))\right\}. \quad (15)$$

The REML estimator $\hat{\alpha}_{REML}$ of α maximizes (15). The REML estimator $\hat{\beta}_{REML}$ of β is obtained by substituting α of (10) with $\hat{\alpha}_{REML}$. Because (15) does not depend on the choice of A, the resulting estimators $\hat{\beta}_{REML}$ and $\hat{\alpha}_{REML}$ are free of the specific linear transformations.

The log-likelihood of $\mathbf{Y}^*$, $\log[L^*(\alpha)]$, differs from the log-likelihood $L(\hat{\beta},\alpha)$ only through a constant, which does not depend on α, and

$$-\frac{1}{2}\log\left|\sum_{i=1}^{n} X_i^T V_i^{-1}(t_i;\alpha)X_i\right|,$$

which does not depend on β. Because both REML and ML methods are based on the likelihood principle, they all have important theoretical properties such as consistency, asymptotic normality and asymptotic efficiency. In practice, neither one is uniformly superior to the other for all the situations. Their numerical values are also computed from different algorithms. For the ML method, the fixed effects and the variance components are estimated simultaneously, while for the REML method, only the variance components are estimated.

2.2.4. *Inferences*

The results established in the previous sections are useful to construct inference procedures for β. We mention here only a few special cases. A more complete account of inferential and diagnostic tools may be found in Diggle,[8] Zeger, Liang and Albert,[43] Diggle, Liang and Zeger[8] or Vonesh and Chinchilli,[35] among others.

Suppose that we have a consistent estimator $\hat{\alpha}$ of α, which may be either the ML estimator $\hat{\alpha}_{ML}$ or the REML estimator $\hat{\alpha}_{REML}$. Substituting α of (12) with $\hat{\alpha}$, the distribution of $\hat{\beta}(\hat{\alpha})$ can be approximated, at least when n

is large, by

$$\hat{\beta}(\hat{\alpha}) \sim \mathbf{N}(\beta, \hat{V}) , \tag{16}$$

where $\hat{V} = [\sum_{i=1}^{n}(X_i^T V_i^{-1}(t_i;\hat{\alpha})X_i)]^{-1}$. Suppose that C is a known $[r \times (k+1)]$ matrix with full rank. It follows immediately from (16) that, when n is sufficiently large, the distribution of $C\hat{\beta}(\hat{\alpha})$ can be approximated by

$$C\hat{\beta}(\hat{\alpha}) \sim \mathbf{N}(C\beta, C\hat{V}C^T) . \tag{17}$$

Consequently, an approximate $100\times(1-a)\%, 0 < a < 1$, confidence interval for $C\beta$ can be given by

$$C\hat{\beta}(\hat{\alpha}) \pm Z_{1-a/2}(C\hat{V}C^T)^{1/2} .$$

Taking C to be the $(k+1)$ row vector with 1 at its lth place and zero elsewhere, an approximate $100 \times (1-a)\%$ confidence interval for β_l can be given by

$$\hat{\beta}_l(\hat{\alpha}) \pm Z_{1-a/2}\sqrt{\hat{V}_l} , \tag{18}$$

where $\hat{V}_l$ is the lth diagonal element of $\hat{V}$.

The approximation in (17) can also be used to construct test statistics for linear statistical hypotheses. For example, suppose that we would like to test the null hypothesis of $C\beta = \theta_0$ for a known vector θ_0 against the general alternative that $C\beta \neq \theta_0$. A natural test statistic would be

$$\hat{T} = [C\hat{\beta}(\hat{\alpha}) - \theta_0]^T (C\hat{V}C^T)^{-1}[C\hat{\beta}(\hat{\alpha}) - \theta_0] , \tag{19}$$

which has approximately a χ^2-distribution with r degrees of freedom, denoted by χ_r^2, under the null hypothesis. A level $(100 \times a)\%$ test based on $\hat{T}$ then rejects the null hypothesis when $\hat{T} > \chi_r^2(a)$ with $\chi_r^2(a)$ being the $[100 \times (1-a)]$th percentile of χ_r^2. For the special case of testing $\beta_l = 0$ versus $\beta_l \neq 0$, a simple procedure equivalent to (19) is to reject the null hypothesis when

$$|\hat{\beta}_l(\hat{\alpha})| > Z_{1-a/2}\sqrt{\hat{V}_l} ,$$

where $Z_{1-a/2}$ and $\hat{V}_l$ are defined in (18).

3. Partially Linear Models

As discussed in Sec. 1.3.3, this class of models has been studied by Zeger and Diggle[42] and Moyeed and Diggle[26] as a means to generalize the

marginal linear models. With further restrictions on the error process, (2) is equivalent to

$$Y(t) = \beta_0(t) + \sum_{l=1}^{k} \beta_l X^{(l)}(t) + \epsilon(t)\,, \tag{20}$$

where $\epsilon(t)$ is a mean zero stochastic process with variance σ^2 and correlation function $\rho(t)$, and $X^{(l)}(t)$, $l = 1, \ldots, k$, and $\epsilon(t)$ are independent. The errors $\epsilon_i(t_{ij})$ specified in (2) are then independent copies of $\epsilon(t)$. A useful way to view $\epsilon_i(t_{ij})$ is through the decomposition

$$\epsilon_i(t_{ij}) = W_i(t_{ij}) + Z_{ij}\,, \tag{21}$$

where $W_i(t)$ are independent copies of a mean zero stationary process $W(t)$ with covariance function $\sigma_W^2\rho(t)$ and Z_{ij} are independent identically distributed measurement errors with mean zero and variance σ_Z^2. The covariance structure of the measurements Y_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ are

$$\mathrm{Cov}(Y_{i_1 j_1}, Y_{i_2 j_2}) = \begin{cases} \sigma_Z^2 + \sigma_W^2\,, & \text{if } i_1 = i_2 \text{ and } j_1 = j_2\,, \\ \sigma_W^2 \rho(t_{i_1 j_1} - t_{i_2 j_2})\,, & \text{if } i_1 = i_2 \text{ and } j_1 \neq j_2\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{22}$$

Although the above models can be classified as a special case of (3), a class of the structural nonparametric models to be discussed in later sections, their estimation methods are quite different, a fact owing to the structural differences between these two classes of models. The rest of this section focuses on an iteration procedure for the estimation of $\beta_0(t), \beta_1, \ldots, \beta_k$. Inferential and alternative estimation methods, which constitute some major research activities in longitudinal analyses, are still not well-understood and warrant considerable effort in further investigation.

3.1. *Smoothing estimators for the mean response*

Suppose for the moment that no covariate other than time is considered in modeling the mean response. The model (20) then reduces to

$$Y(t) = \beta_0(t) + \epsilon(t)\,. \tag{23}$$

Equivalently, with $\epsilon(t)$ defined in (20), $\beta_0(t)$ is the mean response of $Y(t)$ conditioning on time t; that is, $\beta_0(t) = E[Y(t)|t]$.

A natural approach for estimating $\beta_0(t)$ nonparametrically is to borrow smoothing techniques from the classical independent identically distributed

(i.i.d.) setting, while evaluating the statistical performances of the resulting estimators by taking the influences of the intra-subject correlations into account. A simple method is to use kernel smoothing, which amounts to estimate $\beta_0(t)$ through a weighted average using the measurements obtained within a neighborhood of t defined by a kernel function. Let $K(u)$ be a continuous kernel function, usually a continuous probability density function, defined on the real line, and h a positive bandwidth sequence which shrinks to zero as n tends to infinity. A kernel estimator similar to the well-known Nadaraya–Watson type kernel estimators in the i.i.d. setting is

$$\hat{\beta}_0^K(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{Y_{ij}K[(t-t_{ij})/h]\}}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}K[(t-t_{ij})/h]}\,. \tag{24}$$

Here, (24) uses uniform weight on each measurement, hence, makes no distinction between the subjects that have unequal numbers of repeated measurements. Thus subjects with more repeated measurements are used more often than those with fewer repeated measurements. A general formulation is to assign a specific weight to each subject and estimate $\beta_0(t)$ by

$$\hat{\beta}_0^K(t;w) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{Y_{ij}w_iK[(t-t_{ij})/h]\}}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_iK[(t-t_{ij})/h]\}}\,, \tag{25}$$

where the weights, $w = (w_1, \ldots, w_n)$, satisfy $w_i \geq 0$ for all $i = 1, \ldots, n$ with strict inequality for some $1 \leq i \leq n$. Clearly, (25) reduces to (24) when $w_i = 1/N$. An intuitive weight choice other than $w_i = 1/N$ is to uniformly weight each subject, rather than each measurement, so that the resulting kernel estimator is (25) with $w_i = 1/(nn_i)$.

Other approaches for the estimation of (22) have also been studied.[2,14,15,27,31] We omit these methods here and refer to their original articles for details. These methods, including (25) and the above alternative approaches, are essentially based on the fundamental spirit of local smoothing, hence, often lead to similar results in practice. This is in contrast to the smoothing methods to be discussed in the next section, where, because of the model complexity, different smoothing methods often produce very different results.

A crucial step in obtaining an adequate kernel estimator for $\beta_0(t)$ is to select an appropriate bandwidth h, while the choices of kernel functions are relatively less important. For estimation methods other than kernel smoothing, such as splines, this amounts to selecting an appropriate smoothing parameter. Rice and Silverman[31] suggested a simple cross-validation for selecting a data-driven smoothing parameter which does not depend on

the intra-subject correlation structures of the data. Applying their cross-validation to the kernel estimator (25), we first define $\hat{\beta}_0^{(-i,K)}(t;w)$ to be the estimator computed using (25) and the remaining data after deleting the entire set of repeated measurements of the ith subject. Predicting the ith subject's outcome at time t by $\hat{\beta}_0^{(-i,K)}(t;w)$, the cross-validation score of (25) is

$$\mathrm{CV}(h) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \{w_i [Y_{ij} - \hat{\beta}_0^{(-i,K)}(t_{ij};w)]^2\}. \tag{26}$$

Suppose that (26) can be uniquely minimized. The "leave-one-subject-out" cross-validated bandwidth h_{cv} is the minimizer of (26). Heuristically, the use of h_{cv} can be justified because, by minimizing (26), it approximately minimizes an average prediction error of (25). More details for the implementations and generalizations of this cross-validation will be discussed in Sec. 4.7.

Direct calculation of (26) can often be time consuming, as the algorithm repeats itself each time a new subject is deleted. Denote $K_{ij} = K[(t - t_{ij})/h]$,

$$K_{ij}^* = \frac{w_i K[(t-t_{ij})/h]}{\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i K[(t-t_{ij})/h]} \quad \text{and} \quad K_i^* = \sum_{j=1}^{n_i} K_{ij}^*$$

for $i = 1, \ldots, n$. A computationally simpler approach, also suggested by Rice and Silverman,[31] is to compute $[Y_{ij} - \hat{\beta}_0^{(-i,K)}(t_{ij};w)]$ using the following expression:

$$\begin{aligned} Y_{ij} - \hat{\beta}_0^{(-i,K)}(t_{ij};w) &= Y_{ij} - \left[\hat{\beta}_0^K(t_{ij};w) - \sum_{j=1}^{n_i}(Y_{ij}K_{ij}^*)\right]\left(1 + \frac{K_i^*}{1-K_i^*}\right) \\ &= [Y_{ij} - \hat{\beta}_0^K(t_{ij};w)] + \sum_{j=1}^{n_i}(Y_{ij}K_{ij}^*) \\ &\quad - \left[\hat{\beta}_0^K(t_{ij};w) - \sum_{j=1}^{n_i}(Y_{ij}K_{ij}^*)\right]\left(\frac{K_i^*}{1-K_i^*}\right) \\ &= [Y_{ij} - \hat{\beta}_0^K(t_{ij};w)] + \left(\frac{K_i^*}{1-K_i^*}\right) \\ &\quad \times \left[\frac{\sum_{j=1}^{n_i}(Y_{ij}K_{ij}^*)}{K_i^*} - \hat{\beta}_0^K(t_{ij};w)\right]. \end{aligned} \tag{27}$$

The above expression, as currently stated, is specifically targeted to kernel estimators defined in (25). When other smoothing methods, such as splines, are used, we may not get an explicit expression as the right side of (27), hence, direct calculation of (26) has to be carried out by deleting the subjects one at a time.

Large sample inferences of $\hat{\beta}_0^K(t;w)$ can be derived based on the asymptotic expressions of its means and variances and its asymptotic distributions. Because $\hat{\beta}_0^K(t;w)$ is a linear statistic of Y_{ij}, its means and variances can be directly computed and, consequently, its asymptotic distributions can be easily established by checking the triangular array central limit theorem after taking the intra-subject correlations into account; see, for example, Wu, Chiang and Hoover[39] and Wu and Chiang.[38] Because $\hat{\beta}_0^K(t;w)$ is a special case of the kernel estimators of Sec. 4, details of pointwise and simultaneous inferences for $\beta_0(t)$ are discussed in Sec. 5.1.

3.2. *Estimation of covariate effects*

With covariates other than time entered into the model, the estimation of $(\beta_0(t), \beta_1, \ldots, \beta_k)$ can be proceeded by an iteration that combines smoothing with parametric estimation techniques. Suppose that the error terms $\epsilon_i(t)$ of (21) have known variance-covariance matrices $V_i(t_i)$ for $t_i = (t_{i1}, \ldots, t_{in_i})$ and all $i = 1, \ldots, n$. The iteration can be proceeded as follows:

(a) Set $\beta_0(t)$ to zero and calculate an initial estimate of $(\beta_1, \ldots, \beta_k)^T$ using (10), an expression also for the generalized least squares, with $V_i(t_i;\alpha)$ replaced by $V_i(t_i)$.
(b) Based on the current estimate $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$, calculate the residual $r_{ij} = Y_{ij} - \sum_{l=1}^k \hat{\beta}_l X_{ij}^{(l)}$ and compute the kernel estimator $\hat{\beta}_0^K(t;w)$ of $\beta_0(t)$ using (24) with Y_{ij} replaced by r_{ij}.
(c) Based on the current kernel estimator $\hat{\beta}_0^K(t;w)$, calculate the residual $r_{ij} = Y_{ij} - \hat{\beta}_0^K(t_{ij};w)$ and update the estimate of $(\beta_1, \ldots, \beta_k)$ using (10) with $(V_i(t_i;\alpha), Y_{ij})$ replaced by $(V_i(t_i), r_{ij})$.
(d) Repeat steps (b) and (c) until the estimates converge.

This algorithm is a special case of the more general backfitting algorithm described in Hastie and Tibshirani.[17]

The assumption of having a known correlation structure is unrealistic and can be relaxed. Although an incorrectly specified correlation structure may cost the efficiency of the estimators, it generally does not affect

the consistency. When the variance-covariance matrix is parametrized by a parameter α and the error terms are from a mean zero Gaussian stationary process, the above iteration algorithm can be used in conjunction with the likelihood and restricted likelihood methods of the previous section, i.e. the generalized least squares estimators used in Steps (a) and (c) can be replaced by the likelihood based estimators $\hat{\beta}_{ML}$ or $\hat{\beta}(\hat{\alpha}_{REML})$. Further computational details, statistical properties of the resulting estimators and a modified estimation procedure can be found in Zeger and Diggle[42] and Moyeed and Diggle.[26] Inferences based on the resulting estimators have not been systematically investigated, hence, warrants substantial further development.

4. Smoothing for Varying-Coefficient Models

We present in this section a series of different smoothing methods for estimating the coefficient curves $\beta(t) = (\beta_0(t), \ldots, \beta_k(t))^T$ of (3). Inferences based on smoothing estimators of $\beta(t)$ will be discussed in Sec. 5.

4.1. *Some useful expressions*

In observational studies, the covariates are usually random as the subjects are randomly chosen, although they could in principle be either random or fixed. For generality, we assume throughout that $X(t)$ is random and the matrix $E[X(t)X^T(t)] \equiv E_{XX^T}(t)$ exist. With a proper change of the notation, our methods can be modified to accommodate the case of nonrandom covariates. An equivalent expression of (3) is then

$$Y(t) = X^T(t)\beta(t) + \epsilon(t)\,, \tag{28}$$

where $\epsilon(t)$ is a mean zero stochastic process and $\epsilon(t)$ and $X(t)$ are independent. Suppose that $E_{XX^T}(t)$ is invertible and its inverse is $E^{-1}_{XX^T}(t)$. It directly follows from (28) that $\beta(t)$ uniquely minimizes the second moment of $\epsilon(t)$ in the sense that

$$E\{[Y(t) - X^T(t)\beta(t)]^2\} = \inf_{\text{all } b(\cdot)} E\{[Y(t) - X^T(t)b(t)]^2\}\,, \tag{29}$$

and is given by

$$\beta(t) = E^{-1}_{XX^T}(t)E[X(t)Y(t)]\,. \tag{30}$$

When the covariates are time-invariant, we have $X(t) \equiv X$ and $E_{XX^T}(t) \equiv E_{XX^T}$, so that Eq. (30) reduces to

$$\beta_r(t) = E\left[\left(\sum_{l=0}^{k} e_{rl} X^{(l)}\right) Y(t)\right], \tag{31}$$

where e_{rl} is the element of $E_{XX^T}^{-1}$ at the rth row and lth column.

4.2. *Smoothing based on least squares*

4.2.1. *General formulation*

Intuitively, (29) suggests that $\beta(t)$ can be estimated by a method of local least squares using the measurements observed within a neighborhood of t. Assume that, for each l and some integer $p \geq 0$, $\beta_l(t)$ is p times differentiable and its pth derivative is continuous. Approximating $\beta_l(t_{ij})$ by a pth order polynomial $\sum_{r=0}^{p}\{b_{lr}(t)(t_{ij}-t)^r\}$ for all $l = 0, \ldots, k$, a local polynomial estimator of $\beta(t) = (\beta_0(t), \ldots, \beta_k(t))^T$ based on a kernel neighborhood is $\hat{b}_0(t) = (\hat{b}_{00}(t), \ldots, \hat{b}_{k0}(t))^T$, where $\{\hat{b}_{lr}(t); l = 0, \ldots, k, r = 0, \ldots, p\}$ minimizes

$$L_p(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i \left\{ Y_{ij} - \sum_{l=0}^{k}\left[X_{ij}^{(l)} \left(\sum_{r=0}^{p} b_{lr}(t)(t_{ij}-t)^r \right) \right] \right\}^2 \times K\left(\frac{t_{ij}-t}{h}\right), \tag{32}$$

where w_i are the non-negative weights as in (25), $K(\cdot)$ is a kernel function, usually chosen to be a probability density function, and h is a non-negative bandwidth. As a by-product of (32), $(r!)\hat{b}_{lr}(t)$ may be used to estimate the rth derivative $\beta_l^{(r)}(t)$ of $\beta_l(t)$, $r = 1, \ldots, p$.

4.2.2. *Least squares kernel estimators*

The simplest case of (32) is the ordinary least squares kernel estimator, also known as the local constant fit, obtained by minimizing (32) with $p = 0$. Using the matrix representation $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$,

$$X_i = \begin{pmatrix} 1 & X_{i1}^{(1)} & \cdots & X_{i1}^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{in_i}^{(1)} & \cdots & X_{in_i}^{(k)} \end{pmatrix} \quad \text{and} \quad K_i(t) = \begin{pmatrix} K_{i1} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & K_{in_i} \end{pmatrix}$$

with $K_{ij} = K[(t_{ij} - t)/h]$, if $\sum_{i=1}^{n} X_i^T K_i(t) X_i$ is invertible, then (32) with $p = 0$ can be uniquely minimized and its minimizer, the kernel estimator of $\beta(t)$, is given by

$$\hat{\beta}^{LSK}(t) = \left(\sum_{i=1}^{n} w_i X_i^T K_i(t) X_i \right)^{-1} \left(\sum_{i=1}^{n} w_i X_i^T K_i(t) Y_i \right). \tag{33}$$

When the model incorporates no covariate other than time, i.e. $k = 0$, (33) reduces to a Nadaraya–Watson type kernel estimator of the conditional expectation $E[Y(t)|t]$; for example, Härdle.[12]

4.2.3. *Least squares local linear estimators*

Although (33) has a simple mathematical expression, it often leads to significant bias when t is at the boundary of its support. An automatic procedure to reduce such boundary bias is to use higher order local polynomial fits. But, a high order local polynomial fit can be impractical in some applications because it usually requires large sample sizes and may be computationally intensive. A practical approach that provides automatic boundary bias adjustment is to use local linear fit that minimizes (32) with $p = 1$. Denote

$$\mathcal{N}_{lr} = \begin{pmatrix} \sum_{i,j} [w_i X_{ij}^{(l)} X_{ij}^{(r)} K_{ij}] & \sum_{i,j} [w_i X_{ij}^{(l)} X_{ij}^{(r)} (t_{ij} - t) K_{ij}] \\ \sum_{i,j} [w_i X_{ij}^{(l)} X_{ij}^{(r)} (t_{ij} - t) K_{ij}] & \sum_{i,j} [w_i X_{ij}^{(l)} X_{ij}^{(r)} (t_{ij} - t)^2 K_{ij}] \end{pmatrix},$$

$\mathcal{N}_r = (\mathcal{N}_{0r}, \ldots, \mathcal{N}_{kr})$, $\mathcal{N} = (\mathcal{N}_0^T, \ldots, \mathcal{N}_k^T)^T$,

$$\mathcal{M}_r = \left(\sum_{i,j} [w_i X_{ij}^{(r)} Y_{ij} K_{ij}],\ \sum_{i,j} [w_i X_{ij}^{(r)} (t_{ij} - t) Y_{ij} K_{ij}] \right)^T,$$

$\mathcal{M} = (\mathcal{M}_0^T, \ldots, \mathcal{M}_k^T)^T$, $b_l(t) = (b_{l0}(t), b_{l1}(t))^T$ and $b(t) = (b_0(t), \ldots, b_k(t)^T$ for $r, l = 0, \ldots, k$. Setting the partial derivatives of $L_1(t)$ with respect to $b_{lr}(t)$ to zero, the normal equation of (32) with $p = 1$ is

$$\mathcal{N} b(t) = \mathcal{M}. \tag{34}$$

Suppose that the matrix $\mathcal{N}$ is invertible at t. The solution of (34) exists and is uniquely given by $\hat{b}(t) = \mathcal{N}^{-1}\mathcal{M}$. The least squares local linear estimator $\hat{\beta}_l^{LSL}(t)$ of $\beta_l(t)$ is then

$$\hat{\beta}_l^{LSL}(t) = e_{2l+1}^T \hat{b}(t), \tag{35}$$

where e_q is the $[2(k+1)\times 1]$ column vector with 1 at its qth place and zero elsewhere. Explicit expressions for the general higher order least squares local polynomial estimators can be similarly derived; see Hoover *et al.*[18] We omit the details of these general higher order estimators, as a local linear fitting is sufficiently satisfactory in almost all the biomedical studies that have appeared in the literature.

4.2.4. *Least squares with centered covariates*

In some situations, some of the covariates used in (28) can not have values at zero, so that the baseline coefficient curve $\beta_0(t)$ does not have a practical interpretation. Strictly positive covariates appear naturally both in the ASGA Study (Sec. 1.2.1), such as the mother's placental thickness and pre-pregnancy height, and the HIV/CD4 Depletion Data (Sec. 1.2.2), such as the subject's pre-infection CD4 level. A useful remedy when such a situation arises is to use a centered version of the covariates in the model, so that the corresponding baseline coefficient can be interpreted as the conditional mean of $Y(t)$ when the centered covariates are set to zero.

Let $X^{(*l)}(t) = X^{(l)}(t) - E[X^{(l)}(t)]$ be the centered version of $X^{(l)}(t)$ and $X^{(*)}(t)$ be the covariate vector with some or all of its components being centered. An equivalent form of (28) is

$$Y(t) = (X^{(*)}(t))^T \beta^*(t) + \epsilon(t)\,, \tag{36}$$

where $\beta^*(t) = (\beta_0^*(t), \beta_1(t), \ldots, \beta_k(t))^T$. Note that $\beta_0^*(t)$, the baseline coefficient curve of (36), represents the mean of $Y(t)$, when $X^{(*l)}(t)$, rather than $X^{(l)}(t)$, for $l = 1, \ldots, k$ are set to zero. Other coefficient curves of (36) can be interpreted the same way as those of (28).

The estimation of $\beta^*(t)$ can be obtained by first estimating the centered covariates $X_{ij}^{(*l)}$ of $X_{ij}^{(l)}$ and then minimizing (32) with $X_{ij}^{(l)}$ replaced by $X_{ij}^{(*l)}$. If $X^{(l)}(t)$ is a time-dependent covariate, then, using a kernel smoothing, a centered version of $X_{ij}^{(l)}$ can be estimated by $X_{ij}^{(*l)} = X_{ij}^{(l)} - \hat{\mu}_l(t_{ij})$ with

$$\hat{\mu}_l(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \{w_i X_{ij}^{(l)} \Gamma_l[(t-t_{ij})/\gamma_l]\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} \{w_i \Gamma_l[(t-t_{ij})/\gamma_l]\}}\,, \tag{37}$$

where $(\Gamma_l(\cdot), \gamma_l)$ is a set of kernel and bandwidth. On the other hand, if $X^{(l)}(t) \equiv X^{(l)}$ is time-invariant, then $X_{ij}^{(l)} \equiv X_i^{(l)}$ for all $j = 1, \ldots, n_i$, and $X_i^{(*l)}$ can be taken as $X_i^{(l)} - \bar{X}^{(l)}$, where $\bar{X}^{(l)} = n^{-1} \sum_{i=1}^n X_i^{(l)}$ is the

weighted sample mean for $X^{(l)}$. Let $X_i^{(*)}$ be the $n_i \times (k+1)$ centered covariate vector whose jth row is $(1, X_{ij}^{(*1)}, \ldots, X_{ij}^{(*k)})$. A least squares kernel estimator of $\beta^*(t)$ is

$$\hat{\beta}^{*LSK}(t) = \left[\sum_{i=1}^{n} w_i (X_i^{(*)})^T K_i(t) (X_i^{(*)})^T\right]^{-1} \left[\sum_{i=1}^{n} w_i (X_i^{(*)})^T K_i(t) Y_i\right], \tag{38}$$

where $K_i(t)$ and Y_i are defined as in (33).

Wu, Yu and Chiang[40] investigated the large sample properties of $\hat{\beta}^{*LSK}(t)$. Their results suggest that neither $\hat{\beta}^{LSK}(t)$ nor $\hat{\beta}^{*LSK}(t)$ is uniformly superior to the other. In particular, when the covariates are time-invariant, $\hat{\beta}^{LSK}(t)$ and $\hat{\beta}^{*LSK}(t)$ are asymptotically equivalent. However, when $X^{(l)}(t)$ for $l \geq 1$ changes significantly with t, theoretically and practically superior estimators of $\beta_l(t)$ may be obtained by centering $X^{(l)}(t)$.

Of course, after a covariate is centered, the baseline coefficient curve of the model is changed. The decision on whether a covariate should be centered or not primarily depends on the biological interpretations of the corresponding baseline coefficient curve. Such a decision should be made based on the statistical properties of the estimators only if the effects of the covariates, rather than the baseline coefficient curve, is of primary interest in the investigation. Clearly, methods other than kernel smoothing may also be applied to the estimation with centered covariates. But, because of the complication caused by smoothing the covariates, statistical properties for estimators other than (38) have not been investigated in the literature.

4.2.5. *A simple modification*

The estimators mentioned above, both with and without covariate centering, rely on a single bandwidth to estimate all $(k+1)$ coefficient curves. This simple approach may work well when all the curves roughly belong to the same smoothness family. However, such an idealized scenario is often not anticipated in practice. A flexible method which automatically adjusts for the possibly different smoothing needs for different coefficient curves is always preferred.

In the literature, the potential deficiency associated with the use of a single bandwidth has been reported.[10,18,40] These authors have also proposed a number of alternative approaches (see Secs. 4.3–4.6) to overcome

this potential drawback. A simple method suggested by Wu, Yu and Chiang[40] is to use a linear combination of the form

$$\hat{\beta}(t; \mathbf{K}, \mathbf{h}) = \sum_{l=0}^{k} e_{l+1}^T \hat{\beta}(t; K_l, h_l)\,, \tag{39}$$

where $\mathbf{K}(\cdot) = (K_0(\cdot), \ldots, K_k(\cdot))$, $\mathbf{h} = (h_0, \ldots, h_k)$, e_p is the $[(k+1) \times 1]$ vector with 1 at its pth place and zero elsewhere and $\hat{\beta}(t; K_l, h_l)$ is the kernel estimator of $\beta(t)$ or $\beta^*(t)$ obtained from (33) or (38), respectively, using kernel $K_l(\cdot)$ and bandwidth h_l. Intuitively, $\hat{\beta}(t; \mathbf{K}, \mathbf{h})$ relies on a specific pair of kernel and bandwidth to estimate the corresponding component of $\beta(t)$ or $\beta^*(t)$. As a general methodology, (39) is not limited to kernel estimators and may be applied to other local polynomial estimators as well.

4.2.6. *Choices of w_i*

An important factor that affects the theoretical and practical behaviors of the least squares local polynomial estimators of $\beta(t)$ is the choice of w_i in (32). For cross-sectional studies with independent identically distributed data, a uniform weight choice, $w_i \equiv 1/N$, is often desirable. For the current sampling, it is conceivable that a proper choice of w_i may depend on the intra-subject correlation structures and the numbers of repeated measurements n_i. In practice, however, the correlation structures of the data are often completely unknown and may be difficult to estimate, so that subjective choices such as $w_i = 1/N$ and $w_i = 1/(nn_i)$ are often considered. Intuitively, $w_i = 1/N$ assigns equal weight to each observation point, while $w_i = 1/(nn_i)$ assigns equal weight to each subject. Theoretically, the choice of $w_i = 1/N$ may produce inconsistent least squares kernel estimators when some n_i are much larger than the others. On the other hand, the least squares kernel estimators based on $w_i = 1/(nn_i)$ are always consistent regardless the choices of n_i.[18,38]

4.3. *Penalized least squares*

Suppose that all the components of $\beta(t)$ are twice continuously differentiable and have bounded and square integrable second derivatives with respect to t. A natural penalized least squares criterion is to minimize

$$J(\beta, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ Y_{ij} - \sum_{l=0}^{k} X_{ij}^{(l)} \beta_l(t_{ij}) \right\}^2 + \sum_{l=0}^{k} \lambda_l \int [\beta_l''(t)]^2 dt \tag{40}$$

with respect to $\beta_l(t)$, where $\lambda = (\lambda_0, \ldots, \lambda_k)^T$ and λ_l are positive smoothing parameters. The existence and uniqueness of the minimizer of (40) depend on t_{ij} and $X_{ij}^{(l)}$. Suppose that (40) can be uniquely minimized. The penalized least squares estimator $\hat{\beta}^{PLS}(t) = (\hat{\beta}_0^{PLS}(t), \ldots, \hat{\beta}_k^{PLS}(t))^T$ of $\beta(t)$ is then defined to be the unique minimizer of (40). Using similar techniques as in univariate smoothing, it can be shown that $\hat{\beta}_l^{PLS}(t)$ are natural cubic splines with knots at the distinct values of $\{t_{ij} : i = 1, \ldots, n, j = 1, \ldots, n_i\}$ and can be expressed as linear functions of $\{Y_{ij} : i = 1, \ldots, n, j = 1, \ldots, n_i\}$.

One feature that distinguishes $\hat{\beta}^{PLS}(t)$ from the estimators obtained from (32) is the use of multiple smoothing parameters λ_l in the penalty term. In (40), all $(k + 1)$ smoothing parameters λ_l, $l = 0, \ldots, k$, can be adjusted in the penalty term. Numerical results presented in Hoover *et al.*[18] demonstrated that the extra flexibility created by multiple smoothing parameters could indeed lead to better estimators than the least squares local polynomials that rely on a single smoothing parameter. However, because $\hat{\beta}^{PLS}(t)$ has knots at all the distinct time points, it can be extremely computationally intensive when the number of distinct time points is large, a case often happened in unbalanced longitudinal studies.

4.4. *A two-step method*

In an attempt to provide flexible smoothing estimators that are computationally accessible with large longitudinal data, Fan and Zhang[10] proposed to estimate $\beta(t)$ by a two-step smoothing method which uses $(k+1)$ smoothing parameters in a different way from (39) and (40). Their procedure calls for the following two steps:

(i) computing the raw estimates $\hat{\beta}^{RAW}(s)$ of $\beta(s)$ at a set of distinct time points, say $s_1, \ldots, s_m$, where m may depend on n and n_i, $i = 1, \ldots, n$;
(ii) estimating each coefficient curve $\beta_l(t)$ by smoothing the raw estimates $\hat{\beta}_l^{RAW}(s_r)$, $r = 1, \ldots, m$.

Although Fan and Zhang[10] used local polynomials to illustrate the method, other smoothing methods such as splines may in principle be used.

For the special case of balanced longitudinal data where all the subjects are observed at a same set of time points $\{s_j; j = 1, \ldots, m\}$ with $m = n_i$, $i = 1, \ldots, n$, the raw estimates can be computed by fitting linear models between Y_{ij} and X_{ij} at s_j for all $j = 1, \ldots, m$. However, when the design is unbalanced and the numbers of subjects on some time points are sparse, as in most practical situations, it may be necessary to computing the raw

estimates by grouping the observations from the adjacent time points. In particular, we can first compute $\hat{\beta}_l^{RAW}(s_r)$, $l = 0, \ldots, k$, using the local polynomial method (32) with a small bandwidth, and then, treating $\hat{\beta}_l^{RAW}(s_r)$ as the new data, estimate $\beta_l(t)$ by minimizing

$$L_{p,l}^{TS}(t) = \sum_{j=1}^{m} \left\{ \hat{\beta}_l^{RAW}(s_j) - \sum_{r=0}^{p} b_{lr}(t)(s_j - t)^r \right\}^2 K_l\left(\frac{s_j - t}{h_l}\right) \tag{41}$$

with respect to $b_{lr}(t)$, where $(K_l(\cdot), h_l)$ is a set of kernel and bandwidth. Similar to (32), if $\hat{b}_{lr}^{TS}(t)$ for $r = 0, \ldots, k$ uniquely minimize (41), $\hat{b}_{l0}^{TS}(t)$ is the two-step pth order local polynomial estimator of $\beta_l(t)$, while $(r!)\hat{b}_{lr}^{TS}(t)$ can be used to estimate the rth derivative of $\beta_l(t)$.

In contrast to the estimators obtained from (32) where a single bandwidth must be used for all $\beta_l(t)$, the two-step method has in principle the flexibility to adjust for the specific smoothing need of each coefficient curve. However, a main difficulty in current version of two-step smoothing is that it lacks a specific and practical guideline to construct the raw estimates for unbalanced longitudinal data. Certain data-driven bandwidth procedures would be desirable for computing both the raw and the final estimates. Impacts of different raw estimates on the theoretical and practical properties of the final two-step estimators are still not well-understood and require substantial further development.

4.5. *Smoothing with time-invariant covariates*

When the covariates of interest are time-invariant, such as in clinical trials when the treatments are kept fixed throughout the study periods, an effective way motivated by (30) to provide flexible and computational feasible estimators of $\beta(t)$ is to smooth each component of $\beta(t)$ separately.

Let $Z^{(r)}(t) = [\sum_{l=0}^{k} e_{rl} X^{(l)}] Y(t)$, $X_i = (1, X_i^{(1)}, \ldots, X_i^{(k)})^T$ be the covariate vector of the ith subject and $\hat{e}_{rl}$ be the (r, l)th element of the matrix $(\hat{E}_{XX^T})^{-1}$, the inverse of the sample mean $\hat{E}_{XX^T} = (1/n)\sum_{i=0}^{n} X_i X_i^T$. A natural estimator of $Z^{(r)}(t)$ is $Z_{ij}^{(r)} = [\sum_{l=0}^{k} \hat{e}_{rl} X_i^{(l)}] Y_{ij}$. By (30), a componentwise smoothing estimator of $\beta_r(t)$ can be obtained by smoothing $Z_{ij}^{(r)}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Specifically, a local polynomial estimator of $\beta_r(t)$ with order $p \geq 0$ is $\hat{b}_{r0}^{COM}(t)$, such that $\hat{b}_{rl}^{COM}(t)$, $l = 0, \ldots, p$, uniquely minimize

$$L_{p,r}^{COM}(t) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_i \left\{ Z_{ij}^{(r)} - \sum_{l=0}^{p} b_{rl}(t)(t_{ij} - t)^l \right\}^2 K_r\left(\frac{t_{ij} - t}{h_r}\right), \tag{42}$$

with respect to $b_{rl}(t)$. For the local constant fitting with $p = 0$, (42) leads to the componentwise kernel estimator

$$\hat{\beta}_r^{COM}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i Z_{ij}^{(r)} K_r[(t_{ij}-t)/h_r]\}}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i K_r[(t_{ij}-t)/h_r]\}}. \tag{43}$$

Wu and Chiang[38] established the large sample mean squared errors of $\hat{\beta}_r^{COM}(t)$, while Wu, Yu and Yuan[41] developed a procedure for constructing approximate asymptotic pointwise and simultaneous confidence regions for $\beta_r(t)$. These results shed some light on the asymptotic behaviors of the higher order estimators $\hat{b}_{r0}^{COM}(t)$, although specific asymptotic risks and asymptotic distributions have not been established for the case with $p \geq 1$. The results of Wu and Chiang[38] and Wu, Yu and Yuan[41] indicate some clear advantages of $\hat{\beta}_r^{COM}(t)$ over the kernel estimator (33) both in terms of theoretical convergence rates and practical flexibilities. Similar advantages over the least squares method of (32) are also expected for the componentwise local polynomial estimators.

Obviously, minimizing (42) is not the only componentwise smoothing approach. Suppose that the support of the design time points is contained in a compact set $[a, b]$ and $\beta_r(t)$ is twice differentiable with respect to t in $[a, b]$. A viable alternative is to estimate $\beta_r(t)$ by penalized least squares estimator $\tilde{\beta}_r^{COM}(t)$, where $\tilde{\beta}_r^{COM}(t)$ minimizes

$$J_r^{COM}(\beta_r, \lambda_r) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i[Z_{ij}^{(r)} - \beta_r(t_{ij})]^2\} + \lambda_r \int_a^b [\beta_r''(s)]^2 ds\,, \tag{44}$$

with λ_r being a non-negative smoothing parameter. By the same rationale as in Sec. 2.3, it is easy to verify that $\tilde{\beta}_r^{COM}(t)$ is a natural cubic spline with knots at the distinct points of $\{t_{ij}; i = 1, \ldots, n, j = 1, \ldots, n_i\}$. Furthermore, using the approach of equivalent kernels, Chiang, Rice and Wu[4] derived the asymptotic mean squared errors and the asymptotic distributions of $\tilde{\beta}_r^{COM}(t)$. In contrast to the multiple penalized least squares of (40) whose solution is obtained by solving a large linear system involving all $(k+1)$ components, (44) significantly simplifies the computation by solving $(k+1)$ separate linear systems. This computational advantage ensures the practical implementability of (44) in many situations, while the intensive computational needs often make the optimization of (40) impracticable.

4.6. *Smoothing via basis approximations*

All the smoothing methods described above depend on local smoothing in the sense that only the measurements obtained within some neighborhood of t are effectively used to estimate $\beta(t)$. Although local smoothing

works well when all the coefficient curves $\beta_r(t)$ are nonparametric, it is not adequate when some of the coefficient curves have known parametric forms, as in the partially linear model (2).

Compared with local smoothing, estimation using basis approximations has three important advantages. First, it can be used to estimate $\beta(t)$ whether its components are parametric or nonparametric, hence, is suitable for both nonparametric and semiparametric varying-coefficient models. Second, when a random effect is desired, it provides a natural means to incorporate random effects into a nonparametric or semiparametric model. Third, because popular basis estimators, such as truncated polynomials or B-splines, often rely on far fewer knots or approximation terms than smoothing splines, they often enjoy considerable computationally advantage over smoothing splines or even local polynomials. Although estimation with mixed effects is of great interest in various settings, we only discuss here the case of marginal models. Extension to mixed effects models can be found in Rice and Wu.[31]

The main idea is to first approximate $\beta_r(t)$ by a basis function expansion with K_r terms, where K_r may or may not tend to infinity as n tends to infinity, and then estimate $\beta_r(t)$ by estimating the coefficients of this expansion. For each $r = 0, \ldots, k$, let $B_{rs}(t)$, $s = 1, \ldots, K_r$, be a set of basis functions. If $\beta_r(t)$ can be approximated by an expansion based on $B_{rs}(t)$, $s = 1, \ldots, K_r$, there is a set of constants γ_{rs} so that

$$\beta_r(t) \approx \sum_{s=1}^{K_r} \gamma_{rs} B_{rs}(t)\,. \tag{45}$$

Substituting (45) into (3), an approximation of the varying-coefficient model is

$$Y_{ij} \approx \sum_{r=0}^{k} \sum_{s=1}^{K_r} X_{ij}^{(r)} \gamma_{rs} B_{rs}(t) + \epsilon_i(t_{ij})\,. \tag{46}$$

The approximation sign in (46) will be replaced by the equality sign if, for all $r = 0, \ldots, k$, $\beta_r(t)$ belongs to a linear space spanned by $\{B_{rs}(t); s = 1, \ldots, K_r\}$.

Using (46), the least squares estimators $\hat{\gamma}_{rs}$ of γ_{rs} can be obtained by minimizing

$$\ell(\gamma) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ w_i \left[Y_{ij} - \sum_{r=0}^{k} \sum_{s=1}^{K_r} (X_{ij}^{(r)} \gamma_{rs} B_{rs}(t_{ij})) \right]^2 \right\}, \tag{47}$$

where $\gamma = (\gamma_0^T, \ldots, \gamma_k^T)^T$ and $\gamma_r = (\gamma_{r1}, \ldots, \gamma_{rK_r})^T$. If the minimizer of (47) uniquely exists, the basis function estimator of $\beta_r(t)$ is

$$\hat{\beta}_r^{BAS}(t) = \sum_{s=1}^{K_r} [\hat{\gamma}_{rs} B_{rs}(t)] , \tag{48}$$

where K_r may depend on n and n_i, $i = 1, \ldots, n$. Clearly, if K_r is finite and known and $\beta_r(t)$ belongs to the linear model spanned by $B_{rs}(t)$, $s = 1, \ldots, K_r$, then (48) returns a parametric estimator of $\beta_r(t)$. On the other hand, if (45) holds with K_r unknown, a consistent nonparametric estimator produced by (48) may require K_r to be a function of n and n_i, $i = 1, \ldots, n$, which may tend to infinity as n tends to infinity.

Depending on the underlying scientific nature of the data, many different bases may be used to approximate the components of $\beta(t)$. The most popular basis system in the classical linear models is the polynomial basis $\{1, t, \ldots, t^{K_r-1}\}$. A general class of bases that have certain numerical advantages over the above polynomial basis is the class of piecewise polynomials. Examples of piecewise polynomial bases include B-spline bases, such as linear, quadratic or cubic splines, or other types of truncated power series; see de Boor[6] for further details of the explicit expressions of piecewise polynomials and their numerical properties. If $\beta_r(t)$ is believed to exhibit periodicity, Fourier series are often natural basis choices.

Huang, Wu and Zhou[19] recently established the consistency of (48) and studied the practical performance of (48) with B-splines through an intensive simulation. In general, a B-spline estimator requires a smoothing parameter consisted of three aspects: degrees of the polynomials and number and location of the knots. Although generally desired, it is difficult, however, to simultaneously determine all three of these aspects from the data. Rice and Wu[31] showed that the simple approach of using equally spaced knots often works well in practice, a finding also corroborated by the simulation of Huang, Wu and Zhou.[19]

4.7. *A cross-validation procedure*

The most important factor that affects all of the above smoothing methods is the selection of appropriate smoothing parameters, such as the bandwidth, the positive penalty weight λ and the number and location of knots. It is of both theoretical and practical interest to select these values directly from the data.

Selecting data-driven smoothing parameters for nonparametric regression with independent identically distributed data has been a subject of intense investigation in the literature. Under the current context, a widely used method, suggested by Rice and Silverman,[31] is a cross-validation that deletes the entire repeated measurements of a subject, rather than an individual measurement, one at a time. Hart and Wehrly[15] derived the consistency of this cross-validation for a simple nonparametric regression without the presence of covariates other than time. Without loss of generality, we denote ξ to be a vector of smoothing parameters, $\hat{\beta}(t;\xi)$ a smoothing estimator based on ξ and $\hat{\beta}^{(-i)}(t;\xi)$ an estimator computed using the same method as $\hat{\beta}(t;\xi)$ but with the ith subject's measurements deleted. The cross-validation score for $\hat{\beta}(t;\xi)$ is

$$\mathrm{CV}(\xi) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i[Y_{ij} - X_{ij}^T\hat{\beta}^{(-i)}(t;\xi)]^2\}\,, \tag{49}$$

which measures the predictive error of $\hat{\beta}(t;\xi)$. The cross-validated smoothing parameter ξ_{cv} is then the minimizer of $\mathrm{CV}(\xi)$, provided that the unique minimizer of $\mathrm{CV}(\xi)$ exists.

The above cross-validation criterion is directly applicable to all the smoothing methods presented above, except the two-step smoothing of Sec. 2.4. For the estimators of Secs. 2.2, 2.3 and 2.5 and B-splines with equally spaced knots, minimizing the corresponding cross-validation scores would either return a univariate bandwidth or a R^{k+1}-valued vector. An automatic search of the global minima usually requires a sophisticated optimization software. In practice, particularly when the smoothing parameter is multivariate, it is often reasonable to use a smoothing parameter whose cross-validation score is close to the global minima.

There are three intuitive reasons to use the cross-validation criterion (49). First, by deleting the subjects one at a time, it preserves the correlation structure of the data. Second, in contrast to alternatives such as the AIC, the BIC and the generalized cross-validation,[1,32,33,36] (49) does not depend on the structure of the intra-subject correlations, hence, can be implemented in almost all the practical situations. Third, when the number of subjects is sufficiently large, minimizing (49) leads to a smoothing parameter that approximately minimizes the average squared error:

$$\mathrm{ASE}(\hat{\beta}(\cdot;\xi)) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i[X_{ij}^T(\beta(t_{ij}) - \hat{\beta}(t_{ij};\xi))]^2\}\,. \tag{50}$$

The last assertion can be heuristically seen by the decomposition:

$$\begin{aligned}\mathrm{CV}(\xi) = &\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i[Y_{ij} - X_{ij}^T\beta(t_{ij})]^2\} \\ &+ 2\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i[Y_{ij} - X_{ij}^T\beta(t_{ij})][X_{ij}^T(\beta(t_{ij}) - \hat{\beta}^{(-i)}(t_{ij};\xi))]^2\} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{w_i[X_{ij}^T(\beta(t_{ij}) - \hat{\beta}^{(-i)}(t_{ij};\xi))]^2\}\,. \qquad (51)\end{aligned}$$

Here, (50) and the definition of $\hat{\beta}^{(-i)}(t;\xi)$ imply that the third term at the right side of (51) is approximately the same as $\mathrm{ASE}(\hat{\beta}(\cdot;\xi))$. Because the first term at the right side of (51) does not depend on the smoothing parameter and the second term is approximately zero, ξ_{cv} approximately minimizes $\mathrm{ASE}(\hat{\beta}(\cdot;\xi))$.

5. Confidence Regions Based on Smoothing

Confidence statements can be made either based on the asymptotic distributions of the estimators or through a bootstrap procedure. Currently, explicit expressions of asymptotic distributions have only been developed for the kernel estimators (33) and (43). A bootstrap approach that has broader appeal in longitudinal analysis is to resample the subjects of the original data. Although its theoretical properties have not been well-understood, practical performances of this "resampling-subject" bootstrap have been investigated by a number of simulation studies. We present in this section both asymptotic and bootstrap approaches based on the smoothing estimators of Sec. 4.

5.1. *Asymptotic inferences for kernel estimators*

5.1.1. *Pointwise confidence intervals*

For both (33) and (43), their asymptotic distributions have been developed based on two important assumptions. First, the numbers of repeated measurements n_i are non-random and may or may not tend to infinity as n tending to infinity. Second, the time design points t_{ij} are random and independent identically distributed according to an unknown density function $f(\cdot)$. These assumptions are made for practical considerations as well as mathematical tractability.

We first consider the confidence procedures based on (33). Under the above assumptions and some additional mild regularity conditions, Wu, Chiang and Hoover[39] showed that, if $w_i = 1/N$, $h = N^{-1/5}h_0$ and

$$\lim_{n\to\infty} N^{-6/5}\sum_{i=1}^{n} n_i^2 = \theta$$

for some constants $h_0 > 0$ and $0 \le \theta < \infty$, $\hat{\beta}^{LSK}(t)$ has an asymptotically multivariate normal distribution in the sense that

$$(Nh)^{1/2}[\hat{\beta}^{LSK}(t) - \beta(t)] \to \mathbf{N}(B(t), D^*(t))\,, \tag{52}$$

in distribution as $n \to \infty$. The bias, $B(t)$, and the variance-covariance matrix, $D^*(t)$, of (52) are

$$B(t) = [f(t)]^{-1}E^{-1}_{XX^T}(t)(b_0(t), \ldots, b_k(t))^T \tag{53}$$

and

$$D^*(t) = [f(t)]^{-2}E^{-1}_{XX^T}(t)D(t)E^{1}_{XX^T}(t) \tag{54}$$

where $D(t)$ is a $(k+1)\times(k+1)$ matrix whose (l,r)th element is

$$D_{lr}(t) = \sigma^2(t)E[X^{(l)}(t)X^{(r)}(t)]f(t)\left\{\int [K(u)]^2du\right\}$$

$$+\theta h_0\rho_\epsilon(t)E[X^{(l)}(t)X^{(r)}(t)][f(t)]^2\,,$$

$\sigma^2(t) = E[\epsilon^2(t)]$, $\rho_\epsilon(t) = \lim_{a\to 0} E[\epsilon(t+a)\epsilon(t)]$ and

$$b_l(t) = h_0^{3/2}\sum_{c=0}^{k}\left\{\left[\int u^2K(u)du\right]\{\beta_c'(t)[E[X^{(l)}(t)X^{(c)}(t)]]'f(t)\right.$$

$$\left.+\beta_c'(t)E[X^{(l)}(t)X^{(c)}(t)]f'(t) + (1/2)\beta_c''(t)E[X^{(l)}(t)X^{(c)}(t)]f(t)\}\right\}.$$

Then, there are lower and upper end points $L_\alpha(t)$ and $U_\alpha(t)$ given by

$$\{A^T\hat{\beta}^{LSK}(t) - (Nh)^{-1/2}A^TB(t)\} \pm Z_{\alpha/2}(Nh)^{-1/2}[A^TD^*(t)A]^{1/2}\,, \tag{55}$$

where $Z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, so that

$$\lim_{n\to\infty} P\{L_\alpha(t) \le A^T\beta(t) \le U_\alpha(t)\} = 1-\alpha\,. \tag{56}$$

Because $B(t)$ and $D^*(t)$ depend on unknown quantities, (55) is not implementable in practice. If $B(t)$ and $D^*(t)$ can be consistently estimated by

$\hat{B}(t)$ and $\hat{D}^*(t)$, a pointwise $(1-\alpha)$ confidence interval for $A^T\beta(t)$ can be approximated by $(\hat{L}_\alpha(t), \hat{U}_\alpha(t))$ with $\hat{L}_\alpha(t)$ and $\hat{U}_\alpha(t)$ being the lower and upper end points given by

$$\{A^T\hat{\beta}^{LSK}(t) - (Nh)^{-1/2}A^T\hat{B}(t)\} \pm Z_{\alpha/2}(Nh)^{-1/2}[A^T\hat{D}^*(t)A]^{1/2}. \quad (57)$$

Wu, Chiang and Hoover[39] suggested to compute $\hat{B}(t)$ and $\hat{D}^*(t)$ by substituting $f(t)$, $\sigma^2(t)$, $\rho_\epsilon(t)$, $E[X^{(l)}(t)X^{(r)}(t)]$ and the required derivatives in (53) and (54) with their kernel estimators. Suppose that the kernel function $K(\cdot)$ is at least twice continuously differentiable in the interior of its support. These authors proposed to estimate $f(t)$, $\sigma^2(t)$, $\rho_\epsilon(t)$ and $E[X^{(l)}(t)X^{(r)}(t)]$ by

$$\hat{f}(t) = (Nh)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i} K\left(\frac{t_{ij}-t}{h}\right),$$

$$\hat{\sigma}^2(t) = \frac{1}{Nh\hat{f}(t)}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\hat{\epsilon}_i^2(t_{ij})K\left(\frac{t_{ij}-t}{h}\right)\right\},$$

$$\hat{\rho}_\epsilon(t) = \frac{\sum_{i=1}^{n}\sum_{j_1\neq j_2}\{\hat{\epsilon}_i(t_{ij_1})\hat{\epsilon}_i(t_{ij_2})K(\frac{t_{ij}-t}{h})K(\frac{t_{ij}-t}{h})\}}{\sum_{i=1}^{n}\sum_{j_1\neq j_2}\{K(\frac{t_{ij}-t}{h})K(\frac{t_{ij}-t}{h})\}}$$

and

$$\hat{E}[X^{(l)}(t)X^{(r)}(t)] = \frac{1}{Nh\hat{f}(t)}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{X_i^{(l)}(t_{ij})X_i^{(r)}(t_{ij})K\left(\frac{t_{ij}-t}{h}\right)\right\},$$

where $\hat{\epsilon}_i(t_{ij}) = Y_{ij} - X_i^T(t_{ij})\hat{\beta}(t_{ij})$ are the residuals, and to estimate the first and second derivatives of $f(t)$, $\beta_l(t)$ and $E[X^{(l)}(t)X^{(r)}(t)]$ by the corresponding derivatives of $\hat{f}(t)$, $\hat{\beta}_l^{LSK}(t)$ and $\hat{E}[X^{(l)}(t)X^{(r)}(t)]$. Through an intensive simulation, these authors also suggested that the cross-validation bandwidth h_{cv} obtained from (49) may be used to compute all of the above estimators, although, in general, different bandwidths may be used for these estimators.

The above plug-in approach can also be extended to $\hat{\beta}_r^{COM}(t)$ of (43) when the covariates are time-invariant. Wu, Yu and Yuan[41] have derived the explicit expressions of the bias, $\mathrm{B}(\hat{\beta}_r^{COM};t)$, and the standard deviation, $\mathrm{SD}(\hat{\beta}_r^{COM};t)$, of $\hat{\beta}_r^{COM}(t)$, and suggested to use the approximate $(1-\alpha)$ confidence interval for $\beta_r(t)$ with end points

$$\{\hat{\beta}_r^{COM}(t) - \hat{\mathrm{B}}(\hat{\beta}_r^{COM};t)\} \pm Z_{1-\alpha/2}\widehat{\mathrm{SD}}(\hat{\beta}_r^{COM};t),$$

where $\hat{\mathrm{B}}(\hat{\beta}_r^{COM};t)$ and $\widehat{\mathrm{SD}}(\hat{\beta}_r^{COM};t)$ are plug-in estimators of $\mathrm{B}(\hat{\beta}_r^{COM};t)$ and $\mathrm{SD}(\hat{\beta}_r^{COM};t)$. Because of the similarity it shares with $\hat{\beta}^{LSK}(t)$, we omit the details for this case.

The above asymptotic intervals differ from their counterparts with independent identically distributed data in the inclusion of intra-subject correlations in the variance term. When n_i are not negligible relative to n, θ in (54) may not be negligible, so that the contribution of the correlations may not be ignored. For the HIV/CD4 data (Sec. 1.2.2), the numbers of repeated measurements range from 1 to 14, while the number of subjects is 400. Asymptotic results that do not take the intra-subject correlations into account may not lead to adequate approximations. In this case, it is appropriate to estimate the correlations directly from the data. When the numbers of repeated measurements are negligible relative to the numbers of subjects, as in the ASGA data (Sec. 1.2.1), the contribution of the intra-subject correlation structures becomes negligible in the variances of the kernel estimators. The resulting confidence intervals are then similar to that with independent identically distributed samples.

5.1.2. *Simultaneous bands*

In most applications, the main interest of inference lies in the overall confidence regions of $\beta_l(t)$ within a proper range of t values, rather than the confidence intervals at a particular time point. When the data are from independent identically distributed samples, simultaneous confidence regions for regression curves may be constructed using either extreme value theory of Gaussian processes[9] or variability bands bridged by pointwise intervals over a grid points.[11,13,22] For longitudinal samples, analogous asymptotic theory of extreme values has not been developed. This leaves the latter approach to be the only practical simultaneous inferential tool in longitudinal analysis.

To construct a simultaneous band for $A^T\beta(t)$ over $t \in [a,b]$ based on the least squares kernel estimator $\hat{\beta}^{(LSK)}(t)$, we choose a positive integer M and partition $[a,b]$ into M equally spaced intervals with grid points $a = \xi_1 < \cdots < \xi_{M+1} = b$, such that $\xi_{j+1} - \xi_j = (b-a)/M$ for $j = 1,\ldots,M$. A set of approximate $(1-\alpha)$ simultaneous confidence intervals for $A^T\beta(\xi_j)$, $j = 1,\ldots,M+1$, is then the collection of intervals $(\hat{l}_\alpha(\xi_j), \hat{u}_\alpha(\xi_j))$, $j = 1,\ldots,M+1$, which satisfies

$$\lim_{n\to\infty} P\{\hat{l}_\alpha(\xi_j) \le A^T\beta(\xi_j) \le \hat{u}_\alpha(\xi_j) \text{ for all } j = 1,\ldots,M+1\} \ge 1-\alpha\,. \tag{58}$$

The Bonferroni adjustment suggests

$$(\hat{l}_\alpha(\xi_j), \hat{u}_\alpha(\xi_j)) = (\hat{L}_{\alpha/(M+1)}(\xi_j), \hat{U}_{\alpha/(M+1)}(\xi_j))\,, \tag{59}$$

where $(\hat{L}_\alpha(\xi_j), \hat{U}_\alpha(\xi_j))$ are defined in (57).

To establish a band that covers all the points between the grid points ξ_j, $j = 1, \ldots, M+1$, we first consider the interpolation of $A^T\beta(\xi_j)$ defined by

$$(A^T\beta)^{(I)}(t) = \left\{\frac{M(\xi_{j+1} - t)}{b-a}\right\}[A^T\beta(\xi_j)] + \left\{\frac{M(t-\xi_j)}{b-a}\right\}[A^T\beta(\xi_{j+1})]\,, \tag{60}$$

for $t \in [\xi_j, \xi_{j+1}]$. A simultaneous band for $(A^T\beta)^{(I)}(t)$ over $t \in [a, b]$ is $(\hat{l}_\alpha^{(I)}(t), \hat{u}_\alpha^{(I)}(t))$, where $\hat{l}_\alpha^{(I)}(t)$ and $\hat{u}_\alpha^{(I)}(t)$ are the linear interpolations of $\hat{l}_\alpha(\xi_j)$ and $\hat{u}_\alpha(\xi_j)$, similarly defined as in (60). The gaps between the grid points are then bridged by the smoothness conditions of $A^T\beta(t)$. If $A^T\beta(t)$ satisfies

$$\sup_{t\in[a,b]} |(A^T\beta)'(t)| \le c_1\,, \quad \text{for a known constant } c_1 > 0\,, \tag{61}$$

then it follows that

$$|A^T\beta(t) - (A^T\beta)^{(I)}(t)| \le 2c_1\left[\frac{M(\xi_{j+1}-t)(t-\xi_j)}{b-a}\right],$$

for all $t \in [\xi_j, \xi_{j+1}]$, and consequently

$$\left(\hat{l}_\alpha^{(I)}(t) - 2c_1\left[\frac{M(\xi_{j+1}-t)(t-\xi_j)}{b-a}\right],\right.$$
$$\left.\hat{u}_\alpha^{(I)}(t) + 2c_1\left[\frac{M(\xi_{j+1}-t)(t-\xi_j)}{b-a}\right]\right) \tag{62}$$

is an approximate $(1-\alpha)$ confidence band for $A^T\beta(t)$. If $A^T\beta(t)$ satisfies

$$\sup_{t\in[a,b]} |(A^T\beta)''(t)| \le c_2\,, \quad \text{for a known constant } c_2 > 0\,, \tag{63}$$

then

$$|A^T\beta(t) - (A^T\beta)^{(I)}(t)| \le \frac{c_2}{2}\left[\frac{M(\xi_{j+1}-t)(t-\xi_j)}{b-a}\right],$$

for all $t \in [\xi_j, \xi_{j+1}]$, and an approximate $(1-\alpha)$ confidence band can be given by

$$\left(\hat{l}_\alpha^{(I)}(t) - \frac{c_2}{2}\left[\frac{M(\xi_{j+1}-t)(t-\xi_j)}{b-a}\right],\right.$$
$$\left.\hat{u}_\alpha^{(I)}(t) + \frac{c_2}{2}\left[\frac{M(\xi_{j+1}-t)(t-\xi_j)}{b-a}\right]\right). \tag{64}$$

For smoothness conditions other than the ones considered in (61) and (63), the corresponding confidence bands may be similarly established. When the covariates are time-invariant, the same approach can be used to establish simultaneous confidence bands based on $\hat{\beta}^{COM}(t)$; see Wu, Yu and Yuan[41] for details.

5.2. *Bootstrap variability bands*

The above asymptotic inferences subject to two restrictions which, to some degree, limit their applications in longitudinal analysis. First, because the asymptotic distributions have so far only been developed for the two kernel type estimators, $\hat{\beta}^{LSK}(t)$ and $\hat{\beta}^{COM}(t)$, confidence procedures for other estimators are still not available. Given that smoothing methods such as splines and local polynomials have exhibited a number of theoretical and practical advantages over the kernel methods, particularly at the boundary of the support of t, inferential procedures based on these smoothing methods are in demand. Second, because the plug-in estimators require the estimation of the design densities, covariance functions and the other quantities appeared in the bias and variance terms of the estimators, the procedure is usually computationally intensive and may introduce additional errors in its coverage probabilities.

A more appealing inferential procedure that has been suggested in the literature is the "resampling-subject" bootstrap. Let $\hat{\beta}(t) = (\hat{\beta}_0(t), \ldots, \hat{\beta}_k(t))^T$ be an estimator of $\beta(t)$ constructed based on any of the previously mentioned smoothing method. An approximate $(1-\alpha)$ pointwise percentile interval for $A^T E[\hat{\beta}(t)]$ can be constructed by the following steps:

(1) Randomly draw n subjects with replacement from the original dataset and denote the resulting bootstrap sample to be $\{(Y_{ij}^*, t_{ij}^*, X_{ij}^*); i = 1, \ldots, n, j = 1, \ldots, n_i\}$.
(2) Compute the bootstrap estimator $\hat{\beta}^{boot}(t)$, hence $A^T\hat{\beta}^{boot}(t)$, based on the above bootstrap sample and the smoothing method specified for $\hat{\beta}(t)$.
(3) Repeating the above two steps B times, so that B bootstrap estimators $A^T\hat{\beta}^{boot}(t)$ are obtained.
(4) Calculate $L_\alpha^{boot}(t)$ and $U_\alpha^{boot}(t)$, the lower and upper $(\alpha/2)$th percentiles, respectively, of the B bootstrap estimators $A^T\hat{\beta}^{boot}(t)$. The approximate $(1-\alpha)$ bootstrap interval is then $(L_\alpha^{boot}(t), U_\alpha^{boot}(t))$.

When $A^T E[\hat{\beta}(t)]$ satisfies the smoothness conditions (61) or (63), simultaneous confidence bands for $A^T E[\hat{\beta}(t)]$ can be constructed using (62) and (64) with (59) replaced by $(L^{boot}_{\alpha/(M+1)}(\xi_j), U^{boot}_{\alpha/(M+1)}(\xi_j))$.

The main advantages of this bootstrap are its generality and simplicity. It is not limited to kernel type estimators and does not depend on the correlations and designs of the data. Despite its potential, several related theoretical and practical issues have still yet to be resolved. Because the biases of the estimators have not been adjusted, the resulting intervals or bands may not always have desirable coverage probabilities for $A^T\beta(t)$. If a consistent estimator of the bias is also available, improved confidence regions for $A^T\beta(t)$ may be obtained by adjusting the bias appeared in $(L^{boot}_{\alpha/(M+1)}(\xi_j), U^{boot}_{\alpha/(M+1)}(\xi_j))$. Currently, consistent bias estimators can only be obtained on a case-by-case basis, and no general procedure is available. A natural alternative to the percentile end points used in Step 4 is to consider normal approximated intervals with end points $A^T\hat{\beta}(t) \pm z_{(1-\alpha/2)}\widehat{se}^{boot}(t)$, where $\widehat{se}^{boot}(t)$ is the sample standard error of the B bootstrap estimators $A^T\hat{\beta}^{boot}(t)$. Asymptotic properties for both the percentile and the normal approximation bootstrap procedures have not been investigated.

6. Two Examples

6.1. *Alabama fetal growth study*

Normal fetal growth is naturally thought to influence infant survival and proper child development. Our objective is to investigate the effects of maternal risk factors and maternal anthropometric measurements on the patterns of fetal growth. Although the outcomes measured by fetal abdominal circumference, biparietal diameter and femur length are all time-dependent, the covariates of interest may be either time-dependent or time-invariant. A typical time-dependent covariate is the maternal placental thickness measured by ultrasound at each visit. On the other hand, mother's height, weight and body mass index measured at the beginning of pregnancy, are time-invariant. Other variables, such as maternal habits of cigarette smoking and alcohol consumption, may be either time-dependent or time-invariant depending on how these variables are defined. A simple way to define time-invariant maternal smoking and drinking status is to categorize the mothers as smokers (ever smoked cigarettes during the pregnancy) versus non-smokers (never smoked cigarettes during the pregnancy) and non-drinkers/light-drinkers (consumed one beer/one glass of

wine or less per day in average during the pregnancy) versus heavy-drinkers (consumed more than one beer or one glass of wine per day in average during the pregnancy). As in most self-reported questionnaires, the data contain the average numbers of cigarettes smoked and the average amount of alcohol consumed per day per subject. These actual cigarette and alcohol consumptions are clearly time-dependent as some of the participating subjects change their behaviors during the study. Depending on the specific scientific questions, both smoking and drinking categories and the actual consumptions could be considered in the analysis.

For the purpose of illustration, the analysis present here focuses on the effects of maternal smoking/drinking categories and placental thickness on the growth of fetal abdominal circumference. Other covariate and outcome measurements can be similarly investigated, provided that the models have clear and meaningful biological interpretations. Although the general trend of Fig. 1 shows an upward growth pattern, it hardly provides any clue on the relationship between fetal growth and the covariates of interest. A nonparametric analysis with model (3) seems a natural start.

Let $Y(t)$ and $X^{(1)}(t)$ be the fetal abdominal circumference and placental thickness, respectively, at t weeks of gestation; $X^{(2)}$ and $X^{(3)}$ be the mother's drinking and smoking categories defined by

$$X^{(2)} = \begin{cases} 1 & \text{if she is a heavy-drinker}, \\ 0 & \text{if she is a non-drinker/light drinker}, \end{cases}$$

$$X^{(3)} = \begin{cases} 1 & \text{if she is a smoker}, \\ 0 & \text{otherwise}; \end{cases}$$

and $X^{(4)}$ be the mother's height (in centimeters) at the beginning of the pregnancy.

In view that proper placental development may also be affected by drinking and smoking, we first consider the effects of the time-invariant covariate vector $X = (1, X^{(2)}, X^{(3)}, X^{(4)})^T$. Although we can fit model (3) directly with $(Y(t), t, X)$ and describe the covariate effects by $\beta(t) = (\beta_0(t), \beta_2(t), \beta_3(t), \beta_4(t))^T$, a better biological interpretation can be obtained if $X^{(4)}$ were replaced by its centered version $X^{(*4)} = X^{(4)} - E[X^{(4)}]$, so that the covariate effects are characterized by $\beta^*(t) = (\beta_0^*(t), \beta_2(t), \beta_3(t), \beta_4(t))^T$. For the latter case, the baseline coefficient curve $\beta_0^*(t)$ represents the mean abdominal circumference at t weeks of gestation for a non-smoking and non-drinking/light-drinking mother whose height is at average, while, for the former, $\beta_0(t)$ itself does not have a biological interpretation.

To fit model (3) with $(Y(t), t, X^{(*)})$, $X^{(*)} = (1, X^{(2)}, X^{(3)}, X^{(*4)})^T$, we computed $X_i^{(*4)}$, $i = 1, \ldots, 1475$, by subtracting the sample average of $\{X_j^{(4)}; j = 1, \ldots, 1475\}$ from $X_i^{(4)}$. Figure 2 shows the estimated coefficient curves, including the baseline growth curve and the covariate effects characterized by alcohol consumption, cigarette smoking and mother's height, and their corresponding 95% simultaneous confidence bands. These coefficient curves were computed using the componentwise estimators of (43) with the Epanechikov kernel, the cross-validated bandwidths described in (49) and $w_i = 1/(nn_i)$. It is worthwhile noting that in this data set the numbers of repeated measurements, most of which are around 4, are much smaller compared with the number of subjects $n = 1475$. Thus, asymptotic results obtained by assuming n tending to infinity and n_i remaining finite are expected to give adequate approximations. For kernel smoothing estimators, this means that both $w_i = 1/(nn_i)$ and $w_i = 1/N$ lead to very similar estimates, and the intra-subject correlations can be ignored in the asymptotic variances of the estimators. Thus, no covariance estimators are needed in the construction of asymptotically approximate confidence bands. Based on the same kernel and bandwidths used in the coefficient curve estimates, the simultaneous confidence bands were computed using the asymptotic approximation (62) and the Bonferroni adjustment with $M = 40$ and $c_1 = 5$. These graphs suggest an upward linear baseline curve

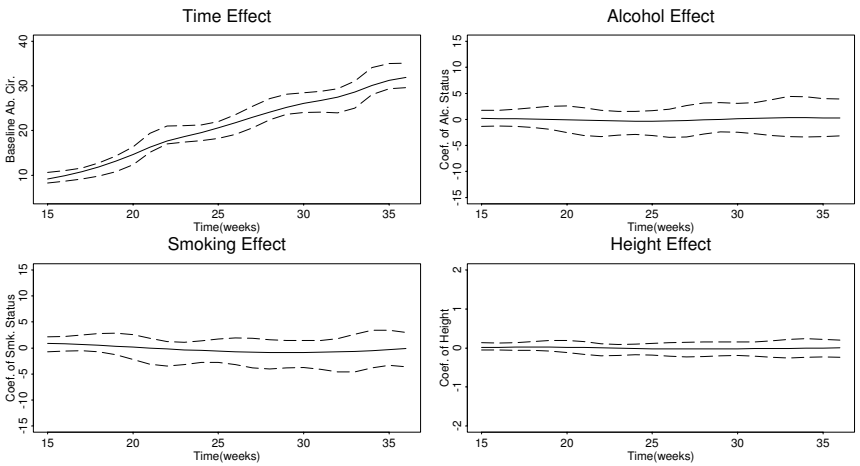


Fig. 2. ————: componentwise kernel estimates of the coefficient curves (covariate effects) computed using the Epanechnikov kernel, the cross-validated bandwidths and $w_i = 1/(nn_i)$. - - - - - - - - -: the 95% Bonferroni-type confidence bands.

Table 1. Parameter estimates and their standard errors computed using the *Mixed-Effects Procedure* in S-plus.

	Parameter Estimate	Standard Error	Z-ratio
β_{00}	−6.5496	0.0614	−106.5880
β_{01}	1.0645	0.0021	496.1262
β_2	0.0026	0.0551	0.0478
β_3	0.1009	0.0516	1.9555
β_4	0.0007	0.0035	0.1996

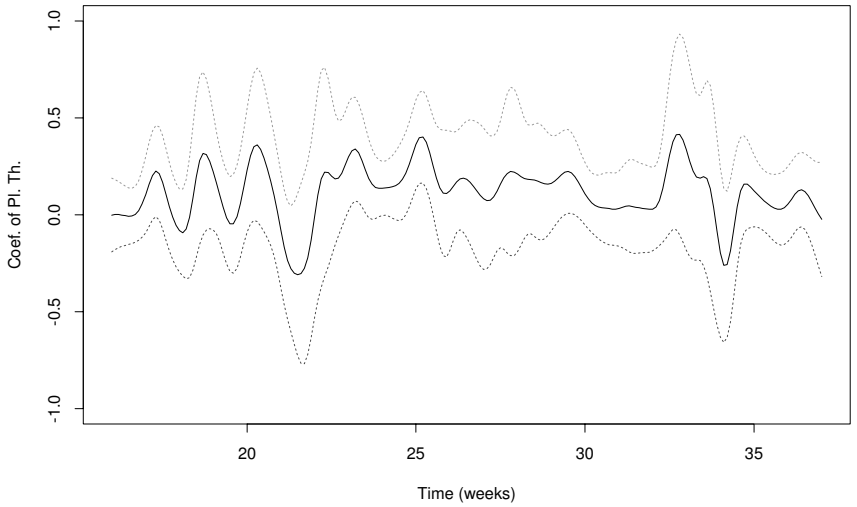
$\beta_0^*(t)$ and undetectable effects from alcohol consumption, cigarette smoking and mother's height. However, because the confidence bands used here tend to be conservative, they may not be sensitive enough to detect small influences of the covariates. We also computed the curve estimates and their corresponding confidence bands using the least squares kernel method of (33). We omit these results from the presentation, because they are similar to the ones shown in Fig. 2.

The above nonparametric results, i.e. graphs shown in Fig. 2, suggest that the relationship between fetal abdominal circumference $Y(t)$, gestational age t, alcohol consumption $X^{(2)}$, cigarette smoking $X^{(3)}$ and centered maternal height $X^{(*4)}$ can be reasonably described by the linear model

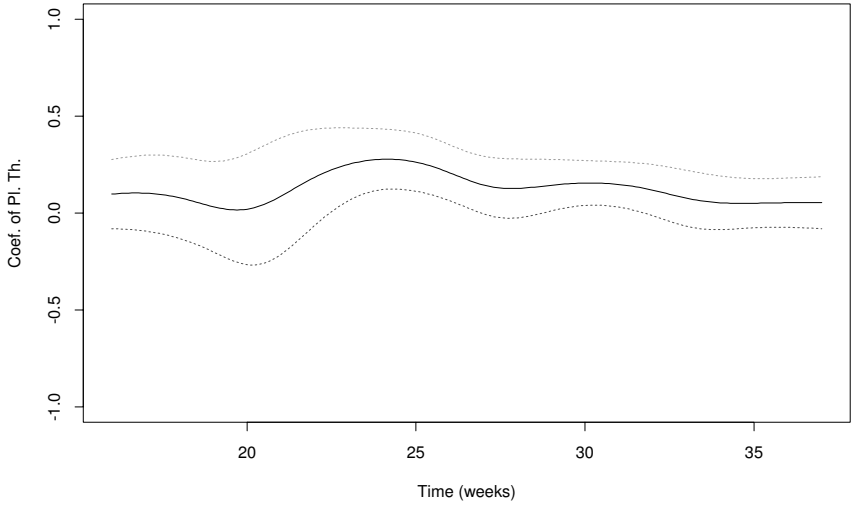
$$Y(t) = \beta_{00} + \beta_{01}t + \beta_2 X^{(2)} + \beta_3 X^{(3)} + \beta_4 X^{(*4)} + \epsilon(t)\,,$$

with unknown parameters $(\beta_{00}, \beta_{01}, \beta_2, \beta_3, \beta_4)$ and a mean zero error process $\epsilon(t)$. This model can be fitted using the *Mixed-Effects Procedure* in S-plus.[3] Table 1 shows the parameter estimates and the corresponding standard errors computed from the above linear model and the S-plus procedure. The results from this linear model suggested clearly non-significant effects for alcohol consumption and maternal height and a very weak, but slightly positive, effect for cigarette smoking. The weak smoking effect shown in this linear analysis is likely caused by the random variations of the data, rather than any substantial association between fetal size and smoking. These results generally agree with the findings obtained from the above nonparametric analysis.

When placental thickness $X^{(1)}(t)$ is added to the model, smoothing has to be carried out with time-dependent covariates. In order to obtain a meaningful biological interpretation for the baseline coefficient curve, we use the centered covariate $X^{(*1)}(t) = X^{(1)}(t) - E[X^{(1)}(t)]$, the difference between a subject's placental thickness at time t and the conditional mean at t. To avoid starting with a model that has too many covariates, we



(a) Placental Effect (CV)



(a) Placental Effect (Subjective)

Fig. 3. ————: estimated coefficient curve (covariate effect) for placental thickness, computed using (38) with the standard Gaussian kernel, $w_i = 1/N$, cross-validated bandwidths (top panel) and bandwidth vector $(\gamma_1, h_0, h_1, h_4) = (1.5, 1.0, 2.0, 1.0)$ (bottom panel). ·····: the 95% pointwise intervals computed using the "resampling-subject" bootstrap percentiles.

consider first fitting (3) with $(t, X^{(*1)}(t), X^{(*4)})$ as the covariate vector. The top panel of Fig. 3 shows the estimated coefficient curve for $X^{(*1)}(t)$ computed using the kernel method of (38) with the standard Gaussian kernel, the cross-validated bandwidths and $w_i = 1/N$. This estimate appears to be undersmoothed, as it can not be explained by a clear biological interpretation. An alternative, perhaps biologically more transparent, estimated coefficient curve of $X^{(*1)}(t)$, shown in the bottom panel of Fig. 3, is computed using the same method except with bandwidth vector $(\gamma_1, h_0, h_1, h_4) = (1.5, 1.0, 2.0, 1.0)$. This bandwidth vector was chosen because its cross-validation score was very close to that of the cross-validated bandwidths. Bootstrap percentile intervals are used to demonstrate the variability of the estimates, while inferences based on asymptotic approximations are still not yet available for this type of estimators.

Figure 3 suggests, at least qualitatively, some positive association between placental thickness and fetal abdominal circumference. The estimated coefficient curve for the centered maternal height $X^{(*4)}(t)$ stays constantly close to zero, suggesting a non-significant effect for the maternal height. The estimated baseline coefficient curve is also very close to the one presented in Fig. 2. Hence, these curves are omitted from the presentation. Also omitted are the analysis with the mother's drinking and smoking status, $X^{(2)}$ and $X^{(3)}$, added to the model, as their effects are very similar to the ones shown in Fig. 2.

6.2. *MACS CD4/HIV study*

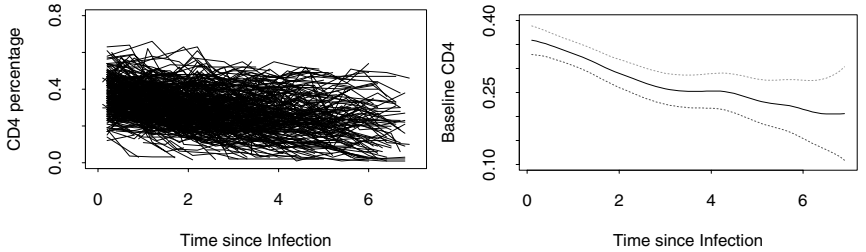
Let t_{ij} denote the ith subject's time length (in years) for his jth measurement since HIV infection. Our objective is to evaluate the effects of two factors, the pre-HIV infection CD4 percent $X^{(1)}$ and the smoking status $X^{(2)}(t)$, on the post-HIV infection depletion of CD4 percent $Y(t)$ over time. The first covariate $X^{(1)}$ does not depend on the time since HIV infection. The second covariate $X^{(2)}(t)$ equals 1 if the subject is classified as a smoker at time t and zero otherwise. Because some of the subjects change their smoking habits during the study, $X^{(2)}(t)$ is a time-dependent variable. Owing to the lack of an existing parametric or semiparametric model that is known to describe the scientific relevance between these variables, it is reasonable to consider an initial analysis with the nonparametric model (3).

The same rationale used in the analysis of the ASGA study suggests that, in terms of biological interpretability, the center variable $X^{(*1)} = X^{(1)} - E[X^{(1)}]$ is more preferable than its uncentered version $X^{(1)}$ in the

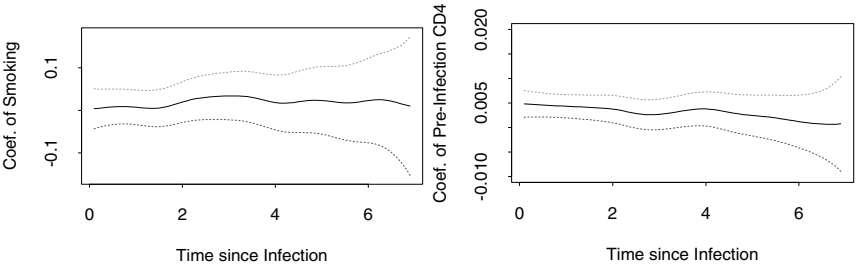
model (3). However, because $X^{(2)}(t)$ is a time-dependent binary variable, it is unnecessary to be centered. Thus, with $X_i^{(*1)}$ estimated by subtracting the corresponding sample mean from $X_i^{(1)}$, the model (3) can be fitted with the data $\{(Y_{ij}, t_{ij}, X_{ij}^*); i = 1, \ldots, 400, j = 1, \ldots, n_i\}$. The baseline coefficient curve $\beta_0^*(t)$ represents the mean CD4 percent at t years after the infection for those who are non-smokers at time t and have average level of CD4 percent before the infection. The effects $\beta_1^*(t)$ and $\beta_2(t)$ of $X^{(*1)}$ and $X^{(2)}(t)$, respectively, can be interpreted the usual way.

Besides the difference in covariate centering, there is another important difference in the estimation and inferences between this and the previous example. The numbers of repeated measurements in this data set can not be simply ignored compared with the number of subjects. Thus, at least for the known case of kernel estimation, the asymptotic approximations assuming n tending to infinity and n_i remaining bounded may not lead to adequate estimators of the variances, although both $w_i = 1/(nn_i)$ and $w_i = 1/N$ seem to be reasonable weight choices. Because the correlation structure of the data is completely unknown and difficult to be estimated accurately, Wu, Chiang and Hoover[39] suggested that it is appropriate in this case to obtain conservative Bonferroni-type bands with the covariance $\rho_\epsilon(t)$ in (55) replaced by the variance $\sigma^2(t)$, an upper bound for $|\rho_\epsilon(t)|$. The graphs in Fig. 4 show the individuals' depletion of CD4 percent over time, the estimated coefficient curves and their corresponding conservative Bonferroni-type 95% asymptotic confidence bands. The estimated coefficient curves were computed using (33) with Epanechnikov kernel, the cross-validated bandwidth and the $w_i = 1/N$ weight. The confidence bands were computed using (57) and (62) with $M = 138$, $c_1 = 3$ and $\rho_\epsilon(t)$ replaced by $\sigma^2(t)$. The same kernel and bandwidth used in computing (33) were also used in computing all the plug-in kernel estimators required in (57).

Figure 4(b) shows a declining baseline CD4 percent curve over time since HIV infection, which coincides with the basic trend suggested by the plot shown in Fig. 4(a). The simultaneous band for the coefficient curve of the pre-infection CD4 percent stays positive at least for the first four years after HIV infection, suggesting strongly the benefit of high pre-infection CD4 level for the initial period since the infection. However, the positive effect of the pre-infection CD4 percent on the post-infection CD4 percent appears tapering down at the later stage of the infection. Although the estimated curve in Fig. 4(c) stays positive throughout the 7-year time range considered in this data set, the confidence band obtained for this curve does not show any significant positive association between cigarette smoking



(a) CD4 VS. Time (b) Time Effect



(c) Smoking Effect (d) Pre-Infection CD4 Effect

Fig. 4. (a) Individuals' CD4 percent versus time (in years) since HIV infection. (b)–(d) Estimated baseline CD4 percent, coefficient curve for smoking and coefficient curve for pre-HIV infection CD4 percent (————) and their corresponding 95% simultaneous confidence bands (· · · · · · · · ·).

and post-infection CD4 level. This may be either caused by the weak association between these two variables or the conservative nature of our confidence bands. Clearly, our findings here only provide some exploratory insights on the data. Biomedical implications and parametric models that provide additional meaningful descriptions of the biological mechanisms have to be further developed and independently confirmed by other studies. Nevertheless, the usefulness of nonparametric regression, particularly the varying-coefficient models, in the initial exploration of longitudinal data is transparent, as was shown in this and the previous examples.

7. Summary and Discussion

This article has presented a series of parametric, semiparametric and nonparametric models and their estimation and inferential methods for

the analysis of longitudinal data. These methods have a wide range of applications in biomedical studies. Theory and methods for parametric models, particularly the linear models, have been extensively studied in the literature. Estimation and inferences based on parametric models can be easily implemented using existing statistical software packages, such as SAS and S-plus. Methods based on semiparametric and nonparametric models, on the other hand, represent the most current progress in this active research field.

The nonparametric estimation and inferential methods introduced here are all based on the general framework of varying-coefficient models. These methods have the advantage of being flexible while applicable to large longitudinal studies. Smoothing methods for these models have been developed using local polynomials and splines, each has its own advantages and disadvantages in practice. Generally speaking, the componentwise smoothing methods are flexible and computationally feasible when the covariates are time-invariant, while methods based on ordinary and penalized least squares and basis approximations can be applied to models with both time-dependent and time-invariant covariates. Pointwise and simultaneous confidence bands for the coefficient curves can be constructed using either asymptotic approximations or the "resampling-subject" bootstrap. The asymptotic confidence procedures have only been developed for the kernel methods. The "resampling-subject" bootstrap may in principle be used with any smoothing estimators. However, despite the usefulness of this bootstrap shown by a number of simulation studies, its theoretical properties have not been investigated. The approach of two-step smoothing appears to be useful to overcome some of the drawbacks of the ordinary least squares. But, in order for this approach to be useful in an unbalanced longitudinal study, further research is needed to establish specific methods for calculating the raw estimates and the asymptotic properties of the final estimators. Finally, a practical consideration is the use of the uniform weight $w_i = 1/N$ versus the uniform subject weight $w_i = 1/(nn_i)$. Although none of these weight uniformly dominates the other in all the longitudinal designs and an ideal weight may depend on the unknown correlation structure and how fast n_i, $i = 1, \ldots, n$, tending to infinity relative to n, simulation studies that have been reported in the literature so far suggest that both weight choices are appropriate when all the subjects have approximately the same numbers of repeated measurements, while the $w_i = 1/(nn_i)$ weight is usually preferred when the numbers of repeated measurements differ from each other significantly.

There are a number of topics that warrant further investigation. First

and foremost, although estimation and confidence tools are important in longitudinal analyses, methods that are enormously useful in biomedical studies are testing procedures that can evaluate the statistical evidence for different hypotheses. Such procedures distinguish a parametric submodel that explains a given scientific hypothesis from the general nonparametric model. The main task of decision making is to determine the distributions of the appropriate test statistics. Another practically important problem is to improve the confidence procedures. The procedures presented in this article are known to be conservative, which often hinders their usefulness in practice. Further work needs to focus on reducing the widths of the bands while maintaining satisfactory coverage probabilities. Finally, in view that the varying-coefficient models are still inadequate for a number of longitudinal settings, there is a need to further extend these models. A useful extension is to consider regression models where the outcome variable depends on the history as well as the current values of the covariates. All the estimation and inference methods will have to be redeveloped for this extension.

Acknowledgment

The first author was partial supported by the National Institute on Drug Abuse (R01 DA10184-01), the National Science Foundation (DMS 0103832) and the Acheson J. Duncan Fund of the Johns Hopkins University. The authors would like to thank Professor Joseph Margolick (Johns Hopkins School of Hygiene and Public Health) for providing the MACS Public Use Data Set Release PO4 (1984–1991) and Ms. Vivian W.-S. Yuan for computing many of the numerical results used in this work.

References

1. Akaike, H. (1970). Statistical predictor identification. *Annals Institute Statistics Mathematics* **22**: 203–217.
2. Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**: 749–759.
3. Bates, D. M. and Pinheiro, J. C. (1999). *Mixed Effects Models in S*, Springer-Verlag, New York.
4. Chiang, C.-T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variable. *Journal of the American Statistical Association* **96**.
5. Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, Chapman Hall, London; New York.
6. De Boor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag, New York.

7. Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**: 959–971.
8. Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Oxford University Press, Oxford, England.
9. Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association* **88**: 1287–1301.
10. Fan, J. Q. and Zhang, J.-T. (2000). Functional linear models for longitudinal data. *Journal of Royal Statistical Society, Series* **B62**: 303–322.
11. Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal Multiple Analysis* **27**: 228–254.
12. Härdle, W. (1990). *Applied Nonparametric Regerssion*, Cambridge University Press, Cambridge, UK.
13. Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics* **19**: 778–796.
14. Hart, T. D. (1991). Kernel regression estimation with time series errors. *Journal of Royal Statistical Society, Series* **B53**: 173–187.
15. Hart, T. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association* **81**: 1080–1088.
16. Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**: 383–385.
17. Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of Royal Statistical Society, Series* **B55**: 757–796.
18. Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**: 809–822.
19. Huang, J., Wu, C. O. and Zhou, L. (2002). Varying coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**: 111–128.
20. Jones, R. H. and Ackerson, L. M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**: 721–731.
21. Jones, R. H. and Boadi-Boteng, F. (1991). Unequally spaced longitudinal data with serial correlation. *Biometrics* **47**: 161–175.
22. Knafl, G., Sacks, J. and Ylvisaker, D. (1985). Confidence bands for regression functions. *Journal of the American Statistical Association* **80**: 683–691.
23. Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. and Rinaldo, C. R. (1987). The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants. *American Journal of Epidemiology* **126**: 310–318.
24. Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**: 963–974.
25. Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
26. Moyeed, R. A. and Diggle, P. J. (1994). Rates of convergence in

semiparametric modeling of longitudinal data. *Australian Journal Statistics* **36**: 75–93.
27. Müller, H.-G. (1988). Nonparametric regression analysis of longitudinal data. *Lecture Notes in Statistics* **46**, Springer-Verlag, Berlin.
28. Pantula, S. G. and Pollock, K. H. (1985). Nested analysis of variance with autocorrelated errors. *Biometrics* **41**: 909–920.
29. Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.
30. Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of Royal Statistical Society, Series* **B53**: 233–243.
31. Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**: 253–259.
32. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**: 461–464.
33. Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**: 45–54.
34. Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.
35. Vonesh, E. F. and Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Marcel Dekker, New York.
36. Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.
37. Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician* **39**: 95–101.
38. Wu, C. O. and Chiang, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica* **10**: 433–456.
39. Wu, C. O., Chiang, C.-T. and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association* **93**: 1388–1402.
40. Wu, C. O., Yu, K. F. and Chiang, C.-T. (2000). A two-step smoothing method for varying-coefficient models with repeated measurements. *Annals Institute Statistics Mathematics* **52**: 519–543.
41. Wu, C. O., Yu, K. F. and Yuan, V. W. S. (2000). Large sample properties and confidence bands for component-wise varying-coefficient regression with longitudinal dependent variable. *Communications in Statistics — Theory Methods* **29**: 1017–1037.
42. Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**: 689–699.
43. Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**: 1049–1060.

About the Author

Colin O. Wu is currently Mathematical Statistician in the Office of Biostatistics Research, Division of Epidemiology and Clinical Applications, National Heart Lung and Blood Institute. He received his BA (1985) in Mathematics from the University of California, Los Angeles, and his PhD (1990) in Statistics from the University of California, Berkeley. His past academic and research experience includes academic appointments at the Department of Statistics, The University of Michigan, Ann Arbor, and the Department of Mathematical Sciences, The Johns Hopkins University, joint appointment at the Department of Biostatistics, the Johns Hopkins Bloomberg School of Hygiene and Public Health, and guest researcher at the Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development. His publications have spanned over a number of areas including asymptotic efficiency for semi-parametric models, asymptotic minimaxity for kernel estimators, density estimation and nonparametric regression for biased sampling data, varying-coefficient models for longitudinal data and nonparametric mixed-effects models. His current research is focused on the theory and methods of basis function approximations for the modeling, estimation and inferences with longitudinal data.

CHAPTER 24

LOCAL MODELING: DENSITY ESTIMATION AND NONPARAMETRIC REGRESSION

JIANQING FAN

Department of Statistics, Chinese University of Hong Kong
Shatin, Hong Kong
Tel: 852-2609-7942; jfan@sta.cuhk.edu.hk

RUNZE LI

Department of Statistics, 326 Joab L Thomas Building,
Penn State University, University Park, PA 16802-2111, USA
Tel: (814) 865-1555; rli@stat.psu.edu

1. Introduction

Local modeling approaches are useful tools for exploring features of data without imposing a parametric model. These approaches have been received increasing attention in last two decades and successfully applied to various scientific disciplines, such as, economics, engineering, medicine, environmental science, health science and social science. There are a vast amount of literature on this topic.[29,34] A comprehensive account of local modeling can be found in the books.[6,28,48,72,75,76,85] see also Fan and Gijbels[29] and Fan and Müller[34] for a brief overview on this topic. In this chapter, we will introduce fundamental ideas of local modeling and illustrate the ideas by real data examples. For ease of presentation, we will omit all technical parts.

This chapter basically consists of two parts: Kernel density estimation and local polynomial fitting. In Sec. 2, the kernel density estimation method will be introduced. Important issues, including bandwidth selection, will be addressed. Real data examples will be used to illustrate the ideas how to implement this type of method. Local polynomial regression will be introduced in Sec. 3. In this section, we also discuss how to decide the amount

of smoothing, and extend the ideas of local polynomial regression to other contexts. The idea is further extended to the local likelihood and local partially likelihood in Sec. 4. Section 5 introduces the ideas of nonparametric smoothing tests. Section 6 summarizes some applications of local modeling, including estimation of conditional quantile functions, conditional variance functions and conditional densities, and change point detection.

2. Density Estimation

Suppose that $X_1, \ldots, X_n$ are an independent and identically distributed sample from a population with an unknown probability density $f(x)$. Of interest is to estimate the density f. In explanatory data analysis, we may construct a histogram for the data. If the resulting histogram has a bell shape, then we may assume that the samples were taken from a normal distribution. In this situation, one may just estimate the population mean and variance using the sample mean and sample variance because a normal distribution is completely determined by its mean and variance. In general, parametric approaches to estimation of a density function assume that the density belongs to a parametric family of distributions, such as normal, gamma or beta family. In order to fully specify the density function, one has to estimate the unknown parameters using, for example, maximum likelihood estimation. One may use prior knowledge or scientific reasons to determine a parametric distribution family. In explanatory data analysis, data analysts frequently construct a histogram based on the sample, and then draw reasonable conclusions on the population density.

2.1. *Histogram*

A histogram is usually formed by partitioning the range of data into equally length intervals, called bins, and then drawing a block over each interval with height being the proportion of the data falling in the bin divided by the width of the bin. Specifically, the histogram estimate at a point x is given by

$$\hat{f}(x, h) = \frac{\text{number of observations in the bin containing } x}{nh},$$

where h is the width of the bins, namely binwidth. For a fixed choice of bins, it can be shown that under some mild conditions, $\hat{f}(\cdot, h)$ is a maximum likelihood estimate of the unknown density f. It is worthwhile to note that

the nonparametric maximum likelihood estimate of the unknown density f without any further restriction does not exist, since

$$\max_{\{f: f \geq 0, \int f = 1\}} \prod_{i=1}^{n} f(X_i) = \infty \,.$$

When one constructs a histogram, one has to choose the binwidth and the centers of bins. Figure 1 depicts four histograms based on the same data set and the same binwidth, but using different locations of bin centers. It can be seen from Fig. 1 that the shapes of the resulting histograms are quite different. This implies that the histogram suffers the "edge" effect. Figure 2 shows four histograms of the lengths of crabs, collected from 1973 to 1986, but with different binwidths. The crab data set is available from the

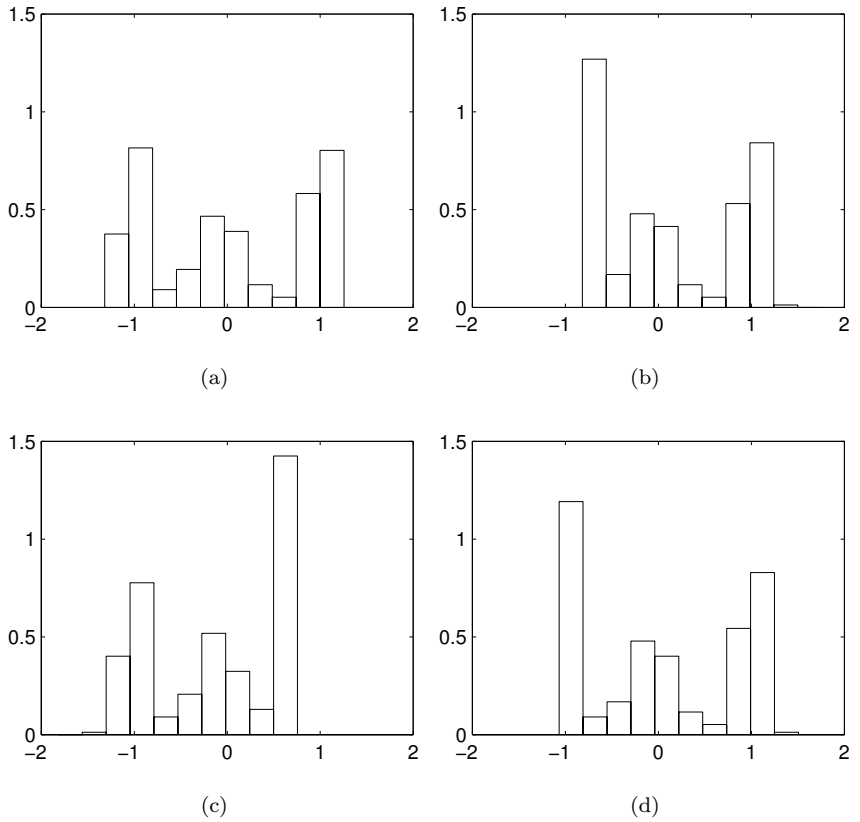


Fig. 1. Histograms of a sample of size 300 from a mixture of normal distribution $1/3N(-1, 0.1^2) + 1/3N(0, 0.25^2) + 1/3N(1, 0.1^2)$.

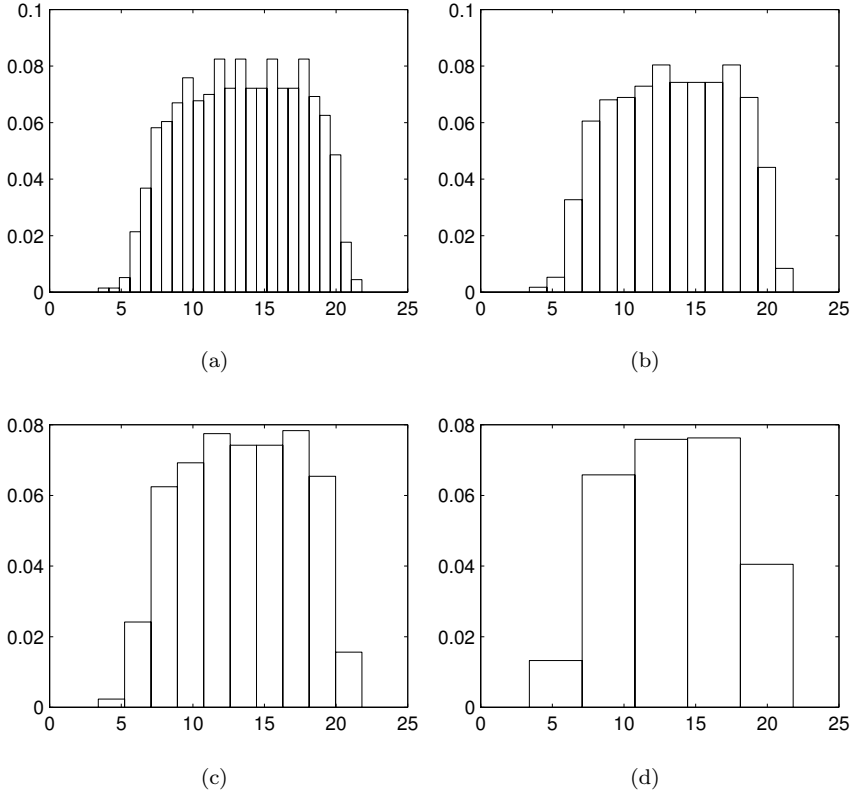


Fig. 2. Histograms for crab sizes. The data is the length of crab (cm).

website of statlib at Carnegie Mellon University at http://lib.stat.cmu.edu. From Fig. 2, if the binwidth h is too small, then the resulting histogram is rough, on the other hand, if the binwidth is too large, then the resulting histogram is too smooth. Thus constructing a histogram actually is not so simple! Usually one may start from an undersmoothed histogram, and then increase gradually the binwidth until getting a satisfactory result.

The histogram is the oldest and most widely used nonparametric estimate of density. The choice of binwidth is a smoothing problem. The edge effect of histograms can be repaired by the kernel density estimation introduced in next section. Furthermore, the kernel estimate will result in a smooth density curve rather than a step function as in histograms. It is an improved technique over the kernel density estimation.

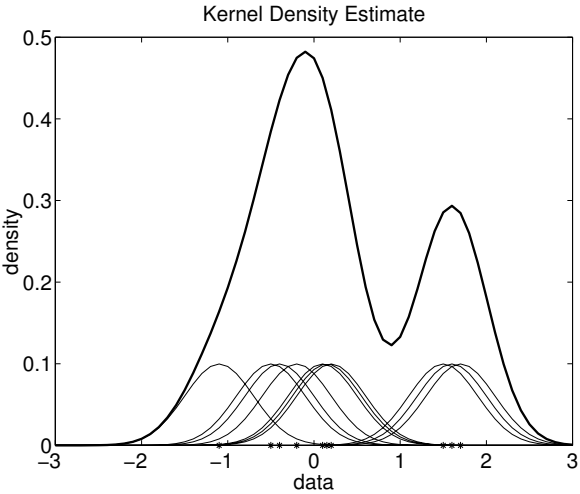


Fig. 3. Kernel density estimate for an hypothetical data set (thick curve). It smoothly redistributes the point mass at X_i by the function $(nh)^{-1}K\{(x - X_i)/h\}$. The small bumps show how point masses are redistributed.

2.2. *Kernel density estimation*

A kernel density estimate is defined as

$$\hat{f}_h = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}\,,$$

where $K(\cdot)$ is a function satisfying $\int K(x)dx = 1$, called a *kernel function* and h is a positive number, called a *bandwidth* or a *smoothing parameter*. A density function such as the plot (thick curve) in Fig. 3 is usually obtained by evaluating the function $\hat{f}_h(x)$ over a few hundred of grid points. From the definition, indeed, the kernel estimate is the average of density functions $h^{-1}K\{(x - X_i)/h\}$, which smoothly redistribute the point mass at the point X_i. Figure 3 depicts the redistribution of point masses. To facilitate notation, let $K_h(t) = \frac{1}{h}K(t/h)$ be a rescaling function of K. This allows us to write

$$\hat{f}_h = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\,. \tag{1}$$

It is well known that the choice of K is not very sensitive, scaled in a canonical form[64] to the estimate $\hat{f}_h(x)$. Thus it is assumed throughout this chapter that the kernel function is a symmetric probability density

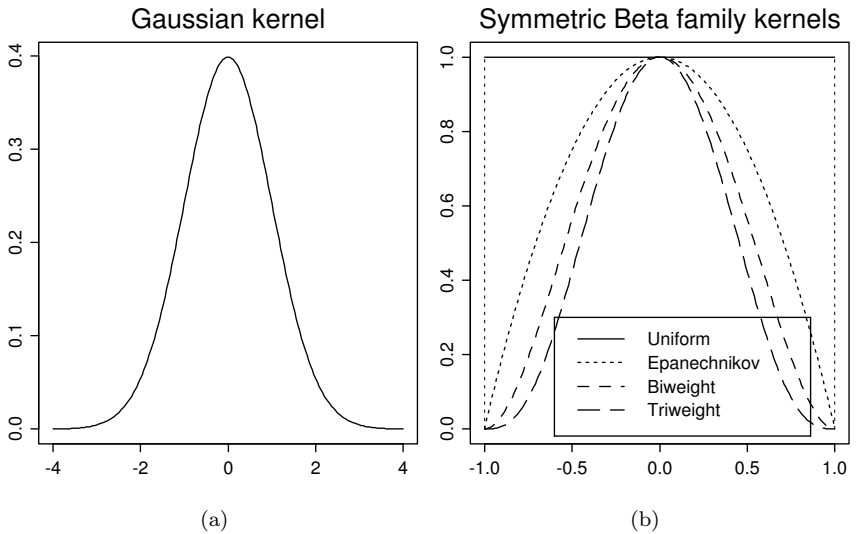


Fig. 4. Commonly-used kernels. (a) Gaussian kernel; (b) Symmetric Beta family of kernels that are renormalized to have maximum height 1.

function. The most commonly used kernel function is the Gaussian density function given by

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\,. \tag{2}$$

Other popular kernel functions include the symmetric beta family

$$K(t) = \frac{1}{\beta(1/2, \gamma+1)}(1-t^2)_+^{\gamma}\,, \quad \gamma = 0, 1, \ldots\,, \tag{3}$$

where $+$ denotes the positive part, which is assumed to be taken before exponentiation, so that the support of K is $[-1, 1]$, and $\beta(\cdot, \cdot)$ is a beta function. The corresponding kernel functions when $\gamma = 0$, 1, 2 and 3 are the uniform, the Epanechnikov, the biweight and the triweight kernel functions. Figure 4 shows these kernel functions.

The smoothing parameter h controls the smoothness of density estimates, acting as the binwidth in histograms. The choice of the bandwidth is of crucial importance. If h is chosen too large, then the resulting estimate misses fine features of the data, while if h is selected too small then spurious sharp structure become visible. See Fig. 6 for example. In fact, it can be shown that under some mild conditions, when $n \to \infty$, $h \to 0$ and $nh \to \infty$,

$$E\hat{f}_h(x) - f(x) = \frac{f''(x)}{2}\mu(K)h^2 + o(h^2) \tag{4}$$

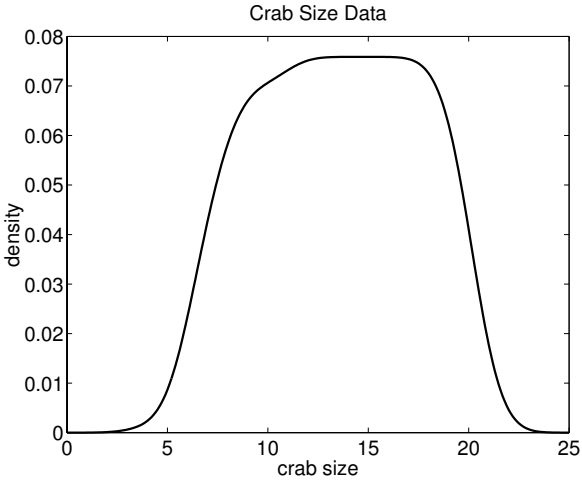


Fig. 5. Automatic kernel density estimates using the bandwidth according the rule of thumb. The data set is the crab size data collected from 1973 to 1986.

and

$$\mathrm{var}\{\hat{f}_h(x)\} = \frac{R(K)f(x)}{nh}(1 + o(1))\,, \tag{5}$$

where $\mu(K) = \int t^2 K(t)dt$ and $R(K) = \int K^2(t)dt$. Thus, from (4) and (5), a large bandwidth h results in a large bias while a small bandwidth produces an estimate with a large variance. A good choice of bandwidth would balance the bias and variance trade-off. This is conveniently assessed by the *Asymptotic Mean Integrated Square Error* (AMISE) which is defined as

$$\mathrm{AMISE}(h) = \frac{\mu^2(K)h^4}{4}\int\{f''(x)\}^2dx + \frac{R(K)}{nh}\,. \tag{6}$$

Minimizing (6) with respect to h gives the ideal bandwidth

$$h_I = \left(\frac{R(K)}{\mu^2(K)\int\{f''(x)\}^2dx}\right)^{1/5} n^{-1/5}\,, \tag{7}$$

which involves the unknown density function, and cannot be directly used in kernel smoothing. Since the choice of bandwidth is critical to kernel density estimation, there has a large literature on this topic. See Jones *et al.*[56,57] for a survey. In practice, we may take the Gaussian density with variance

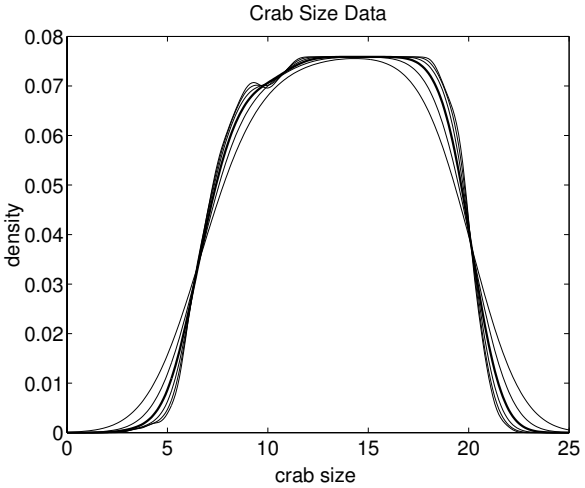


Fig. 6. A family of kernel estimates. The data set is the crab 5size data. The thick curve corresponds to $\hat{h}_I$.

σ^2 as a reference density. In this situation, Eq. (7) becomes

$$h_I = \left(\frac{8\sqrt{\pi} R(K)}{3\mu^2(K)} \right)^{1/5} \sigma n^{-1/5} . \tag{8}$$

Here we focus on a rule of thumb.[75] The rule of thumb of bandwidth selection is to replace σ by the sample standard deviation s_n. Thus, for the Gaussian kernel,

$$\hat{h}_I = 1.06 s_n n^{-1/5} ,$$

and for the symmetric β family

$$\hat{h}_I = \left[\frac{8\sqrt{\pi}\beta(1/2, 2\gamma+1)}{\{\beta(3/2, \gamma+1)\}^2} \right]^{1/5} s_n n^{-1/5} .$$

Figure 5 depicts a kernel density estimate of the length of crab using the bandwidth $\hat{h}_I$ with the Gaussian kernel. From the shape of the estimated density curve, it seems that a normal distribution is not appropriate for modeling the crab size.

While the rule of thumb works well for many data sets, it tends to produce oversmooth estimates as the referenced density is a Gaussian density. Another method to avoid choosing a single optimal bandwidth is the family smoothing approach. This can be done by using a family of estimates

$$\{\hat{f}_h, h = 1.4^j \hat{h}_I, j = -3, -2, -1, 0, 1, 2\} \tag{9}$$

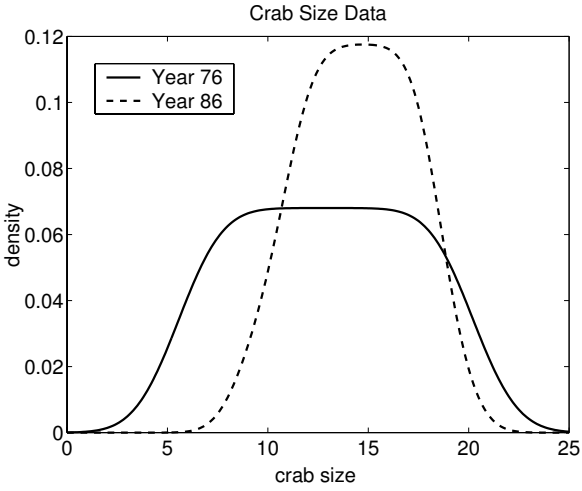


Fig. 7. Comparison of the length of crabs between 1976 and 1986. The sample mean and standard deviation for 1976 are 12.9020 and 4.2905, while the sample mean and standard deviation for 1986 are 14.4494 and 2.6257, respectively.

and then overlaying them in the same plot. The family smoothing approach allows us to explore possible patterns contained in data using different scale of bandwidths. This is closely related to scale space ideas in computer science. Choosing a smaller bandwidth acts as "zoom in", while selecting a larger bandwidth corresponds to "zoom out" in the scale space. These ideas have been further developed in a SiZer map.[13] The SiZer map can detect significant features in estimated curve with different scales. Figure 6 depicts a family smoothing plot for the crab size data.

The density estimation method is also a powerful graphic tool for comparing the results of two experiments. This is related to the classical two-sample mean problem. The advantage of the kernel smoothing approach over the traditional two sample tests is that the smoothing approach can show an overall pattern of the experiments, including the locations of centers and the dispersions of the data. Further, it gives us some ideas of two population distributions. To illustrate the idea, we applied the smoothing techniques for two subsets of the crab size data. One contains the 1976 data set and the other consists of the 1986 data set. The two estimated density curves are depicted in Fig. 7. They have different centers and dispersions.

In this section, the bandwidth remains constant, that is, it depends on neither the location x nor the datum point X_i. This kind of bandwidth is referred as a global bandwidth. From (7), it is desirable to use a larger

bandwidth when changes of curvature is small and use a smaller bandwidth when curvature of underlying density dramatically changes. This leads to studying variable bandwidth selection, which suggests the use of different bandwidth at different location of x. Usually, a global bandwidth is easier to choose than the variable bandwidth. In order to use a constant bandwidth, one may first transform the data by

$$Y_i = g(X_i)\,, \quad i = 1, \ldots, n\,,$$

where g is a given monotone increasing function. The transformation g should be chosen so that the transformed data have a density with more homogeneous degree of smoothness so that a global bandwidth for the transformed data is more appropriate. Then apply the kernel density estimate to the transformed data set and obtain the estimate $\hat{f}_Y(y)$. Finally apply the inverse transform to obtain the density of X:

$$\hat{f}_X(x) = g'(x)\hat{f}_Y(g(x)) = g'(x)n^{-1}\sum_{i=1}^{n} K_h(g(x) - g(X_i)).$$

The performance of this type estimate has been illustrated in Wand *et al.*[86] Marron and Yang (1999) proposed an approach to selecting a good transformation g.

3. Local Polynomial Fitting

Regression is one of the most useful techniques in statistics. Consider the $(d+1)$-dimensional data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, which form an independent and identically distributed sample from a population $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional random vector and Y is a random variable. Of interest is to estimate the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. In other words, the data are regarded as realizations from the model:

$$Y = m(\mathbf{X}) + \varepsilon\,,$$

where ε is a random error with zero mean. For a given data set, one may try to fit the data by using a linear regression model. If a nonlinear pattern appears in the scatter plot of Y against $\mathbf{X}$, one may employ polynomial regression to reduce the modeling bias of linear regression. Consider, for example, the data plotted in Fig. 8, where the relationship between the concentration of nitric oxides in engine exhaust (taken as dependent variable) and the equivalence ratio (taken as independent variable), a measure of the richness of the air/ethanol mix, is depicted for a burning of ethanol in

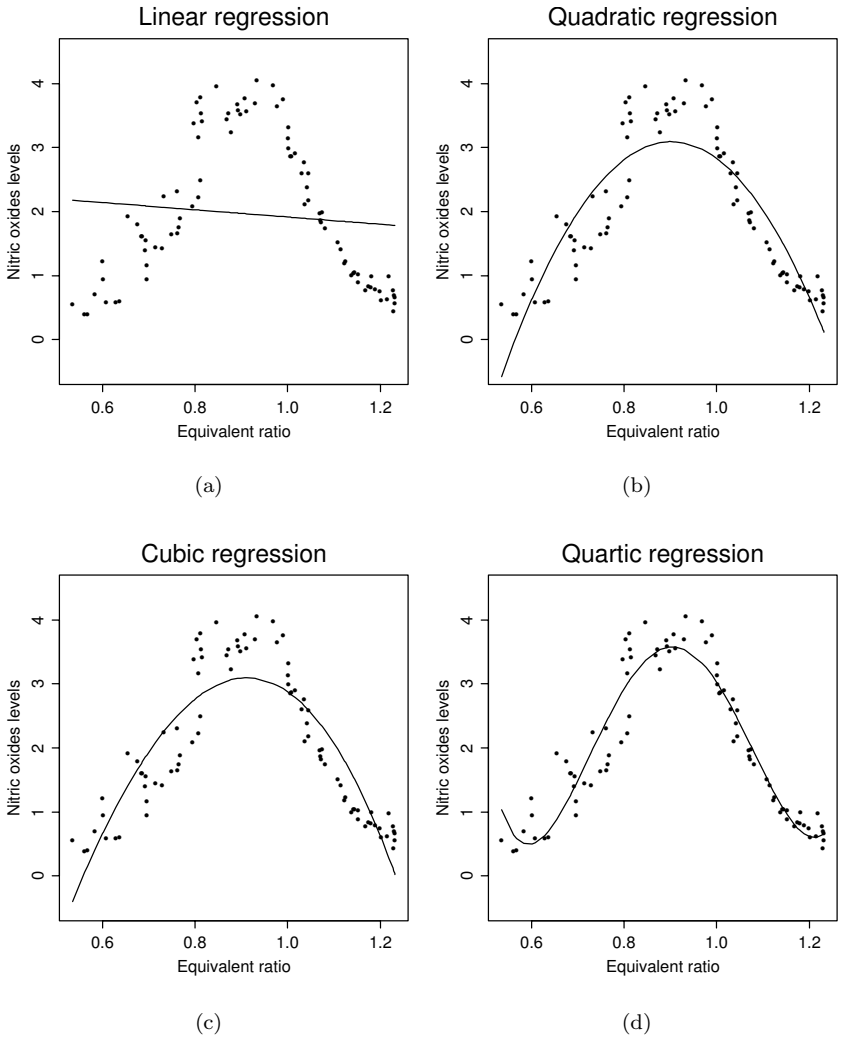


Fig. 8. Polynomial fits to the ethanol data. Presented are the scatter plots of the concentration of nitric oxides against the equivalence ratio along with the fitted polynomial regression functions. Adapted from Fan and Gijbels.[29]

a single-cylinder automobile test engine. From Fig. 8, it can be seen that the relationship between the concentration of nitric oxides and the equivalence ratio is highly nonlinear. Polynomial regression is used to fit the data. Figure 8 presents the resulting fits by using four different degrees of polynomials. One can easily see that all resulting fits have substantial biases.

Because polynomial functions have all orders of derivatives everywhere, and polynomial degree cannot be controlled continuously, polynomial functions are not very flexible in modeling features encountered in practice. Further individual observations can have a large influence on remote parts of the curve in polynomial regression models. Nonparametric regression techniques can be used to repair the drawbacks of polynomial fitting. Fan and Gijbels[28] give detailed background and excellent overview on various nonparametric regression techniques, which can be classified into two categories. One is to approximate the regression function globally and the other one is to parameterize the regression function $m(\mathbf{x})$ locally. Two common methods of global approximation are the *spline approach* and the *orthogonal series method*. In this section, we focus on the techniques of local modeling.

3.1. *Kernel regression*

Consider the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, an i.i.d. sample from the model:

$$Y = m(X) + \varepsilon\,,$$

where ε is random error with $E(\varepsilon|X) = 0$ and $\text{var}(\varepsilon|X = x) = \sigma^2(x)$. The nonparametric regression problem is to estimate the regression function $m(\cdot)$ with imposing a form. Usually, a datum point closer to x carries more information about the value of $m(x)$. Therefore an intuitive estimator for the regression function $m(x)$ is the running local average. An improved version of this is the locally weighted average. That is

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i \Big/ \sum_{i=1}^{n} w_i(x)\,.$$

An alternative interpretation of locally weighted average estimators is that the resulting estimator is the solution to the following weighted least-squares problem:

$$\min_{\theta} \sum_{i=1}^{n} (Y_i - \theta)^2 w_i(x)\,.$$

In other words, the kernel regression estimators are a weighted least squares estimate at the point x using a local constant approximation.

Table 1. Leading terms in the asymptotic biases and variances.[25]

Method	Bias	Variance
NW Estimator	$\left\{m''(x) + \dfrac{2m'(x)f'(x)}{f(x)}\right\} b_n$	V_n
GM Estimator	$m''(x)b_n$	$1.5V_n$
Local Linear	$m''(x)b_n$	V_n

Here $b_n = \dfrac{1}{2}\displaystyle\int_{-\infty}^{+\infty} u^2 K(u)du h^2$ and $V_n = \dfrac{\sigma^2(x)}{f(x)nh}\displaystyle\int_{-\infty}^{+\infty} K^2(u)du$.

Setting the weights $w_i(x) = K_h(X_i - x)$ results in the NW kernel regression estimator, which is given by[68,87]

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K_h(X_i - x)Y_i}{\sum_{i=1}^n K_h(X_i - x)}\,. \tag{10}$$

See Nadaraya[68] and Watson.[87]

Since the denominator in (10) is a random variable, it is inconvenient to take derivatives with respect to x and to derive the asymptotic properties of the estimator. Assume that the data have already been sorted according to the X-variable. Taking the local weights $w_i(x) = \int_{s_{i-1}}^{s_i} K_h(u-x)du$ with $s_i = (X_i + X_{i+1})/2$, $X_0 = -\infty$ and $X_{n+1} = +\infty$, we obtain the GM regression estimator given by

$$\hat{m}_h(x) = \sum_{i=1}^n \int_{s_{i-1}}^{s_i} K_h(u - x)\, du Y_i\,.$$

See Gasser and Müller.[41]

Just like the kernel density estimate, the choice of bandwidth is critical to the quality of the estimate. A too large bandwidth yields an over-smooth estimate, while a too small bandwidth gives a rough estimate. The basic asymptotic properties of the NW and GM regression estimators have been well established. The asymptotic biases and variances of these two estimators are depicted in Table 1.[23] The properties on the GM estimator were established in Mack and Müller[63] and Chu and Marron.[16]

3.2. *Local polynomial regression*

As indicated in the last section, both the NW estimator and the GM estimator are a local constant fit. It is natural to extend this to a local polynomial fit. The idea of local polynomial regression has been around for a long time. Since both a local constant and local polynomial fits use

effectively datum points in a local neighborhood, this idea is referred as local modeling. It appeared in the statistical literature.[17,79] Stone[80,81] shows that local regression achieves optimal rates in a minimax sense. Müller[16] establishes the equivalence between a local polynomial fit and a local constant fit under an equally-spaced design model. Fan[23,24] focus on local linear regression in the random design case and show that it has many advantages, such as simple expression for local bias and variance, spatial adaptation and high minimax efficiency. Fan and Gijbels[28] proved that theoretically the local linear regression estimator adapts automatically to the boundary. This was also empirically observed by Tibshirani and Hastie.[82] Ruppert and Wand[70] extended the results of Fan and Gijbels[28] to the case of local polynomial estimation. A thorough study of this topic can also be found in Chaps. 3 and 4 of Fan and Gijbels.[28]

Suppose that the regression function m is smooth. For z in a neighborhood of x, it follows from using Taylor's expansion that

$$m(z) \approx \sum_{j=1}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=1}^{p} \beta_j (z-x)^j \,. \tag{11}$$

Thus, for X_i close enough to x,

$$m(X_i) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \equiv \mathbf{X}_i^T \boldsymbol{\beta} \,,$$

where $\mathbf{X}_i = (1, (X_i - x_0), \ldots, (X_i - x_0)^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. Intuitively datum points further from x have less information about $m(x)$. This suggests using a locally weighted polynomial regression

$$\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 K_h(X_i - x) \,. \tag{12}$$

Denote by $\hat{\beta}_j (j = 0, \ldots, p)$ the minimizer of (12). The above exposition suggests that an estimator for the regression function $m(x_0)$ is

$$\hat{m}(x_0) = \hat{\beta}_0(x_0) \,. \tag{13}$$

Furthermore, an estimator for the νth order derivative of $m(x_0)$ at x_0 is

$$\hat{m}_\nu(x_0) = \nu! \hat{\beta}_\nu(x_0) \,.$$

In general, local polynomial fitting has certain advantages over the NW and the GM estimators not only for estimating regression curves, but also for derivative estimation.

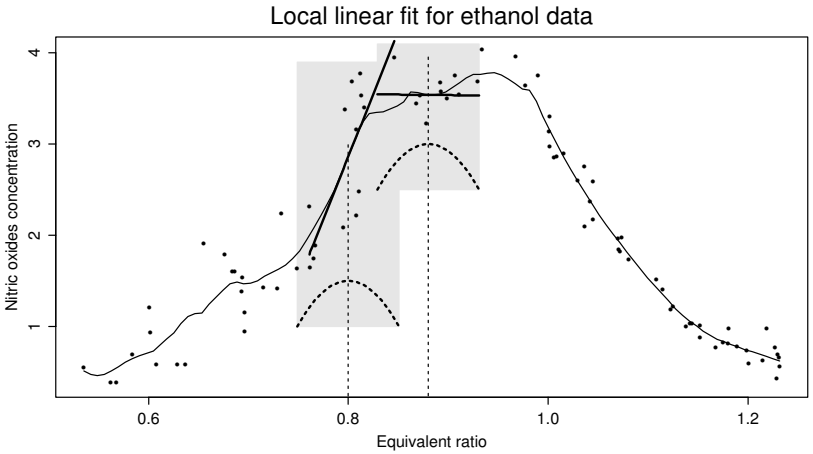


Fig. 9. Illustration of the local linear fit. For each given x_0, a linear regression is fitted through the data contained in the strip $x_0 \pm h$, using the weight function indicated at the bottom of the strip. The interactions of the fitted lines and the short dashed lines are the local linear fits. Adapted from Fan and Gijbels.[29]

To better appreciate the above local polynomial regression, consider the ethanol data presented in Fig. 8. The window size h is taken to be 0.051 and the kernel is the Epanechnikov kernel. To estimate the regression function at the point $x_0 = 0.8$, we use the local data in the strip $x_0 \pm h$ to fit a regression line (c.f. Fig. 9). The local linear estimate at x_0 is simply the intersection of the fitted line and the line $x = x_0$. Suppose that we wish to estimate the regression function at another point $x_0 = 0.88$, another line is fitted using the data in the window 0.88 ± 0.051. The whole curve is obtained by estimating the regression function in a grid of points. Indeed, the curve in Fig. 9 was obtained by 101 local linear regressions, taking the 101 grid points from 0.0535 to 1.232.

The *local linear regression smoother* is particularly simple to implement. Indeed, the estimator has the simple expression

$$\hat{m}_L(x) = \sum_{i=1}^{n} w_i(x) Y_i \,, \tag{14}$$

where with $S_{n,j}(x) = \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j$,

$$w_i(x) = K_h(X_i - x) \{S_{n,2}(x) - (X_i - x)S_{n,1}(x)\}/(S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)) \,. \tag{15}$$

We can either use the explicit formula (15) or a regression package to compute it. It has several nice properties such as high statistical efficiency (in an asymptotic minimax sense), design adaption[24] and good boundary behavior.[28,70] The asymptotic bias and variance for this estimator is

$$E\{\hat{m}_L(x)|X_1,\ldots,X_n\} - m(x) = \mu(K)\frac{m''(x)}{2}h^2 + o(h^2) \tag{16}$$

and

$$\text{var}\{\hat{m}_L(x)|X_1,\ldots,X_n\} = R(K)\frac{\sigma^2(x)}{f(x)nh} + o\left(\frac{1}{nh}\right), \tag{17}$$

provided that the bandwidth h tends to zero in such a manner that $nh \to \infty$, where f is the marginal density of X, namely, *the design density*.[29] Table 1 lists the leading term in the asymptotic bias and variance. By comparing the leading terms in the asymptotic variance, clearly the local linear fit uses locally one extra parameter without increasing its variability. But this extra parameter creates opportunities for significant bias reduction, particularly at the boundary regions and slope regions. This is evidenced by comparing their asymptotic biases.

Local linear fitting requires a choice for the smoothing parameter h and for the kernel function K. It is well known that the choice of the kernel function is of less importance in kernel smoothing. This holds truely for local polynomial regression. It has been shown that the Epanechnikov kernel is optimal in some sense. See Gasser, Müller, and Mamitzsch,[42] Granovsky and Müller[45] and Chap. 3 of Fan and Gijbels.[28]

The bandwidth selection is critical to all nonparametric estimators. A too-large bandwidth creates excessive biases in nonparametric estimates and a too small bandwidth results in a large variance in nonparametric estimate. There are two basic choices of bandwidth: subjective and data-driven. In subjective choices, data analysts use different bandwidths to estimate the regression function and choose the one that visually balances the bias and variance trade-off. Trials-and-errors are needed in this endeavor. Alternatively, one can present the nonparametric estimates using a few different bandwidths (c.f. Fig. 6 for a similar idea). The data-driven bandwidth is to let data themselves choose a bandwidth that balances the bias and variance, via minimizing certain estimated *Mean Integrated Square Errors* (MISE).

We now briefly discuss some data-driven choices of the bandwidth. By (16) and (17), the weighted MISE of the local linear estimator is

$$\frac{\mu(K)^2h^4}{4}\int\{m''(x)\}^2w(x)dx + \frac{R(K)}{nh}\int\frac{\sigma^2(x)}{f(x)}w(x)dx\,.$$

The asymptotic optimal bandwidth, that minimizes the asymptotic weighted MISE of $\hat{m}_L(x)$, is given by

$$h_{\text{opt}} = \left(\frac{R(K) \int \sigma^2(x) f^{-1}(x) w(x) dx}{\mu^2(K) \int \{m''(x)\}^2 w(x) dx} \right)^{1/5} n^{-1/5}, \tag{18}$$

where $w(x)$ is a weight function.

The optimal bandwidth involves the unknown regression function and the unknown density function of X. Hence it cannot be applied directly. There are many references on the topic of bandwidth selection. See Chap. 4 of Fan and Gijbels[28] and references therein. Here, we focus on the cross-validation method, which is conceptually simple, but needs intensive computation. Let $\hat{m}_{h,(-i)}(x)$ be the local linear regression estimator (12) without using the ith-observation (X_i, Y_i). We now analogously validate the "goodness-of-fit" by measuring the "prediction error" $Y_i - \hat{m}_{h,(-i)}(X_i)$. The cross-validation criterion measures the overall "prediction errors", which is defined by

$$CV(h) = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{m}_{h,(-i)}(X_i)\}^2 . \tag{19}$$

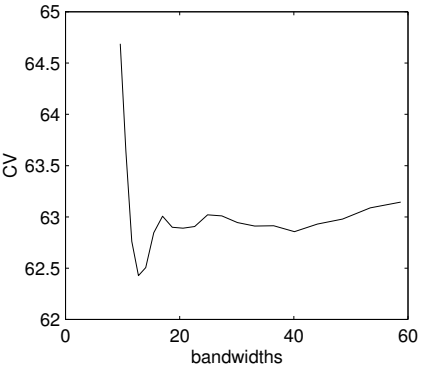
The cross-validation bandwidth selector $\hat{h}_{CV}$ chooses the one that minimizes $CV(h)$.

In what follows, we illustrate the methodology of local linear regression in details by an environmental data set. This data set consists of 612 observations of 15 variables and has been analyzed by Rawlings and Spruill.[69] See Sec. 2 of Rawlings and Spruill[69] for a detailed description. Here, we are interested in how *depth to mottling* (DMOT) of soil affects the increment of diameter growth of some kinds of pine. Thus we take the increment of diameter as a response variable Y and the DMOT of soil as an independent variable X. After excluding the data points with missing values, we have 216 observations. The scatter plot of the data is depicted in Fig. 10.

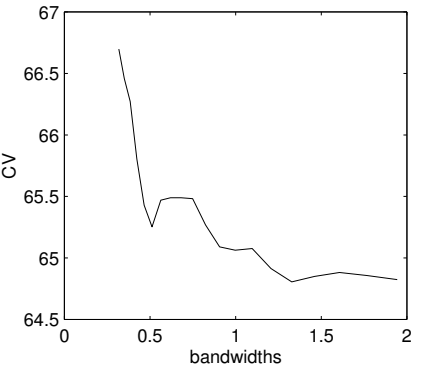
The cross-validation method was used to search a bandwidth over 20 grid points $0.15 * 1.1^j$ multiplying the range of X variable, $j = 0, \ldots, 19$. With the smallest bandwidth 0.15 multiplying the range of X, we used 15% of data around x_0 to estimate $m(x_0)$, while with the largest bandwidth 0.15×1.1^{19} multiple the range of X, we used about 92% of data around x_0 to estimate $m(x_0)$. Here the Epanechnikov kernel was used. The plot of cross-validation scores against candidate bandwidths is depicted in Fig. 10(a). The corresponding $\hat{h}_{CV}$ is 12.776.

With the selected bandwidth, we are able to estimate the regression function. In nonparametric regression, one usually plots the curve of the

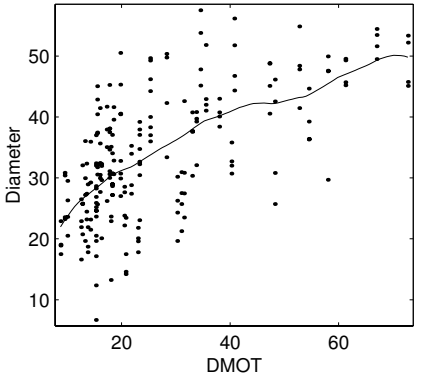
estimated regression function. Thus one has to evaluate the regression function over a grid of points. Usually we take the grid of points evenly distributing over the range of X. A natural question arises here is how many grid points at which the estimate needed to be evaluated. Figure 10(c)–(e) depicts the resulting estimated curve with the number of grid points (Ngrid) being 100, 200 and 400, respectively. The plots shows nonlinear between the increment and the DMOT. From these plots, the estimated curves



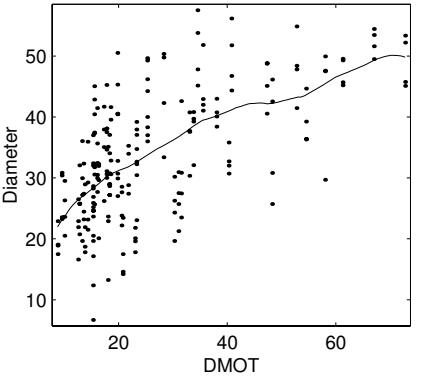
(a) Cross-Validation Scores



(b) Cross-Validation Scores

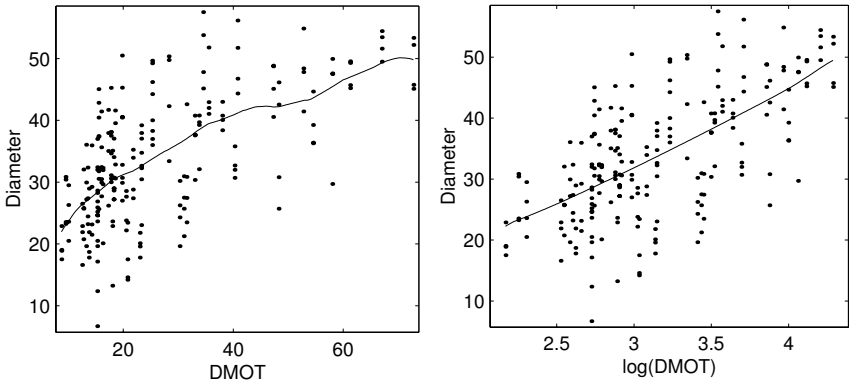


(c) Ngrid = 100



(d) Ngrid = 200

Fig. 10. Estimated regression functions. (a) and (b) are plots of cross-validation scores for increment of diameter versus depth of mottling (DMOT) and for versus log(DMOT), respectively. (c)–(e) are estimated regression function curves E(increment|DMOT) with scatter plot of data, corresponding to the number of grid points 100, 200 and 400, respectively. (h) is the estimated regression curve E(increment|log(DMOT)).



(e) Ngrid = 400 (f) Ngrid = 200

Fig. 10. (*Continued*).

are almost the same, since the underline estimate is relatively smooth. In practice, we recommend using 100 or 200 grid points to evaluate estimated regression functions.

Now we take the natural logarithm of DMOT as the X-variable, and then use the cross-validation method to choose a bandwidth. The CV scores are depicted in Fig. 10(b). This yields $\hat{h}_{CV} = 1.3271$. The estimated curve is depicted in Fig. 10(h). Figure 10(h) shows that increment of diameter growth versus log(DMOT) is nearly linear. For such an implementation, it spent about 2 seconds (using MATLAB on PC Pentium II 450 MHz) to compute the estimated function over 200 grid points, including bandwidth selection using the cross-validation method.

Direct implementation of local polynomial regression for a large data set needs a considerable amount of computation. Fast computation algorithms have been proposed in Fan and Marron.[33] Many computer codes are available through internet. For example, S-plus codes can be downloaded from Matt Wand's homepage at

http://www.biostat.harvard.edu/ mwand/software.html,

while Matlab codes can be downloaded from James S. Marron's homepage at

http://www.stat.unc.edu/faculty/marron/marron_software.html

or through the authors. These codes can be easily implemented by directly plugging-in data. There is also a procedure of kernel smoothing in the latest version of SAS.

4. Local Likelihood and Local Partial Likelihood

The local likelihood approach was first proposed by Tibshirani and Hastie[82] based on the running line smoother. As an extension of the local likelihood approach, local quasi-likelihood estimation using local constant fits, was considered by Severini and Staniwalis.[74] Fan, Heckman and Wand[31] investigated the asymptotic properties of the local quasi-likelihood method using local polynomial modeling. Fan *et al.*[27] addressed the issue of bandwidth selection, bias and variance assessment and constructed confidence intervals in local maximum likelihood estimation. Fan and Chen[26] proposed one-step local quasi-likelihood estimation, and demonstrated that the one-step local quasi-likelihood estimator performs as well as the maximum local quasi-likelihood estimator using the ideal optimal bandwidth. Fan *et al.*[30] extended the idea of the local likelihood approach to local partial likelihood in the context of censored survival data analysis, such as Cox's regression model. The ideas in this section are motivated from Fan *et al.*[30,31] Carroll *et al.*[11] extend the idea further to the likelihood equations.

4.1. *Generalized linear models and local likelihool estimate*

4.1.1. *Generalized linear models*

Generalized linear models introduced by Nelder and Wedderburn in 1972 extend the scope of the traditional least squares fitting of linear models. The relationship between a response variable and a set of covariates is modeled as a linear fit to the transformed conditional mean. A comprehensive account of generalized linear models can be found in McMullagh and Nelder.[65] Suppose that we have n independent observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of random vector $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional real vector of covariates, and Y is a scalar response variable. The conditional density of Y given covariate $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\{[\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)\} \tag{20}$$

for known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. In parametric generalized linear models it is usual to model a transformation of the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ as linear, that is

$$\eta(\mathbf{x}) = g\{m(\mathbf{x})\} = \mathbf{x}^T\boldsymbol{\beta}\,,$$

and g is a known *link* function. If $g = (b')^{-1}$, then g is called the canonical link because it transform the regression function into the canonical parameter: $(b')^{-1}\{m(\mathbf{x})\} = \theta(\mathbf{x})$.

Here are a few examples that illustrate the model (20). The first example is that the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ has a normal distribution with mean $m(\mathbf{x})$ and variance σ^2. The normal density can be rewritten as

$$f_{Y|\mathbf{X}} = \exp\left\{\frac{m(\mathbf{x})y - m^2(\mathbf{x})/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right\}.$$

It can be easily seen that

$$\phi = \sigma^2,\ a(\phi) = \phi,\ b(m) = m^2/2$$

and

$$c(y, \phi) = -y^2/(2\phi) - \log(\sqrt{2\pi\phi}).$$

The canonical link function is the identity link $g(t) = t$. This model is useful for a continuous response with homoscedastic errors.

Suppose that the conditional distributions of Y given $\mathbf{X} = \mathbf{x}$ is a Bernoulli distribution with the probability of success $p(\mathbf{x})$, in which case it can be seen that

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\left(y \log[p(\mathbf{x})/\{1 - p(\mathbf{x})\}] + \log\{1 - p(\mathbf{x})\}\right).$$

The canonical parameter in this example is $\theta(\mathbf{x}) = \text{logit}\{p(\mathbf{x})\}$, and the logit function is the canonical link.

Under model (20), it can be easily shown that the conditional mean and conditional variance are given respectively by $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = b'\{\theta(\mathbf{x})\}$, and $\text{var}(Y|\mathbf{X} = \mathbf{x}) = a(\phi)b''\{\theta(\mathbf{x})\}$. Hence,

$$\theta(\mathbf{x}) = (b')^{-1}\{m(\mathbf{x})\}.$$

Using the definition of $\eta(\cdot)$, we have

$$\theta(\mathbf{x}) = (b')^{-1}\{g^{-1}[\eta(\mathbf{x})]\}. \tag{21}$$

Since our primary interest is to estimate the mean function, without loss of generality, the factors related to the dispersion parameter ϕ are omitted. This leads to the following conditional log-likelihood function

$$\ell\{\theta, y\} = \theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}.$$

By (21), the above log-likelihood can be expressed as

$$\ell\{\theta, y\} = \left[y(b')^{-1} \circ g^{-1}(\eta(\mathbf{x})) - b\{(b')^{-1} \circ g^{-1}(\eta(\mathbf{x}))\}\right], \tag{22}$$

where $\circ$ denotes composition. In particular, when g is the canonical link,

$$\ell\{\theta, y\} = \eta(\mathbf{x})y - b\{\eta(\mathbf{x})\}.$$

4.1.2. *Local likelihood estimate*

It has been of interest to adapt these models to situations where the functional form for the dependence of $g(m(\mathbf{x}))$ on $\mathbf{x}$ is unknown. In what follows, the covariate $\mathbf{X}$ is assumed to be a scalar random variable. If $\eta(x)$ is a smooth function of x, then for X_i close enough to a given point x_0,

$$\eta(X_i) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \equiv \mathbf{X}_i^T \boldsymbol{\beta}\,, \tag{23}$$

where $\mathbf{X}_i = (1, (X_i - x_0), \ldots, (X_i - x_0)^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. Intuitively data points close to x_0 have more information about $\eta(x_0)$ than those away from x_0. Therefore, by (22), the local log-likelihood function based on the random sample $\{(X_i, Y_i)\}_{i=1}^n$ is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i (b')^{-1} \circ g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta}) - b\{(b')^{-1} \circ g^{-1}(\mathbf{X}_i \boldsymbol{\beta})\}] K_h(X_i - x_0)\,. \tag{24}$$

Define the local maximum likelihood estimator of $\boldsymbol{\beta}$ to be

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in R^{p+1}} \ell(\boldsymbol{\beta})\,.$$

Thus $\eta(x_0)$ and the νth derivative of $\eta(x_0)$ can be estimated by

$$\hat{\eta}(x_0) = \hat{\beta} \quad \text{and} \quad \hat{\eta}^{(\nu)} = \nu! \hat{\beta}_\nu$$

respectively, assuming that η has p derivatives. When the canonical like $g = (b')^{-1}$ is used, (24) becomes

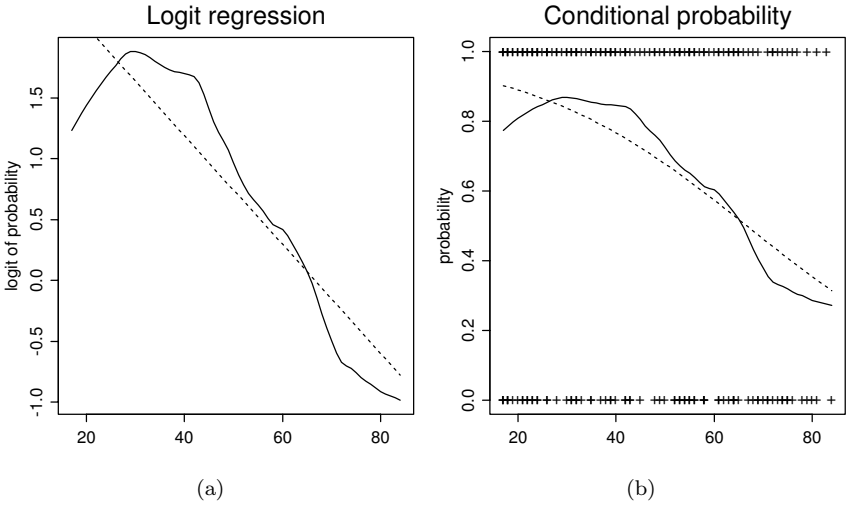
$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i (\mathbf{X}_i^T \boldsymbol{\beta}) - b(\mathbf{X}_i \boldsymbol{\beta})] K_h(X_i - x_0)\,.$$

The log-likelihood function (24) is really a weighted log-likelihood and hence can be computed by using the existing software. In fact, suppose that we want to estimate $\hat{\eta}(\cdot)$ in a given interval. Take a grid of points (200, say) in that interval. For each given grid point x_0, model (24) can be maximized by using existing software packages such as SAS and Splus that contains the parametric Glim function. The whole estimated function is obtained by plotting the estimates obtained at grid points.

The choice of the link function g is not as crucial as for parametric generalized linear models, because the fitting is localized. Indeed it is conceivable to dispense with the link function and just estimate $m(x)$ directly. But there are several drawbacks to having the link equal to the identity. An

identity link may yield a local likelihood that is not convex, allowing for the possibility of multiple maxima, inconsistency and computational problems. Use of the canonical link guarantees convexity. Furthermore the canonical link ensures that the final estimate has the correct range. For example, in the logistic regression context using the logit link leads to an estimate that is always a probability whereas using the identity link does not have. A final reason for preferring the canonical link is that the estimate of $m(x)$ approaches the usual parametric estimate as the bandwidth becomes large. This can be useful as a diagnostic tool.[31]

We now illustrate the local likelihood approach via analyzing the data set: *Burns data*, collected by General Hospital Burn Center at the University of South California. It is of interest to estimate the probability of surviving given the age of victims. Local likelihood estimate was computed over a grid of points with bandwidth 0.4 multiplying the range of X, and the estimated curves are depicted in Fig. 11. Note that the conditional probability function must be monotonic for the parametric linear model, whereas for the local linear model, the conditional probability function can be any curve. The former model can overstate the probability of survival for the younger group and for the senior group. The solid curves in Fig. 11 suggest



(a) (b)

Fig. 11. Illustration of local likelihood approach for the burn data. (a) Estimated logit transform of the conditional probability. (b) Estimated conditional probability. Solid curve — local modeling with about 40% of the data points in each local neighborhood; dashed curve — global parametric logit linear model. Taken from Fan and Gijbels.[29]

that the conditional probability function is unimodal, which is reasonable in the current context.

4.2. *Local partial likelihood estimate*

In this section, we apply the local likelihood techniques to survival data analysis. Let T, C and X be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \leq C)$ be the censoring indicator. It is assumed that T and C are conditionally independent given X and that the censoring mechanism is noninformative. Suppose that $\{(X_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ are an i.i.d. sample from the population (X, Z, δ). For a thorough introduction to survival analysis, see books by Fleming and Harrington[40] and Andersen *et al.*[2]

Let $h(t|x)$ be the conditional hazard rate function. The proportional hazards model assumes that

$$h(t|x) = h_0(t) \exp\{\theta(x)\}. \tag{25}$$

This model indicates that the covariate x inflates or deflates the hazard risk by a factor of $\exp\{\theta(x)\}$. The function $\theta(x)$ is called a hazard regression function, and characterizes the risk contribution of the covariate at value x. See Cox[19] for proportional hazard models with time-dependent covariates.

In the parametric model, a linear form $\theta(x) = \beta x$ is imposed on the hazard regression function. The local modeling methodology aims at removing this restriction and exploring possible nonlinearity, and is applicable to any smooth hazard regression function. For simplicity of discussion, we focus on the univariate cases. For multivariate settings, a dimensionality reduction technique such as additive models should be used.[52]

A commonly-used technique for estimating the hazard regression function is the *partial likelihood* technique introduced by Cox.[20] Let $t_1^o < \cdots < t_N^o$ denote the ordered observed failure times. Let (j) provide the label for the item failing at t_j^o so that the covariates associated with the N failures are $X_{(1)}, \ldots, X_{(N)}$. Denote by $R_j = \{i : Z_i \geq t_j^o\}$, the risk set at time instantaneously before t_j^o. Then, the log-partial likelihood in our context is given by

$$\sum_{j=1}^{N} \left[\theta(X_{(j)}) - \log \left(\sum_{k \in R_j} \exp\{\theta(X_k)\} \right) \right]. \tag{26}$$

See Cox,[20] Fleming and Harriton[40] and Fan and Gijbels.[28] Substituting the parametric form of $\theta(\cdot)$ into (26) yields a maximum partial likelihood estimate of the hazard regression function.

We now apply the local modeling technique to estimate the hazard regression function $\theta(\cdot)$. For a given x_0, approximate $\theta(x)$ by

$$\theta(x) \approx \beta_0 + \cdots + \beta_p(x - x_0)^p \,, \tag{27}$$

for x in a neighborhood of x_0. Let

$$\beta = (\beta_1, \ldots, \beta_p)^T \quad \text{and} \quad \mathbf{X}_j = \{(X_j - x_0), \ldots, (X_j - x_0)^p\}^T \,.$$

Then the *local partial likelihood* is

$$\sum_{j=1}^{N} K_h(X_{(j)} - x_0) \left[\mathbf{X}_{(j)}^T \beta - \log \left\{ \sum_{k \in R_j} \exp(\mathbf{X}_k^T \beta) K_h(X_k - x_0) \right\} \right] . \tag{28}$$

See Fan *et al.*[30] for a derivation of the local partial likelihood (28). When the kernel function is uniform and the bandwidth is of the nearest neighbor type, the local likelihood (28) was introduced by Tibshirani and Hastie.[82] For a related approach based on the local likelihood, see Gentleman and Crowley.[43]

The function value $\theta(x_0)$ is not directly estimable since (28) does not depend on the intercept β_0. However, the derivative functions are directly estimable. Let $\hat{\beta}(x_0)$ be the maximum local log-partial likelihood estimate that maximizes (28). An estimate $\hat{\theta}_\nu(x_0)$ of $\theta^{(\nu)}(x_0)$ is given by $\nu!\hat{\beta}_\nu(x_0)$.

We impose the condition $\theta(0) = 0$ for identifiability. With this extra constraint, the function $\theta(x)$ can be estimated by

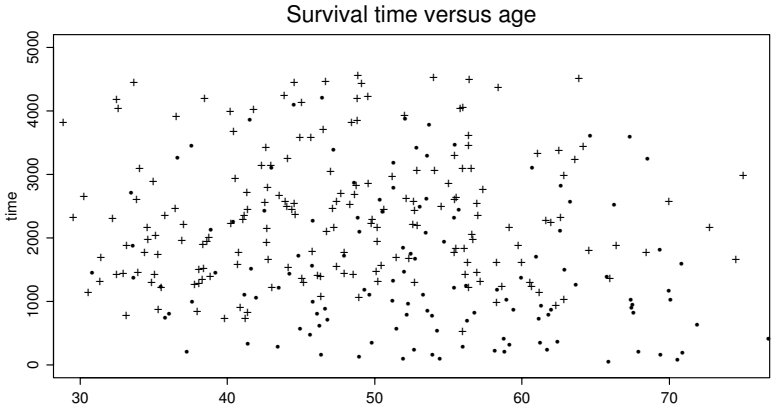
$$\hat{\theta}(x) = \int_0^x \hat{\theta}'(t)dt \,, \tag{29}$$

where $\hat{\theta}'(t) = \hat{\theta}_1(t)$ is the derivative estimator. In practice, the function $\hat{\theta}_1(x)$ is often evaluated at either grid points or the design points. Assume that $\hat{\theta}_1(x_j) = \hat{\beta}_1(x_j)$ are computed at points $\{x_0, \ldots, x_m\}$. Then, $\hat{\theta}(x_i)$ can be approximated by the trapezoidal rule

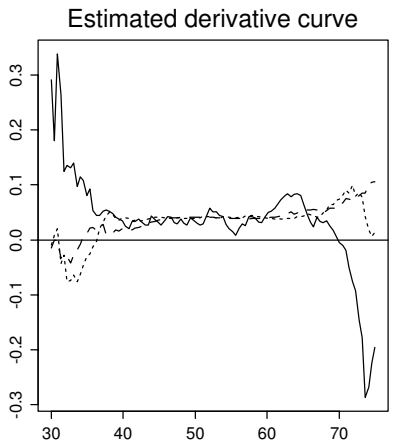
$$\hat{\theta}(x_i) = \sum_{j=1}^{i} (x_j - x_{j-1})(\hat{\beta}_{1,j} + \hat{\beta}_{1,j-1})/2 \,,$$

where $\hat{\beta}_{1,j} = \hat{\beta}_1(x_j)$. The coefficients can simply be computed by using existing software packages for parametric Cox's proportional hazards model.

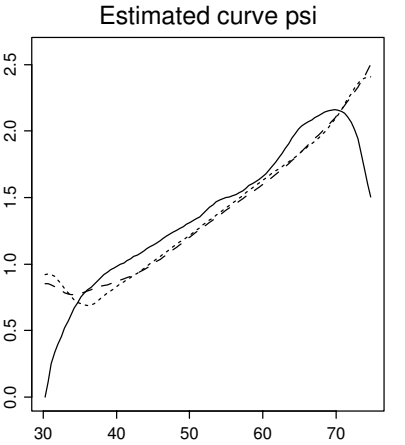
We conclude this section with an analysis of the Primary Biliary Cirrhosis (PBC) data set, which can be found in Fleming and Harrington.[40]



(a)



(b)



(c)

Fig. 12. Local partial likelihood estimation of the hazard regression function. (a) Observed time versus age with "+" indicating censored observations. (b) Estimated derivative function $\theta'(\cdot)$. (c) Estimated hazard regression function $\theta(\cdot)$; solid curve — bandwidth = 10; short-dashed curve — bandwidth = 20; long-dashed curve — bandwidth = 30. From Fan and Gijbels.[29]

PBC is a rare but fatal chronic liver disease of unknown cause. The analysis is here based on the data collected at Mayo Clinic between January 1974 and May 1984. Of 312 patients who participated in the randomized trial, 187 cases were censored. In our analysis, we take the time (in days) between

registration and death, or liver transplantation or the time of the study analysis (July 1986) as response and the ages of the patients as a covariate. The observed data are presented in Fig. 12(a). The local partial likelihood method (28) with $p = 2$ was employed for three different bandwidths $h = 10$, 20 and 30. The estimated hazard regression function and its derivative function are respectively given in Figs. 12(b) and (c). Note that since the hazard regression function is only identifiable within a constant, the curves in Fig. 12(c) are normalized to have the same average height so that they can be better compared. Figure 12(c) reveals the fact that it is reasonable to model linearly the hazard regression function of covariate age.

5. Nonparametric Goodness of Fit Tests

Nonparametric goodness of fit test has received increasing attention recently. A motivating and simple example is to consider a simple nonparametric regression model. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from the nonparametric regression model

$$Y_i = m(X_i) + \varepsilon_i$$

with $E(\varepsilon_i|X_i) = 0$ and $\text{var}(\varepsilon_i|X_i) = \sigma^2$. Of interest is to test the hypothesis

$$H_0 : m(x) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x^p \text{ versus } H_1 : m(x) \neq \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x^p .$$

This testing problem is well known as test of linearity in the context of model diagnostic where the question arises whether a family of parametric models fit adequately the data. It is natural to use the nonparametric model as an alternative hypothesis. On the other hand, it is known that nonparametric regression may yield a complicated model. Thus, after fitting a data set by a nonparametric model, we may check whether the data can be fitted by a less complicated parametric model. This leads to a nonparametric goodness of fit test. Hart[51] gives a comprehensive study and presents many examples on this topic. Fan[25] and Fan and Huang[32] proposed some goodness of fit tests for various parametric models and nonparametric models. Fan *et al.*[37] proposed generalized likelihood ratio tests and established a general framework for nonparametric smoothing tests. Many related literature, are available[1,3,4,21,22,50,55,58] In this section, we illustrate the idea of nonparametric likelihood ratio test by generalized varying coefficient models. Some material of this section was extracted from Cai *et al.*,[8] referred as CFL.

5.1. *Generalized varying coefficient models*

A generalized varying-coefficient model has the form

$$\eta(\mathbf{u}, \mathbf{x}) = g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p} a_j(\mathbf{u}) x_j \tag{30}$$

for some given link function $g(\cdot)$, where $\mathbf{x} = (x_1, \ldots, x_p)^T$, and $m(\mathbf{u}, \mathbf{x})$ is the mean regression function of the response variable Y given the covariates $\mathbf{U} = \mathbf{u}$ and $\mathbf{X} = \mathbf{x}$ with $\mathbf{X} = (X_1, \ldots, X_p)^T$. Clearly, model (30) includes both the parametric generalized linear model[65] and the generalized partially linear model.[9,14,46,77] An advantage of model (30) is that by allowing the coefficients $\{a_j(\cdot)\}$ to depend on $\mathbf{U}$, the modeling bias can be reduced significantly and the "curse of dimensionality" is avoided.

In this section, we focus on the cases in which the response is discrete. For continuous responses, many works have been done. In the least-squares setting, model (30) with the identity link was introduced by Cleveland *et al.*[18] and extended by Hastie and Tibshirani[53] to various aspects. Varying-coefficient models are a simple and useful extension of classical generalized linear models. They are particularly appealing in longitudinal studies where they allow one to explore the extent to which covariates affect responses changing over time. See Hoover *et al.*,[54] Brumback and Rice[7] and Fan and Zhang[38] for details on novel applications of the varying-coefficient models to longitudinal data. For nonlinear time series applications, see Chen and Tsay[15] and Cai *et al.*[9] for statistical inferences based on the functional-coefficient autoregressive models. Kauermann and Tutz[59] used varying coefficient models for model disgnostics.

5.2. *Estimation procedures*

For simplicity, we consider the important case that $\mathbf{u}$ is one-dimensional. Extension to multivariate $\mathbf{u}$ involves no fundamentally new ideas. However, implementations with $\mathbf{u}$ having more than two dimensions may have some difficulties due to the "curse of dimensionality".

In this section, it is assumed that the conditional log-likelihood function $\ell(v, y)$ is known and linear in y for fixed v. This assumption is satisfied for the canonical exponential family, which is the focus of this section. The methods, introduced in this section, are directly applicable to situations in which one cannot specify fully the conditional log-likelihood function $\ell(v, y)$, but can model the relationship between the mean and

variance by $\text{var}(Y|\mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}) = \sigma^2 V\{m(\mathbf{u}, \mathbf{x})\}$ for a known variance function $V(\cdot)$ and unknown σ. In this case, one needs only to replace the log-likelihood function $\ell(v, y)$ by the quasi-likelihood function $Q(\cdot, \cdot)$, defined by $\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}$.

5.2.1. *Local MLE*

Local linear modeling will be used here, though general local polynomial methods are also applicable. Suppose that $a_j(\cdot)$ has a continuous second derivative. For each given point u_0, $a_j(u)$ can be approximated locally by a linear function $a_j(u) \approx a_j + b_j(u - u_0)$ for u in a neighborhood of u_0. Based on a random sample $\{(U_i, \mathbf{X}_i, Y_i)\}_{i=1}^n$, one may use the following local likelihood method to estimate the coefficient functions

$$\ell_n(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left[g^{-1} \left\{ \sum_{j=1}^{p} (a_j + b_j(U_i - u_0)) X_{ij} \right\}, Y_i \right] K_h(U_i - u_0)\,, \tag{31}$$

where $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{b} = (b_1, \ldots, b_p)^T$. Note that a_j and b_j depend on u_0, and so does $\ell_n(\cdot, \cdot)$. Maximizing the local likelihood function $\ell_n(\mathbf{a}, \mathbf{b})$ results in estimates $\hat{\mathbf{a}}(u_0)$ and $\hat{\mathbf{b}}(u_0)$. The components in $\hat{\mathbf{a}}(u_0)$ provide an estimate of $a_1(u_0), \ldots, a_p(u_0)$. For simplicity of notation, let $\boldsymbol{\beta} = \boldsymbol{\beta}(u_0) = (a_1, \ldots, a_p, b_1, \ldots, b_p)^T$, and write the local likelihood function (31) as $\ell_n(\boldsymbol{\beta})$. Likewise, the local MLE is denoted by $\hat{\boldsymbol{\beta}}_{\text{MLE}} = \hat{\boldsymbol{\beta}}_{\text{MLE}}(u_0)$. The sampling properties have been established in CFL.

5.2.2. *One-step local MLE*

Computation for the above local MLE is expensive. We have to maximize the local likelihood (31) for usually hundreds of distinct values of u_0, with each maximization requiring an iterative algorithm, in order to obtain the estimated functions $\{\hat{a}_j(\cdot)\}$. To alleviate this expense, we replace an iterative local MLE by the one-step estimator, which has been frequently used in parametric models.[5,61] The one-step local MLE does not lose any statistical efficiency provided that the initial estimator is good enough. See CFL for theoretic insights.

Let $\ell_n'(\boldsymbol{\beta})$ and $\ell_n''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local log-likelihood $\ell_n(\boldsymbol{\beta})$. Given an initial estimator $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0(u_0) = (\hat{\mathbf{a}}(u_0)^T, \hat{\mathbf{b}}(u_0)^T)^T$, one-step of the Newton-Raphson algorithm updates its

solution by

$$\hat{\boldsymbol{\beta}}_{\mathrm{OS}} = \hat{\boldsymbol{\beta}}_0 - \{\ell_n''(\hat{\boldsymbol{\beta}}_0)\}^{-1}\ell_n'(\hat{\boldsymbol{\beta}}_0)\,, \tag{32}$$

thus featuring the computational expediency of least-squares local polynomial fitting. Furthermore, the sandwich formula can be used as an estimate for standard errors of the resulting estimate

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{OS}}) = \{\ell_n''(\hat{\boldsymbol{\beta}}_0)\}^{-1}\widehat{\mathrm{cov}}\{\ell_n'(\hat{\boldsymbol{\beta}}_0)\}\{\ell_n''(\hat{\boldsymbol{\beta}}_0)\}^{-1}\,.$$

This formula has been tested in CFL to be accuracy enough for most of practical purpose.

In univariate generalized linear models, Fan and Chen[26] carefully studied properties of the local one-step estimator. In that setting, the least-squares estimate serves a natural candidate as an initial estimator. However, in the multivariate setting, it is not clear how an initial estimator can be constructed. The following is proposed in CFL. Suppose that we wish to evaluate the functions $\hat{\mathbf{a}}(\cdot)$ at grid points $u_j, j = 1, \ldots, n_{\mathrm{grid}}$. Our idea of finding initial estimators is as follows. Take a point u_{i_0}, usually the center of the grid points. Compute the local MLE $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{i_0})$. Use this estimate as the initial estimate for the point u_{i_0+1} and apply (32) to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0+1})$. Now, use $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0+1})$ as the initial estimate at the point u_{i_0+2} and apply (32) to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0+2})$ and so on. Likewise, we can compute $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0-1})$, $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0-2})$, etc. In this way, we obtain our estimates at all grid points.

A refine alternative of the above proposal is to calculate a fresh local MLE as a new initial value after iterating along the grid points for a while. For example, if we wish to evaluate the functions at 200 grid points and are willing to compute the local maximum likelihood at five distinct points. A sensible placement of these points is $u_{20}, u_{60}, u_{100}, u_{140}$ and u_{180}. Use for example $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{60})$ along with the idea in the last paragraph to compute $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_i)$ for $i = 40, \ldots, 79$, and use $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{100})$ to compute $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_i)$ for $i = 80, \ldots, 119$, and so on.

Note that $\ell_n''(\hat{\boldsymbol{\beta}}_0)$ can be nearly singular for certain u_0, due to possible data sparsity in certain local regions. Seifert and Gasser[73] and Fan and Chen[26] explored the use of the ridge regression as an approach to handling such problems in the univariate setting. See CFL[8] for details.

5.3. *Hypothesis testing*

When fitting a varying-coefficient model, it is natural to ask whether the coefficient functions are actually varying or whether any particular covariate

Table 2. Normalization constant r_K.

Kernel	**Uniform**	**Epanechnikov**	**Biweight**	**Triweight**	**Gaussian**
r_K	1.2000	2.1153	2.3061	2.3797	2.5375

is significant in the model. For simplicity of description, we only consider the first hypothesis testing problem

$$H_0 : a_1(u) \equiv a_1, \ldots, a_p(u) \equiv a_p \,, \tag{33}$$

though the technique also applies to other testing problems. A useful procedure is based on the nonparametric likelihood ratio test statistic

$$T = 2\{\ell(H_1) - \ell(H_0)\} \,, \tag{34}$$

where $\ell(H_0)$ and $\ell(H_1)$ are respectively the log-likelihood functions computed under the null and alternative hypotheses. Note that the normalization constant in (34) does not change the testing procedure. However, in order for it to possess a χ^2 distribution, it needs to be normalized as in Ref. 37

$$T_K = r_K\{\ell(H_1) - \ell(H_0)\} \,, \tag{35}$$

where

$$r_K = \frac{K(0) - \frac{1}{2}\int K^2(t)dt}{\int (K(t) - \frac{1}{2}K * K(t))^2 dt} \,.$$

Table 2 gives the value of r_K for a few commonly used kernels.

For parametric models, it is well known that the likelihood ratio statistic follows asymptotically a χ^2-distribution. The asymptotic null distribution is independent of nuisance parameters under the null hypothesis. This is the Wilks type of phenomenon. Fan *et al.*[37] has shown the Wilks phenomenon still holds for the nonparametric likelihood ratio tests. Furthermore, they showed that the null distribution of the nonparametric likelihood ratio test is a χ^2-distribution in some sense and does not depend on the values of $a_1, \ldots, a_p$. Thus one may use the following *conditional bootstrap* to construct the null distribution of T_K and hence the P-value. Let $\{\hat{a}_j\}$ be the MLE under the null hypothesis. Given the covariates $(U_i, \mathbf{X}_i)$, generate a bootstrap sample Y_i^* from the given distribution of Y with the estimated linear predictor $\hat{\eta}(U_i, \mathbf{X}_i) = \sum_{j=1}^p \hat{a}_j X_{ij}$ and compute the test statistic T_K^* in (34). Use the distribution of T_K^* as an approximation to the distribution of T_K.

Note that the above conditional bootstrap method applies readily to setting without presence of dispersion parameter, such as the Poisson and Bernoulli distributions. It is really a simulation approximation to the conditional distribution of T_K given observed covariates under the particular null hypothesis: $H_0 : a_j(u) = \hat{a}_j$ $(j = 1, \ldots, p)$. As pointed out above, this approximation is valid under both H_0 and H_1 as the null distribution does not asymptotically depend on the values of $\{a_j\}$. In the case where model (30) involves a dispersion parameter (e.g. the Gaussian model), the dispersion parameter should be estimated based on the residuals from the *alternative* hypothesis.

It is also of interest to investigate whether some covariates are significant. For example, we want to check whether the covariate X_p can be excluded from the model. This is equivalent to testing the hypothesis $H_0 : a_p(\cdot) = 0$, the above conditional bootstrap idea can be employed to obtain the null distribution of T_K under the model (30) and the generalized likelihood ratio statistics continue to apply. In this case, the data should be generated from the mean function $g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p-1} \hat{a}_j(\mathbf{u}) x_j$, where $\hat{a}_j(\cdot)$ is an estimate under the alternative hypothesis.

5.4. *An application*

We conclude this section via illustrating the proposed methodology to analyze the *Burn Data* set. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise, and covariates $X_1 = age$, $X_2 = sex$, $X_3 = \log(\text{burn area} + 1)$ and binary variable $X_4 = Oxygen$ (0 if oxygen supply is normal, 1 otherwise) are considered. Of interest is to study how burn areas and the other variables affect the survival probabilities for victims at different age groups. This naturally leads to the following varying-coefficient model

$$\begin{aligned} \text{logit}\{p(x_1, x_2, x_3, x_4)\} &= a_1(x_1) + a_2(x_1)x_2 \\ &\quad + a_3(x_1)x_3 + a_4(x_1)x_4 \,. \end{aligned} \tag{36}$$

Figure 13 presents the estimated coefficients for model (36) via the one-step approach with bandwidth $h = 65.7882$, selected by a cross-validation method. See CFL[8] for details.

A natural question arises whether the coefficients in (36) are actually varying. To see this, we consider the parametric logistic regression model

$$\text{logit}\{p(x_1, x_2, x_3, x_4)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{37}$$

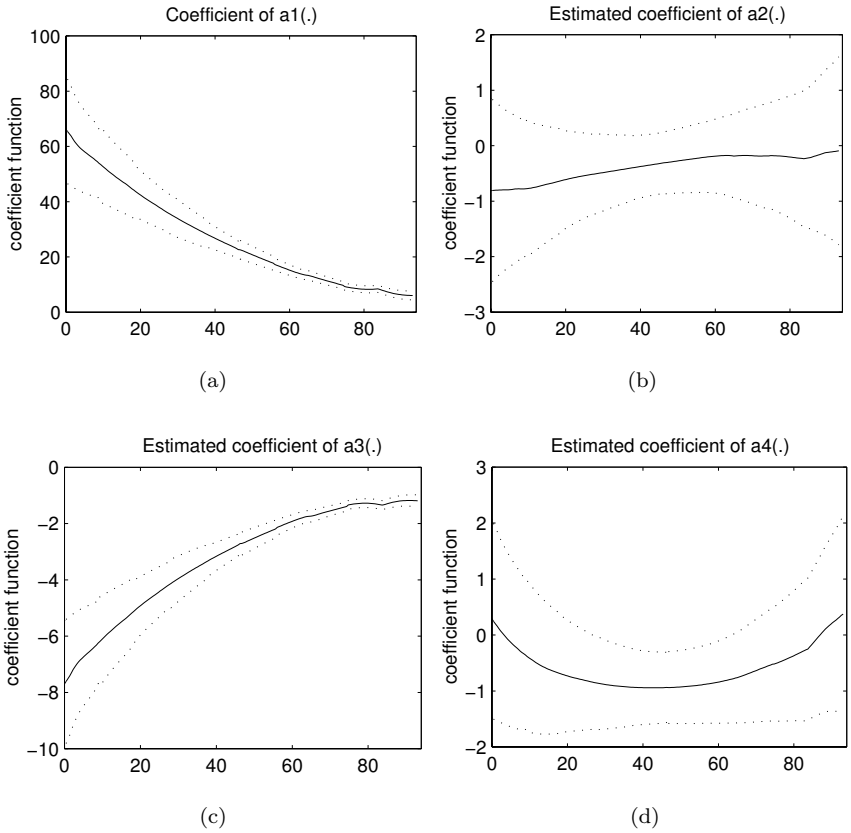


Fig. 13. The estimated coefficient functions (the solid curves) via the one-step approach with bandwidth chosen by the CV. The dot curves are the estimated functions plus/minus twice estimated standard errors. Adapted from Cai, Fan and Li.[8]

as the null model. As a result, the MLE of $(\beta_0, \ldots, \beta_4)$ in model (37) and its standard deviation are $(23.2213, -6.1485, -0.4661, -2.4496, -0.9683)$ and (1.9180, 0.6647, 0.2825, 0.2206, 0.2900), respectively. The likelihood ratio test T_K is 58.1284 with p-value 0.000, based on 1000 bootstrap samples (the sample mean and variance of T_K^* are 6.3201 and 11.98023, respectively). This implies that the varying-coefficient logistic regression model fits the data much better than the parametric fit. It also allows us to examine the extent to which the regression coefficients vary over different ages. The estimated density of T_K^* is depicted in Fig. 14, from which we can seen that the null distribution is well approximated by a χ^2 distribution with 6.5 degrees of freedom (a gamma distribution).

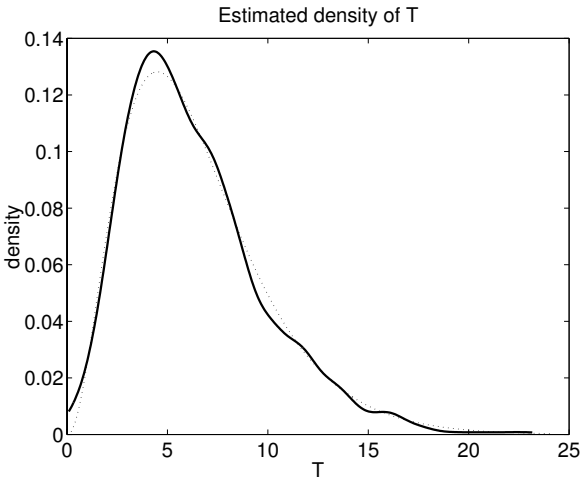


Fig. 14. The estimated density of T_K by Monte Carlo simulation. The solid curve is the estimated density, and the dashed curve stands for the density of chi-squared distribution (gamma distribution) with 6.5 degrees of freedom.

To examine whether there is any gender gap for different age groups or if the variable X_4 affects the survival probabilities for different age of burn victims, we consider testing the null hypothesis H_0 : both $a_2(\cdot)$ and $a_4(\cdot)$ are constant under model (36). The corresponding test statistic T_K is 3.4567 with p-value 0.7050 based on 1000 bootstrap samples. This in turn suggests that the coefficient functions $a_2(\cdot)$ and $a_4(\cdot)$ are independent of age and indicates that there are no gender differences for different age groups.

Finally, we examine whether both covariates *sex* and *oxygen* are statistically significant in model (36). The likelihood ratio test for this problem is $T_K = 11.9256$ with p-value 0.0860, based on 1000 bootstrap samples (the sample mean and variance of T_K^* are 5.5915 and 10.9211, respectively). Both covariates *sex* and *oxygen* are not significant at level 0.05. This suggests that gender and oxygen do not play a significant role in determining the survival probability of a victim.

6. Other Applications

There are many other applications of local modeling methods. This section briefly introduces some of them and gives some relevant references for those who wish for more details. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from a population (X, Y). We are interested in estimating

a population parameter function θ. The function $\theta(\cdot)$ can be, for example, the conditional mean function $E(Y|X)$ and the conditional quantile function. In parametric settings, we model $\theta(x)$ using a parametric family $\theta(x) = g(x;\beta)$. To get an estimator of β, we optimize (either minimize or maximize) an objective function

$$L(\beta) = \sum_{i=1}^{n} \ell\{X_i, Y_i, g(X_i,\beta)\}\,. \tag{38}$$

Here ℓ is a discrepancy loss function or the log-likelihood function of an individual observation. For example, the L_2-loss function leads to a least squares estimate, while the L_1-loss function corresponds to a robust linear regression.

The local modeling method can be used to relax the global parametric model assumption and to significantly reduce the modeling bias. For a given point x_0, we replace the objective function by its local version

$$L\{\beta(x_0)\} = \sum_{i=1}^{n} \ell\{X_i, Y_i, g(X_i,\beta(x_0))\}K_h(X_i - x_0)\,. \tag{39}$$

Optimizing (39) yields an estimate $\hat{\beta}(x_0)$, just like the local likelihood estimate discussed in the last section. Thus, an estimate of the function $\theta(\cdot)$ by $\hat{\theta}(x_0) = g\{x_0;\hat{\beta}(x_0)\}$. Since the local estimate $\hat{\beta}(x_0)$ optimizes (39), the estimate $g\{x_0,\hat{\beta}(x_0)\}$ should converge to its population version. Therefore the estimate $\hat{\theta}(x_0)$ is a consistent estimator of the function $\theta(x_0)$ if $h \to 0$ in such a way that $nh \to \infty$.

For a given x_0, by Taylor's expansion, we can parametrize the function in a local neighborhood of x_0 as

$$g(x;\beta) = \beta_0(x_0) + \beta_1(x_0)(x - x_0) + \cdots + \beta_p(x_0)(x - x_0)^p\,. \tag{40}$$

With suppressing the dependence of β's on x_0, (39) can be rewritten as

$$L\{\beta(x_0)\} = \sum_{i=1}^{n} \ell\{X_i, Y_i, \beta_0 + \beta_1(X_i - x_0) + \cdots + \beta_p(X_i - x_0)^p\}$$
$$\times K_h(X_i - x_0)\,. \tag{41}$$

Let $\hat{\beta}_j$ $(j = 0, 1, \ldots, p)$ optimize (41). Then as in last section,

$$\hat{\theta}(x_0) = \hat{\beta}_0$$

and

$$\hat{\theta}_\nu(x_0) = \nu!\hat{\beta}_\nu\,, \quad \nu = 1, \ldots, p$$

estimates the νth derivative of the function $\theta(x)$ at $x = x_0$.

It is clear that local polynomial regression and local likelihood approach are special cases hereof. An extension of the ideas for estimating bias and variance can be found in Fan *et al.*,[27] in which methods for selecting bandwidths and constructing confidence intervals are also proposed. A closely related framework is the local estimating equation method introduced by Carroll *et al.*[11] and the kernel generalized estimating equation (GEE) proposed by Lin and Carroll.[62]

6.1. *Estimation of conditonal quantiles and median*

In explanatory data analysis, quantiles provide us informative summary of a population. In regression analysis, conditional quantiles have important applications for constructing predictive intervals and detecting heteroscedasticity. When the error distribution is asymmetric, the conditional median regression function is more informative than the conditional mean regression.

Take the loss function in (38) to be $\ell(x, y, \theta) = \ell_\alpha(y - \theta)$ with

$$\ell_\alpha(t) = |t| + (2\alpha - 1)t\,. \tag{42}$$

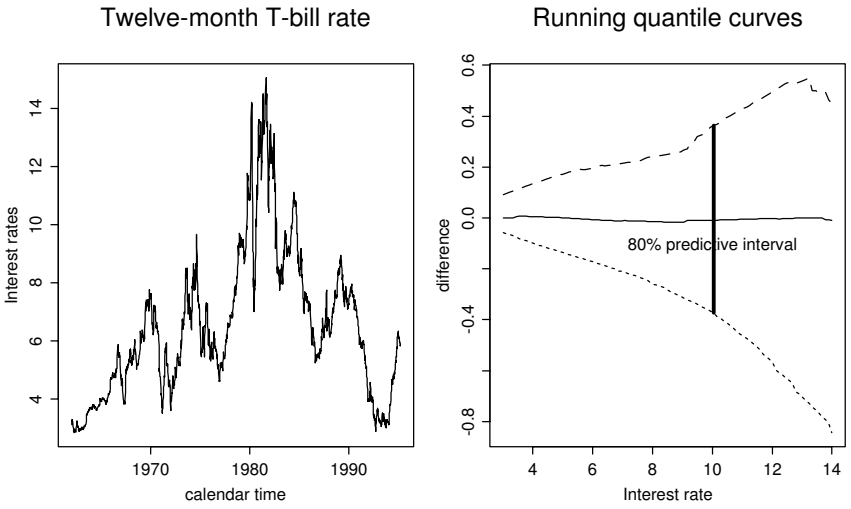
The minimizer of $E\ell(X, Y, \theta)$ in this situation is the conditional α-quantile function $\xi_\alpha(x) = G^{-1}(\alpha|x)$, where $G^{-1}(y|x)$ is the conditional distribution of Y given $X = x$.

Now we apply the local modeling approach to estimate the conditional quantile function. Minimize

$$\sum_{i=1}^{n} \ell_\alpha\{Y_i - \beta_0 - \beta_1(X_i - x_0) - \cdots - \beta_p(X_i - x_0)^p\}K_h(X_i - x_0) \tag{43}$$

and the resulting estimator for $\xi_\alpha(x_0)$ is simply $\hat{\beta}_0$.

Now we apply the proposed approach to the 12-month Treasury bill data presented in Fig. 15(a). Figure 15(b) depicts the estimated conditional median, the conditional 10th percentile and the conditional 90th percentile. The fan shape of the conditional quantiles shows that the variability gets larger as the interest rate gets higher. The intervals sandwiched by conditional 10th and 90th percentiles are 80%-predictive intervals. For example, given the current interest rate being 10%, with probability 80% the difference of the next week's rate and this week's rate falls in the interval $[-0.373\%, 0.363\%]$.



(a) (b)

Fig. 15. Quantile regression. (a) The yields of 12-month Treasury bill. (b) Conditional quantiles ----- : $\alpha = 0.1$, solid curve — $\alpha = 0.5$, – – – – – : $\alpha = 0.9$. The vertical bar indicates the 80%-predictive interval at the point $x = 10$. Taken from Fan and Gijbels.[29]

For robust estimation of the regression function, one can simply replace the loss function in (42) by an outlier-resistant loss function such as

$$\ell(t) = \begin{cases} t^2/2 & \text{when } |t| \le c \\ c|t| - c^2/2 & \text{when } |t| > c\,, \end{cases}$$

namely, taking the derivative of $\ell(t)$ to be Huber's ψ-function: $\psi_c(t) = \max\{-c, \min(c, t)\}$. When the conditional distribution of Y given $X = x$ is symmetric about the regression function $m(x)$, the resulting estimates are consistent for all $c \ge 0$. Another useful robust procedure is LOWESS, introduced by Cleveland,[17] which reduces the influence of outliers by an iterative reweighted least-squares scheme with weights proportional to the residuals from the previous iteration.

There is a large literature on nonparametric quantile regression and robust regression. Härdle and Gasser[49] and Tsybakov[84] considered respectively local constant and local polynomial fitting. Other contributions in this area are also available.[12,28,47,60,83]

6.2. *Estimation of conditional variance*

Conditional variance functions have many statistical applications, particularly in finance. Because of their important applications in finance in which data are often dependent, we formulate the problems in stochastic setup.

Let $\{(X_i, Y_i)\}$ be a two-dimensional strictly stationary process having the same joint distribution as (X, Y). Let $m(x) = E(Y|X = x)$ and $\sigma^2(x) = \text{var}(Y|X = x)$ be respectively the regression function and the conditional variance function. Our approach is based on the residuals of the local fit. Let $\hat{m}_{h_1,K}(\cdot)$ be the local fit of $m(\cdot)$ using a kernel K and a bandwidth h_1. Consider the squared residuals

$$\hat{r}_i = \{Y_i - \hat{m}_{h_1,K}(X_i)\}^2 . \tag{44}$$

Note that the conditional variance function can be expressed as

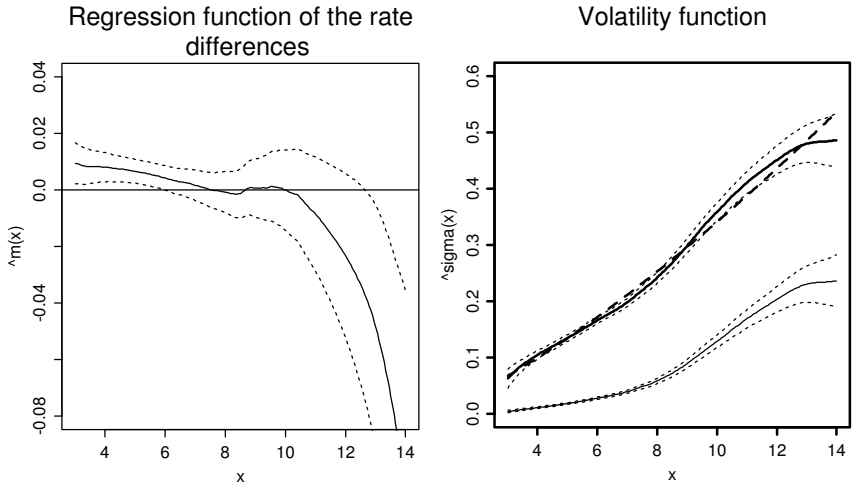
$$\sigma^2(x) = E[\{Y - m(X)\}^2 | X = x] ,$$

which is the regression function of the squared residuals. Therefore, a natural procedure is to run a local fit on the squared residuals. Let $\hat{\sigma}^2_{h_2,W}(x)$ be the local fit based on the data $\{(X_i, \hat{r}_i), i = 1, \ldots, n\}$, using a bandwidth h_2 and a kernel W. Then, it was shown by Fan and Yao,[35] Ruppert *et al.*[71] that the estimator $\hat{\sigma}^2_{h_2,W}(x)$ performs as well as the ideal estimator, which is a local linear fit to the true squared residuals

$$\{(X_i, \{Y_i - m(X_i)\}^2),\ i = 1, \ldots, n\}$$

using the same bandwidth h_2 and the same kernel W. They also obtained the order of bias and variance of the resulting estimators. Their results suggest that if the bandwidth h_1 is of order $n^{-1/5}$, then the residual-based *conditional variance estimator* performs asymptotically as well as the ideal one. In particular, the optimal bandwidth for estimating the mean regression function is permitted to be used for computing the residuals. Thus a data-driven procedure can be established.[35]

To illustrate the usefulness of the above automatic method, consider the yields of 12-month Treasury bill. The refined global bandwidth selector of Fan and Gijbels[28] and the Epanechnikov kernel. Figure 16(a) gives the estimated mean regression function. The bandwidth $\hat{h}_1 = 3.99$ was chosen by the software. Figure 16(b) depicts the estimated conditional standard deviation (the volatility function) and the conditional variance function. The bandwidth $\hat{h}_2 = 3.63$ was selected by the software. Visual inspection suggests that the volatility function should be a power function. Indeed,



(a) (b)

Fig. 16. The regression function and the volatility function for the 12-month Treasury bill data. (a) The estimated mean regression function and, (b) Estimated volatility function (thick curve) and the estimated conditional variance function (thin curve). The two dashed curves around a solid one indicate one standard error above and below the estimated mean regression function. From Fan and Gijbels.[29]

the correlation coefficient between $\{\log(x_j)\}$ and $\{\log(\hat{\sigma}(x_j))\}$ is 0.997!, where $x_j, (j = 1, \ldots, 201)$ are grid points in the interval [3, 14]. Fitting a line through the data $\{(\log(x_j), \log\{\hat{\sigma}(x_j)\}), j = 1, \ldots, 201\}$, we obtain the estimate

$$\hat{\sigma}(x) = 0.0154x^{1.3347}\,.$$

This estimate is presented as a thick-dashed curve in Fig. 16(b). This is an example where the nonparametric analyses yield a good parametric model $\sigma(x) = \alpha x^{\beta}$. Based on the linear regression on the data $(\log(X_i), \log(\hat{r}_i))$, one can also obtain directly an estimate of α and β.

6.3. *Estimation of conditional density*

It is well known that probability density function is much more informative than the mean and the variance. Similarly, in regression settings, the conditional probability density function provided more information about the population than the conditional regression function. The probability density function plots can show us about the center as well as the spreadness

of the population. The shape of conditional probability density function tells us whether it is symmetric. This provides a guidance for us to summarize the population via the conditional mean regression function or the conditional median regression function.

Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from the population (X, Y) with the conditional density $g(y|x)$. Note that

$$E\{K_{h_2}(Y - y)|X = x\} \approx g(y|x)\,, \quad \text{as } h_2 \to 0\,. \tag{45}$$

Thus, $g(y|x)$ can be regarded approximately as the regression function of the variable $K_{h_2}(Y - y)$ on X. Considerations of this nature lead to the following local polynomial regression problem:

$$\sum_{i=1}^{n} \left\{ K_{h_2}(Y_i - y) - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right\}^2 W_{h_1}(X_i - x)\,, \tag{46}$$

for a given bandwidth h_1 and a kernel function W. Let $\{\hat{\beta}_j(x, y), j = 0, \ldots, p\}$ be the solution of the least-squares problem. Then an estimator of $g^{(\nu)}(y|x) = \frac{\partial^\nu g(y|x)}{\partial x^\nu}$ is $\nu!\hat{\beta}_\nu(x, y)$. We write $\hat{g}(y|x) = \hat{\beta}_0(x, y)$ as the estimator of the conditional density.

To apply the proposed approach of conditional density estimation, we have to choose two bandwidths h_1 and h_2. The method for constructing a data-driven bandwidth for local polynomial regression can be used to compute a bandwidth for h_1, and the method for choosing a bandwidth for kernel density estimation can be employed to find a bandwidth h_2.[36]

With the estimated conditional density function, one can derive many statistical estimators. For example, the mean regression function can simply be estimated by

$$\hat{m}(x) = \int y\hat{g}(y|x)dy\,.$$

It can be shown that this estimator is the same as the local polynomial regression estimator when the kernel function K has mean zero. Similarly, one can derive estimates for the conditional variance and conditional quantile functions.

6.4. *Change point detection*

Change point detection is useful in medical monitoring and quality control. For example, when the treatment effects change suddenly without warning

or planning, *jump points* arise. The statistical problem can be formulated as follows.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a population (X, Y) with conditional mean function m, which is smooth except for a few number of jump discontinuities. For simplicity, we assume that there is only one single discontinuity point, also called a *change point*.

One may regard the change point as the location where the derivative function $|m'(\cdot)|$ is maximized. Thus, a naive method is to first estimate the derivative curve and then find the maximizer of the absolute value of the estimated derivative function. Let $D(x, h)$ be a derivative estimator resulting from a local polynomial fit of order p with bandwidth h and kernel K. For simplicity, assume that the support of K is $[-1, 1]$. The above idea translates into the following estimating scheme: plot the function $|D(\cdot, h)|$ for a range of values of h and identify the jump as the point x in the vicinity of which $|D(x, h)|$ is consistently large for a *range of values of* h. More precisely, let $\tilde{x}(h)$ be the global maximum of the function $|D(\cdot, h)|$. Put

$$\tilde{x}_-(h) = \sup_{h_1 \in [h, \eta_n]} \{\tilde{x}(h_1) - 2h_1\}, \quad \tilde{x}_+(h) = \inf_{h_1 \in [h, \eta_n]} \{\tilde{x}(h_1) + 2h_1\}, \quad (47)$$

for $h \leq \eta_n$, where $\eta_n > 0$ is a prescribed number, tending to zero more slowly than $n^{-1} \log n$. Let $\tilde{h}$ denote the infimum of values h such that $\tilde{x}_-(h) \leq \tilde{x}_+(h)$. The proposed jump point estimator is $\hat{x}_0 = \tilde{x}(\tilde{h})$.[44]

Gijbels *et al.*[44] also propose a further refinement of the above idea. For a given bandwidth h, pretend the change point lies in the interval $\tilde{x}(h) \pm 2h$ and the regression function is a step function on this interval. Then, find the unknown location of the jump such that it minimizes the residual sum of squares, using only the data in the strip $\tilde{x}(h) \pm 2h$. The resulting estimator is a refinement of the estimator $\tilde{x}(h)$. In particular, we can take $h = \tilde{h}$ to yield a refinement of $\hat{x}_0$.

Müller[67] proposed an alternative method based on a one-sided kernel approach. The idea can be extended to the local polynomial setting as follows. Denote by K_- a kernel function supported on $[-1, 0]$ and $\hat{m}_-(x, h)$ a local polynomial fit using the bandwidth h and the kernel K_-. Note that the estimator $\hat{m}_-(x, h)$ uses only the local data on the left-hand side of the point x. Analogously, let K_+ be a kernel function supported on [0, 1] and $\hat{m}_+(x, h)$ be a local polynomial fit using the bandwidth h and the kernel K_+. Then, $\hat{m}_+(x, h)$ uses only the data on the right-hand side of the point x. At the smooth locations, the estimates $\hat{m}_-(x, h)$ and $\hat{m}_+(x, h)$ are about the same, since both are consistent estimates of $m(x)$. At the discontinuity

point, however, they estimate respectively the left-limit and the right-limit of the function m at the point x. Thus, a natural estimator is the location such that the difference function $|\hat{m}_+(x,h) - \hat{m}_-(x,h)|$ is maximized. The bandwidth for detecting the change point is typically much smaller than the optimal bandwidth for curve estimation. Müller[67] and Gijbels *et al.*[44] also gave some interesting examples.

References

1. Aerts, M., Claeskens, G. and Hart, J. D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association* **94**: 869–879.
2. Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
3. Azzalini, A. and Bowman, A. N. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society Series* **B55**: 549–557.
4. Azzalini, A., Bowman, A. N. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**: 1–11.
5. Bickel, P. J. (1975). One-step Huber estimates in linear models. *Journal of the American Statistical Association* **70**: 428–433.
6. Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis, The Kernel Approach with S-Plus Illustrations*, Oxford Science Publications, Oxford.
7. Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**: 961–976.
8. Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* **95**: 888–902.
9. Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**: 941–956.
10. Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**: 477–489.
11. Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Local estimating equations. *Journal of the American Statistical Association* **93**: 214–227.
12. Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics* **19**: 760–777.
13. Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**: 807–822.
14. Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics* **16**: 136–146.
15. Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**: 298–308.

16. Chu, C. K. and Marron, J. S. (1991). Choosing a kernel regression estimator (with discussions). *Statistical Sciences* **6**: 404–436.
17. Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**: 829–836.
18. Cleveland, W. S., Grosse, E. and Shyu, W. M. (1992). Local regression models. In *Statistical Models in S*, eds. J. M. Chambers and T. J. Hastie, Wadsworth and Brooks, California, 309–376.
19. Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* **B34**: 187–220.
20. Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**: 269–276.
21. Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics* **20**: 1412–1425.
22. Eubank, R. L. and LaRiccia, V. M. (1992). Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *The Annals of Statistics* **20**: 2071–86.
23. Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association* **87**: 998—1004.
24. Fan, J. (1993). Local linear regression smoothers and their minimax. *The Annals of Statistics* **21**: 196–216.
25. Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association* **91**: 674–688.
26. Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society* **B61**: 927–943.
27. Fan, J., Farmen, M. and Gijbels, I. (1998). A blueprint of local maximum likelihood estimation. *Journal of the Royal Statistical Society* **B60**: 591–608.
28. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.
29. Fan, J. and Gijbels, I. (2000). Local polynomial fitting, *Smoothing and Regression. Approaches, Computation and Application*, ed. M. G. Schimek, John Wiley and Sons, 228–275.
30. Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics* **25**: 1661–1690.
31. Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**: 141–150.
32. Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models. *Journal of the Americal Statistical Association*, to appear.
33. Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* **3**: 35–56.
34. Fan, J. and Müller, M. (1995). Density and regression smoothing. In *XploRe: An Interactive Statistical Computing Environment*, eds. W. Härdle, S. Klinke and B. A. Turlach, Springer, Berlin, 77–99.
35. Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**: 645–660.

36. Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**: 189–206.
37. Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics* **29**, to appear.
38. Fan, J. and Zhang, J. (2000). Functional linear models for longitudinal data. *Journal of the Royal Statistical Society* **B62**: 303–332.
39. Fan, J. and Zhang, W. (1999). Statistical estimation in varying-coefficient models. *The Annals of Statistics* **27**: 1491–1518.
40. Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, Wiley, New York.
41. Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics* **11**: 171–185.
42. Gasser, T., Müller, H.-G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society* **B47**: 238–252.
43. Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* **47**: 1283–1296.
44. Gijbels, I., Hall, P. and Kneip, A. (1995). On the estimation of jump points in smooth curves. Discussion Paper #9515, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.
45. Granovsky, B. L. and Müller, H.-G. (1991). Optimizing kernel methods: A unifying variational principle. *International Statistical Review* **59**: 373–388.
46. Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*: *A Robust Penalty Approach*, Chapman and Hall, London.
47. Hall, P. and Jones, M. C. (1990). Adaptive M-estimation in nonparametric regression. *The Annals of Statistics* **18**: 1712–1728.
48. Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Boston.
49. Härdle, W. and Gasser, T. (1984). Robust non-parametric function fitting. *Journal of the Royal Statistical Society* **B46**: 42–51.
50. Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* **21**: 1926–47.
51. Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*, Springer, New York.
52. Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.
53. Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society* **B55**: 757–796.
54. Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**: 809–822.
55. Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1994). Power approximations to and power comparison of smooth goodness-of-fit tests. *Scandinavian Journal of Statistics* **21**: 131–45.

56. Jones, M. C., Marron, J. S. and Sheater, S. J. (1996a). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**: 401–407.
57. Jones, M. C., Marron, J. S. and Sheater, S. J. (1996b). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* **11**: 337–381.
58. Kallenberg, W. C. M. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association* **92**: 1094–1104.
59. Kauermann, G. and Tutz, G. (1999). On model diagnostics using varying coefficient models. *Biometrika* **86**: 119–128.
60. Koenker, R., Portnoy, S. and Ng, P. (1992). Nonparametric estimation of conditional quantile function. In *Proceedings of the conference on L_1 — Statistical Analysis and Related Methods*, ed. Y. Dodge, Elsevier, 217–229.
61. Lehmann, E. L. (1983). *Theory of Point Estimation*, Pacific Grove, Wadsworth and Brooks/Cole, California.
62. Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* **95**: 520–534.
63. Mack, Y. P. and Müller, H. G. (1989). Convolution type estimators for nonparametric regression estimation. *Statistics and Probability Letters* **7**: 229–239.
64. Marron, J. S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics and Probability Letters* **7**: 195–199.
65. McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.
66. Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association* **82**: 231–238.
67. Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *The Annals of Statistics* **20**: 737–761.
68. Nadaraya, E. A. (1964). On estimating regression. *Theory Probability Applied* **9**: 141–142.
69. Rawlings, J. O. and Spruill, S. E. (1994). Estimating pine seedling response to ozone and acidic rain. In *Case Studies in Biometry*, eds. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest and J. Greenhouse, Wiley, New York, 81–106.
70. Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *The Annals of Statistics* **22**: 1346–1370.
71. Ruppert, D., Wand, M. P., Holst, U. and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**: 262–73.
72. Scott, D. W. (1992). *Multivariate Density Estimation*: *Theory*, *Practice*, *and Visualization*, John Wiley and Sons, New York.
73. Seifert, B. and Gasser, T. (1996). Finite-sample variance of local polynomials: Analysis and solutions. *Journal of the American Statistical Association* **91**: 267–275.

74. Severini, T. A. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* **89**: 501–511.
75. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, London.
76. Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.
77. Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society* **B50**: 413–436.
78. Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics* **24**: 2477–2498.
79. Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics* **5**: 595–645.
80. Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics* **8**: 1348–1360.
81. Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10**: 1040–1053.
82. Tibshirani, R. and Hastie, T. J. (1987). Local likelihood estimation. *Journal of the Americal Statistical Association* **82**: 559–567.
83. Truong, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *The Annals of Statistics* **17**: 606–617.
84. Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission* **22**: 133–146.
85. Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, London.
86. Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformations in density estimation, *Journal of the Americal Statistical Association* **86**: 343–361.
87. Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Series* **A26**: 359–372.
88. Yang, L. and Marron, J. S. (1999). Iterated transformation — Kernel density estimation. *Journal of the American Statistical Association* **94**: 580–589.

About the Author

Jianqing Fan is currently chair professor and chairman of Department of Statistics, Chinese University of Hong Kong. He is also a professor of Statistics at The University of North Carolina at Chapel Hill and a former professor at The University of California at Los Angeles. He is an elected fellow of the American Statistical Association and the Institute of Mathematical Statistics, and an associate editor of *The Annals of Statistics*, *The Journal of American Statistical Association* and *Statistica Sinica*. He obtained BS in Mathematics from Fudan University, and MS degrees in Statistics from The Chinese Academy of Sciences, and PhD in Statistics from University

of California, Berkeley. He has published many papers in various aspects of applied, computational and theoretical statistics. His published work has been recognized by the Hettleman Prize by the University of North Carolina, and the Presidents' Award of the Committee of Presidents of Statistical Societies. His current research interests include nonparametric modeling, statistical methods in finance, nonlinear time series, analysis of longitudinal data, model selections, wavelets, survival analysis, generalized linear models, among others.

This page intentionally left blank

CHAPTER 25

BAYESIAN METHODS

MING-HUI CHEN

Department of Statistics, University of Connecticut,
215 Glenbrook Rd, U-4120, Stoors, CT 06269-4120, USA
Tel: 860-486-6984; mhchen@stat.uconn.edu

KEYING YE

Department of Statistics,
Virginia Polytechnic Institute and State University, USA

1. Introduction

In the practice of applied statistics and data analysis, summarizing data points, making inference to the unknowns, fitting probability models and predicting the future are all important elements to be considered. Many statistical methodologies have been developed in modern days to deal with different problems in the world full of randomness. Two schools of statistics are popular nowadays. One of them is called *classic* or *frequentist* statistics. The other one, which has progressed rapidly in the last decade and which has been more and more used in common practice, is called *Bayesian statistics.* Due to the current advances in computing technology and the development of efficient computational algorithms, Bayesian statistics are now becoming more popular in many applied fields such as agriculture, medicine, biology, public health, and epidemiology.

The Bayesian paradigm is based on specifying a probability model for the observed data $D = (n, \boldsymbol{y}, X)$, where n is the sample size, $\boldsymbol{y}$ is the $n \times 1$ response vector, and X is the $n \times p$ matrix of covariates, given a vector of unknown parameters $\boldsymbol{\theta}$, leading to the likelihood function $L(\boldsymbol{\theta}|D)$. Then we assume that $\boldsymbol{\theta}$ is random and has a *prior* distribution denoted by $\pi(\boldsymbol{\theta})$. Inference concerning $\boldsymbol{\theta}$ is then based on the *posterior* distribution, which is

obtained by Bayes' theorem. The posterior distribution is given by

$$\pi(\boldsymbol{\theta}|D) = \frac{L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})}{\int_\Omega L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}}\,, \tag{1}$$

where Ω denotes the parameter space of $\boldsymbol{\theta}$. From (1), it is clear that $\pi(\boldsymbol{\theta}|D)$ is *proportional* to the likelihood multiplied by the prior,

$$\pi(\boldsymbol{\theta}|D) \propto L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})\,,$$

and thus it involves a contribution from the observed data through $L(\boldsymbol{\theta}|D)$, and a contribution from prior information quantified through $\pi(\boldsymbol{\theta})$. The quantity $m(D) = \int_\Omega L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}$ is the *normalizing constant* of $\pi(\boldsymbol{\theta}|D)$, and is often called the *marginal* distribution of the data or the prior predictive distribution.

Notice that other than the data D to be random, from a Bayesian point of view, the parameter $\boldsymbol{\theta}$ is also random. Foundationally speaking, to gain information of an unknown, say $\boldsymbol{\theta}$, it would be natural if the knowledge of $\boldsymbol{\theta}$ can be described by using a form of statistical distribution. The more the information about the unknown through data is obtained, the better knowledge of the unknown is gained. The posterior in (1) can also be viewed as a prior distribution for future experimental observations, if any. Hence, Bayesian thinking requires a sequential learning process that leads to understanding unknowns in the scientifical world.

In this chapter, we are going to describe a few aspects of Bayesian statistics, including posterior inference (Sec. 2), prior elicitation (Sec. 3), Bayesian computations (Sec. 4), and applicational examples (Sec. 5).

2. Posterior Inference

2.1. *Summary of posterior distributions*

In Bayesian data analysis, many posterior quantities are of the form

$$E[h(\boldsymbol{\theta})|D] = \int_{R^p} h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta}\,, \tag{2}$$

where $h(\cdot)$ is a real-valued function of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$. We call (2) an integral-type posterior quantity, or the posterior expectation of $h(\boldsymbol{\theta})$. In (2), we assume that

$$E(|h(\boldsymbol{\theta})|\ |\ D) = \int_{R^p} |h(\boldsymbol{\theta})|\pi(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta} < \infty\,.$$

Integral-type posterior quantities include posterior means, posterior variances, covariances, higher-order moments, and probabilities of sets by taking appropriate functional forms of h. For example, (2) reduces to:

(a) the posterior mean of $\boldsymbol{\theta}$ when $h(\boldsymbol{\theta}) = \boldsymbol{\theta}$;
(b) the posterior covariance of θ_j and θ_{j^*} if $h(\boldsymbol{\theta}) = (\theta_j - E(\theta_j|D))(\theta_{j^*} - E(\theta_{j^*}|D))'$, where $E(\theta_j|D) = \int_{R^p} \theta_j \pi(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$;
(c) the posterior predictive density when $h(\boldsymbol{\theta}) = f(z|\boldsymbol{\theta})$, where $f(z|\boldsymbol{\theta})$ is the predictive density given the parameter $\boldsymbol{\theta}$; and
(d) the posterior probability of a set A if $h(\boldsymbol{\theta}) = 1\{\boldsymbol{\theta} \in A\}$, where $1\{\boldsymbol{\theta} \in A\}$ denotes the indicator function.

In (d), the posterior probability leads to a Bayesian p-value[72] by taking an appropriate form of A.

Some other posterior quantities such as normalizing constants, Bayes factors, and posterior model probabilities, may not simply be written in the form of (2). However, they are actually functions of integral-type posterior quantities. Posterior quantiles, Bayesian credible intervals, and Bayesian Highest Posterior Density (HPD) intervals are often viewed as nonintegral-type posterior quantities. Even for these types of posterior quantities, we can express them as functions of integral-type posterior quantities under certain conditions. For example, let $\xi = h(\boldsymbol{\theta})$, and $\xi_{1-\alpha}$ be the $(1-\alpha)$th posterior quantile of ξ with respect to $\pi(\boldsymbol{\theta}|D)$, where $0 < \alpha < 1$ and $h(\cdot)$ is a real-valued function. Then, $\xi_{1-\alpha}$ is the solution of the following equation:

$$\int_{R^p} 1\{h(\boldsymbol{\theta}) \le t\}\pi(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} = 1 - \alpha\,.$$

Therefore, the posterior quantile is a function of the posterior expectation of $1\{h(\boldsymbol{\theta}) \le t\}$.

2.2. *Predictive distributions*

A major aspect of the Bayesian paradigm is prediction. Prediction is often an important goal in regression problems, and usually plays an important role in model selection problems. The *posterior predictive* distribution of a future observation vector z given the data D is defined as

$$\pi(z|D) = \int_{\Omega} f(z|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}\,, \tag{3}$$

where $f(z|\boldsymbol{\theta})$ denotes the sampling density of z, and $\pi(\boldsymbol{\theta}|D)$ is the posterior distribution of $\boldsymbol{\theta}$. We see that (3) is just the posterior expectation of $f(z|\boldsymbol{\theta})$,

and thus sampling from (3) is easily accomplished via the Gibbs sampler (See Sec. 4 for detail) from $\pi(\boldsymbol{\theta}|D)$. This is a nice feature of the Bayesian paradigm since Eq. (3) shows that predictions and predictive distributions are easily computed once samples from $\pi(\boldsymbol{\theta}|D)$ are available.

Regarding the complementary roles of the predictive and posterior distributions in Bayesian data analysis, Box[12] notes that the posterior distribution provides a basis for "estimation of parameters conditional on the adequacy of the entertained model" while the predictive distribution enables "criticism of the entertained model in light of current data". In this spirit, Gelfand *et al.*[37] consider a cross-validation approach, in which the predictive distribution is used in various ways to assess model adequacy. The main idea of this cross-validation approach is to validate conditional predictive distributions arising from single observation deletion against observed responses.

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ denote the $n \times 1$ vector of the observed responses. Let X denote the $n \times p$ matrix of covariates whose ith row $\boldsymbol{x}_i'$ is associated with y_i. Then, the observed data can be written as $D = (n, \boldsymbol{y}, X)$. Also let $\boldsymbol{y}^{(-i)}$ denote the $(n-1)\times 1$ response vector with y_i deleted, let $X^{(-i)}$ denote the $(n-1) \times p$ matrix that is X with the ith row $\boldsymbol{x}_i'$ deleted, and the resulting observed data are written as $D^{(-i)} = ((n-1), \boldsymbol{y}^{(-i)}, X^{(-i)})$. In addition, let $\boldsymbol{\theta}$ be the vector of model parameters. We assume that $y_i \sim f(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i)$ and we let $\pi(\theta)$ denote the prior distribution of θ. Then, the posterior distribution of $\boldsymbol{\theta}$ based on the data D is given by

$$\pi(\boldsymbol{\theta}|D) \propto \left[\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i)\right] \pi(\theta)\,, \tag{4}$$

and the posterior distribution of $\boldsymbol{\theta}$ based on the data $D^{(-i)}$ is given by

$$\pi(\boldsymbol{\theta}|D^{(-i)}) \propto \left[\prod_{j\neq i} f(y_j|\boldsymbol{\theta}, \boldsymbol{x}_j)\right] \pi(\theta)\,. \tag{5}$$

Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)'$ denote future values of a replicate experiment. Also let $\pi(z_i|\boldsymbol{x}_i, D^{(-i)})$ denote the conditional density of z_i given $\boldsymbol{x}_i$ and $D^{(-i)}$ defined as

$$\pi(z_i|\boldsymbol{x}_i, D^{(-i)}) = \int f(z_i|\boldsymbol{\theta}, \boldsymbol{x}_i)\pi(\theta|D^{(-i)})\; d\boldsymbol{\theta}\,, \tag{6}$$

for $i = 1, 2, \ldots, n$. The conditional predictive density $\pi(z_i|\boldsymbol{x}_i, D^{(-i)})$ is also called the cross-validated predictive density. This density is to be checked against y_i, for $i = 1, 2, \ldots, n$ in the sense that, if the model holds, y_i may be viewed as a random observation from $\pi(z_i|\boldsymbol{x}_i, D^{(-i)})$. To do this, we

take $z_i = y_i$ in (6) and then we obtain the Conditional Predictive Ordinate (CPO):

$$\mathrm{CPO}_i = \pi(y_i|\boldsymbol{x}_i, D^{(-i)})\,. \tag{7}$$

CPO_i, which was proposed by Geisser[36] and further discussed in Gelfand *et al.*,[37] is a very useful quantity for model checking, since it describes how much the ith observation supports the model. Large CPO values indicate a good fit.

Another application of the predictive distribution to construct the Bayesian standardized residual. Similar to the Studentized residuals with the current observation deleted, the Bayesian standardized residual can be computed as

$$d_i = E[g(z_i, y_i)|\boldsymbol{x}_i, D^{(-i)}] = \frac{y_i - E(z_i|\boldsymbol{x}_i, D^{(-i)})}{\sqrt{\mathrm{var}(z_i|\boldsymbol{x}_i, D^{(-i)})}}\,, \tag{8}$$

where $\mathrm{var}(z_i|\boldsymbol{x}_i, D^{(-i)})$ is the variance of z_i with respect to the predictive distribution $\pi(z_i|\boldsymbol{x}_i, D^{(-i)})$ given by (6). Large $|d_i|$'s cast doubt upon the model but retaining the sign of d_i allows patterns of under or over fitting to be revealed.

Example 1. Estimating apple production y in New Zealand.[17]

Let β_j denote the average number of cartons per tree, conditional on the age of the tree, j, for $j = 1, 2, \ldots, 10$. These averages are combined using the linear model

$$y = \sum_{j=1}^{10} \beta_j x_j + \epsilon\,, \tag{9}$$

where x_j is the number of trees of age j and and $\epsilon \sim N(0, \sigma^2)$. Younger trees are known to produce fewer apples on average, so the model is subject to the constraints

$$0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_{10}\,. \tag{10}$$

Given data D on the number of trees and production by year and by orchard, Chen and Deely[17] choose a noninformative prior for $\beta_1, \ldots, \beta_9$, and σ^2 as well as a proper prior for β_{10}, which allow them to derive the full joint posterior density

$$\pi(\boldsymbol{\beta}, \sigma^2|D) = \frac{\exp\left\{-\frac{(\beta_{10}-\mu_{10})^2}{2\sigma_{10}^2}\right\}}{c(D)\sigma^{N+1}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{10}\beta_j x_{ij}\right)^2\right\}, \tag{11}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{10})'$, $c(D)$ is the normalizing constant, $0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_{10}$, and $\sigma^2 > 0$. For the New Zealand apple data, $N = 207$, $\mu_{10} = 0.998$, and $\sigma_{10}^2 = 0.0891$, where μ_{10} and σ_{10}^2 are specified using method-of-moments estimates from the growers' data for trees of age 10.

For the model given in (9), we have

$$f(z_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right\},$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i,10})'$. Thus,

$$E(z_i|\boldsymbol{x}_i, D^{(-i)}) = \int (\boldsymbol{x}_i'\boldsymbol{\beta}) \pi(\boldsymbol{\beta}, \sigma^2 | D^{(-i)}) \, d\boldsymbol{\beta} \, d\sigma^2 .$$

Note that CPO_i given by (7) can be rewritten as

$$\mathrm{CPO}_i = f(y_i|\boldsymbol{x}_i, D^{(-i)}) = \left(\int \frac{1}{f(y_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i)} \pi(\boldsymbol{\beta}, \sigma^2|D) \, d\boldsymbol{\beta} \right)^{-1} .$$

Let $\{(\boldsymbol{\beta}_l, \sigma_l^2), l = 1, 2, \ldots, L\}$ denote a Gibbs sample from $\pi(\boldsymbol{\beta}, \sigma^2|D)$ using the Gibbs sampler given in Sec. 4. Then, the Monte Carlo estimate of CPO_i is given by

$$\widehat{\mathrm{CPO}}_i = L \left[\sum_{l=1}^{L} (f(y_i|\boldsymbol{\beta}_l, \sigma_l^2, \boldsymbol{x}_i))^{-1} \right]^{-1} , \tag{12}$$

and the Monte Carlo estimates of $E(z_i|\boldsymbol{x}_i, D^{(-i)})$ and $\mathrm{var}(z_i|\boldsymbol{x}_i, D^{(-i)})$ are given by

$$\hat{E}(z_i|\boldsymbol{x}_i, D^{(-i)})) = \widehat{\mathrm{CPO}}_i L^{-1} \sum_{l=1}^{L} \frac{\boldsymbol{x}_i'\boldsymbol{\beta}_l}{f(y_i|\boldsymbol{\beta}_l, \sigma_l^2, \boldsymbol{x}_i)} , \tag{13}$$

and

$$\begin{aligned} \widehat{\mathrm{var}}(z_i|\boldsymbol{x}_i, D^{(-i)})) &= \hat{E}(z_i^2|\boldsymbol{x}_i, D^{(-i)}) - [\hat{E}(z_i|\boldsymbol{x}_i, D^{(-i)})]^2 \\ &= \widehat{\mathrm{CPO}}_i L^{-1} \sum_{l=1}^{L} \frac{\sigma_l^2 + (\boldsymbol{x}_i'\boldsymbol{\beta}_l)^2}{f(y_i|\boldsymbol{\beta}_l, \sigma_l^2, \boldsymbol{x}_i)} \\ &\quad - [\hat{E}(z_i|\boldsymbol{x}_i, D^{(-i)})]^2 , \end{aligned} \tag{14}$$

respectively. Using (13) and (14), the Monte Carlo estimate of the Bayesian standardized residual d_i is

$$\hat{d}_i = \frac{y_i - \hat{E}(z_i|\boldsymbol{x}_i, D^{(-i)})}{\sqrt{\widehat{\mathrm{var}}(z_i|\boldsymbol{x}_i, D^{(-i)})}} . \tag{15}$$

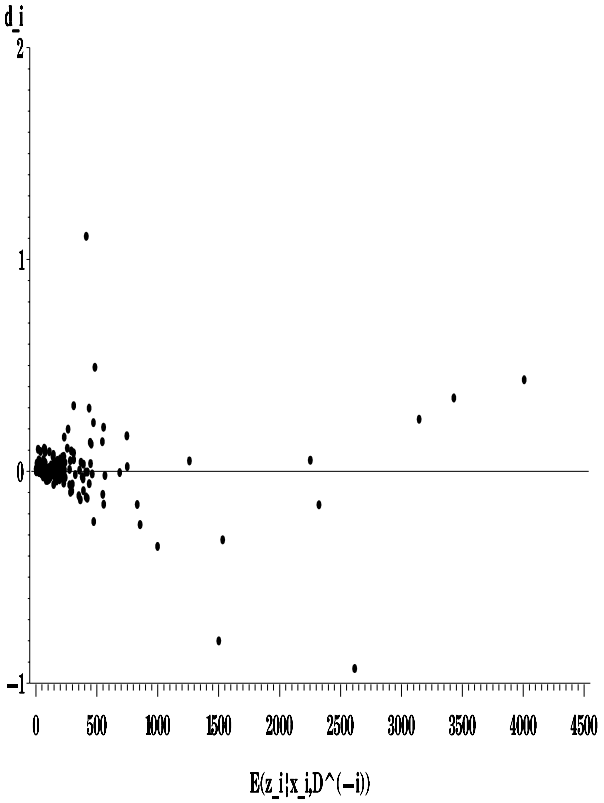


Fig. 1. The Bayesian standardized residual plot.

For the New Zealand apple data, Chen and Deely[17] use 50,000 Gibbs iterations to obtain the $\hat{d}_i$'s, and the results are displayed in Fig. 1.

From Fig. 1, it can be seen that: (i) the $\hat{d}_i$'s are small when the $\hat{E}(z_i|\boldsymbol{x}_i, D^{(-i)})$'s are small; and (ii) the $\hat{d}_i$'s are roughly symmetric about zero, which implies that the model is neither over-fitted nor under-fitted. Chen and Deely[17] also check the distribution of $\hat{d}_i$ and find that the $\hat{d}_i$'s roughly follow a Student t distribution. Noting that $f(y_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i)$ is a normal distribution and $\hat{f}(y_i|\boldsymbol{x}_i, D^{(-i)})$ in (12) is a finite mixture of normal distributions, it follows from a result of Johnson and Geisser[65] that $f(y_i|\boldsymbol{x}_i, D^{(-i)})$ is approximately a Student t distribution. Hence the results obtained by Chen and Deely[17] are consistent with the theoretical result of Johnson and Geisser,[65] and give further support that the normal assumption of the error terms in the constrained multiple linear regression model is appropriate.

2.3. *Marginal distributions*

In Bayesian inference, a joint posterior distribution is available through the likelihood function and a prior distribution. One purpose of Bayesian inference is to calculate and display marginal posterior densities because the marginal posterior densities provide complete information about parameters of interest.

Let

$$\boldsymbol{\theta}^{(j)} = (\theta_1, \ldots, \theta_j)' \quad \text{and} \quad \boldsymbol{\theta}^{(-j)} = (\theta_{j+1}, \ldots, \theta_p)'$$

be the first j and last $p - j$ components of $\boldsymbol{\theta}$, respectively. The support of the conditional joint marginal posterior density of $\boldsymbol{\theta}^{(j)}$ given $\boldsymbol{\theta}^{(-j)}$ is denoted by

$$\Omega_j(\boldsymbol{\theta}^{(-j)}) = \{(\theta_1, \ldots, \theta_j)' : (\theta_1, \ldots, \theta_j, \theta_{j+1}, \ldots, \theta_p)' \in \Omega\}\,, \tag{16}$$

and the subspace of Ω, given the first j components $\boldsymbol{\theta}^{*(j)} = (\theta_1^*, \ldots, \theta_j^*)'$, is denoted by

$$\Omega_{-j}(\boldsymbol{\theta}^{*(j)}) = \{(\theta_{j+1}, \ldots, \theta_p)' : (\theta_1^*, \ldots, \theta_j^*, \theta_{j+1}, \ldots, \theta_p)' \in \Omega\}\,. \tag{17}$$

Then the marginal posterior density of $\boldsymbol{\theta}^{(j)}$ evaluated at $\boldsymbol{\theta}^{*(j)}$ has the form

$$\pi(\boldsymbol{\theta}^{*(j)}|D) = \int_{\Omega_{-j}(\boldsymbol{\theta}^{*(j)})} \pi(\boldsymbol{\theta}^{*(j)}, \boldsymbol{\theta}^{(-j)}|D)\; d\boldsymbol{\theta}^{(-j)}\,. \tag{18}$$

In general, the analytical evaluation of $\pi(\boldsymbol{\theta}^{*(j)}|D)$ is not available. Thus, a Monte Carlo method is much needed to estimate it. There are several Monte Carlo methods available. These include the kernel density estimation, the conditional density estimation, and the importance weighted marginal density estimation (IWMDE) of Chen.[16] Here, we briefly describe how IWMDE works, and we refer the interesting readers to Chen *et al.*[26] for detailed discussion of other methods.

Consider the following identity:

$$\pi(\boldsymbol{\theta}^{*(j)}|D) = \int_{\Omega} \frac{w(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(-j)})\pi(\boldsymbol{\theta}^{*(j)}, \boldsymbol{\theta}^{(-j)}|D)}{\pi(\boldsymbol{\theta}|D)} \pi(\boldsymbol{\theta}|D)\; d\boldsymbol{\theta}\,, \tag{19}$$

where $w(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(-j)})$ is a completely known conditional density whose support is contained in, or equal to, the support, $\Omega_j(\boldsymbol{\theta}^{(-j)})$, of the conditional density $\pi(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(-j)}, D)$. Here, "*completely* known" means that $w(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(-j)})$ can be evaluated at any point of $(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(-j)})$. In other words, the kernel *and* the normalizing constant of this conditional density are

available in closed form. Using the identity (19), the IWMDE of $\pi(\boldsymbol{\theta}^{*(j)}|D)$ is defined by

$$\hat{\pi}(\boldsymbol{\theta}^{*(j)}|D) = \frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{\theta}_i^{(j)}|\boldsymbol{\theta}_i^{(-j)}) \frac{\pi(\boldsymbol{\theta}^{*(j)}, \boldsymbol{\theta}_i^{(-j)}|D)}{\pi(\boldsymbol{\theta}_i^{(j)}, \boldsymbol{\theta}_i^{(-j)}|D)}\,, \tag{20}$$

where $\{\boldsymbol{\theta}_i = (\boldsymbol{\theta}_i^{(j)}, \boldsymbol{\theta}_i^{(-j)}),\ i = 1, 2, \ldots, n\}$ is an MCMC sample from $\pi(\boldsymbol{\theta}|D)$. In (20), w plays the role of a weight function. Further, $\hat{\pi}(\boldsymbol{\theta}^{*(j)}|D)$ does not depend on the unknown normalizing constant $c(D)$, since $c(D)$ cancels in the ratio $\pi(\boldsymbol{\theta}^{*(j)}, \boldsymbol{\theta}_i^{(-j)}|D)/\pi(\boldsymbol{\theta}_i^{(j)}, \boldsymbol{\theta}_i^{(-j)}|D)$. In fact, using (1), we can rewrite (20) as

$$\hat{\pi}(\boldsymbol{\theta}^{*(j)}|D) = \frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{\theta}_i^{(j)}|\boldsymbol{\theta}_i^{(-j)}) \frac{L(\boldsymbol{\theta}^{*(j)}, \boldsymbol{\theta}_i^{(-j)}|D)\pi(\boldsymbol{\theta}^{*(j)}, \boldsymbol{\theta}_i^{(-j)})}{L(\boldsymbol{\theta}_i|D)\pi(\boldsymbol{\theta}_i)}\,.$$

The choice of w and the properties of $\hat{\pi}(\boldsymbol{\theta}^{*(j)}|D)$ can be found in Chen.[16] Thus, the detail is omitted here for brevity.

2.4. *Posterior Model Probabilities*

Suppose there are $\mathcal{K}$ models under consideration. Assume model m has a vector $\boldsymbol{\theta}^{(m)}$ of unknown parameters, with dimension p_m, which may vary from model to model, for $m = 1, 2, \ldots, \mathcal{K}$. Under model m, the posterior distribution of $\boldsymbol{\theta}^{(m)}$ takes the form

$$\pi(\boldsymbol{\theta}^{(m)}|D, m) \propto \pi^*(\boldsymbol{\theta}^{(m)}|D, m) = L(\boldsymbol{\theta}^{(m)}|D, m)\pi(\boldsymbol{\theta}^{(m)}|m)\,, \tag{21}$$

where $L(\boldsymbol{\theta}^{(m)}|D, m)$ is the likelihood function, D denotes the data, $\pi(\boldsymbol{\theta}^{(m)}|m)$ is the prior distribution, and $\pi^*(\boldsymbol{\theta}^{(m)}|D, m)$ is the unnormalized posterior density. Let $p(m)$ denote the prior probability of model m. Then, using Bayes' theorem, the posterior probability of model m can be written as

$$p(m|D) = \frac{p(D|m)p(m)}{\sum_{j=1}^{\mathcal{K}} p(D|j)p(j)}\,, \tag{22}$$

where

$$\begin{aligned} p(D|m) &= \int L(\boldsymbol{\theta}^{(m)}|D, m)\ \pi(\boldsymbol{\theta}^{(m)}|m)\ d\boldsymbol{\theta}^{(m)} \\ &= \int \pi^*(\boldsymbol{\theta}^{(m)}|D, m)\, d\boldsymbol{\theta}^{(m)} \end{aligned} \tag{23}$$

denotes the marginal distribution of the data D under model m. The marginal density $p(D|m)$ is precisely the normalizing constant of the joint

posterior density of $\boldsymbol{\theta}^{(m)}$. We choose the model with the largest posterior model probability $p(m|D)$.

Model selection, in particular variable selection, is one of the most frequently encountered problems in statistical data analysis. In cancer or AIDS clinical trials, for example, one often wishes to assess the importance of certain prognostic factors such as treatment, age, gender, or race in predicting survival outcome. Bayesian approach to model selection is more attractive than a criterion-based classical method such as the Akaike Information Criterion (AIC)[2] or Bayesian Information Criterion (BIC),[90] since available prior information can be incorporated into the posterior model probability via $p(m)$ and $\pi(\boldsymbol{\theta}^{(m)}|m)$ and thus more power can be achieved in order to identify the correct model. However, Bayesian model selection is often difficult to carry out because of the challenge in specifying prior distributions for the regression parameters for all possible models; specifying a prior distribution on the model space; and computations.

Other than focusing on a particular model selection using the largest posterior probability, one may also use the probability in (22) as a weight function to incorporate model uncertainty in a prediction. Such a criterion is called *Bayesian Model Averaging* or simply *BMA*. Suppose that one is interested in predicting certain quantity Δ such as predicting a future observation or a coefficient estimation for a regression problem. Instead of using just one model in prediction, one makes a prediction by average all feasible models through a weighted average whereas the weights are calculated from the posterior probabilities of the models. The posterior distribution given data D is

$$p(\Delta|D) = \sum_{\text{model } m} p(\Delta|m, D)p(m|D)\,. \tag{24}$$

The models used in the above calculation are the ones with significant posterior probabilities. The purpose of this model averaging is to avoid any risk of believing that the data belongs to a particular model. Instead, it accounts for the model uncertainty in predictions. More details of *BMA* can be found in Raftery *et al.*[83] and Hoeting *et al.*[57] and the references therein. As an illustration, we look at the following example.

Example 2. Assessment of health using body fat data.

A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. A data set used in Penrose *et al.*[80] studied the predictive equations of human's body fat with other variables such as age, weight, height, neck circumference,

Table 1. Regular model selection results for the body fat data.

Method	Variable Names	R^2	s^2
Stepwise	Age, Weight, Neck, Abdomen, Thigh, Forearm, Wrist	0.7445	18.41
Adjusted R^2	Age, Weight, Neck, Abdomen, Hip, Thigh, Biceps, Forearm, Wrist	0.7447	18.32

Table 2. Posterior probabilities of models for the body fat data.

Model	Variable Names	R^2	$p(m\|D)$
1	Weight, Abdomen, Forearm, Wrist	0.7350	0.47043
2	Weight, Abdomen, Wrist	0.7277	0.24656
3	Weight, Abdomen, Biceps, Wrist	0.7328	0.16415
4	Weight, Neck, Abdomen, Forearm, Wrist	0.7379	0.11885

chest circumference, abdomen 2 circumference, hip circumference, t high circumference, knee circumference, ankle circumference, biceps (extended) circumference, forearm circumference and wrist circumference. Apparently a multiple linear regression model can be used here. However, since there are a lot of independent variables, it is quite natural to use certain model selection techniques to obtain a "best" model. The two commonly used methods, namely stepwise regression and adjusted R^2 method come up with two different models as follows.

From Table 1, it seems both methods yielded quite comparative results in terms of variation explanation and regression accuracy. Suppose that we want to predict somebody's body fat at the values Age = 36, Weight = 226.75, Height = 71.75, Neck = 41.5, Chest = 115.3, Abdomen = 108.8, Hip = 114.4, Thigh = 69.2, Knee = 42.4, Ankle = 24, Biceps = 35.4, Forearm = 21 and Wrist = 20.1. The stepwise regression gave an estimate of 21.39 with a prediction variance 22.73, while the adjusted R^2 resulted in an estimate of 21.73 with a prediction variance 22.81. Note that those variances are the variances of the predictions under given models. If the prediction model is not a correct one, then the variance in prediction would be very different.

On the other hand, we may use the method of *BMA* to deal with this data set. The models with significant posterior probabilities are given in Table 2.

In Table 2, it clearly shows that both models selected by classical sequential model selection methods are not with high posterior probabilities. To predict the person's body fat for the same values as above, the prediction

is at 24.26 with a prediction variance of 26.82. However, this is the variance over all plausible models and it can be decomposed by two parts. The first part is 21.37 which is quite comparable to the variances we derived using sequential methods. This part is called pooled variances of all the models used in *BMA*. The second part, which is 5.45, is the part due to model uncertainty.

Bayesian approaches are now feasible due to recent advances in computing technology and the development of efficient computational algorithms. In particular, Chen *et al.*[21] and Ibrahim *et al.*[61] propose informative prior distributions $\pi(\boldsymbol{\theta}^{(m)}|m)$ and $p(m)$ for the parameter $\boldsymbol{\theta}^{(m)}$ and model m, and develop novel methods for computing the marginal distribution of the data. In addition, the stochastic search variable selection of George and McCulloch[45] and a novel reversible jump MCMC algorithm proposed by Green[52] make the computation of posterior model probabilities possible when $\mathcal{K}$ is large. In Sec. 6 below, we present a real data example from a series of animal toxicological experiments performed in the Department of Biology at the University of Waterloo to illustrate Bayesian model selection using informative priors.

3. Prior Elicitation

Prior distribution is one of the most important elements in Bayesian methodology. As the matter of fact, it is the most challenging element to practitioners. Rather than having a large amount of data when people can use large sample theory, most of the experiments consist of small to moderate sample size data sets. Thus, prior distribution plays a very important role in Bayesian analysis. We insist that whenever a practitioner can summarize historical or subjective information on an unknown, an informative prior should be elicited. The difficulty of seeking an informative prior is how one can connect the known or subjective information to a prior distribution. Conjugate priors are most commonly sought before because of the simplicity of the distributional forms and computational reason. However, the recent development in Bayesian computation overcome much of the difficulty using non-conjugate priors. Hence, to a practitioner, it is important to summarize all the information about an unknown to an approximate distribution form and use such a distribution as a prior.

On the other hand, many times, either historical or subjective knowledge of the unknown is not available, or there are too many parameters whose prior distributions need to be specified, noninformative priors are constantly

used as alternatives. However, as contrary to its name, all noninformative priors are actually informative. They are usually based on different criteria people use to generate prior distributions serving different purposes.

In subsequent subsections, we discuss in more details about informative and noninformative priors.

3.1. *Informative priors*

Informative priors are useful in applied research settings where the investigator has access to previous studies measuring the same response and covariates as the current study. For example, in many cancer and AIDS clinical trials, current studies often use treatments that are very similar or slight modifications of treatments used in previous studies. We refer to data arising from previous similar studies as *historical data*. In carcinogenicity studies, for example, large historical databases exist for the control animals from previous experiments. In all of these situations, it is natural to incorporate the historical data into the current study by quantifying it with a suitable prior distribution on the model parameters. The methodology discussed here can be applied to each of these situations as well as in other applications that involve historical data.

From a Bayesian perspective, historical data from past similar studies can be very helpful in interpreting the results of the current study. For example, historical control data can be very helpful in interpreting the results of a carcinogenicity study. According to Haseman *et al.*,[55] historical data can be useful when control tumor rates are low and when marginal significance levels are obtained in a test for dose effects. Suppose, for example, that 4 of 50 animals in an exposed group develop a specific tumor, compared with 0 of 50 in a control group. This difference is not statistically significant ($p = 0.12$, based on Fisher's exact test). However, the difference may be biologically significant if the observed tumor type is known to be extremely rare in the particular animal strain being studied. By specifying a suitable prior distribution on the control response rates that reflect the observed rates of a particular defect over a large series of past studies, one can derive a modified test statistic that incorporates historical information. If the defect is rare enough in the historical series, then even the difference of 4/50 versus 0/50 will be statistically significant based on a method that appropriately incorporates historical information.

To fix ideas, suppose we have historical data from a similar previous study, denoted by $D_0 = (n_0, y_0, X_0)$ where n_0 is the sample size of the

historical data, y_0 is the $n_0 \times 1$ response vector, and X_0 is the $n_0 \times p$ matrix of covariates based on the historical data. Chen *et al.*[19] and Ibrahim and Chen[58] proposed the power prior to incorporate historical information. The power prior is defined to be the likelihood function based on the historical data D_0, raised to a power a_0, where $0 \leq a_0 \leq 1$ is a scalar parameter that it controls the influence of the historical data on the current data. One of the most useful applications of the power prior is for model selection problems, since these priors inherently automate the informative prior specification for all possible models in the model space. They are quite attractive in this context, since specifying meaningful informative prior distributions for the parameters in each model is a difficult task requiring contextual interpretations of a large number of parameters. In variable subset selection, for example, the prior distributions for all possible subset models are automatically determined once the historical data D_0, and a_0 are specified. Berger and Mallows[8] refer to such priors as "semi-automatic" in their discussion of Mitchell and Beauchamp.[76] Chen *et al.*[23] use the power prior for heritability estimates from human twin data. Chen *et al.*[22] demonstrate the use of the power prior in variable selection contexts for logistic regression. Ibrahim *et al.*[61] and Chen *et al.*[20] develop the power prior for the class of generalized linear mixed models. Ibrahim and Chen,[59] Ibrahim *et al.*,[60] Chen *et al.*[18,21] develop the power prior for various types of models for survival data.

Let $\pi_0(\boldsymbol{\theta})$ denote the prior distribution for $\boldsymbol{\theta}$ before the historical data D_0 is observed. We shall call $\pi_0(\boldsymbol{\theta})$ the *initial prior* distribution for $\boldsymbol{\theta}$. Given a_0, we define the *power prior* distribution of $\boldsymbol{\theta}$ for the current study as

$$\pi(\boldsymbol{\theta}|D_0, a_0) \propto L(\boldsymbol{\theta}|D_0)^{a_0}\ \pi_0(\boldsymbol{\theta})\,, \tag{25}$$

where a_0 is a scalar prior parameter that weights the historical data relative to the likelihood of the current study. The parameter a_0 can be interpreted as a precision parameter for the historical data. It is reasonable to restrict the range of a_0 to be between 0 and 1, and thus we take $0 \leq a_0 \leq 1$. One of the main roles of a_0 is that it controls the heaviness of the tails of the prior for $\boldsymbol{\theta}$. As a_0 becomes smaller, the tails of (25) become heavier. Setting $a_0 = 1$, (25) corresponds to the update of $\pi_0(\boldsymbol{\theta}|c_0)$ using Bayes theorem. That is, with $a_0 = 1$, (25) corresponds to the posterior distribution of $\boldsymbol{\theta}$ from the previous study. When $a_0 = 0$, then the prior does not depend on the historical data, and in this case, $\pi(\boldsymbol{\theta}|D_0, a_0 = 0) \equiv \pi_0(\boldsymbol{\theta})$. Thus, $a_0 = 0$ is equivalent to prior specification with no incorporation of historical data.

Therefore, (25) can be viewed as a generalization of the usual Bayesian update of $\pi_0(\boldsymbol{\theta})$. The parameter a_0 allows the investigator to control the influence of the historical data on the current study. Such control is important in cases where there is heterogeneity between the previous and current study, or when the sample sizes of the two studies are quite different.

The hierarchical power prior specification is completed by specifying a (proper) prior distribution for a_0. Thus we propose a joint power prior distribution for (θ, a_0) of the form

$$\pi(\boldsymbol{\theta}, a_0|D_0) \propto L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta}) = \pi(a_0|\boldsymbol{\gamma}_0)\,, \tag{26}$$

where $\boldsymbol{\gamma}_0$ is a specified hyperparameter vector. A natural choice for $\pi(a_0|\boldsymbol{\gamma}_0)$ is a beta prior. However, other choices, including a truncated gamma prior or a truncated normal prior can be used. These three priors for a_0 have similar theoretical properties, and our experience shows that they have similar computational properties. In practice, they yield similar results when the hyperparameters are appropriately chosen. Thus, for a clear focus and exposition, we will use a *beta* distribution for $\pi(a_0|\gamma_0)$, which takes the form

$$\pi(a_0|\boldsymbol{\gamma}_0) \propto a_0^{\delta_0-1}(1-a_0)^{\lambda_0-1}\,,$$

where $\boldsymbol{\gamma}_0 = (\delta_0, \lambda_0)$. The beta prior for a_0 appears to be the most natural prior to use and leads to the most natural elicitation scheme. The prior in (26) does not have a closed form in general, but it has several attractive theoretical and computational properties for the classes of models considered here. One attractive feature of (26) is that it creates heavier tails for the marginal prior of $\boldsymbol{\theta}$ than the prior in (25), which assumes that a_0 is a fixed value. This is a desirable feature since it gives the investigator more flexibility in weighting the historical data. In addition, the construction of (26) is quite general, with various possibilities for $\pi_0(\boldsymbol{\theta})$. If $\pi_0(\boldsymbol{\theta})$ is proper, then (26) is guaranteed to be proper. Further, (26) can be proper even if $\pi_0(\boldsymbol{\theta})$ is an improper uniform prior. Specifically, Ibrahim *et al.*[62] and Chen *et al.*[22] characterize the propriety of (26) for generalized linear models, and also show that for fixed a_0, the prior converges to a multivariate normal distribution as $n_0 \to \infty$. For the class of generalized linear mixed models, Ibrahim *et al.*,[61] Chen *et al.*[18,19] characterize the propriety of (26) and derive various other theoretical properties of the power prior. Ibrahim *et al.*,[60] and Ibrahim and Chen[59] characterize various properties of (26) for proportional hazards models, and Chen *et al.*[21] examine various theoretical properties of (26) for a class of cure rate models.

3.1.1. *Example 3. An analysis of the AIDS study ACTG036 using the data from ACTG019 as historical data*

The ACTG019 study was a double blind placebo-controlled clinical trial comparing zidovudine (AZT) to placebo in persons with CD4 counts less than 500. The results of this study were published in Volberding *et al.*[100] The sample size for this study, excluding cases with missing data, was $n_0 = 823$. The response variable (y_0) for these data is binary with a 1 indicating death, development of AIDS, or AIDS related complex (ARC), and a 0 indicates otherwise. Several covariates were also measured. The ACTG036 study was also a placebo-controlled clinical trial comparing AZT to placebo in patients with hereditary coagulation disorders. The results of this study have been published by Merigen *et al.*[74] The sample size in this study, excluding cases with missing data, was $n = 183$. The response variable (y) for these data is binary with a 1 indicating death, development of AIDS, or AIDS related complex (ARC), and a 0 indicates otherwise. Several covariates were measured for these data. A summary of both data sets can be found in Chen *et al.*[22] Therefore, we let D_0 denote the data from the ACTG019 study and D denote the data from the ACTG036 study.

Chen *et al.*[22] use the priors given by (26) and the logistic regression model to carry out variable subset selection, which yields the model containing an intercept, CD4 count (cell count per mm^3 of serum), age, and treatment as the best model. For this model, we use the power prior (26) to obtain posterior estimates of the regression coefficients for various choices of $(\mu_{a_0}, \sigma_{a_0})$, where $\mu_{a_0} = \frac{\delta_0}{\delta_0+\lambda_0}$ and $\sigma^2_{a_0} = \mu_{a_0}(1-\mu_{a_0})(\delta_0+\lambda_0+1)^{-1}$. The results based on the standardized covariates and the logit model with an improper uniform prior for the regression coefficients are given in Table 3. The values of $(\mu_{a_0}, \sigma_{a_0})$ and the corresponding values of (δ_0, λ_0) are also reported in the table. We used 50,000 Gibbs iterations for all posterior computations and the Monte Carlo method of Chen and Shao[24] to calculate 95% highest probability density (HPD) intervals for the parameters of interest. From Table 3, we see that as the weight for ACTG019 study increases, the posterior mean of a_0 (denoted $E(a_0|D, D_0)$) increases, the posterior standard deviations (SD) for all parameters decrease, and the 95% HPD intervals get narrower. Most noticeably, when $(\delta_0, \lambda_0) = (100, 1)$, none of the HPD intervals for the regression coefficients contain 0. Table 3 also indicates that the HPD intervals are not too sensitive for moderate changes in $(\mu_{a_0}, \sigma_{a_0})$. This is a comforting feature, since it implies that the HPD intervals are fairly robust with respect to the hyperparameters

Table 3. Posterior estimates for AIDS data.

(δ_0, λ_0)	$(\mu_{a_0}, \sigma_{a_0})$	$E(a_0\|D, D_0)$	Variable	Posterior Mean	Posterior SD	95% HPD Interval
(5, 5)	(0.50, 0.151)	0.02	Intercept	−4.389	0.725	(−5.836, −3.055)
			CD4 count	−1.437	0.394	(−2.238, −0.711)
			Age	0.135	0.221	(−0.314, 0.556)
			Treatment	−0.120	0.354	(−0.817, 0.570)
(20, 20)	(0.50, 0.078)	0.09	Intercept	−3.803	0.511	(−4.834, −2.868)
			CD4 count	−1.129	0.300	(−1.723, −0.559)
			Age	0.176	0.195	(−0.214, 0.552)
			Treatment	−0.223	0.300	(−0.821, 0.364)
(30, 30)	(0.50, 0.064)	0.13	Intercept	−3.621	0.436	(−4.489, −2.809)
			CD4 count	−1.028	0.265	(−1.551, −0.515)
			Age	0.194	0.185	(−0.170, 0.557)
			Treatment	−0.259	0.278	(−0.805, 0.288)
(50, 1)	(0.98, 0.019)	0.26	Intercept	−3.337	0.323	(−3.978, −2.715)
			CD4 count	−0.865	0.211	(−1.276, −0.448)
			Age	0.233	0.160	(−0.081, 0.548)
			Treatment	−0.314	0.230	(−0.766, 0.138)
(100, 1)	(0.99, 0.010)	0.53	Intercept	−3.144	0.231	(−3.601, −2.705)
			CD4 count	−0.746	0.161	(−1.058, −0.429)
			Age	0.271	0.135	(0.001, 0.529)
			Treatment	−0.356	0.181	(−0.717, −0.011)

of a_0. This same robustness feature is also exhibited in posterior model probability calculations.[22]

3.2. *Conjugate priors*

Conjugate priors were quite popular before the powerful breakthrough of the Bayesian computational techniques. Suppose $\mathcal{F}$ is a class of prior distributions for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the parameter, $\mathcal{P}$ a class of sampling distributions $f(y|\boldsymbol{\theta})$. The class $\mathcal{F}$ is conjugate for $\mathcal{P}$ for any $f(y|\boldsymbol{\theta})$ in $\mathcal{P}$ if the prior $\pi(\boldsymbol{\theta})$ and the posterior $\pi(\boldsymbol{\theta}|y)$ all from the same family $\mathcal{F}$. Since the posterior distributions have the same form as the prior, the closed form of the posteriors can be derived accordingly. Hence, it does not cause much trouble in computing the posterior as described in Sec. 1.

Example 4.

Suppose that a random variable Y follows a Poisson distribution with mean λ. The density function of Y given λ can be written as

$$f(y|\lambda) = e^{-\lambda}\frac{\lambda^y}{y!}, \quad \text{for } y = 0, 1, \ldots . \tag{27}$$

Further suppose that the prior distribution of λ is a Gamma distribution with parameters α and β, as follows

$$\pi(\lambda) \propto \lambda^{\alpha} e^{-\beta\lambda}, \quad \text{for } \lambda > 0\,. \tag{28}$$

The posterior distribution of λ thus can be calculated as

$$\pi(\lambda|y) \propto \pi(\lambda) f(y|\lambda) \propto \lambda^{\alpha} e^{-\beta\lambda} \lambda^{y} = e^{-\lambda} \propto \lambda^{\alpha+y} e^{-(\beta+1)\lambda}, \tag{29}$$

which is another Gamma distribution. Clearly, the prior (28) and the posterior (29) all belong to the same Gamma distribution family for any Poisson family sampling distribution.

The practical advantage of the conjugate prior distributions is obvious. This is the reason why it is still popular when practitioners often use it if they believe that their priors may be specified as conjugate priors. Although, it is flexible to elicit a conjugate prior due to changing the hyper-parameter values, many times it is not accurate to decide a prior knowledge by only choosing one or two parameter values. On the other hand, sometimes a finite mixture of conjugate priors may be a good idea to overcome this difficulty since the mixture of conjugate priors is also a conjugate prior.[5]

We have to note here that the conjugacy depends on the family $\mathcal{F}$ and $\mathcal{P}$ one chooses. Also, it depends on dimensions of the parameter space. For instance, normal priors on the mean of normal sampling distribution when the variance is known constitute a conjugate prior family, while inverse gamma priors on the variance of normal sampling distribution when the mean is assumed known constitute another conjugate prior family. However, normal priors on the mean and inverse gamma priors on the variance of the normal sampling distribution do not constitute a conjugate family since the marginal posteriors of the mean parameter is no longer normal. This suggests that people have to be careful when examining conjugacy for multi-dimensional parameter problems.

Another point about conjugacy we like to point out is that it is often quite useful to have conditionally conjugacy to parameters for a multi-dimensional parameter model. Conditional conjugacy or sometimes called semi-conjugacy means that the conditional posterior distributions of a set of parameters given others and the prior distribution of the same set of parameters belong to the same distributional family, for instance, to the example we just mentioned above about normal mean and normal variance. Although the marginal posterior distribution of the normal mean is no longer normal, the conditional posterior distribution of the normal mean

given that the normal variance is still normal. Likewise, the conditional posterior distribution of the normal variance given the normal mean is still inverse gamma. The advantage of this semi-conjugacy can be used in full conditional distributions in Gibbs sampling (see Sec. 4.1).

3.3. *Noninformative priors*

As described in the opening of this section, none of the noninformative priors are non-informative. The derivations of those so called noninformative priors all depend on certain informative criteria. One of the earliest methods of defining noninformative priors was based on the principle of insufficient reason. This method, sometimes referred to as *Laplace's rule*, prescribes a uniform prior on the parameter space $\boldsymbol{\Theta}$. Laplace used uniform priors on the probabilities of two binomial populations. Laplace's rule and the principle of insufficient reason are intuitively appealing. The reasoning is that if no prior information is available that favors certain parameter values over others, then all parameter values should be considered equally likely. However, the immediate criticism of this uniform prior is that it does not follow probability law in the sense of invariance in parameter transformation.

Example 5.

Suppose that Y follows a Binomial distribution with parameters n and p, while n is known. Using Laplace's argument, if there is no subjective prior information available for p, a uniform prior $\pi(p) = 1$, for $0 < p < 1$ should be used. Now assume that we are interested in the parameter $q = 1/(1 + p)$. Since we still do not have information about q, we *have to* assume the prior distribution of q as uniform, i.e. $\pi(q) = 2$, for $1/2 < q < 1$. Since q is a variable transformation of p, with the Jacobian $1/q^2$, following the probability law it follows that $\pi(q) = \pi(p(q))/q^2 = 1/q^2$ for $1/2 < q < 1$. However, if both uniform distributions are used, the above equality cannot hold. This implies that the uniform prior is not invariant under transformation.

There is also another issue that has been discussed extensively in Bayesian school. Suppose we are going to use Laplace's rule and assign a uniform prior on the parameter. If the parameter space is finite, the uniform distribution is proper (that is, its integration over its domain is finite). However, if the parameter space is infinite such as the mean of a normal distribution, a uniform prior is improper (not integrable). Such a phenomenon does not only happen for a uniform prior, it may happen to

many other noninformative priors we will discuss later. As a matter of fact, many of the noninformative priors we use are improper.

Although an improper prior is not supported by Bayes' rule, it does not necessarily lead to problems in Bayesian analysis as long as the posterior distributions are proper. For more discussions of improper priors, readers are referred to Berger,[5] Bernardo and Smith,[10] and Kass and Wasserman.[66] Once a posterior distribution is integrable, after normalization to a probability distribution, the final posterior distribution still represents a post-knowledge of the unknowns. Therefore, if an improper noninformative prior is used in practice, it is important to verify that the posterior distribution is integrable before making posterior inferences. This is even more important if simulation methods (see Sec. 4) are used to draw posterior inference because even if the posterior distribution is improper, sometimes it cannot be detected by using simulation. Therefore, it may lead to inappropriate conclusion when the true posterior distribution is actually improper.

Since the pioneer work of Laplace to use Bayesian methodology into applied statistics, there have been a lot of attempts to seek *default* or *automatic* prior distributions. In the following subsections, we will discuss Jeffreys priors, the reference priors and the probability matching priors.

3.4. *Jeffreys priors*

To overcome the difficulty of the non-invariant uniform prior criterion, Jeffreys[64] derived a prior using invariance of parameter transformations. Before we present the Jeffreys prior, one term called expected Fisher information needs to be defined. Suppose that a random vector $\boldsymbol{Y}$, given $\boldsymbol{\theta}$, has a probability density function $f(\boldsymbol{y}|\boldsymbol{\theta})$ which is twice differentiable with respect to $\boldsymbol{\theta}$. The expected Fisher information matrix, denoted by $I(\boldsymbol{\theta}) = \{I_{ij}(\boldsymbol{\theta})\}$, is defined as

$$I_{ij}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\delta^2}{\delta\theta_i\delta\theta_j}\log(f(\boldsymbol{y}|\boldsymbol{\theta}))\right]. \tag{30}$$

Once a sampling distribution is known with the density satisfying the existence of the Fisher information matrix, the Jeffreys prior is simply

$$\pi_J(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}\,, \tag{31}$$

where det stands for a determinant. For any one-to-one transformation between two parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, the two priors for $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$ calculated using (31) will not cause any ambiguous results, i.e. the method used here

is invariant through parameter transformation. On the other hand, like a uniform prior, Jeffreys prior may be improper.

Example 6.

Suppose that Y follows a Binomial distribution with unknown parameter p and known n. The density function of this distribution is $f(y|p) \propto p^y(1-p)^{1-y}$. Taking second derivative to $-\log(f)$, with respect to p, yields $y/p^2 + (1-y)/(1-p)^2$. The expectation of this form becomes $n/p(1-p)$. Thus, the Jeffreys prior is proportional to $1/\sqrt{p(1-p)}$.

Example 7.

Suppose $y_1, \ldots, y_n$ form a random sample from a normal population $N(\mu, \sigma^2)$. The density function is given as $f(y|\mu, \sigma) \propto e^{-(y-\mu)^2/2\sigma^2}/\sigma$. To calculate the Fisher information matrix, note that $I_{11}(\mu, \sigma) = E(1/\sigma^2) = 1/\sigma^2$, $I_{12} = I_{21} = 0$ and $I_{22} = 2/\sigma^2$. Hence the Jeffreys prior for (μ, σ) is $1/\sigma^2$.

Note that while the prior distribution in Example 6 is proper, the prior in Example 7 is improper. Yet, in the later example, the posterior distributions are usually proper except in certain degenerate cases.

Jeffreys priors are very commonly used for many different models. Even in many other developments of noninformative priors, one can always trace them back to Jeffreys priors in some sense. One property of the Jeffreys priors is the invariance under parameter transformations. However, the use of Jeffreys prior is not quite appealing in multi-dimensional situations. For instance, Jeffreys prior in Example 7 is $1/\sigma^2$. Of making inference about the mean variable, even Jeffreys himself pointed out that this prior does not yield satisfactory results. Instead, in this case a prior $1/\sigma$ is usually used for the parameterization (μ, σ) which is the product of the Jeffreys prior of μ alone (uniform) and that of σ alone $(1/\sigma)$. Here alone means that when the Jeffreys prior is calculated for one parameter, the other parameter is treated as fixed. In this case, both individual priors do not depend on the fixed parameter. When the product of the two priors is used, it means that "independence" of the prior knowledge of those two parameters is assumed. However, once the data is used, in the posterior analysis, those two parameters are rarely independent.

The above discussion raises a question that what kind of prior is good for multi-dimensional parameter problems. In most of the applied statistical problems, there are more than one parameters and some of them are treated as very important and the others are treated as nuisance. To find a noninformative prior for such kinds of models, Berger and Bernardo[6] developed an

iterative algorithm to calculate noninformative priors for multi-parameter problems. Such priors are called the reference prior which will be briefly described in the next subsection.

3.5. *The reference priors*

The reference prior method, introduced by Bernardo[9] and further developed by Berger and Bernardo,[6,7] is motivated by the notion of maximizing the expected amount of information about the parameter $\boldsymbol{\theta}$ provided by the data $\boldsymbol{y}$. The amount of information provided by the experiment is quantified by the *Kullback–Liebler divergence*, which is defined by

$$D(g,h) = \int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta}) \log\left(\frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}\right) \, d\boldsymbol{\theta}\,,$$

for two densities g and h. The expected information about $\boldsymbol{\theta}$ provided by the data can be naturally defined as

$$E_{\boldsymbol{Y}}(D(\pi(\boldsymbol{\theta}|\boldsymbol{y}), \pi(\boldsymbol{\theta})))\,, \tag{32}$$

where $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ are prior and posterior distributions, respectively. Theoretically, the reference prior approach is to find a prior such that the quantity (32) is maximized. However, the actual process of this maximization involves a modification of the form (32) and asymptotic process using infinitely many independent replications of the experiments would be used. Now we briefly mention the idea and procedure of the algorithm developed by Berger and Bernardo.[7]

To derive the reference priors for an experiment, one has to decompose the parameter space by ordered groups in the order of importance of the groups: $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \ldots, \boldsymbol{\theta}_{(m)}$, where each group $\boldsymbol{\theta}_{(j)}$ contains one or more of the scalar parameter in $\boldsymbol{\theta}$. The reference prior is developed iteratively by first computing the marginal prior for $\boldsymbol{\theta}_{(m)}$, then the conditional prior for $\boldsymbol{\theta}_{(m-1)}$ given $\boldsymbol{\theta}_{(m)}$, then the conditional prior for $\boldsymbol{\theta}_{(m-2)}$ given $\boldsymbol{\theta}_{(m-1)}$ and $\boldsymbol{\theta}_{(m)}$, etc. Finally, a reference prior can be obtained by multiplying all the priors above together. In the derivation, the parameter spaces should be truncated to compact sets and certain limiting procedures may be used. The detailed algorithm can be found in Berger and Bernardo.[7]

Note that in addition to being divided into groups, the parameters in $\boldsymbol{\theta}$ are also ordered. The order of the importance of the groups may be different by different users, although the parameter of interests stay the same. Berger and Bernardo recommended single group ordering, which means that there is only one parameter in each group.[105,107] This recommendation is based on

their experience in applying the reference prior method to various applied problem. Since different groupings may yield different preference priors,[7,106] it is possible that there exist different reference priors for the same model.

Berger and Bernardo also state, concerning the ordering of the parameters in terms of inferential importance, that "... beyond putting the "parameters of interest" first, it is too vague to be of much use." They recommend that, if possible, all reference priors for which the parameters of interest are placed first in the ordering should be computed. This provides a set of prior distributions which can be compared, to assess the sensitivity of the resulting analyses to the choice of prior distribution.

Finally, we want to point it out that interestingly, under certain regular conditions, the reference prior is the same as the Jeffreys prior which means that using the Jeffreys prior, the expected information about the parameter coming from data only is maximized. More studies of the reference priors can be found in Refs. 19, 46, 66, 94 and 105. In Sec. 5.3, we will discuss the reference priors for a statistical calibration model.

3.6. *Probability matching priors*

From inference point of view, confidence interval is quite frequently used in practice. Although, the concept of confidence interval creates a lot of confusing in interpretation, this interval however, gives quite important information in accuracy of an estimate. In the frequentist domain such that for many runs of experiments, the probability associated with a confidence interval provides coverage probability of the random intervals covering the true unknown. On the other hand, one can also derive a credible interval in Bayesian study. Such an interval is quite similar to a confidence interval, except that the probability of this credible interval implies the probability that the unknown parameter belonging to two fixed numbers. This is actually how people usually interpret a confidence interval. The advantage in this interpretation comes from the fact that the parameter in consideration is random and the posterior probability is calculated under the parameter domain, not in the frequentist domain anymore.

Getting an *automatic* prior is one purpose of developing a noninformative prior method. This means that the prior derived using this method can be applied to any data created from the statistical model in study. Obviously, not only a Bayesian credible interval is of interest, but also the confidence interval. Probability matching priors are those priors to have the property that posterior probabilities of the posterior quantiles from

the resulting Bayesian analysis match frequentist coverage probabilities of the same quantiles, at least asymptotically.

Suppose that $\boldsymbol{\theta}$ is a parameter of interest and an interval $(\phi(\boldsymbol{y}) \leq \boldsymbol{\theta})$ has the posterior probability $\alpha = P(\{\phi(\boldsymbol{y}) \leq \boldsymbol{\theta})|\boldsymbol{y})$ (αth posterior quantile). On the other hand, if we treat $\boldsymbol{\theta}$ as fixed, the frequentist coverage probability of this interval can also be calculated as $P(\phi(\boldsymbol{y}) \leq \boldsymbol{\theta}|\boldsymbol{\theta})$. If a prior can be obtained such that

$$\alpha = P(\phi(\boldsymbol{y}) \leq \boldsymbol{\theta}|\boldsymbol{y}) \approx P(\phi(\boldsymbol{y}) \leq \boldsymbol{\theta}|\boldsymbol{\theta})\,, \tag{33}$$

for all $\boldsymbol{y}$ and $\boldsymbol{\theta}$, asymptotically, we say the prior a probability matching prior.

Welch and Peers[101] are the first ones to study such kind of priors. In one-dimensional case, they found that the Jeffreys priors satisfies this equality in the order of $O(1/\sqrt{n})$, which means that when n goes to infinity, the rate of the difference between the two probabilities in (33) goes to zero in the rate same as $1/\sqrt{n}$. This is called the first order matching. Stein[93] and Tibshirani[97] extended their work and used differential equations to obtain more first order matching priors.

The probability matching priors have played certain justification rules to many of the noninformative priors. For instance, in many models, the reference priors are matching priors. However, in a few occasions, they are not. Since there are usually many priors satisfying the differential equations in deriving the probability matching priors, only using this method does not lead to a single prior which may be satisfactory.

4. Bayesian Computation

There are two major challenges involved in advanced Bayesian computation. These are how to sample from posterior distributions and how to compute posterior quantities of interest using Markov chain Monte Carlo (MCMC) samples. Several books,[26,34,48,85,95] cover the development of MCMC sampling and advanced Monte Carlo (MC) methods for computing posterior quantities using the samples from the posterior distribution.

4.1. *Sampling from posterior distribution*

During the last decade, Monte Carlo (MC) based sampling methods for evaluating high-dimensional posterior integrals have been rapidly developing. Those sampling methods include MC importance sampling,[44,54,84,102]

Gibbs sampling,[41,43] Metropolis–Hastings sampling[52,56,75] and many other hybrid algorithms.

4.1.1. *Basic Gibbs sampler*

The Gibbs sampler may be one of the best known MCMC sampling algorithms in the Bayesian computational literature. As discussed in Besag and Green,[11] the Gibbs sampler is founded on the ideas of Grenander,[53] while the formal term is introduced by Geman and Geman.[43] The primary bibliographical landmark for Gibbs sampling in problems of Bayesian inference is Gelfand and Smith.[41] A similar idea termed as *data augmentation* is introduced by Tanner and Wong.[96] Casella and George[15] provide an excellent tutorial on the Gibbs sampler.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ be a p-dimensional vector of parameters and let $\pi(\boldsymbol{\theta}|D)$ be its posterior distribution given the data D. Then, the basic scheme of the Gibbs sampler is given as follows:

Step 0. Choose an arbitrary starting point $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{p,0})'$, and set $i = 0$.

Step 1. Generate $\boldsymbol{\theta}_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, \ldots, = \theta_{p,i+1})'$ as follows:

- Generate $\theta_{1,i+1} \sim \pi(\theta_1|\theta_{2,i}, \ldots, \theta_{p,i}, D)$;
- Generate $\theta_{2,i+1} \sim \pi(\theta_2|\theta_{1,i+1}, \theta_{3,i}, \ldots, \theta_{p,i}, D)$;
  ... ...
- Generate $\theta_{p,i+1} \sim \pi(\theta_p|\theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p-1,i+1}, D)$.

Step 2. Set $i = i + 1$, and go to Step 1.

Thus each component of $\boldsymbol{\theta}$ is visited in the natural order and a cycle in this scheme requires generation of p random variates. Gelfand and Smith[41] show that under certain regularity conditions, the vector sequence $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots\}$ has a stationary distribution $\pi(\boldsymbol{\theta}|D)$. Schervish and Carlin[88] provide a sufficient condition that guarantees geometric convergence. Other properties regarding geometric convergence are discussed in Roberts and Polson.[87]

Example 8.

For the constrained linear model considered in Example 1, the posterior distribution for $(\boldsymbol{\beta}, \sigma^2)$ based on the New Zealand apple data D is given by (11). The Gibbs sampler can be implemented by taking

$$\beta_j|\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{10}, \sigma^2\,, \quad D \sim N(\theta_j, \delta_j^2)\,, \tag{34}$$

subject to $\beta_{j-1} \le \beta_j \le \beta_{j+1}$ $(\beta_0 = 0)$ for $j = 1, 2, \ldots, 9$,

$$\beta_{10}|\beta_1, \ldots, \beta_9, \sigma^2\,, \quad D \sim N(\psi\theta_{10} + (1-\psi)\mu_{10}, (1-\psi)\sigma_{10}^2)\,, \tag{35}$$

subject to $\beta_{10} \ge \beta_9$ and

$$\sigma^2|\boldsymbol{\beta}\,, \quad D \sim \mathcal{IG}\left(\frac{n}{2},\ \frac{\sum_{i=1}^n (y_i - \sum_{j=1}^{10} x_{ij}\beta_j)^2}{2}\right), \tag{36}$$

where in (34) and (35), $\psi = \sigma_{10}^2/(\sigma_{10}^2 + \delta_{10}^2)$,

$$\theta_j = \left(\sum_{i=1}^n x_{ij}^2\right)^{-1} \left[\sum_{i=1}^n \left(y_i - \sum_{l \ne j} x_{il}\beta_l\right) x_{ij}\right], \tag{37}$$

and

$$\delta_j^2 = \left(\sum_{i=1}^n x_{ij}^2\right)^{-1} \sigma^2\,, \tag{38}$$

for $j = 1, \ldots, 10$, and $\mathcal{IG}(\xi, \eta)$ denotes the inverse gamma distribution. Inverse gamma distribution with parameters (ξ, η), whose density is given by

$$\pi(\sigma^2|\xi, \eta) \propto (\sigma^2)^{-(\xi+1)} e^{-\eta/\sigma^2}\,.$$

4.1.2. *Metropolis–Hastings algorithm*

The Metropolis–Hastings algorithm is developed by Metropolis *et al.*[75] and subsequently generalized by Hastings.[56] Tierney[98] gives a comprehensive theoretical exposition of this algorithm, and Chib and Greenberg[27] provide an excellent tutorial on this topic.

Let $q(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ be a proposal density, which is also termed as a *candidate-generating density* by Chib and Greenberg,[27] such that

$$\int q(\boldsymbol{\theta}, \boldsymbol{\vartheta})\ d\boldsymbol{\vartheta} = 1\,.$$

Also let $U(0, 1)$ denote the uniform distribution over (0, 1). Then, a general version of the Metropolis–Hastings algorithm for sampling from the posterior distribution $\pi(\boldsymbol{\theta}|D)$ can be described as follows:

Step 0. Choose an arbitrary starting point $\boldsymbol{\theta}_0$ and set $i = 0$.
Step 1. Generate a candidate point $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}_i, \cdot)$ and u from $U(0, 1)$.

Step 2. Set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ if $u \leq a(\theta_i, \boldsymbol{\theta}^*)$ and $\theta_{i+1} = \theta_i$ otherwise, where the acceptance probability is given by

$$a(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \min \left\{ \frac{\pi(\boldsymbol{\vartheta}|D) q(\boldsymbol{\vartheta}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|D) q(\boldsymbol{\theta}, \boldsymbol{\vartheta})},\ 1 \right\}. \tag{39}$$

Step 3. Set $i = i + 1$, and go to Step 1.

The performance of a Metropolis–Hastings algorithm depends on the choice of a proposal density q. In the context of the random walk proposal density, which is of the form $q(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = q_1(\boldsymbol{\vartheta} - \boldsymbol{\theta})$, where $q_1(\cdot)$ is a multivariate density, Roberts *et al.*,[86] show that if the target and proposal densities are normal, then the scale of the latter should be tuned so that the acceptance rate is approximately 0.45 in one-dimensional problems and approximately 0.23 as the number of dimensions approaches infinity, with the optimal acceptance rate being around 0.25 in six dimensions. For the *independence chain*, in which we take $q(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = q(\boldsymbol{\vartheta})$, it is important to ensure that the tails of the proposal density $q(\boldsymbol{\vartheta})$ dominate those of the target density $\pi(\boldsymbol{\theta}|D)$, which is similar to a requirement on the importance sampling function in Monte Carlo integration with importance sampling.

Example 9. Consider a Poisson mixed model:

$$y_i \sim \mathcal{P}(\mu_i)\,,$$

where $\mu_i = \exp(x_i'\boldsymbol{\beta} + \epsilon_i)$ for $i = 1, 2, \ldots, n$, $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. We assume the random effects

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)' \sim N(0, \Sigma)\,,$$

where

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}.$$

Assume that a noninformative prior for $(\boldsymbol{\beta}, \sigma^2, \rho)$ has the form

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho) \propto (\sigma^2)^{-(\delta_0+1)} \exp(-\sigma^{-2}\gamma_0)\,,$$

where the hyperparameters $\delta_0 > 0$ and $\gamma_0 > 0$ are prespecified. Then, the joint posterior distribution for $(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon})$ is given by

$$\pi(\boldsymbol{\beta}, \rho, \sigma^2, \boldsymbol{\epsilon}|D) \propto \exp\left\{y'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) - J_n' Q(\boldsymbol{\beta}, \boldsymbol{\epsilon}) - \frac{1}{2}\boldsymbol{\epsilon}'\Sigma^{-1}\boldsymbol{\epsilon}\right\}$$
$$\times \frac{1}{\sigma^n(1-\rho^2)^{\frac{n-1}{2}}} \times (\sigma^2)^{-(\delta_0+1)} \exp\left(-\frac{\lambda_0}{\sigma^2}\right), \qquad (40)$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $J_n = (1, 1, \ldots, 1)'$, $Q(\boldsymbol{\beta}, \boldsymbol{\epsilon}) = (q_1, q_2, \ldots, q_n)'$, $q_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i) + \log(y_i!)$, X is the covariate matrix with the ith row equal to $\boldsymbol{x}_i'$, and $D = (n, \boldsymbol{y}, X)$.

To obtain a more efficient MCMC sampling algorithm, we consider a hierarchically centered reparameterization, which is given by

$$\boldsymbol{\eta} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}\,.$$

Using (40), the reparameterized posterior for $(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta})$ is written as

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta}|D) \propto \exp\{\boldsymbol{y}'\boldsymbol{\eta} - J_n' Q(\boldsymbol{\eta}) - J_n' C(\boldsymbol{y})\}$$
$$\times (2\pi\sigma^2)^{-n/2}(1-\rho^2)^{-(n-1)/2}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\eta} - X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta} - X\boldsymbol{\beta})\right\}, \qquad (41)$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n)'$, and $Q(\boldsymbol{\eta})$ is an $n \times 1$ vector with the tth element equal to $q_i = \exp(\boldsymbol{\eta}_i)$. We note that the hierarchical centering method of Gelfand *et al.*[39,40] is a tool to improve convergence of MCMC sampling. As discussed in Chen *et al.*,[26] this technique is particularly useful for the Poisson mixed model.

To sample from the reparameterized posterior $\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta}|D)$, the following steps are required:

Step 1. Draw $\boldsymbol{\eta}$ from its conditional posterior distribution

$$\pi(\boldsymbol{\eta}|\boldsymbol{\beta}, \sigma^2, \rho, D)$$
$$\propto \exp\left\{\boldsymbol{y}'\boldsymbol{\eta} - J_n' Q(\boldsymbol{\eta}) - \frac{(\boldsymbol{\eta} - X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta} - X\boldsymbol{\beta})}{2\sigma^2}\right\}. \qquad (42)$$

Step 2. Draw $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|\boldsymbol{\eta}, \sigma^2, \rho,\ D \sim N_8((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\boldsymbol{\eta}, \sigma^2(X'\Sigma^{-1}X)^{-1})\,.$$

Step 3. Draw σ^2 from its conditional posterior

$$\sigma^2|\boldsymbol{\beta}, \rho, \boldsymbol{\eta}, D \sim \mathcal{IG}(\delta^*, \gamma^*)\,,$$

where $\delta^* = \delta_0 + n/2$, $\gamma^* = \gamma_0 + \frac{1}{2}(\boldsymbol{\eta} - X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta} - X\boldsymbol{\beta})$, and $\mathcal{IG}(\delta^*, \gamma^*)$ is an inverse gamma distribution.

Step 4. Draw ρ from its conditional posterior

$$\pi(\rho|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D) \propto (1-\rho^2)^{-(n-1)/2} \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\eta} - X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta} - X\boldsymbol{\beta})\right\}. \quad (43)$$

In Step 1, it can be shown that $\pi(\boldsymbol{\eta}|\boldsymbol{\beta}, \sigma^2, \rho, D)$ is log-concave in each component of $\boldsymbol{\eta}$. Thus $\boldsymbol{\eta}$ can be drawn using the adaptive rejection sampling algorithm of Gilks and Wild.[49] The implementation of Steps 2 and 3 is straightforward, which may be a bonus of hierarchical centering, since sampling $\boldsymbol{\beta}$ is much more expensive before the reparameterization. In Step 4, we use a so-called "*Localized Metropolis*" algorithm, which was introduced in Chen *et al.*[26]

The *Localized Metropolis* algorithm requires the following transformation:

$$\rho = \frac{-1 + e^{\xi}}{1 + e^{\xi}}, \quad -\infty < \xi < \infty.$$

Using (43), we have

$$\pi(\xi|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D) = \pi(\rho|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D)\frac{2e^{\xi}}{(1+e^{\xi})^2}.$$

Now, we generate ξ by using a normal proposal $N(\hat{\xi}, \hat{\sigma}^2_{\hat{\xi}})$, where $\hat{\xi}$ is a maximizer of the logarithm of $\pi(\xi|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\epsilon}, D)$, which can be obtained by, for example, the Newton-Raphson algorithm, and $\hat{\sigma}^2_{\hat{\xi}}$ is the minus of the inverse of the second derivative of $\log \pi(\xi|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D)$ evaluated at $\xi = \hat{\xi}$, i.e.

$$\hat{\sigma}^{-2}_{\hat{\xi}} = -\frac{d^2 \log \pi(\xi|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D)}{d\xi^2}\bigg|_{\xi=\hat{\xi}}.$$

The algorithm to generate ξ operates as follows:

(1) Let ξ be the current value.
(2) Generate a proposal value ξ^* from $N(\hat{\xi}, \hat{\sigma}^2_{\hat{\xi}})$.
(3) A move from ξ to ξ^* is made with probability

$$\min\left\{\frac{\pi(\xi^*|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D)\phi\left(\frac{\xi - \hat{\xi}}{\hat{\sigma}_{\hat{\xi}}}\right)}{\pi(\xi|\sigma^2, \boldsymbol{\beta}, \boldsymbol{\eta}, D)\phi\left(\frac{\xi^* - \hat{\xi}}{\hat{\sigma}_{\hat{\xi}}}\right)}, \ 1\right\},$$

where ϕ is the $N(0,1)$ probability density function.

Note that the proposal $(\hat{\xi}, \hat{\sigma}^2_{\xi})$ does not depend on the current value of ξ, which will typically produce a small autocorrelation among ξ's.

Recently, several Bayesian software packages have been developed. These include BUGS for analyzing general hierarchical models via MCMC (http://www.mrc-bsu.cam.ac.uk/bugs/), BATS for Bayesian time series analysis (http://www.stat.duke.edu/~mw/bats.html), Matlab and Minitab Bayesian computational algorithms for introductory Bayesian analysis (http://www-math.bgsu.edu/~albert/), and many others. A more complete listing and description of pre-1990 Bayesian software can be found in Goel.[50] A listing of some of the Bayesian software developed since 1990 is given in Berger.[4]

4.2. *Computing posterior quantities*

In Bayesian inference, MC methods are often used to compute the posterior expectation $E(h(\boldsymbol{\theta})|D)$, since the analytical evaluation of $E(h(\boldsymbol{\theta})|D)$ is typically not available. Assuming that $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots, n\}$ is an MCMC sample from $\pi(\boldsymbol{\theta}|D)$, the MC estimator of $E(h(\boldsymbol{\theta})|D)$ is given by

$$\hat{E}(h) = \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i)\,. \tag{44}$$

Asymptotic or small sample properties of $\hat{E}(h)$ depend on the algorithm used to generate the sample $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots, n\}$. Under certain regularity conditions such as *ergodicity*, the MC estimator $\hat{E}(h)$ is consistent.

Since $\hat{E}(h)$ is a random quantity, it is important to compute the simulation standard error of $\hat{E}(h)$, as it provides the magnitude of the simulation accuracy of the estimator $\hat{E}(h)$. Let $\mathrm{var}(\hat{E}(h))$ be the variance of $\hat{E}(h)$, and let $\widehat{\mathrm{var}}(\hat{E}(h))$ be an estimate of $\mathrm{var}(\hat{E}(h))$. Then, the simulation standard error of $\hat{E}(h)$ is defined as

$$\mathrm{se}(\hat{E}(h)) = [\widehat{\mathrm{var}}(\hat{E}(h))]^{1/2}\,, \tag{45}$$

which is the square root of the estimated variance of the MC estimator $\hat{E}(h)$. Since the sample generated by an MCMC sampling algorithm is often dependent, a complication that arises from the autocorrelation is that $\mathrm{var}(\hat{E}(h))$ is difficult to obtain. A variety of methods for obtaining a dependent sample based estimate of $\mathrm{var}(\hat{E}(h))$ are discussed in system simulation textbooks.[13,71,84] In this subsection, we briefly discuss a general overlapping batch statistics (obs) method considered in Schmeiser *et al.*[89] for computing $\widehat{\mathrm{var}}(\hat{E}(h))$.

Suppose that $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots, n\}$ is a dependent sample, from which a point estimator $\hat{\xi}$ of the posterior quantity of interest is computed. (Here, $\hat{\xi} = \hat{E}(h)$.) The obs estimate of the variance of $\hat{\xi}$ is

$$\hat{V}(m) = \left[\frac{m}{n-m}\right] \frac{\sum_{j=1}^{n-m+1}(\hat{\xi}_j - \hat{\xi})^2}{(n-m+1)}, \tag{46}$$

where $\hat{\xi}_j$ is defined analogously to $\hat{\xi}$, but is a function of only $\boldsymbol{\theta}_j$, $\boldsymbol{\theta}_{j+1}$, …, $\boldsymbol{\theta}_{j+m-1}$. Sufficient conditions for obs estimators to be unbiased and have variance inversely proportional to n are given in Schmeiser *et al.*[89] Using (46), the simulation standard error of $\hat{\xi}$ is $\text{se}(\hat{\xi}) = \sqrt{\hat{V}(m)}$. The primary difficulty in using the obs estimator is the choice of the batch size m to balance bias and variance, since no optimal batch size formula is known for general obs estimators. Limiting behavior for $\hat{V}(m)$ for some special obs estimators is discussed.[51,92] For many situations, choosing m so that $10 \leq n/m \leq 20$ is reasonable.

During last several years, many other efficient Monte Carlo methods have been developed for computing posterior quantities other than $E(h(\boldsymbol{\theta})|D)$. These include the bridge sampling method of Meng and Wong,[73] the path sampling method of Gelman and Meng[42] and the ratio importance sampling method of Chen and Shao[25] for computing normalizing constants and Bayes factors; and the MC methods of Chen and Shao[24] for calculating HPD intervals. The detailed description and discussion of these methods can be found in Chen *et al.*[26]

5. Applications and Examples

5.1. *Bayesian analysis for survival data with a cure fraction*

The cure rate model is needed for modelling time-to-event data for various types of cancers, including breast cancer, non-Hodgkins lymphoma, leukemia, prostate cancer, melanoma, and head and neck cancer, where for these diseases, a significant proportion of patients are "cured". To demonstrate such a phenomenon, we consider a recent phase III clinical trial in malignant melanoma (E1684) undertaken by the Eastern Cooperative Oncology Group (ECOG). The graph in Fig. 2 gives the Kaplan–Meier survival curve for 284 patients in E1684, with the survival time given in years. We see from Fig. 2 that a plateau in the curve occurs at approximately 0.36, suggesting that 36% fraction of patients are "cured" after sufficient follow-up.

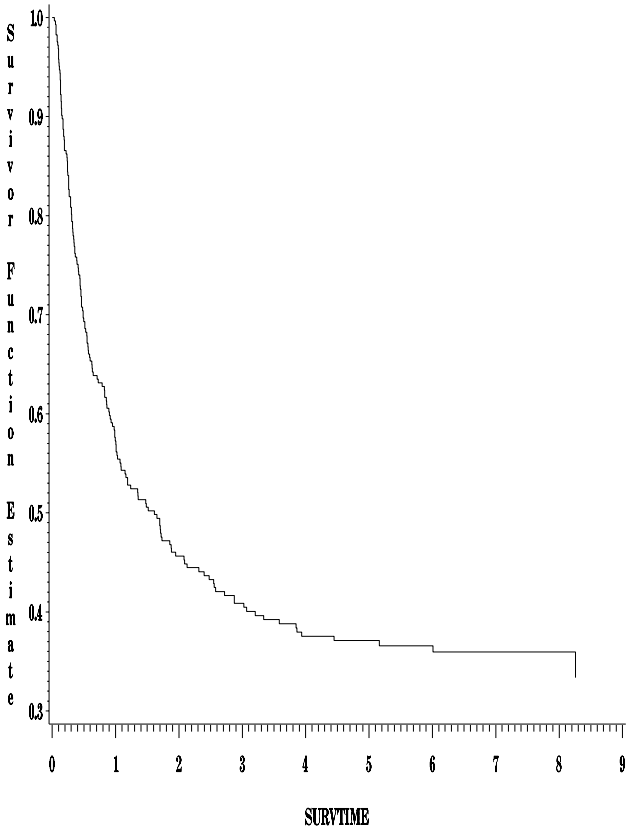


Fig. 2. Kaplan–Meier plot for E1684 data.

An important issue with cure rate modelling is model comparison. It will be of interest to compare various cure models to the Cox model. It will also be of interest to compare various semi-parametric models to obtain the most parsimonious and best fitting semi-parametric model. For model comparisons, Bayes factors require proper priors, and criterion based statistics such as the L measure[63] will not be well defined for cure rate models since the cure rate model does not have proper probability density. As a result, we need to turn to other measures to carry out model comparisons. Here, we use the Conditional Predictive Ordinate (CPO) as a goodness of fit statistic that is well defined for these models and will allow us to do formal model comparisons.

We compare three types of models for modelling time-to-event data.

5.1.1. *Cox model*

A proportional hazards model is defined by a hazard function of the form

$$h(t, \boldsymbol{x}) = h_0(t) \exp(\boldsymbol{x}'\boldsymbol{\beta}) \,, \tag{47}$$

where $h_0(t)$ denotes the baseline hazard function at time t, $\boldsymbol{x}$ denotes the covariate vector for an arbitrary individual in the population, and $\boldsymbol{\beta}$ denotes a vector of regression coefficients. Suppose we have n subjects, and let $y_1, \ldots, y_n$ denote the observed failure times or censoring times for the individuals, and ν_i is the indicator variable taking on the value 1 if y_i is a failure time, and 0 if it is a censoring time. Our semi-parametric development for this model is based on a piece-wise constant hazard. We construct a finite partition of the time axis, $0 < s_1 < \cdots < s_J$, with $s_J > y_i$ for all $i = 1, 2, \ldots, n$. Thus, we have the J intervals $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J]$. In the jth interval, we assume a constant hazard λ_j. Throughout, we let $D = (n, \boldsymbol{y}, X, \boldsymbol{\nu})$ denote the observed data for the current study, where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$, and X is the $n \times p$ matrix of covariates with ith row $\boldsymbol{x}_i'$. Letting $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)'$, we can write the likelihood function of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ for all n subjects as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda} | D) = \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_j \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{\delta_{ij}\boldsymbol{\nu}_i} \\ \times \exp\left\{ -\delta_{ij} \left[\lambda_j (y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \right] \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \right\}, \end{aligned} \tag{48}$$

where $\delta_{ij} = 1$ if the ith subject failed or was censored in the jth interval, and 0 otherwise, $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ denotes the $p \times 1$ vector of covariates for the ith subject, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)'$ is the corresponding vector of regression coefficients. The indicator δ_{ij} is needed to properly define the likelihood over the J intervals for the semi-parametric models. The semi-parametric model in (48), sometimes referred to as a piecewise exponential model, is quite general and can accommodate various shapes of the baseline hazard over the intervals. Moreover, we note that if $J = 1$, then the model reduces to a parametric exponential model with failure rate parameter $\lambda \equiv \lambda_j$, $j = 1, 2, \ldots, J$. This semi-parametric proportional hazards model is a useful and simple model for modelling survival data. It serves as the benchmark for comparisons with other semi-parametric or fully parametric models for survival data.

5.1.2. *Parametric cure rate model*

We present a version of the cure rate model studied.[21,104] Suppose that for an individual in the population, we let N denote the number of *metastatic-competent* tumor cells for that individual left active after the initial treatment. A metastatic-competent tumor cell is a tumor cell which has the potential of metastasizing. Further, we assume that N has a Poisson distribution with mean θ. We let Z_i denote the random time for the ith metastatic-competent tumor cell to produce detectable metastatic disease. That is, Z_i can be viewed as an incubation time for the ith tumor cell. The variables Z_i, $i = 1, 2, \ldots$, are assumed to be independent and identically distributed with a common distribution function $F(t) = 1 - S(t)$ and are independent of N. The time to relapse of cancer can be defined by the random variable $T = \min\{Z_i, 0 \le i \le N\}$, where $P(Z_0 = \infty) = 1$ and N is independent of the sequence $Z_1, Z_2, \ldots$. The survival function for T, and hence the survival function for the population, is given by

$$\begin{aligned} S_p(t) &= P(\text{no metastatic cancer by time } t) \\ &= P(N = 0) + P(Z_1 > t, \ldots, Z_N > t, N \ge 1) \\ &= \exp(-\theta) + \sum_{k=1}^{\infty} S(t)^k \, \frac{\theta^k}{k!} \exp(-\theta) = \exp(-\theta + \theta S(t)) \\ &= \exp(-\theta F(t)) \,. \end{aligned} \tag{49}$$

Since $S_p(\infty) = \exp(-\theta) > 0$, (49) is not a proper survival function. We see that (49) shows explicitly the contribution to the failure time of two distinct characteristics of tumor growth: the initial number of metastatic-competent cells and the rate of their progression. Thus the model incorporates parameters bearing clear biological meaning. The model in (49) is quite different from the standard mixture cure rate model proposed by Berkson and Gage,[3] and has several attractive properties. For a detailed discussion of the various properties of (49), we refer the reader to Yakovlev and Tsodikov.[104] Aside from the biological motivation, the model in (49) is suitable for any type of failure-time data with a surviving fraction. Thus, failure-time data which do not "fit" the biological definition given above can still certainly be modeled by (49) as long as the data has a surviving fraction and can be thought of as being generated by an unknown number N of latent independent competing risks (Z_i's). Yakovlev *et al.*[103] discuss a similar modeling technique for tumor latency, but do not consider a Bayesian formulation of the model.

We also see from (49) that the cure fraction (i.e. cure rate) is given by

$$S_p(\infty) \equiv P(N = 0) = \exp(-\theta)\,. \tag{50}$$

As $\theta \to \infty$, the cure fraction tends to 0, whereas as $\theta \to 0$, the cure fraction tends to 1. The sub-density corresponding to (49) is given by

$$f_p(t) = \theta f(t) \exp(-\theta F(t))\,, \tag{51}$$

where $f(t) = \frac{d}{dt}F(t)$ is a proper probability density function. The hazard function is given by

$$h_p(t) = \theta f(t)\,. \tag{52}$$

Note that $h_p(t)$ is *not* a hazard function corresponding to a probability distribution since $S_p(t)$ is not a proper survival function.

Suppose we have n subjects and for the ith subject, let y_i denote the observed survival time, let ν_i be the censoring indicator that equals 1 if y_i is a failure time and 0 if it is right censored, and also let N_i denote the number of metastatic-competent tumor cells. Further, we assume that the N_i's are i.i.d. Poisson random variables with mean θ_i, which is related to the covariates by $\theta_i \equiv \theta(\boldsymbol{x}_i'\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Letting $\boldsymbol{N} = (N_1, \ldots, N_n)'$, the "complete data" is given by $D_{\text{comp}} = (n, \boldsymbol{y}, \boldsymbol{\nu}, \boldsymbol{N})$, where $\boldsymbol{N}$ is an unobserved vector of latent variables. Then, we can write the complete data likelihood of $(\boldsymbol{\beta}, \lambda)$ as

$$\begin{aligned} L(\boldsymbol{\beta}, \lambda | D_{\text{comp}}) = &\left(\prod_{i=1}^{n} S(y_i|\lambda)^{N_i - \nu_i} \; (N_i f(y_i|\lambda))^{\nu_i} \right) \\ &\times \exp\left\{ \sum_{i=1}^{n} [N_i x_i'\beta - \log(N_i!) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})] \right\}, \end{aligned} \tag{53}$$

where $f(y_i|\lambda)$ is exponential density given above, and $S(y_i|\lambda) = 1 - F(y_i|\lambda) = \exp(-\lambda y_i)$. Since the latent vector $\boldsymbol{N}$ is not observed, the likelihood function based on the observed data $D = (n, \boldsymbol{y}, X, \boldsymbol{\nu})$ is obtained by summing (53) over $\boldsymbol{N}$, leading to

$$L(\boldsymbol{\beta}, \lambda | D) = \sum_{\boldsymbol{N}} L(\boldsymbol{\beta}, \lambda | D_{\text{comp}})\,. \tag{54}$$

5.1.3. *Semi-parametric cure rate model*

We consider a semi-parametric version of the parametric cure rate model in (53) by considering a piecewise constant hazard model, and thus assume

that the hazard is equal to λ_j for the jth interval, $j = 1, \ldots, J$. With this assumption, the complete data likelihood can be written as

$$
\begin{aligned}
&L(\boldsymbol{\beta}, \boldsymbol{\lambda} | D_{\text{comp}}) \\
&\quad = \prod_{i=1}^{n} \prod_{j=1}^{J} \exp\left\{ -(N_i - \nu_i)\delta_{ij} \left[\lambda_j (y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \right] \right\} \\
&\qquad \times \prod_{i=1}^{n} \prod_{j=1}^{J} (N_i \lambda_j)^{\delta_{ij}\nu_i} \exp\left\{ -\nu_i \delta_{ij} \left[\lambda_j (y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \right] \right\} \\
&\qquad \times \exp\left\{ \sum_{i=1}^{n} [N_i \boldsymbol{x}_i' \boldsymbol{\beta} - \log(N_i!) - \exp(\boldsymbol{x}_i' \boldsymbol{\beta})] \right\}, \qquad (55)
\end{aligned}
$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)'$. The model in (55) is a semi-parametric version of (53). If we take $J = 1$ in (55), then the model reduces to the fully parametric model given in (53). There are several attractive features of the model in (55). First, we note the degree of the non-parametricity is controlled by J. The larger the J, the more non-parametric the model is. However, by picking a small to moderate J, we get more of a parametric shape for the survival function. This is an important aspect for the cure rate model, since the estimation of the cure rate parameter θ could be highly affected by the non-parametric nature of the survival function. For this reason, it may be desirable to choose small to moderate values of J for cure rate modelling. In practice, we recommend doing analyses for several values of J to see the sensitivity of the posterior estimates of the regression coefficients. We recommend doing sensitivity analyses for small, moderate, and large values of J. Thus, the semi-parametric cure rate model (55) is quite flexible, as it allows us to model general shapes of the hazard function, as well as choose the degree of parametricity through suitable choices of J. Again, since $\boldsymbol{N}$ is not observed, the observed data likelihood, $L(\boldsymbol{\beta}, \boldsymbol{\lambda} | D)$ is obtained by summing out N from (55) as in (54).

5.1.4. *Prior distributions*

First, we consider a noninformative prior. Take

$$
\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \propto \prod_{j=1}^{J} \lambda_j^{\zeta_0 - 1} \exp(-\tau_0 \lambda_j), \qquad (56)
$$

where $\zeta_0 \geq 0$ and $\tau_0 \geq 0$, so that β has an improper uniform prior and $\lambda_j \sim \text{gamma}(\zeta_0, \tau_0)$. The prior given in (56) includes two special cases: (i) Jeffreys's prior for λ when $\zeta_0 = \tau_0 = 0$, and (ii) uniform prior for λ when $\zeta_0 = 1$ and $\tau_0 = 0$.

Second, we consider an informative prior. We use the *power prior* to formally construct an informative prior distribution from *historical data* D_0. Let n_0 denote the sample size for the historical data, $\boldsymbol{y}_0 = (y_{01}, y_{02}, \ldots, y_{0n_0})'$ be an $n_0 \times 1$ vector of right censored failure times for the historical data with censoring indicators $\boldsymbol{\nu}_0 = (\nu_{01}, \nu_{02}, \ldots, \nu_{0n_0})'$, and X_0 is an $n_0 \times k$ matrix of covariates with ith row $\boldsymbol{x}'_{0i}$. Let $D_0 = (n_0, \boldsymbol{y}_0, X_0, \boldsymbol{\nu}_0)$ denote the observed historical data. The power prior given by (26) has the form

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, a_0 | D_0) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda} | D_0)^{a_0} \pi_0(\boldsymbol{\beta}, \boldsymbol{\lambda}) a_0^{\alpha_0 - 1} (1 - a_0)^{\lambda_0 - 1}, \tag{57}$$

where $L(\beta, \lambda | D_0)$ is the likelihood function based on the observed historical data, and α_0 and λ_0 are prespecified hyperparameters. The quantity $\pi_0(\boldsymbol{\beta}, \lambda)$ is the initial prior for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, which is (56).

It is well known that with insufficient follow-up or with too few events, the estimate of the cure rate can be quite unreliable and unstable. In addition, the model itself may not be identifiable or nearly identifiable if there is insufficient follow-up and there are too few events. The use of informative prior distributions can help overcome such difficulties, which can provide better estimates of the cure rate and make the model identifiable.

5.1.5. *Model assessment*

We use a summary statistic of the CPO_i's given by (7), the logarithm of the *pseudo-Bayes factor*, for model assessment. The CPO_i given by (7) has the form

$$\text{CPO}_i = f(y_i | y_{(i)}) = \int f(y_i | \boldsymbol{\beta}, \boldsymbol{\lambda}) \; \pi(\boldsymbol{\beta}, \boldsymbol{\lambda} | y_{(i)}) \, d\boldsymbol{\beta} \, d\boldsymbol{\lambda}, \tag{58}$$

where y_i denotes the response variable for case i, and $y_{(i)}$ denotes the entire response vector with the ith case deleted. Then, the logarithm of the pseudo-Bayes factor is defined as

$$B = \sum_{i=1}^{n} \log(\text{CPO}_i). \tag{59}$$

In the context of survival data, the statistic B has been discussed before.[30,38,91]

We see from (58) that B is always well defined as long as the posterior predictive density is proper. Thus, B is well defined under improper priors, and in addition, it is very computationally stable. Therefore, B has a clear advantage over the Bayes factor as a model assessment tool, since it is well known that the Bayes factor is not well defined with improper priors, and is generally quite sensitive to vague proper priors. Thus, the Bayes factor is not applicable for many of our models here, since we consider several models involving improper priors. In addition, the B statistic also has clear advantages over other model selection criteria, such as the *L measure.*[63,70] The L measure is a Bayesian criterion requiring finite second moments of the sampling distribution of y_i, whereas the B statistic does not require existence of any moments. Since the cure rate models in (53) and (55) have improper survival functions, no moments of the sampling distribution exist, and therefore the L measure is not well defined for these models. Thus, for the models considered here, the B statistic is well motivated.

5.1.6. *E1690 melanoma study*

To demonstrate the methodologies, we consider the second ECOG trial, E1690, in malignant melanoma. This study had $n = 427$ patients on the high dose interferon arm and observation arm combined. ECOG initiated this trial, right after the completion of E1684, to attempt to confirm the results of E1684 and to study the benefit of Interferon Alpha-2b (IFN) given at a lower dose. The E1690 trial accrued patients from 1991 until 1995, and was unblinded in 1998. The E1690 trial was designed for exactly the same patient population as E1684, and the high dose IFN arm in E1690 was identical to that of E1684. See Kirkwood *et al.*[67,68] for detail.

We carry out a Bayesian analysis of E1690 using E1684 as historical data using relapse-free survival (RFS) as the response variable with the treatment as a covariate. Since E1684 has longer follow-up than E1684, the use of E1684 as a historical may help to improve the accuracy in the estimates of cure rates based on E1690. But, in this example, we solely focus on model comparisons. We consider Cox model, parametric cure rate (PCR) model, and semiparametric cure rate (SPCR) model with $J = 1$, $J = 5$, and $J = 10$ for noninformative and informative priors. Table 4 shows the results of Pseudo-Bayes Factors (B's) when $a_0 = 0$ with probability 1, $E(a_0|D) = 0.05$, 0.20, 0.30, and 0.60, and $a_0 = 1$ with probability 1. We note that $a_0 = 0$ implies the use of noninformative prior, which is equivalent to a prior specification with no incorporation of historical data, while with $a_0 = 1$, we simply combine D and D_0 together.

Table 4. Pseudo-Bayes factors (B's).

Model		$a_0 = 0$	$E(a_0\|D) \simeq$ 0.05	0.20	0.30	0.60	$a_0 = 1$
Cox	$J = 1$	−575.60	−575.45	−575.23	−575.13	−574.95	−574.64
	$J = 5$	−522.30	−522.05	−521.67	−521.59	−521.61	−522.24
	$J = 10$	−523.62	−523.20	−522.39	−522.12	−522.02	−522.71
SPCR	$J = 1$	−519.75	−519.61	−519.39	−519.34	−519.40	−519.67
	$J = 5$	−520.24	−519.89	−519.43	−519.31	−519.67	−520.16
	$J = 10$	−524.42	−523.82	−522.83	−522.53	−522.56	−522.97

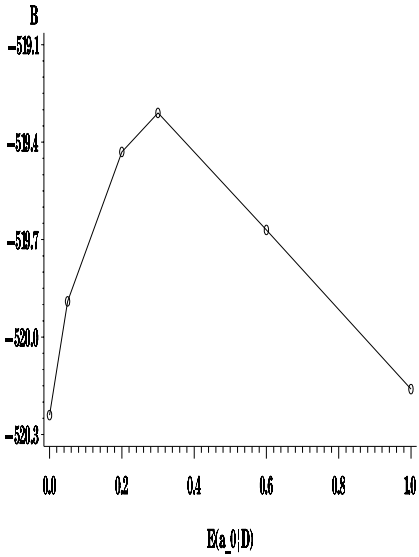


Fig. 3. Plot of pseudo-Bayes factors for SPCR with $J = 5$.

Table 4 is quite informative. First, for the degree of parametricity, $J = 5$ is better than $J = 1$ or $J = 10$. However, for SPCR $J = 1$ and $J = 5$ are fairly close. Second, for both $J = 1$ or $J = 5$, the cure rate model yields a better fit than the Cox model. Third, the incorporation of E1684 in the analysis improves the model fit over the exclusion of historical data. Fourth, for all the cases, B is a concave function of $E(a_0|D)$, see Fig. 3 for an illustration. This is an interesting feature in B in that it demonstrates that there is an "optimal" weight for the historical data with respect to the statistic B, and thus this property is potentially very useful in selecting a model and the prior weight a_0.

5.2. *Bayesian model selection for multivariate mortality data with large families*

To illustrate Bayesian model selection, we consider an analysis of the multivariate mortality data with large families from a series of animal toxicological experiments performed in the Department of Biology at the University of Waterloo. One of these experiments to study the toxic effect of potassium thiocyanate (KSCN) on the mortality of trout fish eggs. In this experiment, each of the six levels of KSCN were added to different tanks each containing many trout fish eggs (61 to 179 eggs per vial). Half of the tanks were water hardened before the application of the KSCN ($x_1 = 1$) and other half were water hardened after the application of the KSCN ($x_1 = 0$). Another covariate is the continuous variable x_2, which is defined as the natural logarithm of the level of KSCN. Each experimental condition was replicated 4 times, so there were in total 48 tanks. Another similar experiment was conducted in the same laboratory with a different toxicant, sodium thiocyanate (NaSCN). For the KSCN data, mortality counts for each tank were taken at 5, 11, 19, 31 and 35 days after the application of KSCN, while for the NaSCN data mortality counts for each tank were taken at 1, 6, 13, 20, and 27 days after the application of the NaSCN. We refer the reader to O'Hara Hines[77] and O'Hara Hines and Lawless[78] for the more detailed description of these two experiments. Since NaSCN is a toxicant similar to KSCN and both experiments were similar in design and purpose, the data from the NaSCN experiment will be used to build an informative prior distribution for the KSCN study.

For the KSCN data, there are $K = 48$ tanks (families) of fishes (subjects). Suppose that the observation times for each tank are $t_0 = 0 < t_1 < \cdots < t_m$ ($m = 5$). Moreover, it is given that $t_1 = 5, t_2 = 11, t_3 = 19, t_4 = 31$ and $t_5 = 35$. The cumulative mortality counts (the observed data) in each tank at these five days, can be summarized as $\{(d_{kj}, r_{kj}) : j = 1, \ldots, 5;\ k = 1, \ldots, 48\}$ where d_{kj} is the number of fishes dying and r_{kj} is the number of fishes at risk in the kth tank during the time interval $(t_{j-1}, t_j] = I_j$. Let h_{kj} be the mortality rate, which can be also interpreted as the discrete hazard rate, for a fish from the kth tank at the time interval I_j:

$$h_{kj} = P(T_i \in I_j | i \in R_{kj}, \boldsymbol{x}_i)\,,$$

where T_i is the time of death for a fish from the kth tank, and R_{kj} is the set of fishes still "at risk" (alive) in the kth tank at the beginning of the time interval $(t_{j-1}, t_j]$ (for $j = 1, 2, \ldots, 5$ and $k = 1, 2, \ldots, K$). There are two

covariates x_{1k} and x_{2k} for each tank, where x_{2k} is the primary covariate, the natural logarithm of the KSCN level applied to the kth tank, and x_{1k} is the water hardening of the kth tank. Thus, the number of fishes dying in the kth tank during I_j is distributed as Binomial with success probability h_{kj} and the number of trials is equal to the number of fishes at risk during that interval in the kth tank, i.e. $d_{kj} \sim Binomial(r_{kj}; h_{kj})$.

Chen *et al.*[18] assume that for a logit link function,

$$\text{logit}(h_{kj}) = \log\left(\frac{h_{kj}}{1-h_{kj}}\right) = \boldsymbol{x}_k'\boldsymbol{\beta} + g_\gamma(t_j) + e_{kj}\,, \tag{60}$$

where g_γ is a known function with unknown parameters γ, possibly vector valued, $\boldsymbol{x}_k = (1, x_{1k}, x_{2k})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is a 3×1 vector of the regression coefficients with respect to covariates $\boldsymbol{x}_k$. In (60), e_{kj}'s are the random effects for the hazard rate of the fishes in the kth tank during the time interval I_j. Let $\boldsymbol{e}_k = (e_{k1}, e_{k2}, \ldots, e_{k,m})'$ for $k = 1, 2, \ldots, K$. We assume that these random effects, e_{kj}'s, are dependent on each other within the same tank (family) but independent between any two different tanks. Also let $\boldsymbol{A} = \text{diag}(a_1, a_2, \ldots, a_m)$, where $a_j = (t_j - t_{j-1})^{1/2}$ is the square root of the length of consecutive time intervals for $j = 1, 2, \ldots, m$. We build the dependence structure using a m-dimensional multivariate normal distribution within each tank, so that

$$\boldsymbol{e}_k \sim N_m(0, \sigma^2\boldsymbol{\Sigma})\,, \tag{61}$$

where $\boldsymbol{\Sigma}$ is a $m \times m$ matrix defined by

$$\boldsymbol{\Sigma} = \boldsymbol{A}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{m-1} \\ \rho & 1 & \rho & \cdots & \rho^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \cdots & 1 \end{pmatrix}\boldsymbol{A}\,. \tag{62}$$

The function $g_\gamma(t)$ in (60) is a function of time known up to the unknown parameter γ. The choice of g_γ depends on the entertained response-time model of these kinds of mortality data. The most simple form of $g_\gamma(t)$ is linear, i.e. $g_\gamma(t) = \gamma_1 t$, the constant term is absent to preserve the identifiability of the model. There are many other possible choices for the function $g_\gamma(t)$ such as quadratic, $g_\gamma(t) = \gamma_1 t + \gamma_2 t^2$, and even more complex one such as a spline function of known order with a known number of knots but, with unknown knot positions. In this example, we use Bayesian variable selection approach to explore a suitable form for $g_\gamma(t)$ from the data.

Here, we assume that $g_\gamma(t)$ takes the form

$$g_\gamma(t) = \gamma_1 g_1(t) + \gamma_2 g_2(t) + \cdots + \gamma_q g_q(t)\,, \tag{63}$$

where the $g_j(t)$ are the known functions of t and $q \geq 0$. For the notational convenience, we denote $g_\gamma(t) \equiv 0$ when $q = 0$. We further denote $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$.

Let D denote the complete data, that is, $D = ((\boldsymbol{d}_k, \boldsymbol{r}_k, \boldsymbol{x}_k, g_1(t_{k1}), \ldots, g_q(t_{km})), k = 1, \ldots, 48)$, where $\boldsymbol{d}_k = (d_{kj}, j = 1, 2, \ldots, m)$ and $\boldsymbol{r}_k = (r_{kj}, j = 1, 2, \ldots, m)$. Also let $\phi_m(\boldsymbol{e}_k|\mu, \sigma^2\Sigma)$ denote the m-dimensional normal density of the random effect $\boldsymbol{e}_k$ with mean μ and covariance matrix $\sigma^2\Sigma$, i.e.

$$\phi_m(\boldsymbol{e}_k|\mu, \sigma^2\Sigma) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{e}_k - \mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{e}_k - \mu)\right). \tag{64}$$

Then, it can be shown that the likelihood function is given by

$$L(\boldsymbol{\beta}, \gamma, \sigma^2, \rho|D) = \prod_{k=1}^{K} \left\{ \int \left[\prod_{j=1}^{m} (h_{kj})^{d_{kj}} (1 - h_{kj})^{r_{kj} - d_{kj}} \right] \times \phi_m(\boldsymbol{e}_k|0, \sigma^2\Sigma) d\boldsymbol{e}_k \right\}, \tag{65}$$

where h_{kj} and e_{kj} are given in (60). Notice that from (62), we have

$$|\boldsymbol{\Sigma}| = (1 - \rho^2)^{m-1} \times \prod_{j=1}^{m} a_j\,.$$

We use the NaSCN added fish tank data[77] to build our prior distributions. Let D_0 denote the complete NaSCN data, that is, $D_0 = ((\boldsymbol{d}_{0k}, \boldsymbol{r}_{0k}, \boldsymbol{x}_{0k}, \ g_1(t_{01}), \ldots, g_q(t_{0m_0})), \ k = 1, 2, \ldots, K_0 = 36)$, where $\boldsymbol{x}_{0k} = (1, x_{01k}, x_{02k})'$, $\boldsymbol{d}_{0k} = (d_{0kj}, j = 1, 2, \ldots, m_0 = 5)$ and $\boldsymbol{r}_{0k} = (r_{0kj}, j = 1, 2, \ldots, m_0 = 5)$. In the NaSCN data, d_{0kj} is the number of fishes dying and r_{0kj} is the number of fishes at risk in the kth tank during the time interval $I_{0j} = (t_{0,j-1}, t_{0j})$ for $j = 1, 2, \ldots, m_0 = 5$, where $t_{01} = 1$, $t_{02} = 6$, $t_{03} = 13$, $t_{04} = 20$, $t_{05} = 27$, and $m_0 = 5$ for $k = 1, 2, \ldots, K_0$ ($K_0 = 36$).

To determine the form of g_γ, let $\mathcal{M}$ denote the model space with each model containing $\boldsymbol{\beta}$ and a specific choice of covariates $g_l(t_j)$. The full model is defined here as the model containing all of the available covariates in the toxicity experiment. Also, let $\boldsymbol{\theta}^{(\mathcal{Q})} = (\boldsymbol{\beta}', \gamma_1, \gamma_2, \ldots, \gamma_q)'$ and let $\boldsymbol{\theta}^{(m)}$ denote

a $q_m \times 1$ vector of regression coefficients for model m with $\boldsymbol{\beta}$, and a specific choice of $q_m - 3$ covariates $g_l(t_j)$. We write $\boldsymbol{\theta}^{(\mathcal{Q})} = (\boldsymbol{\theta}^{(m)'}, \boldsymbol{\theta}^{(-m)'})'$, where $\boldsymbol{\theta}^{(-m)}$ is $\boldsymbol{\theta}^{(\mathcal{Q})}$ with $\boldsymbol{\theta}^{(m)}$ deleted.

Using the power prior given by (26), we construct the prior distribution for $(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho)$ under model m as

$$\begin{aligned}\pi(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho|m) &\propto \pi_0^*(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho|D_0, m) \\ &= \int_0^1 \prod_{k=1}^{K_0} \left\{ \prod_{j=1}^{m_0} \int \frac{\exp\{a_0 d_{0kj}((\boldsymbol{x}_{0kj}^{*(m)})'\boldsymbol{\theta}^{(m)} + e_{0kj})\}}{[1+\exp\{(\boldsymbol{x}_{0kj}^{*(m)})'\boldsymbol{\theta}^{(m)} + e_{0kj}\}]^{a_0 r_{0kj}}} \right. \\ &\quad \left. \times \phi_{m_0}(\boldsymbol{e}_{0k}|0, \sigma^2\Sigma_0) d\boldsymbol{e}_{0k} \right\} \\ &\quad \times \pi_1(\sigma^2)\ \pi_2(\rho)\ \pi_3(a_0) da_0\,, \end{aligned} \tag{66}$$

where $\boldsymbol{e}_{0k}$ is a $m_0 \times 1$ vector of random effects, and $\boldsymbol{x}_{0kj}^{*(m)}$ is a $q_m \times 1$ vector of covariates corresponding to $\boldsymbol{\theta}^{(m)}$. Note that under the full model, $\boldsymbol{x}_{0kj}^* = (1, x_{01k}, x_{02k}, g_1(t_{0kj}), \ldots, g_q(t_{0kj}))'$. In (66), we specify an inverse gamma prior for σ^2 given as $\pi_1(\sigma^2) \propto (\sigma^2)^{-(\delta_0+1)} \exp(-\sigma^{-2}\zeta_0)$, a scaled beta prior for ρ given as $\pi_2(\rho) \propto (1+\rho)^{\nu_0-1}(1-\rho)^{\psi_0-1}$, and independent beta priors for each a_0 given as $\pi_3(a_0) \propto a_0^{\alpha_0-1}(1-a_0)^{\lambda_0-1}$. Let $L(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho|D, m)$ denote the likelihood function given by (65) under model m. Then, the marginal likelihood given in (23) has the following expression:

$$p(D|m) = \int L(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho|D, m)\pi(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho|m)\, d\boldsymbol{\theta}^{(m)}\, d\sigma^2\, d\rho\,. \tag{67}$$

To completely determine the posterior probability of model m given by (22), we elicit the prior model probability $p(m)$ as:

$$p(m) = \frac{\int \pi_0^*(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho|D_0, m)\, d\boldsymbol{\theta}^{(m)}\, d\sigma^2\, d\rho}{\sum_{j=1}^{\mathcal{Q}} \int \pi_0^*(\boldsymbol{\theta}^{(j)}, \sigma^2, \rho|D_0, j)\, d\boldsymbol{\theta}^{(j)}\, d\sigma^2\, d\rho}\,. \tag{68}$$

The choice for $p(m)$ in (68) is a natural one since the numerator is just the normalizing constant of the joint prior of $(\boldsymbol{\theta}^{(m)}, \sigma^2, \rho)$ under model m. The prior model probabilities in (68) are based on coherent Bayesian updating. It can be shown that $p(m)$ in (68) corresponds to the posterior probability of model m based on the data D_0 using a uniform prior on the model space for the previous study, $p_0(m) = 2^{-q}$ for $m \in \mathcal{M}$ as $\alpha_0 \to \infty$. That is, $p(m) \propto p(m|D_0^{(m)})$, and thus $p(m)$ corresponds to the usual Bayesian update of $p_0(m)$ using D_0 as the data.

Next, we briefly discuss how to compute the posterior model probability $p(m|D)$. From (67) and (68), it can be seen that $p(m|D)$ is a function of the ratios of analytically intractable prior and posterior normalizing constants, which are expensive to compute. However, the following result can greatly ease such computational burden. Using (22), (67) and (68) along with the Savage–Dicky ratio,[99] it directly follows from Ibrahim *et al.*[61] that the posterior probability $p(m|D)$ in (22) of model m reduces to

$$p(m|D) = \frac{\pi(\boldsymbol{\theta}^{(-m)} = 0|D, \mathcal{Q})}{\sum_{j=1}^{\mathcal{Q}} \pi(\boldsymbol{\theta}^{(-j)} = 0|D, \mathcal{Q})}, \tag{69}$$

$m = 1, \ldots, \mathcal{Q}$, where $\pi(\boldsymbol{\theta}^{(-m)} = 0|D, \mathcal{Q})$ is the marginal posterior density of $\boldsymbol{\theta}^{(-m)}$ evaluated at $\boldsymbol{\theta}^{(-m)} = 0$ under the full model. In (69), for notational convenience we assume that $\pi(\boldsymbol{\theta}^{(-\mathcal{Q})} = 0|D, \mathcal{Q}) = 1$. Note that the joint posterior distribution of $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|D, \mathcal{Q}) \propto \int L(\boldsymbol{\theta}, \sigma^2, \rho|D)\pi(\boldsymbol{\theta}, \sigma^2, \rho|\mathcal{Q})d\sigma^2\, d\rho\,.$$

The result in (69) is attractive since it shows that the posterior model probability $p(m|D)$ is simply a function of the marginal posterior density functions of $\boldsymbol{\theta}^{(-m)}$ for the full model evaluated at $\boldsymbol{\theta}^{(-m)} = 0$. This formula does not algebraically depend on the prior model probability $p(m)$ since it cancels out in the derivation due to the structure of $p(m)$. This is an important feature since it allows us to compute the posterior model probabilities directly *without* numerically computing the prior model probabilities. This has a clear computational advantage and as a result, allows us to compute posterior model probabilities efficiently. Although the analytical evaluation of $\pi(\boldsymbol{\theta}^{(-m)} = 0|D, \mathcal{Q})$ does not appear possible due to the complexity of our model, it can be easily computed by using the IWMDE method discussed in Sec. 2.3.

We implement the above Bayesian variable subset selection to determine the form of g_γ. The terms $t, t^2, \log(t)$ were previously used in the toxicological mortality estimation literature by some authors[77,78] to model the effect of time. We do not know which form of $g(t)$ fits the data best before our analysis, and therefore, we use the formal Bayesian model selection procedure allowing the possibility of including all of these coefficients in the selected model. So, we consider that the full model for $g_\gamma(t_j)$ contains t_j, t_j^2, and $\ln t_j$, that is, $g_\gamma(t_j) = \gamma_1 t_j + \gamma_2 t_j^2 + \gamma_3 \ln t_j$. Thus, the model space $\mathcal{M}$ has a dimension of $\mathcal{Q} = 2^3 = 8$. We specify

Table 5. Posterior model probabilities for the fish tank data.

(μ_0, σ_0)	Model	$p(m\|D)$
(0.50, 0.109)	(x_1, x_2, T)	0.486
	(x_1, x_2, T^2)	0.352
	$(x_1, x_2, \ln T)$	0.153
(0.50, 0.050)	(x_1, x_2, T)	0.484
	(x_1, x_2, T^2)	0.351
	$(x_1, x_2, \ln T)$	0.153
(0.91, 0.027)	(x_1, x_2, T)	0.484
	(x_1, x_2, T^2)	0.350
	$(x_1, x_2, \ln T)$	0.153
(0.98, 0.006)	(x_1, x_2, T)	0.483
	(x_1, x_2, T^2)	0.350
	$(x_1, x_2, \ln T)$	0.154

noninformative priors for ρ and σ^2. Specifically, we take a uniform prior for ρ on $[-1, 1]$ (i.e. $\nu_0 = \psi_0 = 1$) and take an inverse gamma prior for σ^2 given as $\pi_1(\sigma^2) \propto (\sigma^2)^{-(\delta_0+1)} \exp(-\sigma^{-2}\zeta_0)$ with $\delta_0 = \zeta_0 = 0.005$. Then, we consider several choices of hyperparameters (α_0, λ_0) for a_0 to perform a small scale sensitivity study. Similar to Example 3, we let $\mu_{a_0} = \alpha_0/(\alpha_0 + \lambda_0)$, and $\sigma_{a_0} = (\mu_0(1 - \mu_0)(\alpha_0 + \lambda_0 + 1)^{-1})^{1/2}$. We generate 50,000 Gibbs iterations from the posterior distribution under the full model to obtain IWMDE. Table 5 gives results for the top three models based on several values of $(\mu_{a_0}, \sigma_{a_0})$. In Table 5, we let T, T^2, and $\ln T$ denote time, time square, and the natural logarithm of time. It can be seen that (i) for all choices of $(\mu_{a_0}, \sigma_{a_0})$, the order of the top three models does not change while model (x_1, x_2, T) is clearly the top model, and (ii) the posterior model probabilities for all top three models are almost the same for all choices of $(\mu_{a_0}, \sigma_{a_0})$ despite the one with a strong prior on a_0 (see $(\mu_{a_0}, \sigma_{a_0}) = (0.98, 0.006)$). Therefore, model choice is reasonably robust to the choice of $(\mu_{a_0}, \sigma_{a_0})$. In addition, the total sum of the posterior model probabilities for the top three models is close to 1, for example, this sum equals 0.988 for $(\mu_{a_0}, \sigma_{a_0}) = (0.50, 0.050)$. This result implies that for the purpose of posterior prediction, it suffices to use these three models to apply model averaging techniques[82] to incorporate model uncertainty in posterior densities for parameters. Also, from the principle of parsimony and from the result of Bayesian variable selection, the best model to use is $g(t) = \gamma_1 + \gamma_2 t$.

5.3. *Reference prior analysis to a statistical calibration model*

A statistical calibration problem (or, more precisely, an *absolute* statistical calibration problem), bears a resemblance to a regression problem. It is still assumed that the variables x and y are related through a function of specified form. However, in calibration, interest centers on the estimation of an unknown value of x, corresponding to an observed value of y. Inferences are based on two samples of data. At a first stage of data collection, n pairs (x_i, y_i) are observed, with the x values fixed at known levels. At a second stage, c replications of the response variable y are observed, corresponding to an unknown value of the regressor x_0; estimation of this regressor value is of primary interest. A review of statistical calibration is given by Osbourne.[29]

Noninformative priors for the linear calibration problem have been presented by several authors.[31,47,69,81] Also, Eno and Ye[32] studied the reference priors for the polynomial calibration models, as well as the probability matching prior for an extended calibration problem.[33]

5.3.1. *An example of polynomial calibration*

The data set was presented by Aitchison and Dunsmore.[1] These data resulted from an assay of an antibiotic, based on the "clearance circle" technique. The goal of such an experiment is to estimate the concentration of the active constituent in a particular test preparation of the antibiotic.

In a clearance circle assay, the regressor variable x is a precise measure of the concentration of active constituent in a preparation of antibiotic. This concentration is controlled in a laboratory experiment, where it is set at several different values by diluting a known full-strength antibiotic preparation to varying degrees. Each response y_i is obtained by placing a drop of antibiotic solution (of a specified volume) on a petri dish which is uniformly infected with bacteria. The actual response variable y_i is the measured diameter of the circle which has been disinfected by the antibiotic preparation after a specified period of time. It is expected that the diameter of this clearance circle depends on the concentration of the active constituent in an antibiotic solution. Based on the known dilutions of the standard preparation, this regression relationship can be estimated.

At the same time that the clearance circles corresponding to the known antibiotic concentrations are measured, clearance circles are also measured for the test preparation whose unknown antibiotic concentration is

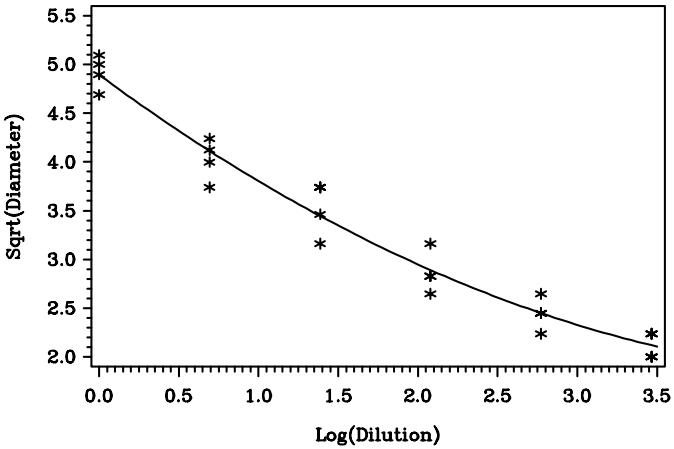


Fig. 4. Scatterplot of the transformed data in clearance circle assay.

of interest. What we want to do here is to use the reference prior Bayesian analysis to estimate this unknown antibiotic concentration. The data are plotted in Fig. 4.

In the plot (Fig. 4), the response is transformed via the square root transformation and the regressor is transformed via the log transformation. The plot really fits a quadratic model well.

5.3.2. *The model and the reference priors*

The polynomial calibration problem can be formally stated as follows. Data in the form of n pairs (x_i, y_i) are collected. In addition to these n data pairs, we observe c values of the response, $y_{n+1}, y_{n+2}, \ldots, y_{n+c}$, which correspond to a single unknown value of the regressor x_0. The response variable y is assumed to be related to the regressor x via a polynomial function of order p:

$$y_i = \alpha + \boldsymbol{\beta}_1 x_i + \boldsymbol{\beta}_2 x_i^2 + \cdots + \boldsymbol{\beta}_p x_i^p + \boldsymbol{\epsilon}_i\,, \quad \text{for} \quad i = 1, 2, \ldots, n + c\,.$$

For convenience, we have written x_i in place of x_0, for $i = n + 1, n + 2, \ldots, n + c$. We assume that the errors $\boldsymbol{\epsilon}_i$ are independent and identically distributed normal deviates, with mean 0 and standard deviation σ.

Of primary interest in this problem is the estimation of the unknown regressor value x_0. A feature that distinguishes the polynomial calibration problem from the linear calibration problem is that, since a polynomial

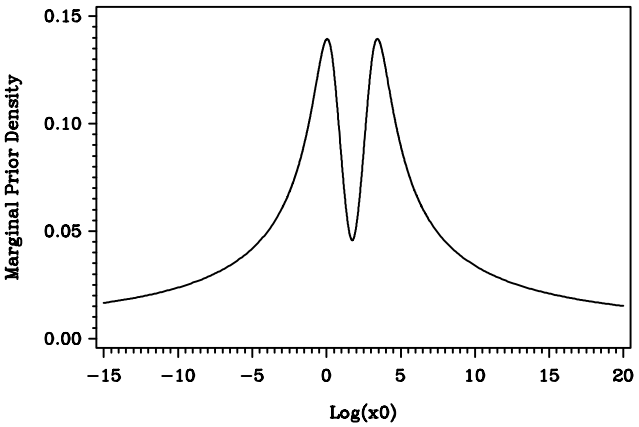


Fig. 5. Reference prior for the quadratic model, for the clearance circle assay example.

function need not be monotonic, more than one value of x_0 may give rise to a particular mean response $\bar{y}_0$. This issue in the context of the clearance circle assay described above has been discussed and addressed in Eno and Ye.[32]

Reference priors for the univariate polynomial calibration problem, as described above, are given as follows.

$$\pi_k(x_0, \alpha, \boldsymbol{\beta}, \sigma) \propto \sigma^{-k} \left(\frac{\zeta_0' \zeta_0}{1 + c\xi_0'(\mathbf{X}_{\alpha,1}' \mathbf{X}_{\alpha,1})^{-1} \xi_0} \right)^{\frac{1}{2}}$$

$$= \sigma^{-k} (\zeta_0' \zeta_0)^{\frac{1}{2}} \left(1 + \frac{cn}{n+c} (\mathbf{x}_0 - \bar{\mathbf{x}})' (\mathbf{X}_1' \mathbf{X}_1)^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}) \right)^{-\frac{1}{2}}, \tag{70}$$

where $\mathbf{X}_1$ is the $n \times p$ matrix whose ith row is $\mathbf{x}_i'$, the vector of regressor at ith observation, $\mathbf{X}_{\alpha,1}$ is the $n \times (s+1)$ matrix whose ith row is $(1, \mathbf{x}_i')$, $\boldsymbol{\zeta}_0' = (1, 2x_0, 3x_0^2, \ldots, px_0^{p-1})$ is a vector of derivative terms of $\mathbf{x}_0$, and k is the number of parameters in the group involving σ in the implementation of the reference prior algorithm.

Clearly the prior in (70) is improper. As we noted in Sec. 3.3, it is necessary to check the propriety of the posteriors once an improper prior is used. In Eno and Ye,[32] integrability of the reference prior in form (70) was proven. The prior function is shown in Fig. 5 for the clearance circle assay problem.

5.3.3. *Posterior results*

Applying the reference prior (70) to the clearance circle assay problem, we are ready to estimate the antibiotic concentration level for the observed responses. The marginal posterior distribution of the log(*concentration*) is shown in Fig. 6. It is quite clear to see the bi-modal properties of both the prior and the posterior since the model considered here is quadratic. Furthermore, it can easily seen that the small bump in the posterior distribution reflects to x_0 value way beyond the region of original regressor. It is conceivable that this local mode does not belong to this data set.

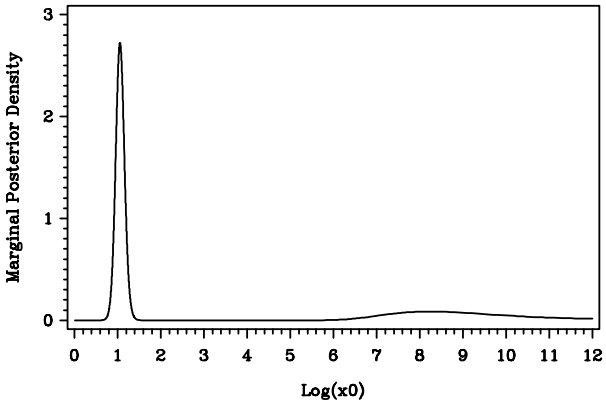


Fig. 6. Marginal reference posterior for $\log(x_0)$.

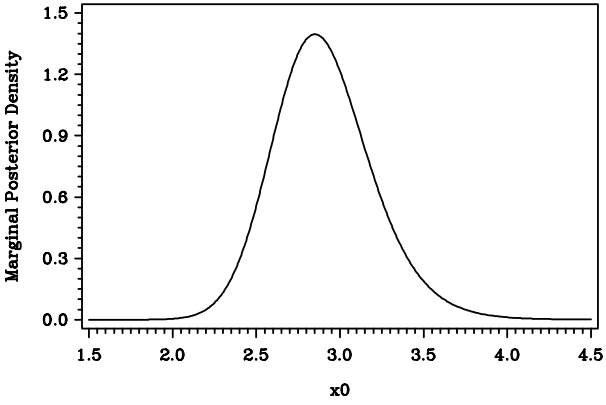


Fig. 7. Marginal reference posterior for x_0.

Since the model is not likely to be reliable outside the range of the controlled regressor values, we truncate the range of the posterior density and transform the logarithm back to the original scale. Figure 7 shows the marginal posterior distribution of x_0.

References

1. Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*, Cambridge University Press.
2. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Akademia Kiado, Budapest, 267–281.
3. Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **47**: 501–515.
4. Berger, J. O. (1999). Bayesian analysis today and tomorrow. *Technical Report 99-30*. Institute of Statistics and Decision Sciences, Duke University.
5. Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edn., Wiley, New York.
6. Berger, J. O. and Bernardo, J. (1989). Estimating a product of normal means: Bayesian analysis with reference priors. *Journal of the American Statistical Association* **84**: 200–207.
7. Berger, J. O. and Bernardo, J. (1992). On the development of the reference prior method. In *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, London, 35–60.
8. Berger, J. O. and Mallows, C. L. (1988). Discussion of Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**: 1033–1034.
9. Bernardo, J. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society* **B41**: 113–147.
10. Bernardo, J. M. and Smith, A. F. M. (1995). *Bayesian Theory*, John Wiley and Sons, New York.
11. Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series* **B55**: 25–37.
12. Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series* **A143**: 383–430.
13. Bratley, P., Fox, B. L. and Schrage, L. E. (1987). *A Guide to Simulation*, 2nd edn., Springer-Verlag, New York.
14. Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn, Chapman and Hall, New York.
15. Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**: 167–174.
16. Chen, M.-H. (1994). Importance weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association* **89**: 818–824.

17. Chen, M.-H. and Deely, J. J. (1996). Bayesian analysis for a constrained linear multiple regression problem for predicting the new crop of apples. *Journal of Agricultural, Biological and Environmental Statistics* **1**: 467–89.
18. Chen, M.-H., Dey, D. K. and Sinha, D. (2000). Bayesian analysis of multivariate mortality data with large families. *Applied Statistics* **49**: 129–144.
19. Chen, M.-H., Ibrahim, J. G. and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference* **41**: 121–137.
20. Chen, M.-H., Ibrahim, J. G., Shao, Q.-M. and Weiss, R. E. (2002). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, in print.
21. Chen, M.-H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**: 909–919.
22. Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior elicitation and Bayesian computation for logistic regression models with applications to variable selection. *Journal of the Royal Statistical Society, Series* **B61**: 223–242.
23. Chen, M.-H., Manatunga, A. K. and Williams, C. J. (1998). Heritability estimates from human twin data by incorporating historical prior information. *Biometrics* **54**: 1348–1362.
24. Chen, M.-H. and Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* **8**: 69–92.
25. Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics* **25**: 1563–1594.
26. Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag, New York.
27. Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician* **49**: 327–335.
28. Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society* **49**: 1–39.
29. Datta, G. S. and Ghosh, M. (1996). On the invariance of noninformative priors. *Technical Report 94-20*, Department of Statistics, University of Georgia.
30. Dey, D. K., Chen, M.-H. and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics* **53**: 1239–1252.
31. du Plessis, J. L., van der Merwe, A. J. and Groenewald, P. C. N. (1995). Reference priors for the multivariate calibration problem. *South African Statistical Journal* **29**: 155–168.
32. Eno, D. R. and Ye, K. (2000). Bayesian reference prior analysis for polynomial calibration models. *Test* **9**: 191–208.
33. Eno, D. R. and Ye, K. (2001). Probability matching priors for an extended statistical calibration model. *Canadian Journal of Statistics* **29**: 19–35.

34. Gamerman, D. (1997). *Markov Chain Monte Carlo*, Chapman and Hall, London.
35. Geisser, S. (1993). *Predictive Inference*: *An Introduction*, Chapman and Hall, London.
36. Geisser, S (1980). In discussion of G. E. P. Box. *Journal of the Royal Statistical Society, Series* **A143**: 416–417.
37. Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determinating using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics* **4**, eds. J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, 147–167.
38. Gelfand, A. E. and Mallick, B. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics* **51**: 843–852.
39. Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996). Efficient parametrisations for generalized linear mixed models (with discussion). In *Bayesian Statistics* **5**, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, 165–180.
40. Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika* **82**: 479–488.
41. Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409.
42. Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13**: 163–185.
43. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.
44. Geweke, J. (1989). Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica* **57**, 1317–1340.
45. George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**: 881–889.
46. Ghosh, J. K. and Mukerjee, R. (1992). Noninformative priors. In *Bayesian Statistic 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, London, 195–210.
47. Ghosh, M., Carlin, B. P. and Srivastava, M. S. (1998). Probability matching priors for linear calibration. *Test* **4**: 333–357.
48. Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.
49. Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**: 337–348.
50. Goel, P. K. (1988). Software for Bayesian analysis: Current status and additional needs. In *Bayesian Statistics 3*. Eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, Oxford University Press, 173–188.
51. Goldsman, D. and Meketon, M. S. (1986). A comparison of several variance estimators. *Technical Report J-85-12*. School of Industrial and Systems Engineering, Georgia Institute of Technology.

52. Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.
53. Grenander, U. (1983). Tutorial in pattern theorey. *Technical Report*. Division of Applied Mathematics, Brown University, Providence, RI.
54. Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*, Methuen, London.
55. Haseman, J. K., Huff, J. and Boorman, G. A. (1984). Use of historical control data in carcinogenicity studies in rodents. *Toxocologic Pathology* **12**: 126–135.
56. Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.
57. Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**: 382–417.
58. Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Sciences* **15**, 46–60.
59. Ibrahim, J. G. and Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazards models. *Sankhyā, Series* **B60**: 48–64.
60. Ibrahim, J. G., Chen, M.-H. and MacEachern, S. N. (1999). Bayesian variable selection for proportional hazards models. *The Canadian Journal of Statistics* **27**: 701–717.
61. Ibrahim, J. G., Chen, M.-H. and Ryan, L.-M. (2000). Bayesian variable selection for time series count data. *Statistica Sinica* **10**: 971–987.
62. Ibrahim, J. G., Ryan, L.-M. and Chen, M.-H. (1998). Use of historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association* **93**: 1282–1293.
63. Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association* **89**: 309–319.
64. Jeffreys, H. (1961). *Theory of Probability*, Oxford University Press.
65. Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of American Statistical Association* **78**: 137–144.
66. Kass, R. E. and Wasserman, L. A. (1996). Formal rules for selecting prior distributions. *Journal of the American Statistical Association* **91**: 1343–1370.
67. Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M. and Blum, R. H. (1999). The role of high- and low-dose interferon Alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S9111/C9190. *Journal of Clinical Oncology* **18**: 2444–2458.
68. Kirkwood, J. M., Strawderman, M. H., Ernstoff, M. S., Smith, T. J., Borden, E. C. and Blum, R. H. (1996). Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: The Eastern Cooperative Oncology Group trial EST 1684. *Journal of Clinical Oncology* **14**: 7–17.
69. Kubokawa, T. and Robert, C. P. (1994). New perspectives on linear calibration. *Journal of Multivariate Analysis* **51**: 178–200.

70. Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series* **B57**: 247–262.
71. Law, A. M. and Kelton, W. D. (1991). *Simulation Modeling and Analysis*. 2nd edn., McGraw-Hill, New York.
72. Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**: 1142–1160.
73. Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**: 831–860.
74. Merigan, T. C., Amato, D. A., Balsley, J., Power, M., Price, W. A., Benoit, S., Perez-Michael, A., Brownstein, A., Kramer, A. S., Brettler, D., Aledort, L., Ragni, M. V., Andes, A. W., Gill, J. C., Goldsmith, J., Stabler, S., Sanders, N., Gjerset, G., Lusher, J. and the NHF-ACTG036 Study Group (1991). Placebo-controlled trial to evaluate zidovudine in treatment of human immunodeficiency virus infection in asymptomatic patients with hemophilia. *Blood* **78**: 900–906.
75. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1092.
76. Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* **83**: 1023–1036.
77. O'Hara Hines, R. J. (1989). Some methods for the analysis of toxicological mortality data grouped over time. Unpublished PhD Thesis, Department of Statistics and Actuarial Science, University of Waterloo, Canada.
78. O'Hara Hines, R. J. and Lawless, J. F. (1993), Modelling overdispersion in toxicological mortality data grouped over time. *Biometrics* **49**: 107–122.
79. Osbourne, C. (1991). Statistical calibration: A review. *Journal of Statistical Review* **59**: 309–336.
80. Penrose, K., Nelson, A. and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (Abstract). *Medicine and Science in Sports and Exercise* **17**: 189.
81. Philippe, A. and Robert, C. P. (1998). A note on the confidence properties of reference priors for the calibration model. *Test* **7**: 147–160.
82. Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**: 251–266.
83. Raftery, A. E., Madigan, D. M. and Hoeting, J. (1997). Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association* **92**: 179–191.
84. Ripley, B. D. (1987). *Stochastic Simulation*, Wiley, New York.
85. Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New York.
86. Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability* **7**: 110–120.

87. Roberts, G. O. and Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society, Series* **B56**: 377–384.
88. Schervish, M. J. and Carlin, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1**: 111–127.
89. Schmeiser, B. W., Avramidis, A. N. and Hashem, S. (1990). Overlapping batch statistics. In *Proceedings of the 1990 Winter Simulation Conference*, 395–398.
90. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**: 461–464.
91. Sinha, D. and Dey, D. K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**: 1195–1212.
92. Song, W.-M. T. and Schmeiser, B. W. (1995). Optimal mean-squared-error batch sizes. *Management Science* **41**: 110–123.
93. Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*, ed. R. Zielinski, PWN-Polish Scientific Publishers, Warsaw, 485–514.
94. Sun, D. and Ye, K. (1995). Reference prior Bayesian analysis for normal mean products. *Journal of American Statistical Association* **90**: 589–597.
95. Tanner, M. A. (1996). *Tools for Statistical Inference*, 3rd edn., Springer-Verlag, New York.
96. Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–549.
97. Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**: 604–608.
98. Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *The Annals of Statistics* **22**: 1701–1762.
99. Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association* **90**: 614–618.
100. Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour, H. H., Reichman, R. C., Bartlett, J. A., Hirsch, M. S., Murphy, R. L., Hardy, D., Soeiro, R., Fischl, M. A., Bartlett, J. G., Merigan, T. C., Hylsop, N. E., Richman, D. D., Valentine, F. T., Corey, L. and the AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine* **322**: 941–949.
101. Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society* **25**: 318–329.
102. Wolpert, R. L. (1991). Monte Carlo importance sampling in Bayesian statistics. In *Statistical Multiple Integration*, eds. N. Flournoy and R. Tsutakawa, *Contemporary Mathematics* **116**: 101–115.

103. Yakovlev, A. Y., Asselain, B., Bardou V. J., Fourquet, A., Hoang, T. Rochefediere, A. and Tsodikov, A. D. (1993). A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer. In *Biometrie et Analyse de Donnees Spatio-Temporelles*, # 12, eds. B. Asselain, M. Boniface, C. Duby, C. Lopez, J. P. Masson and J. Tranchefort, Rennes, France, 66–82.
104. Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, New Jersey.
105. Ye, K. (1993). Reference priors when the stopping rule depends on the parameter of interest. *Journal of American Statistical Association* **88**: 360–363.
106. Ye, K. (1994). Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model. *Journal of Statistical Planning and Inference* **41**: 267–280.
107. Ye, K. and Berger, J. O. (1989). Noninformative priors for inferences in exponential regression models. *Biometrika* **78**: 645–656.

About the Authors

Ming-Hui Chen is currently an associate professor at Department of Statistics, University of Connecticut. He has been an associate professor at Department of Mathematical Sciences, Worcester Polytechnic Institute before he joined the current position. He obtained BS in Mathematics from Hangzhou University, MS in Applied Probability from Shanghai Jiao Tong University, and MS in Applied Statistics and PhD in Statistics from Purdue University. He has coauthored the books: *Applied Statistics for Engineers* (Prentice-Hall, INC., 1999), *Monte Carlo Methods in Bayesian Computation* (Springer-Verlag, 2000) and *Bayesian Survival Analysis* (Springer-Verlag, 2001). His current research interests include Bayesian statistical methodology, Bayesian computation, categorical data analysis, Monte Carlo methodology, prior elicitation, missing data analysis, statistical analysis for prostate cancer study, variable selection, and survival models.

Keying Ye is currently an associate professor at Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. Dr. Ye received his BS in Mathematics, Fudan University, MS in Mathematics, Institute of Applied Mathematics, Academia Sinica, in the People's Republic of China, and PhD in Statistics from Purdue University, USA. His research interests are Bayesian methodologies in applications to various fields such as environmental and ecological systems, gene expressions,

industrial and quality improvement. Recent works involve Bayesian hierarchical modeling, prediction incorporating model uncertainty, experimental designs, multivariate statistical analysis in environmental data, spatial effects in data analysis and response surface experimental designs with model uncertainty.

This page intentionally left blank

CHAPTER 26

STOCHASTIC PROCESSES AND THEIR APPLICATIONS IN MEDICAL SCIENCE

JIQIAN FANG* and LI CAIXIA

Department of Medical Statistics, School of Public Health, Sun Yat-sen University, 74 Zhongshan Road II, Guangzhou 510080, PR China

*Tel: 86-20-87330671; *fangjq@gzsums.edu.cn*

A stochastic process $X = \{X(t), t \in T\}$ is a t-indexed collection of random variables. i.e. for any $t \in T$, $X(t)$ is a random variable and t is a parameter. We often interpret t as time. If the set T is a countable set, we call the process a discrete-time stochastic process, usually denoted by $\{X_n, n = 1, 2, \ldots\}$. And if T is continuum, we call it a continuous-time stochastic process, usually denoted by $\{X(t), t \geq 0\}$. $X(t)$ is called the state of process at time t. The collection of possible values of $X(t)$ is called state space.

Stochastic processes have ever applied in many fields. Now we introduce some important stochastic processes and their applications in medical science.

1. Markov Chains

1.1. *Discrete-time Markov chains*

Suppose that we roll a six-sided dice. The probability of rolling 1 is denoted $p_1 (0 < p_1 < 1)$. Now consider a sequence of consecutive rolls. Suppose that they are all independent. If we let X_n denote the accumulative number of rolling 1 after n consecutive rolls, it is easy to see that the variables $\{X_n, n = 1, 2, \ldots\}$ are not independent. However, If the value of X_n is given, for example $X_n = i$, we can see X_{n+1} takes either the value i (with probability $1-p_1$) or the value $i+1$ (with probability p_1). In other words, the process $\{X_n, n = 1, 2, \ldots\}$ shows the property that conditional distribution of the future state X_{n+1}, given the present state X_n, depends only on the present state and is independent of the past states of $X_1, X_2, \ldots, X_{n-1}$.

This property is called Markovian property. Markov chains are discrete-state stochastic processes with Markovian property.

Definition.[2,19,21] Consider a stochastic process $\{X_n, n = 0, 1, 2, \ldots\}$ that takes on a finite or countable values. $\{X_n, n = 0, 1, 2, \ldots\}$ is said to be Markov chain if

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\}. \quad (1)$$

For all states $i_0, i_1, \ldots, i_{n-1}, i, j$ and all $n \geq 0$.

$P\{X_{n+1} = j | X_n = i\}$ in Eq. (1) is associated with a transition taking place in one step, so it is called (one-step) transition probability and is denoted as $p_{i,j}(n, n+1)$. A Markov chain is said to be homogeneous if $p_{i,j}(n, n+1)$ is independent of n. Then $p_{i,j}(n, n+1)$ can be denoted as p_{ij}. Let P denote the matrix of transition probability p_{ij}, so that

$$P \widehat{=} (p_{ij}) = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

It is obvious that $\sum_j p_{ij} = 1$ for any i.

1.1.1. *A simple example — simple random walk*

A particle make a random walk on the integer points. Wherever it is, it will either goes up one step (with probability p) or down one step (with probability $1 - p$). Let X_n denote the site of the particle after n steps.

It is easy to see the simple random walk is a Markov chain. Its transition probability

$$p_{ij} = \begin{cases} p & j = i + 1 \\ 1 - p & j = i - 1 \\ 0 & \text{else} \end{cases} \quad i = 0, \pm 1, \ldots$$

As similar as one-step transition probability p_{ij}. m-step ($m > 1$) transition probability is

$$p_{ij}^{(m)} \widehat{=} P\{X_{n+m} = j | X_n = i\}.$$

Let $P^{(m)}$ denote the matrix of $p_{ij}^{(m)}$, i.e. $\boldsymbol{P}^{(m)} = (\boldsymbol{p}_{ij}^{(m)})$.

1.1.2. *Chapman–Kolmogorov equationm (C–K equation)*

For any m, n,

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)} , \tag{2}$$

or in terms of the transition probability matrices:

$$P^{(m+n)} = P^{(m)} \cdot P^{(n)} , \tag{3}$$

especially,

$$P^{(n)} = [P^{(1)}]^n .$$

So C–K equation can be used to derive higher order transition probability from one-step transition probability. Chiang[11] pointed out that

$$p_{ij}^{(n)} = \sum_{l=1}^{s} \frac{A'_{ij}(\lambda_l)\lambda_l^n}{\prod_{\substack{m=1 \\ m\neq l}}^{s}(\lambda_l - \lambda_m)} , \quad i,j = 1,2,\ldots,s , \tag{4}$$

when $s \times s$ transition probability matrix P has s distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$, where the matrix $A'(\lambda_l) = (\lambda_l I - P)'$.

The probability distribution of X_n is

$$p_j^{(n)} \mathrel{\widehat{=}} P\{X_n = j\} = \sum_i P\{X_0 = i\} P\{X_n = j | X_0 = i\}$$

$$= \sum_i P\{X_0 = i\} p_{ij}^{(n)} .$$

And in terms of the matrices,

$$P_{X_n} = (p_j^{(n)}) = P_{X_0} \cdot P^{(n)} . \tag{5}$$

So, the probability distribution of a Markov chain can be derived from transition probability matrix and initial distribution.

1.1.3. *Example*: *Hardy–Weinberg law of equilibrium in genetics*[2,11]

Consider a biological population. Each individual in the population is assumed to have a genotype AA or Aa or aa, where A and a are two alleles. Suppose that the initial genotype frequency composition (AA, Aa, aa) equals to $(d, 2h, r)$, where $d + 2h + r = 1$. Then the gene frequencies of A and a are p and q, where $p = d + h$, $q = r + h$ and $p + q = 1$. We can use Markov chain to describe the heredity process. We number the three genotypes AA, Aa and aa by 1, 2, 3 and denote by p_{ij} the probability that

an offspring has genotype j given that a specified parent has genotype i. For example,

$$\begin{aligned} p_{12} &= P\{\text{a child has genotype } Aa | \text{his mother has genotype } AA\} \\ &= P\{\text{his father has gene } a | \text{his mother has genotype } AA\}. \end{aligned}$$

Under random mating assumption,

$$\begin{aligned} &P\{\text{his father has gene } a | \text{his mother has genotype } AA\} \\ &\quad = P\{\text{his father has gene } a\}. \end{aligned}$$

So $p_{12} = q$. Similar computations yield the other transition probabilities. The one-step transition probability matrix is

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{pmatrix}.$$

Let $p_i^{(k)}$ denote the probability that the kth generation has genotype i. The initial genotype distribution of the 0th generation

$$(p_1^{(0)}, p_2^{(0)}, p_3^{(0)}) = (d, 2h, r).$$

And then the genotype distribution of the first generation

$$\begin{aligned} (p_1^{(1)}, p_2^{(1)}, p_3^{(1)}) &= (p_1^{(0)}, p_2^{(0)}, p_3^{(0)})P \\ &= (d, 2h, r) \begin{pmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{pmatrix} \\ &= ((d+h)p, dq + h + rp, (h+r)q) \\ &= (p^2, 2pq, q^2). \end{aligned}$$

The genotype distribution of the second generation

$$\begin{aligned} (p_1^{(2)}, p_2^{(2)}, p_3^{(2)}) &= (p_1^{(1)}, p_2^{(1)}, p_3^{(1)})P \\ &= (p^2, 2pq, q^2) \begin{pmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{pmatrix} \end{aligned}$$

$$= (p^2(p+q), pq(p+q+1), q^2(p+q))$$
$$= (p^2, 2pq, q^2)$$

and has the same distribution as that of the first generation. Similar computations show the distributions of the 3rd, 4th, ... are all same and still are $(p^2, 2pq, q^2)$. This is Hardy–Weinberg law of equilibrium. That is, whatever the parent genotype frequency compositions $(d, 2h, r)$ may be, under random mating assumption, the first generation progenies will have the genotype composition $(p^2, 2pq, q^2)$ and this composition will remain in equilibrium forever.

1.2. *Stationary distribution and limiting distribution*

State j is said to be accessible[2] from state i if for some $n \geq 0$, $p_{ij}^{(n)} > 0$. Two states accessible to each other are said to communicate.[2] We say that the Markov chain is irreducible[2] if all states communicate with each other.

State i is said to have period d if $p_{ii}^{(n)} = 0$ whenever n is not divisible by d and d is the greatest integer with the property. A state with period 1 is called aperiodic.[20]

A probability distribution $\{\pi_j\}$ related to a Markov chain is called stationary if it satisfied the relation

$$\pi_j = \sum_i \pi_i p_{ij} \,. \tag{6}$$

If the initial distribution $\{P(X_0 = i)\}$ is stationary distribution, then

$$P\{X_1 = j\} = \sum_i P\{X_0 = i\} P\{X_1 = j | X_0 = i\} = \sum_i \pi_i p_{ij} = \pi_j$$

and by induction, the probability $P\{X_n = j\} = \pi_j$. Therefore the distribution of X_n is independent of n (time) and the corresponding process is in a statistical equilibrium. In the example of Hardy–Weinberg law of equilibrium, the stationary distribution of the Markov chain is $(p^2, 2pq, q^2)$.

A Markov chain is called finite[2] if the chain has finite states. There must exist unique stationary distribution $\{\pi_j\}$ in a finite and irreducible Markov chain,[2] the $\pi_j, j \geq 0$, are the unique solution of Eq. (6) and $\sum_j \pi_j = 1$.

If there is a distribution $\{\pi_j\}$ such that

$$\lim_{n\to\infty} \sum_i \pi_j p_{ij}^{(n)} = \pi_j \quad \text{for any } i, j \tag{7}$$

$\{\pi_j\}$ is called long-run distribution (or limiting distribution).

If a Markov chain has long-run distribution $\{\pi_j\}$, the chain has asymptotic distribution $\{\pi_j\}$ no matter what the initial distribution is. A long-run distribution must be a stationary distribution. If a finite Markov chain is aperiodic, its stationary distribution is long-run distribution.[2,18,20]

1.2.1. *Example: Social status change*[2,18]

In sociology, there is a question about how much effect be made on son's social status by father's social status. We take one's occupation indicates his social status. Now, consider the conditional probability distribution for son's occupation. In a research report about social status change, probability distribution is provided as following:

Table 1.

Father's occupation	Son's occupation		
	good	median	bad
good	0.448	0.484	0.068
median	0.054	0.699	0.247
bad	0.011	0.503	0.486

We consider social status's change as transition between states. If Markovian property is satisfied in the states, we can use a finite (three states) Markov chain to describe the social status's change. This chain is irreducible and aperiodic, and there must be long-run distribution (π_1, π_2, π_3) satisfied

$$(\pi_1, \pi_2, \pi_3)P = \pi_1, \pi_2, \pi_3 \,.$$

So we can get

$$\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \begin{pmatrix} 0.067 \\ 0.624 \\ 0.309 \end{pmatrix} .$$

We can say, if social status's change is a Markov chain with above transition probability, the social status take asymptotically the proportions: 6.7 for good, 62.4 for median, 30.9 for bad.

1.3. *Continuous-time Markov chain*

Definition. For all states i, j, x_u and all $s, t \geq 0$, if the equation

$$P\{X_{t+s} = j | X_s = i, X_u = x_u, 0 \leq u < s\} = P\{X_{t+s} = j | X_s = i\} \qquad (8)$$

is satisfied, the discrete process $\{X_t, t \geq 0\}$ is called continuous-time Markov chain. $P\{X_{t+s} = j | X_s = i\}$ in the equation is also called transition probability, denoted by $p_{i,j}(s, s+t)$. A continuous-time Markov chain is said to be homogeneous if $p_{i,j}(s, s+t)$ is independent of (denoted by $p_{ij}(t)$ here).

If the state is i at time t, the chain transform into state j with the probability

$$p_{ij}(\Delta t) = P\{X(t+\Delta t) = j | X(t) = i\}$$

after Δt.

Let

$$\delta_{ij} \begin{cases} 0 & j \neq i \\ 1 & j = i \end{cases} .$$

$q_{ij} \widehat{=} \lim_{\Delta t \to 0+} \frac{p_{ij}(\Delta t) - \delta_{ij}}{\Delta t}$ is said to be transition intensity.[20] The matrix $Q \widehat{=} (q_{ij})$ is called transition intensity matrix.

$$q_{ij} dt = P\{X(t+dt) = j | X(t) = i\}, \quad j \neq i$$

$$q_{ii} dt = P\{X(t+dt) = i | X(t) = i\} - 1 = -P\{X(t+dt) \neq i | X(t) = i\}.$$

So

$$\sum_j q_{ij} = 0 .$$

The continuous-time Markov chain with intensity matrix $Q = (q_{ij})$.

(1) The sojourn time of state i have exponential distribution with the mean $-q_{ii}$.
(2) The chain step into state $j (j \neq i)$ with the probability $p_{ij} = -\frac{q_{ij}}{q_{ii}}$ after leaving state i.

The transition probability satisfies Chapman–Kolmogorov equation

$$p_{ij}(t+s) = \sum_k p_{ik}(t) p_{kj}(s)$$

and two Chapman–Kolmogorov differential equations which are Chapman–Kolmogorov forward equation

$$p'_{ij}(t) = \sum_k p_{ik}(t) q_{kj} , \quad \text{i.e. } P'(t) = P(t)Q \tag{9}$$

and Chapman–Kolmogorov backward equation

$$p'_{ij}(t) = \sum_k q_{ik} p_{kj}(t) , \quad \text{i.e. } P'(t) = QP(t) \tag{10}$$

for all i, j and $t \geq 0$.

We can obtain transition probabilities from the differential equations.

Example. Now consider a two-state continuous-time Markov chain. The sojourn time of state 0 has exponential distribution with rate λ and the sojourn time of state 1 has exponential distribution with rate u. Therefore the intensity matrix is

$$\begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

From forward equation

$$\begin{aligned} p'_{00}(t) &= -\lambda p_{00}(t) + \mu p_{01}(t) = -\lambda p_{00}(t) + \mu(1 - p_{00}(t)) \\ &= -(\lambda + \mu)p_{00}(t) + \mu \end{aligned}$$

we have

$$p_{00}(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} \exp(-(\lambda + \mu)t)\,.$$

Similarly,

$$p_{11}(t) = \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} \exp(-(\lambda + \mu)t)\,.$$

Hence, transition probability matrix

$$P(t) = \begin{pmatrix} \dfrac{\mu}{\lambda + \mu} + \dfrac{\lambda}{\lambda + \mu} \exp(-(\lambda + \mu)t) & \dfrac{\lambda}{\lambda + \mu} - \dfrac{\lambda}{\lambda + \mu} \exp(-(\lambda + \mu)t) \\ \dfrac{\mu}{\lambda + \mu} - \dfrac{\mu}{\lambda + \mu} \exp(-(\lambda + \mu)t) & \dfrac{\lambda}{\lambda + \mu} + \dfrac{\mu}{\lambda + \mu} \exp(-(\lambda + \mu)t) \end{pmatrix}.$$

Generally, for s-state chain, when the intensity matrix has single eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$, Chiang[11] presented the solution of Chapman–Kolmogorov differential equations

$$P_{ij}(t) = \sum_{l=1}^{s} \frac{A'_{ij}(\lambda_l) \exp(\lambda_l t)}{\prod_{\substack{m=1 \\ m \neq l}}^{s} (\lambda_l - \lambda_m)}, \quad i, j = 1, 2, \ldots, s \tag{11}$$

where $A'(\lambda_l) = (\lambda_l I - Q)'$.

2. Applications of Markov Chains

Markov chain is usually used to describe a systems with Markovian property. For example, We usually divide a certain disease into several states in medical science. Under Markovian property assumption, we can get the transition information among states.

Table 2. Frequence of tree species in 1983 and 1988.

1983	1988						Total
	1	2	3	4	5	6	
1	83	8	0	0	1	12	104
2	8	81	0	0	0	18	107
3	12	2	116	2	1	6	139
4	2	0	6	75	0	1	84
5	7	0	3	0	35	4	49
6	39	27	44	5	4	362	481
Total	151	118	169	82	41	403	964

2.1. *Example* 1: *Predict the structure of future system*

In the paper of Chen Jianzhong *et al.*[1] analysis the data investigated in 964 areas of NanPing in 1983 and 1988, a Markov model is built to predict forest resources with tree species. The tree species include cunninghamia lanceolata (state 1), pinus massoniana (state 2), broad-leaved trees (state 3), phyllostachys pubescens (state 4), economic trees (state 5) and others (state 6). The data from two investigations see Table 2.

One step (five years) transition probability matrix P is estimated as follows:

$$P = \begin{pmatrix} 79.81 & 7.69 & 0 & 0 & 0.96 & 11.54 \\ 7.48 & 75.70 & 0 & 0 & 0 & 16.82 \\ 8.63 & 1.44 & 83.45 & 1.44 & 0.72 & 4.32 \\ 2.38 & 0 & 7.14 & 89.29 & 0 & 1.19 \\ 14.29 & 0 & 6.12 & 0 & 71.43 & 8.16 \\ 8.11 & 5.61 & 9.15 & 1.04 & 0.83 & 75.26 \end{pmatrix}.$$

The chain with transition matrix P has stationary distribution. Based on the initial distribution in 1988 and transition matrix P, the distributions of trees from 1993 to 2023 are computed in Table 3.

The stationary distribution shows the structure at present is unreasonable and needs to be adjusted in accordance to future structure. The transition probability matrix after adjustment becomes

$$P = \begin{pmatrix} 79.81 & 7.69 & 0 & 0 & 0.96 & 11.54 \\ 7.48 & 75.70 & 0 & 0 & 0 & 16.82 \\ 2.88 & 1.44 & 89.21 & 1.44 & 0.72 & 4.32 \\ 2.38 & 0 & 7.14 & 89.29 & 0 & 1.19 \\ 14.29 & 0 & 6.12 & 0 & 71.43 & 8.16 \\ 3.95 & 5.61 & 7.28 & 3.12 & 4.78 & 75.26 \end{pmatrix}.$$

Table 3. Prediction for occupied % of tree species in different years.

Year	1	2	3	4	5	6
1983	0.1079	0.1111	0.1442	0.0871	0.0508	0.4990
1988	0.1566	0.1224	0.1753	0.0851	0.0425	0.4180
1993	0.1913	0.1307	0.1932	0.0828	0.0366	0.3653
1998	0.2159	0.1369	0.2028	0.0805	0.0324	0.3318
2003	0.2335	0.1418	0.2073	0.0783	0.0295	0.3097
2008	0.2460	0.1457	0.2087	0.0761	0.0273	0.2961
2013	0.2550	0.1488	0.2084	0.0740	0.0259	0.2879
2018	0.2614	0.1514	0.2071	0.0721	0.0248	0.2831
2023	0.2661	0.1536	0.2054	0.0703	0.0241	0.2805
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Stationary	0.2805	0.1663	0.1901	0.0534	0.0226	0.2871

Table 4. Prediction for occupied % of tree species in different years after adjustment.

Year	1	2	3	4	5	6
1983	0.1079	0.1111	0.1442	0.0871	0.0508	0.4990
1988	0.1276	0.1224	0.1743	0.0954	0.0622	0.4180
1993	0.1437	0.1285	0.1965	0.1008	0.0669	0.3637
1998	0.1562	0.1315	0.2131	0.1041	0.0680	0.3270
2003	0.1658	0.1330	0.2255	0.1062	0.0672	0.3023
2008	0.1728	0.1337	0.2348	0.1075	0.0657	0.2855
2013	0.1779	0.1339	0.2420	0.1083	0.0639	0.2741
2018	0.1815	0.1339	0.2474	0.1087	0.0622	0.2663
2023	0.1840	0.1338	0.2572	0.1089	0.0607	0.2609
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Stationary	0.1872	0.1324	0.2698	0.1084	0.0545	0.2477

The distributions of trees from 1993 to 2023 are computed similarly in Table 4.

2.2. *Example* 2: *Decision analysis and cost-effectiveness analysis*

Helicobacter pylori (HP) infection is a factor on tummy cancer. A markov model is provided for cost analysis in Wang Qian *et al.*[9] Four states in the chain are without HP infection (state 1), HP infection (state 2), cancer (state 3) and death (state 4). The transition probabilities are given in Fig. 1.

Suppose that is 50% individuals in the population is HP infectious and the cancer incidence is $27/10^5$. For cost analysis, assume that the heath

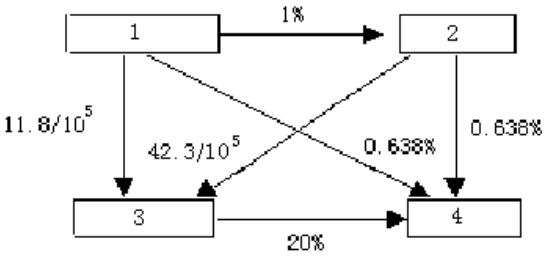


Fig. 1. Transition probabilities.

Table 5. The cost in the population (10,000 individuals) without screening.

T	States				Communication value					
	1	2	3	4	S	Q	Cost	S	Q	Cost
0	0.5	0.5	0	0	0	0	0	0	0	0
1	0.49175	0.50160	0.00027	0.00638	9936	9684	27050	9936	9684	27050
2	0.48364	0.50310	0.00049	0.01277	9872	9617	48660	19808	19301	75710
3	0.47566	0.50452	0.00066	0.01916	9808	9551	65916	29617	28852	141627
4	0.46781	0.50584	0.00080	0.02555	9745	9486	79687	39361	38338	221314
5	0.46009	0.50708	0.00091	0.03192	9681	9421	90667	49042	47759	311981
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	0.31382	0.51532	0.00129	0.16957	8304	8038	129002	254820	247422	3181863
29	0.30864	0.51496	0.00129	0.17512	8249	7982	128703	263069	255404	3310566
30	0.30355	0.51454	0.00128	0.18063	8194	7927	128387	271263	263332	3438953

T = time S = survival time Q = quality survival time

values of four states are 1, 0.95, 0.3, 0, respectively. The cost of a patient with cancer is $\$10^4$ per year. The transitions and costs for the population are given in Table 5.

Suppose that the sensitivity and specificity are both 90% in the screening test and the cost of the screening test is \$25 per individual. The cure rate of HP infection is 80% and the cost for cure is \$300. Similarly, the transitions and costs for the population with screening can be computed. They are given in Table 6.

There is a contrast between them and see Table 7.

2.3. *Example* 3: *Using the transition dependent on covariates to analysis the factors for illness*

In the paper of Fang *et al.*,[3] two non-homogeneous Markov chains were used to study a two-stage model with time-dependent covariates for latent

Table 6. The cost in the population (10,000 individuals) with screening.

T	States				Communication value					
	1	2	3	4	S	Q	Cost	S	Q	Cost
0	0.86	0.14	0	0						
1	0.84581	0.14765	0.00016	0.00638	9936	9861	16070	9936	9816	1766070
2	0.83186	0.15510	0.00029	0.01275	9873	9793	29082	19809	19654	1795152
3	0.81813	0.16236	0.00040	0.01911	9809	9725	39642	29618	29379	1834794
4	0.80464	0.16944	0.00048	0.02544	9746	9658	48236	39363	39037	1883030
5	0.79136	0.17634	0.00055	0.03175	9682	9590	55251	49046	48627	1938281
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	0.53977	0.29162	0.00092	0.16769	8323	8171	91973	255075	251730	3848064
29	0.53086	0.29504	0.00092	0.17318	8268	8114	92283	263343	259844	3940347
30	0.52210	0.29834	0.00093	0.17863	8214	8085	92571	271557	267902	4032918

Table 7. The costs between two populations.

	Screening	Non-Screening	Difference
S (year)	271557	271263	294
Q (year)	267902	263332	4570
Cancer frequency	55	82	−27
Summary cost ($)	4032918	3438953	593965
Screening cost ($)	1750000	—	1750000
Cost for cancer ($)	2282918	3438953	−1156035

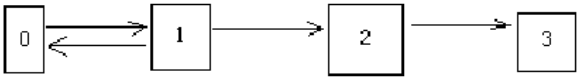


Fig. 2. The transitions between four states.

period of cancer. There are four states in the chains, which are inapparent illness (state 0), soakage stage (state 1), non-soakage stage (state 2) and observable clinic state (state 3). The transitions between the states are in Fig. 2.

2.3.1. *Model* 1. *A non-homogeneous discrete-time Markov model*

Let

$$p_{ij}(t) = P\{X(t+1) = j | X(t) = i\} \text{ and } Z(t) = (Z_1(t), Z_2(t), \ldots, Z_p(t))',$$

where $Z_1(t), Z_2(t), \ldots, Z_p(t)$ are p covariates. The transition probability matrix is

$$P(t) = \begin{pmatrix} p_{00}(t) & p_{01}(t) & 0 & 0 \\ p_{10}(t) & p_{11}(t) & p_{12}(t) & 0 \\ 0 & 0 & p_{22}(t) & p_{23}(t) \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where

$$p_{01}(t) = a_{01} \cdot \theta(t)\,, \quad p_{00}(t) = 1 - p_{01}(t)\,,$$
$$p_{10}(t) = a_{10} \cdot (1 - \theta(t))\,, \quad p_{12}(t) = a_{12} \cdot \theta(t)\,, \quad p_{11}(t) = 1 - p_{10}(t) - p_{12}(t)\,,$$
$$p_{23}(t) = a_{23} \cdot \theta(t)\,, \quad p_{22}(t) = 1 - p_{23}(t)\,,$$
$$\theta(t) = 1 - \exp(-C'Z(t))\,, \quad C = (r_1, r_1, \ldots, r_p)'\,.$$

2.3.2. *Model* 2. *A non-homogeneous continuous-time Markov model*

In the model, the transition intensities

$$\lambda_{ij}(t)dt = P\{X(t + dt) = j | X(t) = i\}$$

Let

$$\lambda_{01}(t) = A_0 + A_1 Z_1(t) + \cdots + A_p Z_p(t)\,,$$
$$\lambda_{10}(t) = B_0 + B_1 Z_1(t) + \cdots + B_p Z_p(t)\,,$$
$$\lambda_{12}(t) = C_0 + C_1 Z_1(t) + \cdots + C_p Z_p(t)\,,$$
$$\lambda_{23}(t) = D_0 + D_1 Z_1(t) + \cdots + D_p Z_p(t)\,.$$

The two models were applied to analyze a set of 12-year and 6-run screening data of cervical cancer in Jingan county, Jiangxi Province, China. The covariates are sex disorder, sex health, age, age-square and cervicitis.

In model 1, the estimation of parameter $C = (r_1, r_1, \ldots, r_p)'$ is

$$\hat{r}_1 = 0.7095\,,\ \hat{r}_2 = 0.0189\,,\ \hat{r}_3 = 0.0152\,,\ \hat{r}_4 = 2.23 \times 10^{-4}\,,\ \hat{r}_5 = 0.631\,,$$

i.e. $\hat{C} = (0.7095,\ 0.0189,\ 0.0152,\ 2.23 \times 10^{-4},\ 0.631)$.

In the likelihood ratio tests for the hypothesis $r_i = 0$, the χ_2 statistics are 58.65, 59.62, 22.97, 39.72 and 77.38 respectively. $P < 0.01$ for all.

The results show the 5 covariates have effect on the transition.

In model 2, let $C_i = 0$, $D_i = 0 (i = 1, 2, 3, 4, 5)$. The estimations of other parameters are

$$\hat{A}_0 = 2.81 \times 10^{-5}, \quad \hat{A}_1 = 0.765, \quad \hat{A}_2 = 0.963,$$
$$\hat{A}_3 = 4.16 \times 10^{-4}, \quad \hat{A}_4 = 7.61 \times 10^{-4}, \quad \hat{A}_5 = 1.333,$$
$$\hat{B}_0 = 8.25 \times 10^{-3}, \quad \hat{B}_1 = 0.019, \quad \hat{B}_2 = 0.116,$$
$$\hat{B}_3 = 2.32 \times 10^{-3}, \quad \hat{B}_4 = 3.12 \times 10^{-3}, \quad \hat{B}_5 = 0.109,$$
$$\hat{C}_0 = 7.03 \times 10^{-4}, \quad \hat{D}_0 = 0.0279.$$

After hypothesis test, except the covariates age and age-square, the others have effect on the transition.

2.4. *Example* 4: *Modeling for sequence with short-term memory*

In a sequence $\{X_t\}$, the value of X_s usually has effect on the value of X_{s+t}, i.e. $Cov(X_s, X_{s+1}) \neq 0$. The effect attenuates gradually when the time length t gets longer. The sequence is said to have short-term memory if the attenuation is fast and said to have long-term memory if the attenuation is slow. Markov chain is a sequence with short-term.

Ion-channels sometime is open and sometime is close. The single ion-channel patch-clamp recordings recorded by patch-clamp are shown in Fig. 3.

Fang *et al.*[15] proposed two-state Markov model to study quantitatively memory existing in ion-channels. A two-state Markov process with constant transition intensities well fitted the short-term memory and a two-state Markov process within a kind of random environment well fitted the long-term memory. In the short-term memory model, the auto-correlation function is $\exp[-(\lambda + \mu)t]$.

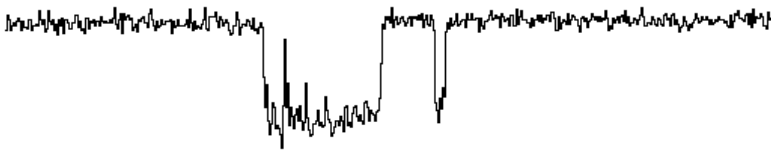


Fig. 3. Single ion-channel patch-clamp recordings.

3. Generalized Markov Chains

3.1. *Markov chains in random environment*

A Markov chain in random environment[15] is called if the transition intensities are random variables.

Example. Considering a two-state continuous-time Markov chain $\{X(t), t \geq 0\}$, the transition intensities $\lambda = q_{01}$, $\mu = q_{10}$ are two independent random variables. The probability density functions are $f(\lambda)$, $g(\mu)$ respectively.

When λ, μ are given, the transition matrix of $X(t)$ is

$$P(t) = \begin{pmatrix} \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}\exp(-(\lambda+\mu)t) & \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{\lambda+\mu}\exp(-(\lambda+\mu)t) \\ \frac{\mu}{\lambda+\mu} - \frac{\mu}{\lambda+\mu}\exp(-(\lambda+\mu)t) & \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}\exp(-(\lambda+\mu)t) \end{pmatrix}.$$

The conditional distribution $\{p_{j|\lambda,\mu}(t), j = 0, 1\}$ of $X(t)$ satisfies

$$\begin{aligned} p_{j|\lambda,\mu}(t) &\widehat{=} \{X(t) = j|\lambda, \mu\} \\ &= \sum_{i=0}^{1} P\{X(0) = i|\lambda, \mu\} P\{X(t) = j|X(0) = i, \lambda, \mu\}. \end{aligned}$$

So the distribution of $X(t)$ becomes

$$p_j(t) \widehat{=} P\{X(t) = j\} = \int_0^\infty P\{X(t) = j|\lambda, \mu\} f(\lambda) g(\mu) d\lambda d\mu.$$

The long-term memory model for ion-channels proposed by Fang *et al.*[15] is Markov model in random environment. The distributions of transition intensities are Γ distributions, $\Gamma(\alpha_1, \beta), \Gamma(\alpha_2, \beta)$. The auto-correlation attenuates as $(\beta t + 1)^{-(\alpha_1+\alpha_2)}$. The attenuation shows long-term memory.

3.2. *Semi-Markov processes*

A continuous-time Markov chain $\{X(t), t \geq 0\}$. The time τ_i spending for the transition between two successive states $i \to j (j \neq i)$ have exponential distribution. The distribution is dependent with state i and independent with state j. The chain is called semi-Markov process if the distribution of τ_i is arbitrary distribution and the distribution is dependent with state i and state j. Markov chain is a semi-Markov process.

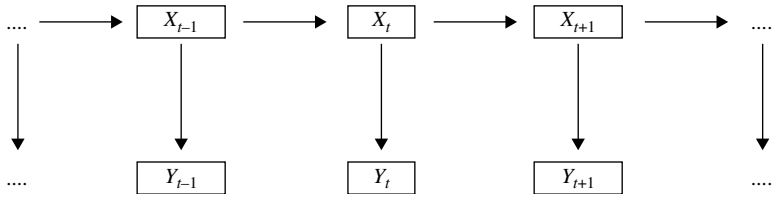


Fig. 4. The relation between $\{X_t\}$ and $\{Y_t\}$.

3.3. *Hidden Markov chains*

A hidden Markov chain[19] $\{X_t, Y_t\}$ is composed by two processes. One is a underline finite-state Markov chain $\{X_t\}$. The other is observed processes $\{Y_t\}$ and the distribution of Y_t is dependent with X_t.

For example, the ion-channel sometimes is open and sometimes is close. The open (state 1) or close (state 0) cannot be observed directly because there are noise in the channel. Hidden Markov model can be used here. Let $\{X_t\}$ denote the sequence of states and $\{Y_t\}$ denote the observed sequence recorded by patch-clamp. Suppose that $\{X_t\}$ has Markov property. Y_t has normal distribution $N(\mu_0, \sigma_0^2)$ when $X_t = 0$ and has normal distribution $N(\mu_1, \sigma_1^2)$ when $X_t = 1$. If $\{X_t\}$ is given, there is conditional independence among Y_t s. The relation between $\{X_t\}$ and $\{Y_t\}$ as shown in Fig. 4.

At present, hidden Markov model is also used for biological sequence analysis.

Example.[19] Let $\{Y_t, t = 1, 2, \ldots, n\}$, $Y_t \in \{a, c, g, t\} \widehat{=} \{1, 2, 3, 4\}$ is a DNA sequence. $S_t, S_t \in \{1, 2, \ldots, r\}$ denote the type of homogeneous segment at position t in the sequence. $S_t, S_t \in \{1, 2, \ldots, r\}$ is unobservable. A hidden Markov processes modeled for $\{S_t, Y_t\}$. Supposed $\{Y_t\}$ is a Markov chain, the transition probability

$$\begin{aligned} P\{Y_t = y_t | S_t = s_t, y_1, y_2, \ldots, y_{t-1}\} &= P\{Y_t = y_t | S_t = s_t, y_{t-1}\} \\ &= p_{y_{t-1} y_t}^{(s_t)} , \end{aligned}$$

dependent with the segment type S_t.

Suppose that the conjugate prior distribution for the row vectors $p_i^{(k)}$ of the transition matrix is Dirichlet distribution. Then the posterior distribution is still Dirichlet distribution. The segment type can be decided by the conditional probability $P(S_t = k|y)$ $(k = 1, 2, \ldots, r)$ and the probability $P(S_t = k, S_{t+1} \neq k|y)$.

3.4. *Time-reversible Markov chain*

Consider a stationary Markov chain with transition matrix $P = (p_{ij})$. If the initial distribution is the stationary distribution $\{\pi_i\}$, the distribution of the chain at any time will remain the same. Now, we trace the sequence of states backwards in time. The reversible sequence $X_n, X_{n-1}, \ldots$, is still a Markov chain. The transition probability

$$\begin{aligned} p_{ij}^* &= P\{X_m = j | X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k\} \\ &= \frac{P\{X_m = j, X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k\}}{P\{X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k\}} \\ &= \frac{\begin{array}{r} P\{X_m = j\} P\{X_{m+1} = i | X_m = j\} \\ \times P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_m = j, X_{m+1} = i\} \end{array}}{P\{X_{m+1} = i\} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i\}} \\ &= \frac{\pi_j p_{ji} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i\}}{\pi_i P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i\}} \\ &= \frac{\pi_j}{\pi_i} p_{ji} \,. \end{aligned}$$

If $p_{ij}^* = p_{ij}$, for all i, j, the Markov chain is called time-reversible.[20] If a Markov chain is time-reversible, then

$$\pi_i p_{ij} = \pi_j p_{ij} \,, \quad \text{for all } i, j \,. \tag{12}$$

4. Applications in Statistic Computation

4.1. *MCMC*

In statistic computation, we usually compute the expectation of a function $f(x)$:

$$E_\pi f = \int f(x)\pi(x)dx \,,$$

where $x = (x_1, x_2, \ldots, x_k)$ is k-dimensional vector and $\pi(x)$ is a density function. Markov chain Monte Carlo (MCMC)[8] methods are applied for computation.

At present, there are two different definitions about Markov chain .In MCMC, the Markov chain is a discrete-state Markov process. There are transition probabilities for discrete-time chain and transition intensities for continuous-time chain. The transition probabilities and transition intensities are called transition kernel for discrete-state Markov process.

In MCMC, an aperiodic irreducible Markov chain $\{X^{(0)}, X^{(1)}, X^{(2)}, \ldots\}$ with stationary distribution $\pi(x)$ was built. If the initial state $X^{(0)}$ has distribution $\pi(x)$, $X^{(t)}$ will have the same distribution $\pi(x)$. An aperiodic irreducible Markov chain with stationary distribution $\pi(x)$ has limiting distribution which is also $\pi(x)$. The distribution of $X^{(m)}$ has little effect on the initial state $X^{(0)}$ and gets close to $\pi(x)$ when m is large enough. Therefore, the $n-m$ state, $X^{(m+1)}, X^{(m+2)}, \ldots, X^{(n)}$, is used for computation.

The steps in MCMC methods as follows.

Step 1. Building a Markov chain with stationary distribution $\pi(x)$.
Step 2. Getting a sample $X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(n)}$.
Step 3. Taking $m, n(m < n)$ and estimating of $E_\pi f$:

$$\hat{E}_\pi f = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X^{(t)}) \,. \tag{13}$$

In MCMC methods, the transition kernel $p(x, x')$ is very important, where $x' = (x'_1, x'_2, \ldots, x'_k)$. In the different MCMC methods, the kernel is different.

In MCMC methods, the full conditional distribution $\pi(x_T|x_{-T})$ are used, where $T \subset \{1, 2, \ldots, k\}$, $x_T = \{x_i, i \in T\}$, $x_{-T} = \{x_i, i \notin T\}$.

$$\pi(x_T|x_{-T}) = \frac{\pi(x_T)}{\int \pi(x) dx_T} \propto \pi(x) \,.$$

Example. Suppose the joint distribution of X_1, X_2 is

$$\pi(x_1, x_2) \propto \exp\left\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\right\}.$$

Then the full conditional distribution

$$\begin{aligned}
\pi(x_1|x_2) &\propto \pi(x_1, x_2) \propto \exp\left\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\right\} \\
&= N(1, (x_2 - 1)^{-2}) \\
\pi(x_2|x_1) &\propto \pi(x_1, x_2) \propto \exp\left\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\right\} \\
&= N(1, (x_1 - 1)^{-2})
\end{aligned}$$

Two important MCMC methods, Gibbs method and Metropolis-Hastings method, are introduced respectively.

4.1.1. *Gibbs sampling method*

Gibbs sampling method is proposed by Geman S. and Geman D. (1984).[21]

Let $x'_{-T} = x_{-T}$. Consider the conditional distribution of $X_T|X_{-T}$. Let the transition kernel $p(x_T, x'_T|x_{-T}) \widehat{=} \pi(x'_T|x_{-T})$. So x'_T can be gotten from the distribution $\pi(\bullet|x_{-T})$. It can be proved that $X' = (X'_T, X'_{-T})$ has distribution $\pi(x')$.

The simple Gibbs method is called if only one element is in T. Sampling from full conditional distribution becomes simple. The steps are given as follows.

Suppose that initial value $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_k^{(0)})$ is given and the $(t-1)$th iterative value is $x^{(t-1)}$. The tth iterative value $x^{(t)}$ is gotten as follow k steps.

Step 1. Sampling $x_1^{(t)}$ from full conditional distribution $\pi(x_1|x_2^{(t-1)}, \ldots, x_k^{(t-1)})$;
. . . .

Step i. Sampling $x_i^{(t)}$ from full conditional distribution $\pi(x_i|x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_k^{(t-1)})$;
. . . .

Step k. Sampling $x_k^{(t)}$ from full conditional distribution $\pi(x_k|x_1^{(t)}, \ldots, x_{k-1}^{(t)})$. Let $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_k^{(t)})$.

4.1.2. *Metropolis–Hastings sampling method*

Metropolis–Hastings sampling method is proposed by Metropolis (1953)[22] and Hastings (1970).[23] The transition kernel $p(x, x')$ is built as follows.

Suppose that $q(x, x')$ is an irreducible transition kernel. Let

$$p(x, x') = q(x, x')\alpha(x, x'), \quad x \neq x'$$

where $\alpha(x, x')$ is a function and $0 < \alpha(x, x') \leq 1$.

When the current state is x, i.e. $X^{(t)} = x$. We propose a transition $x \to x'$ with intensity $q(x, x')$. The proposal is not automatically accepted. The probability of acceptance is $\alpha(x, x')$. Therefore, the successive state will be changed as x' with probability $\alpha(x, x')$ and not changed with probability $1 - \alpha(x, x')$. That is to say,

$$x^{(t+1)} = \begin{cases} x' & u \leq \alpha(x, x') \\ x & u > \alpha(x, x') \end{cases},$$

where u is a random number from [0, 1] uniform distribution.

Table 8. The capture-recapture data when $k = 3$.

		3rd sample			
		Observed		Non-observed	
		2nd sample		2nd sample	
		Observed	Non-observed	Observed	Non-observed
1st sample	Observed	x_{123}	$x_{12\bar{3}}$	$x_{1\bar{2}3}$	$x_{1\bar{2}\bar{3}}$
	Non-observed	$x_{\bar{1}23}$	$x_{\bar{1}2\bar{3}}$	$x_{\bar{1}\bar{2}3}$	$x_{\bar{1}\bar{2}\bar{3}}$

$q(x, \cdot)$ is a probability function or density function. It is called proposal distribution in Metropolis–Hastings sampling. $\alpha(x, x')$ is called acceptance probability. An expression for $\alpha(x, x')$ is derived to ensure the chain with stationary distribution $\pi(x)$.

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}\right\}. \tag{14}$$

For full conditional distribution, let $x'_{-T} = x_{-T}$. The proposal distribution is $q(x_T, x'_T|x_{-T})$ and acceptance probability

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x'_T|x_{-T})q(x'_T, x_T|x_{-T})}{\pi(x_T|x_{-T})q(x_T, x'_T|x_{-T})}\right\}. \tag{15}$$

Gibbs sampling is a special Metropolis–Hastings sampling, where proposal distribution is $\pi(x'_T|x_{-T})$ and acceptance probability is constant 1.

In the dissertation of Gao[6] about Bayesian analysis of capture-recapture data, an application for MCMC is provided.

In capture-recapture model, the size of a closed population is N. N is unknown and need to be estimated by k samples from the population. Suppose that the size of the ith sample is n_i and $n (n < N)$ individuals are observed in all k samples and. Every individual was captured with probability p_i in the ith sample. When $k = 3$, the capture-recapture data as following Table 8, where $x_{\bar{1}\bar{2}\bar{3}}$ is unknown. Let $u = x_{\bar{1}\bar{2}\bar{3}}$. Then $N = n + u$.

Let $\{\omega\} = \{123, 12\bar{3}, 1\bar{2}3, \bar{1}23, 1\bar{2}\bar{3}, \bar{1}2\bar{3}, \bar{1}\bar{2}3\}$ and $\{\omega'\} = \{\omega\} + \{\bar{1}\bar{2}\bar{3}\}$.

When the three samples are independent, the likelihood function is

$$p(\{x_\omega\}|N, \{p_i\}) = \frac{N!}{(N-n)! \prod_\omega x_\omega!} \prod_1^3 p_i^{n_i}(1-p_i)^{N-n_i}.$$

The posterior distribution $(N, \{p_i\})$ is

$$p(N, \{p_i\}|x_{\omega'}) \propto \frac{N!}{(N-n)! \prod_\omega x_\omega!} \prod_\omega (p_i)^{x_\omega}(1-p_i)^{N-n_i}\pi(N, \{p_i\}),$$

where $\pi(N, \{p_i\})$ is prior distribution of $(N, \{p_i\})$. If $\pi(N, \{p_i\})$ is uniform distribution, the full conditional distribution of $(N, \{p_i\})$ is

$$p(p_i|u, p_{-i}, \{x_\omega\}) \propto p_i^{n_i}(1-p_i)^{u+n-n_i} \sim \beta(n_i+1, u+n-n_i+1)\,,$$

where $\beta(n_i+1, u+n-n_i+1)$ is β distribution, and

$$p(u|\{p_i\}, \{x_\omega\}) = \begin{pmatrix} u+n \\ u \end{pmatrix} (p^*)^u(1-p^*)^n = b(n+u, p^*)\,,$$

where $p^* = \prod_1^3(1-p_i)$.

4.2. *Reversible jump MCMC computation*

In above MCMC methods, the dimension k of vector x is known and fixed. They are not available when k is not fixed. Peter J. Green[16] proposed reversible jump MCMC samplers that jump between parameter subspaces of differing dimensionality.

4.2.1. *The general case*

Suppose that we have a countable collection of candidates model $\{M_k, k \in K\}$. Model M_k has a vector $\theta^{(k)}$ of unknown parameters, where $\theta^{(k)} \in R^{n_k}$ and is a n_k dimensional vector. Let $x = (k, \theta^{(k)})$, $\Omega_k = \{k\} \times R^{n_k}$ and $\Omega = \bigcup_k \Omega_k$, $x \in \Omega_k$. For a given k, $\pi(x)$ is the joint posterior distribution of k and $\theta^{(k)}$, i.e.

$$\pi(x) = p(k, \theta^{(k)}|y) = p(k|y)p(\theta^{(k)}|k, y)\,,$$

where y is an observed sample.

When the current state is x, we propose a move $x \to dx'$ of type m with probability $q_m(x, dx')$. Thus the successive state is not changed with probability $1 - \sum_m q_m(x, \Omega)$, where $\sum_m q_m(x, \Omega) \le 1$. Let $\alpha_m(x, x')$ is the acceptance probability of the move of type m. The transition kernel is

$$P(x, B) = \sum_m \int_B q_m(x, dx')\alpha_m(x, x') + s(x)I(x \in B)\,,$$

where $B \subset \Omega$, $I(\cdot)$ is indicator function and

$$s(x) = \sum_m \int_B q_m(x, dx')\{1 - \alpha_m(x, x')\} + 1 - \sum_m q_m(x, \Omega)$$

is the probability of not moving from x.

$\alpha_m(x, x')$ given by Peter J. Green[16] is

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')q_m(x'dx)}{\pi(dx)q_m(x, dx')}\right\}. \tag{16}$$

Suppose that $\pi(dx)q_m(x, dx')$ has a finite density $f_m(x, x')$. Then

$$\alpha_m(x, x') = \min\left\{1, \frac{f_m(x', x)}{f_m(x, x')}\right\}. \tag{17}$$

4.2.2. *Switching between two simple subspaces*

A simple example is given first. Let two subspaces $\Omega_1 = \{1\} \times R$, $\Omega_2 = \{2\} \times R^2$ $x = (1, \theta) \in \Omega_1$ when $k = 1$ and $x = (2, \theta_1, \theta_2) \in \Omega_2$ when $k = 2$. Consider a move between Ω_1 and Ω_2. A move from Ω_1 to Ω_2 is defined as $(1, \theta) \to (2, \theta + u, \theta - u)$, where u and θ are independent random variables. Then the reversible move from Ω_2 to Ω_1 is $(2, \theta_1, \theta_2) \to (1, \frac{1}{2}(\theta_1 + \theta_2))$. For dimensional matching of (θ, u) and (θ_1, θ_2), there is a bijection between $(\theta, u)$and (θ_1, θ_2).

In general, $\Omega_1 = \{1\} \times R^{n_1}, \Omega_2 = \{2\} \times R^{n_2}, \Omega = \bigcup_k \Omega_k, x = (k, \theta^{(k)})$. For a given k, $x \in \Omega_k, k = 1, 2)$. Consider just one move type between Ω_1 and Ω_2. The proposal distribution is $q(x, dx')$ and this move is chosen with probability $j(x)$. From Ω_1 to Ω_2, a m_1 dimension random vector $u^{(1)}$ independent with $\theta^{(1)}$ is generated. Set $\theta^{(2)}$ to be some function of $\theta^{(1)}$ and $u^{(1)}$. Then the move $(1, \theta^{(1)}) \to (2, \theta^{(2)})$ is defined. Similarly, to switch back, a m_2 dimension random vector $u^{(2)}$ independent with $\theta^{(2)}$ is generated and set $\theta^{(1)}$ to be some function of $\theta^{(2)}$ and $u^{(2)}$. Then there is a bijection between $(\theta^{(1)}, u^{(1)})$ and $(\theta^{(2)}, u^{(2)})$. For dimensional matching, m_1 and m_2 must satisfy $n_1 + m_1 = n_2 + m_2$. Suppose that the densities of $u^{(1)}$ and $u^{(2)}$ are $q_1(u^{(1)})$ and $q_2(u^{(2)})$ respectively. Let $x = (1, \theta^{(1)}, \mu^{(1)}) \in \Omega$, $x' = (2, \theta^{(2)}, \mu^{(2)}) \in \Omega$ and

$$f(x, x') = p(1, \theta^{(1)}|y)j(1, \theta^{(1)})q_1(u^{(1)}),$$

$$f(x', x) = p(2, \theta^{(2)}|y)j(2, \theta^{(2)})q_2(u^{(2)}) \left|\frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right|.$$

Then the acceptance probability

$$\alpha(x, x') = \min\left\{1, \frac{p(2, \theta^{(2)}|y)j(2, \theta^{(2)})q_2(u^{(2)})}{p(1, \theta^{(1)}|y)j(1, \theta^{(1)})q_1(u^{(1)})} \left|\frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right|\right\}. \tag{18}$$

Sometimes, m_1 or m_2 is 0. For example, when m_2 is 0, Eq. (18) becomes

$$\alpha(x, x') = \min\left\{1, \frac{p(2, \theta^{(2)}|y)j(2, \theta^{(2)})}{p(1, \theta^{(1)}|y)j(1, \theta^{(1)})q_1(u^{(1)})} \left|\frac{\partial(\theta^{(2)})}{\partial(\theta^{(1)}, u^{(1)})}\right|\right\}.$$

5. Branching Processes

5.1. *Branching processes*

Branching Processes were studied by Galton and Watson in 1874. Consider a population, in which the individuals can produce offspring. Each individual can produce k new offspring with probability $p_k, k = 0, 1, 2, \ldots$, independently of the others' producing. That is to say, all individuals' producings have independent identical distribution (i.i.d.). Suppose that the numbers of initial individuals is X_0, which is called the size of the 0th generation. The size of the first generation, which is constituted by all offspring of the 0th generate, is denoted by $X_1, \ldots$. Let $Z_j^{(n)}$ denote the number of the offspring produced by the jth individual in the nth generation. Then,

$$X_n = Z_1^{(n-1)} + Z_2^{(n-1)} + \cdots + Z_{X_{n-1}}^{(n-1)} = \sum_{j=1}^{X_{n-1}} Z_j^{(n-1)} .$$

It shows that X_n is a sum of X_{n-1} random variables with i.i.d. $\{p_k, k = 0, 1, 2, \ldots\}$. The process $\{X_n\}$ is called branching processes.[18]

The Branching Processes is a Markov Chain and its transition probability is

$$p_{ij} = P\{X_{n+1} = j | X_n = i\} = P\left(\sum_{k=1}^{i} Z_k^{(n)} = j\right).$$

Suppose that there are x_0 individuals in the zeroth generation, i.e. $X_0 = x_0$. Let $E(Z_j^{(n)}) = \sum_{k=0}^{\infty} kp_k = \mu$ and $\mathrm{var}(Z_j^{(n)}) = \sum_{k=0}^{\infty}(k-\mu)^2 p_k = \sigma^2$. Then it is easy to see

$$E(X_n) = x_0\mu^n ,$$

$$\mathrm{var}(X_n) = \begin{cases} x_0^2\mu^{n-1}\sigma^2\dfrac{\mu^{n-1}}{\mu-1} & \mu \neq 1 \\ nx_0^2\sigma^2 & \mu = 1 \end{cases} .$$

Now, we can see that the expectation and variance of the size will increase when $\mu > 1$ and will decrease when $\mu < 1$.

In branching processes, the probability π_0 that the population dies out is shown in the following theorem.

Theorem. *Suppose that* $p_0 > 0$ *and* $p_0 + p_1 < 1$. *Then*

(1) $\pi_0 = q^{x_0}$ *if* $\mu > 1$, *where* q *is the smallest positive number satisfying the equation*

$$x = \sum_{k=0}^{\infty} p_k x^k \,. \tag{19}$$

(2) $\pi_0 = 1$ *if* $\mu \le 1$.

Example. Suppose that each individual in a population can produce 0, 1, 2 and 3 offspring with probabilities 1/8, 3/8, 3/8 and 1/8, respectively. Then the mean number of offspring per individual is

$$\mu = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5 > 1 \,.$$

If the size of 0th generation is 1, i.e. $x_0 = 1$, the probability π_0 that the population dies out satisfies

$$x = x^0 \times \frac{1}{8} + x^1 \times \frac{3}{8} + x^2 \times \frac{3}{8} + x^3 \times \frac{1}{8} \,,$$

i.e. $x^3 + 3x^2 - 5x + 1 = 0$.
This equation has 3 roots, 1, $-\sqrt{5} - 2$ and $\sqrt{5} - 2$. So $q = \sqrt{5} - 2$.

Lotka[2,18] proposed a branching process model to study the white American family. Suppose that a white generates k sons ($k = 0, 1, 2, \ldots$) with probability

$$p_k = \begin{cases} bc^{k-1} & k \neq 0 \\ 1 - \displaystyle\sum_{k=1}^{\infty} bc^{k-1} & k = 0 \,, \end{cases}$$

where b, c is positive number and $b + c < 1$.

In this branching process,

$$\mu = \sum_{k=0}^{\infty} k p_k = \frac{b}{1 - c^2} \,.$$

The equation

$$x = \sum_{k=0}^{\infty} p_k x^k$$

has two roots, i.e. 1 and $q = \frac{1-(b+c)}{c(1-c)}$. Lotka got $b = 0.2126$, $c = 0.5893$ according to a data collected in 1920. So we will obtain

$$\mu = 1.26 > 1 \,, \quad q = 0.819 \,.$$

That is to say, given $x_0 = 1$, the probability that the population will eventually die out is 0.819.

In addition, branching processes has ever used in genes mutation, genetics and epidemiology.

5.2. *Generalized branching processes*

Branching processes introduced above is in very simple cases.

(1) Each individual in certain generation produces offspring with the identical probability distribution.
(2) The probability distribution above is independent of generation.
(3) Any two individuals' producing is independent each other.
(4) Each individual will not die before it produces offspring.
(5) The population is close.

In this simple case, the size of the population will eventually become 0 or infinity. In fact, the above assumptions usually come into broken before the size become infinity. There are some cases to generalize simple branching processes.

(i) Suppose that each individual survives with probability r before producing.
(ii) The distribution $\{p_k, k = 0, 1, 2, \ldots\}$ is dependent on generation n.
(iii) The population is not close.

Suppose that the population in simple branching processes is not close when the generation n is produced, there are Y_n individuals of the same kind immigrating and the Y_n individuals' producing independently. The size of population which is not close can be stationary if the immigration satisfies some certain conditions.

Suppose that survival times of each individual in simple branching processes have independent identical distribution F. Before dying, each individual will have produced k new offspring with probability p_k. Let $X(t)$ denote the number of living individuals at time t. For the process $\{X(t), t \geq 0\}$, when t is large enough,

$$E(X(t)) \approx \frac{(\mu - 1)e^{\alpha t}}{\mu^2 \alpha \int_0^\infty x e^{-\alpha x} dF(x)},$$

where $\mu = \sum_{k=0}^{\infty} kp_k$ is expectation of each individual's offspring. α is the positive number which satisfies the follow equation

$$\int_0^{\infty} e^{-\alpha x} dF(x) = \frac{1}{\mu}.$$

Lucas[17] used branching processes to study plasmodia's producing.

After a plasmodium comes into the cell in the liver, it propagates rapidly. When the number of plasmodia is large enough, the cell will be broken and plasmodia will go into red cells in blood. And with the number of plasmodia produced increasing, the red cell is broken and go into other red cells. The red cell's broken is periodical. Every period is about 48–72 hours.

A branching process is used here. The initial plasmodia which go into red cells constitute 0th generation. The offspring of 0th generation is called 1st generation, and so on. Two models were considered.

In model 1, each plasmodium's survival probability is r before producing. It is easy to see

$$E(X_n) = x_0 r^{n-1} \mu^n .$$

So we can conclude that the expectation of the number of plasmodia increase if $r\mu > 1$, decrease if $r\mu < 1$, and is fixed if $r\mu = 1$. This model cannot fit well the data given in Table 9.

In model 2, suppose that the survive probability is dependent on generation n and denotes r_n. Then

$$E(X_n) = x_0 r_1 r_2 \cdots r_{n-1} \mu^n .$$

Therefore $E(X_{n+1}) = r_n \mu E(X_N)$. If μ is given, r_n can be estimated from this equation.

When $\mu = 10$, the estimations are given in Table 10.

Table 9. The number of invaded from 10^6 red cells every 48 hours.

Date	Time	Invaded number
10/24	22:15	5600
10/26	22:30	1220
10/28	22:00	1330
10/30	22:30	1200
11/1	22:30	1560
11/3	22:30	1440
11/5	22:30	2000
11/7	23:00	1370
11/9	23:15	161

Table 10. The estimations of r_n when $\mu = 10$.

Date	Invaded number	$r_n\mu$	r_n
10/24	5600		
10/26	1220	0.218	0.0218
10/28	1330	1.090	0.1090
10/30	1200	0.902	0.0902
11/1	1560	1.300	0.1300
11/3	1440	0.923	0.0923
11/5	2000	1.389	0.1389
11/7	1370	0.685	0.0685
11/9	161	0.118	0.0118

6. Birth–Death Processes

6.1. *Birth–death process*

Birth–death processes $\{N(t), t \geq 0\}$ were discussed in the processes that a population grow and decline. Birth–death processes are Markov chains with transition probabilities satisfying $q_{ij} = 0$ if $|i - j| \geq 2$. Let

$$\lambda_i = q_{i,i+1}\,, \quad \mu_i = q_{i,i-1}\,.$$

The transition probability matrix for Birth–death processes is

$$\begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \mu_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \mu_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Let $N(t)$ denote the size of a population at time t. We say that 1 birth (or 1 death) occurs in the population when the size increase by 1 (decrease by 1, respectively). λ_i and μ_i are called birth rate and death rate respectively.

According to the C–K forward equation, the distribution $p_k(t) \mathrel{\widehat{=}} P\{N(t) = k\}$ satisfies the equation

$$p_0'(t) = -(\lambda_0 + \mu_0)p_0(t) + \mu_1 p_1(t)\,,$$

$$p_k'(t) = -(\lambda_k + \mu_k)p_k(t) + \lambda_{k-1}p_{k-1}(t) + \mu_{\mu+1}p_{k+1}(t) \quad k \geq 1$$

when $N(0) = 0$.

Example.[20] There are M mice in a cage, and there are infinity foods to provide them to eat. A mouse will stop eating at time $t+h$ with probability $\mu h + o(h)$ if it is eating at time t and will have been eating before time $t+h$

with probability $\lambda h + o(h)$ if it is not eating at time t. Each mice's eating is independent of others' eating. Let $N(t)$ denote the number of mice which are eating at time t. $\{N(t)\}$ is a birth–death process, and

$$P\{N(t+h) = i+1 | N(t) = i\} = (M-i)\lambda h + o(h)\,,$$

$$P\{N(t+h) = i-1 | N(t) = i\} = i\mu h + o(h)\,.$$

Therefore, the birth rate is $\lambda_i = (M-i)\lambda$, the death rate is $\mu_i = i\mu$.

The processes discussed above are homogeneous. A birth–death process is called homogeneous when the birth rate and death rate depend on time.

6.2. *Pure birth process*

A birth–death process is called pure birth process if $\mu_i = 0$ for all i. Therefore, the distribution $\{p_k(t)\}$ of pure birth process satisfies

$$p_0'(t) = -\lambda_0 p_0(t)\,,$$

$$p_k'(t) = -\lambda_k p_k(t) + \lambda_{k-1} p_{k-1}(t) \quad k \geq 1\,.$$

6.2.1. *Example*[20] *Mckendrick model*

There is one population which is constituted by 1 infected and $N-1$ susceptible individuals. The infected state is a absorbing state. Suppose that any given infected individual will cause, with probability $\beta k + o(h)$, any given susceptible individual infected in time interval $(t, t+h)$, where β is called infection rate. Let $X(t)$ denote the number of the infected individuals at time t. Then $\{X(t)\}$ is a pure birth process with birth rate

$$\lambda_n(t) = (N-n)n\beta\,.$$

This epidemic model was proposed by A. M. Mckendrick in 1926.

Let T denote the time until all individuals in the population are infected and T_i denote the time from i infective to $i+1$ infective. Then T_i has exponential distribution with mean $\frac{1}{\lambda_i} = \frac{1}{(N-i)i\beta}$. Therefore

$$ET = E\left(\sum_{i=1}^{N-1} T_i\right) = \frac{1}{\beta}\sum_{i=1}^{N-1}\frac{1}{i(N-i)}\,.$$

6.2.2. *Example: M. J. Faddy and J. S. Fenlon*[13]

Some stochastic models based on pure birth processes are constructed to describe the invasion process of nematodes in fly larvae. Let $X(t)$ denote

the number of nematodes which have invaded the host at time t. Then $\{X(t)\}$ is a pure birth process with birth rate λ_n, i.e.

$$P\{X(t+\Delta t) = n+1 | X(t) = n\} = \lambda_n \Delta t + o(\Delta t)\,,$$

$$\lambda_n = (N-n)a_n\,,$$

where N is the number of nematodes outside the host at time 0 and $X(0) = 0$.

Five models in which appropriate forms for λ_n are given constructed as follows.

Let $\lambda_n = (N-n)a_n$.

Model 1. Let $a_n = a$, where a is a constant. From the differential equations, we can know that $X(t)$ has binomial distribution

$$P\{X(t) = n\} = \binom{N}{n} (1 - \exp(-at))^n \exp(-at)^{N-n}\,.$$

However, in practice, such a model is unlikely to be adequate.

Model 2. Let

$$a_n = \begin{cases} a_0 & n = 0 \\ a_1 & n \geq 1 \end{cases}, \quad \text{where } a_1 > a_0\,.$$

Model 3. Let $a_n = \exp(a + bn)$, where $b > 0$.

Model 4. Let

$$a_n = \frac{a}{1 + \exp(b + cn)}\,, \quad \text{where } c < 0\,.$$

Model 5. Let

$$a_n = \frac{a \exp(-dN)}{1 + \exp(b + cn)}\,, \quad \text{where } c < 0,\ d > 0\,.$$

The solution of the four differential equations for the latter 4 models can be calculated numerically using MATLAB software. Three data sets are given in Tables 11, 12 and 13 respectively.

Three data sets are analyzed. All models fitted to these data resulted in a log-likelihood. Let L_i denote the log-likelihood value for model i.

For Table 11, $L_1 = -626.80$, $L_2 = -602.57$, $L_3 = -588.88$, $L_4 = -588.46$ and $L_5 = -588.21$. Model 3 is good enough. In model 3, $\hat{a} = -1.17(0.05)$, $\hat{b} = 0.25(0.03)$ where the values in parentheses are standard

Table 11. Numbers of invading nematodes for various N.

	Number of larvae with the following numbers of invading nematodes										
N	0	1	2	3	4	5	6	7	8	9	10
10	1	8	12	11	11	6	9	6	6	2	0
7	9	14	27	15	6	3	1	0			
4	28	18	17	7	3						
2	44	26	6								
1	158	60									

Table 12. Numbers of invading nematodes for various N.

	Number of larvae with the following numbers of invading nematodes										
N	0	1	2	3	4	5	6	7	8	9	10
10	4	11	15	10	10	11	8	3	0	0	0
7	12	21	17	12	7	5	0	0			
4	32	22	15	6	0						
2	35	26	17	2							
1	165	59									

Table 13. Numbers of invading nematodes for various N.

	Number of larvae with the following numbers of invading nematodes										
N	0	1	2	3	4	5	6	7	8	9	10
10	21	13	11	11	9	4	2	2	1	0	0
7	34	15	13	1	2	3	1	1			
4	35	19	12	3	2						
2	45	26	3								
1	186	40									

errors. After combining some of the entries in Table 11 with low counts, a χ^2 goodness-of-fit statistic is calculated. $\chi^2 = 17.42$, degree of freedom $df = 16$ and p-value $p \approx 0.36$.

For Table 12, $L_1 = -559.28$, $L_2 = -545.04$, $L_3 = -542.12$, $L_4 = -540.041$ and $L_5 = -540.039$. Model 4 is better. In model 4, the estimators of parameters are $\hat{a} = 0.72(0.20)$, $\hat{b} = 0.60(0.42)$, $\hat{c} = -0.61(0.25)$. After combining some of the entries in Table 12 with low counts, $\chi^2 = 7.75$, $df = 13$ and $p \approx 0.86$.

For Table 13, $L_1 = -554.81$, $L_2 = -517.60$, $L_3 = -512.39$, $L_4 = -509.58$ and $L_5 = -503.71$. Model 5 is the best. In model 5, the estimators

of parameters are $\hat{a} = 1.94(0.78)$, $\hat{b} = 2.07(0.41)$, $\hat{c} = -0.87(0.19)$, $\hat{d} = 0.062(0.018)$. After combining some of the entries in Table 13 with low counts, $\chi^2 = 19.22$, $df = 13$ and $p \approx 0.12$.

When the birth rate λ_i in homogeneous process or $\lambda_i(t)$ in non-homogeneous process takes appropriate forms, some special processes are gotten.

6.3. *Poisson process*

6.3.1. *Poisson process — $\lambda_i = \lambda$*

A Poisson process $\{X(t)\}$ is a pure birth process with constant birth rate. The solution of the differential equations in pure birth processes gives the distribution of $X(t)$

$$p_k(t) \triangleq P\{X(t) = k\} = \frac{\exp(-\lambda t)(\lambda t)^k}{k!}\,.$$

It is a Poisson distribution with mean λt.

6.3.2. *Non-homogeneous Poisson process — $\lambda_i(t) = \lambda(t)$*

Non-homogeneous Poisson processes $\{X(t)\}$ are time dependent Poisson processes. The distribution of $X(t)$ is Poisson distribution

$$p_k(t) \triangleq P\{X(t) = k\} = \frac{\exp(-\int_0^t \lambda(s)ds)(\int_0^t \lambda(s)ds)^k}{k!}\,,$$

with mean $\int_0^t \lambda(s)ds$.

6.3.3. *Weighted Poisson process*[11] *— λ is a random variable*

The Poisson process describe the frequency of occurrence of an event to an individual with risk parameter λ. The variability of individuals with respect to this risk takes into account. That is to say, we permit λ is various with the density function $f(\lambda)$. The conditional distribution $p_{k|\lambda}(t)$ of $X(t)$ given λ is

$$p_{k|\lambda}(t) \triangleq P\{X(t) = k|\lambda\} = \frac{\exp(-\lambda t)(\lambda t)^k}{k!}\,.$$

Thus the distribution of $X(t)$ is

$$p_k(t) = \int_0^\infty \frac{\exp(-\lambda t)(\lambda t)^k}{k!} f(\lambda)d\lambda\,.$$

For example, if λ has a gamma distribution, we can get that the distribution of $X(t)$ is a negative binomial distribution.

Weighted Poisson model which incorporates variability of risk has been found useful in studies of accident proneness.

6.3.4. *Compound Poisson process*

A stochastic process $\{X(t), t \geq 0\}$ is called a compound Poisson process if it is represented by

$$X(t) = \sum_{n=1}^{N(t)} Y_n \,,$$

where $\{N(t), t \geq 0\}$ is a Poisson process, $\{Y_n, n = 1, 2, \ldots\}$ is a collection of random variables with independent and identically distribution (i.i.d.), and $\{N(t), t \geq 0\}$ and $\{Y_n, n = 1, 2, \ldots\}$ are independent.

Example.[2] Suppose that the insurants in a insurance company die at time $\tau_1, \tau_2, \ldots, (0 < \tau_1 < \tau_2 < \cdots)$. Their deaths are Poisson events with rate λ. The insurance is Y_n when the death occurs at time τ_n. Let $X(t)$ denote the total amount of insurance by time t. Then,

$$X(t) = \sum_{n=1}^{N(t)} Y_n \,,$$

where $N(t)$ is the number of deaths by time t and it is a Poisson process with rate λ. $\{X(t)\}$ is a compound Poisson process.

6.4. *Yule process*

6.4.1. *Yule process — $\lambda_i = i\lambda$*

The Yule process is a pure birth process with linear birth rate. Suppose that all individuals alive at time t give birth to another individual with the same rate λ and that individuals give birth independently of each other. Let $N(t)$ denote the total number of the population at time t. Then $N(t)$ is a Yule process.

The Yule process, starting from i individuals, has a negative binomial distribution

$$p_{ij}(t) \,\widehat{=}\, P\{N(t) = j | N(0) = i\} = \binom{j-1}{i-1} \exp(-\lambda t)^i (1 - \exp(-\lambda t))^{j-i} \,.$$

Yule[18] used this process to study evolution. Let $N(t)$ is the number of animal or plant species in a certain genus at time t. Suppose that each species would not be extinct when it come into being. In interval $(t, t+h)$, a new species is generated with probability $\lambda N(t)h$. Suppose that the new genera generated at time $\tau_1, \tau_2, \ldots, (0 < \tau_1 < \tau_2 < \cdots)$ with non-homogeneous Poisson rate. The mean function of the Poisson process is

$$m(t) = N_0 e^{at},$$

where N_0 and a are positive constant. Let $X^{(n)}(T)$ denote be the numbers of genera in which n species are contained at the given time T. Then $X^{(n)}(T)$ can be represented by

$$X^{(n)}(T) = \sum_{m=1}^{\infty} W_m^{(n)}(T, \tau_m),$$

where

$$W_m^{(n)}(T, \tau_m) = \begin{cases} 1 & \text{if the genus generated at time } \tau_m \text{ contains } n \text{ species at time } T \\ 0 & \text{else} \end{cases}.$$

We can get

$$E[W^{(n)}(t, \tau)] = p_{1,n}(t-\tau) = \exp(-\lambda(t-\tau))\{1 - \exp(-\lambda(t-\tau))\}^{n-1},$$

and

$$E[X^{(n)}(T)] \approx \int_0^T W^{(n)}(T, \tau) dm(\tau).$$

When T is large enough,

$$E[X^{(n)}(T)] \approx c \int_0^1 (1-y)^{n-1} y^{a/\lambda} dy,$$

where c is a constant independent of n. Therefore

$$\frac{E[X^{(1)}(T)]}{\sum_{n=1}^{\infty} E[X^{(n)}(T)]} = \frac{\int_0^1 y^{a/\lambda} dy}{\int_0^1 y^{a/\lambda - 1} dy} = \frac{1}{1 + \lambda/a}.$$

Let M denote the total number of genera and M_1 denote the number of the genera in which only one species is contained. If M and M_1 are large enough, we can estimate λ/a from the equation

$$\frac{M_1}{M} = \frac{1}{1 + \lambda/a}.$$

The estimator of λ/a is

$$\frac{\lambda}{a} = \frac{M - M_1}{M_1}\,.$$

6.4.2. *Non-homogeneous Yule process — $\lambda_i(t) = i\lambda(t)$*

Non-homogeneous Yule process $\{N(t)\}$ is a time dependent Yule process. Similarly, $N(t)$ has probability distribution

$$\begin{aligned} p_{ij}(t) &\mathrel{\widehat{=}} P\{N(t) = j | N(0) = i\} \\ &= \binom{j-1}{i-1} \exp[-(-\rho(s))]^i \{1 - \exp[-(\rho(t) - \rho(s))]\}^{j-i}\,, \end{aligned}$$

where $\rho(t) = \int_0^t \lambda(\tau) d\tau$.

6.5. *Pure death process*

A birth–death process $\{N(t)\}$ is said to be a pure death process if birth rates $\lambda_i = 0$ for all i. The pure death process is exactly analogous to the pure birth process.

In the usual applications, $N(t)$ is the number of individuals alive at time t and time t is interpreted as age. Let $\mu(t)$ denote the intensity that an individual alive at time t will die in the interval $(t, t + \Delta t)$. $\mu(t)$ is known as force of mortality, intensity of risk of dying, or failure rate. When $N(t) = i$, one death event occur in the interval $(t, t + \Delta t)$ with probability $i\mu(t)\Delta t + o(\Delta t)$. As we can see, $\{N(t)\}$ is a pure death process with death rates $i\mu(t)$. And $N(t)$ has binomial distribution

$$\begin{aligned} p_{ni}(t) &\mathrel{\widehat{=}} P\{N(t) = i | N(0) = n\} \\ &= \binom{n}{i} \exp\left(-\int_0^t \mu(\tau) d\tau\right)^i \left(1 - \exp\left(-\int_0^t \mu(\tau) d\tau\right)\right)^{n-i} \\ & i = 0, 1, \ldots, n\,. \end{aligned}$$

Suppose that i individuals are independent and have the same force of mortality. Let T denote the individual's survival time. The survival function is defined by

$$S(t) \mathrel{\widehat{=}} P\{T > t\} = 1 - F(t)\,,$$

where $F(t)$ is distribution of T. It is easy to show

$$S(t) = \exp\left(-\int_0^t \mu(\tau)d\tau\right),$$

$$f(t) = \mu(t)\exp\left(-\int_0^t \mu(\tau)d\tau\right),$$

$$\mu(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{f_t^{+\infty} f(s)ds}.$$

For example, when T has Weibull distribution

$$f(t) = \mu\gamma t^{\gamma-1}\exp(-\mu t^{\gamma-1}).$$

Then we can calculate

$$\mu(t) = \mu\gamma t^{\gamma-1},$$

and

$$S(t) = \exp(-\mu t^{\gamma-1}).$$

7. Counting Processes and Regression Models for Survival Data

7.1. *Life table*

There are two kinds of life tables, current life table and cohort life table, working for two different kinds of studies, cross-sectional study and follow-up study. For current life table, Chiang[12] proposed a method to calculate the probability of death.

$$q_i = \frac{n_i M_i}{1 + (1 - a_i)n_i M_i},$$

where q_i is age specific probability of death, the probability in $(x_i, x_i + n_i)$, and M_i is age specific death rate.

Based on stage processes, a new life table is constructed by Chiang.[12] In this new table, probability of death is not only dependent on age, but also dependent on stage of disease. The stage process is usually used to describe the development of chronic diseases. Generally, chronic diseases advance with time from mild through intermediate stages to death. The process often is irreversible but a patient may die while being in any one of stages. For example, evolution of cancer is always a stage process. There are many staging phenomena in many other areas, birth order and child spacing, engagement-marrige-divorce in demography, and so on.

7.2. *Counting process*

A process $\{N(t), t \geq 0\}$ is called a counting process if $N(t)$ represents the total number of events that occurred in $(0, t]$. A counting process must be a non-negative integer valued process. Let τ_i denotes the time of the ith event and it is said to be the arrival time of the ith event. $\tau_1, \tau_2, \ldots$ are random variables and $0 < \tau_1 < \tau_2 < \cdots$. Let

$$T_1 = \tau_1,\ T_2 = \tau_2 - \tau_1, \ldots,\ T_n = \tau_n - \tau_{n-1}, \ldots .$$

$\{T_i, i = 1, 2, \ldots\}$ is called the sequence of interarrival times.

A counting process is called a renewal process if the interarrival times have independent and identically distribution. It is called Poisson process if the distribution is exponential distribution.

Now consider k different kinds of events may occur. Let $N_i(t)$ denote the total number of the ith kind of events that occurred in $(0, t)$. $N(t) = (N_1(t), N_2(t), \ldots, N_k(t))$ is called multiple counting process with k dimensions.

Let $X_1, X_2, \ldots, X_n$ denote survival times of n individuals. They are independent and have the same survival function $S(t)$. Let

$$N(t) = \#(i : X_i \leq t) = \sum_{i=1}^{n} I(X_i \leq t)\,,$$

where $\#(\cdot)$ is a counting function and $I(\cdot)$ is a indicator function. Then $N(t)$ is the total number of death that occurs in $(0, t]$ and $\{N(t), t \geq 0\}$ is a counting process.

In survival analysis problems, compete data is not possible. We can observe that $(\tilde{X}_i, D_i)$, $i = 1, 2, \ldots, n$, where D_i is a censoring indicator. Then

$$X_i = \tilde{X}_i\,, \quad \text{if } D_i = 1\,,$$

$$X_i > \tilde{X}_i\,, \quad \text{if } D_i = 0\,.$$

Let

$$N(t) = \#\{i : \tilde{X}_i \leq t, D_i = 1\}\,.$$

7.3. *Kaplan–Meier estimator*

Kaplan–Meier estimator is a non-parametric estimator for survival function. It is also called product-limit estimator:

$$\widehat{S}(t) = \prod_{s \leq t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right),$$

where $\Delta N(s) = N(s) - N(s-)$, $Y(s) = \#\{i : \tilde{X}_i \geq t\}$ is the number at risk just before time t.

7.4. *Cox regression*

In above assumption, survival times of n individuals have identical distribution. However, survival time usually depend on some covariates. If the values of covariates are different for individuals, survival times of individuals have different distribution. In other words, survival function depends on the covariates. Regression model can be used here. When the distribution is known as a certain distribution, for example, exponential distribution, a parametric regression model can be used. However, the distribution usually is unknown. Some semiparametric models are considered. Cox regression model is a semiparametric model. In Cox regression model, the intensity of hazard is

$$\mu_i(t, z) = \mu_0(t) \exp(\beta' Z_i)\,,$$

where $Z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})'$ is covariates vector, β is regression coefficient vector and $\mu_0(t)$ is an intensity of risk independent of individuals. This model is also called proportional hazard model.

Because $\mu_0(t)$ is unknown, the estimators of parameters are based on partial likelihood function. Suppose that we observed d individuals, denoted by (1), (2), ..., (d), dead. Let $X_{(i)}$ denote the survive time of individual (i). And $X_{(1)} < X_{(2)} < \cdots < X_{(d)}$. Let $R_{(i)} = \{j : \tilde{X}_j \geq X_{(i)}\}$, the number at risk just before the time $X_{(i)}$. Then the partial likelihood function is

$$L = \prod_{i=1}^{d} P \left\{ \begin{array}{l} \text{individual } (i) \text{ die at time } X_{(i)} | \text{ one} \\ \text{individual in } R_{(i)} \text{ die at time } X_{(i)} \end{array} \right\}$$

$$= \prod_{i=1}^{d} \frac{\exp(\beta' Z_{(i)})}{\sum_{j \in R_{(i)}} \exp(\beta' Z_{(j)})}\,.$$

Then logarithm partial likelihood function is

$$\ln L = \sum_{i=1}^{d} \left\{ \beta' Z_{(i)} - \ln \left[\sum_{j \in R_{(i)}} \exp(\beta' Z_{(j)}) \right] \right\}.$$

Therefore, the maximal partial likelihood estimator of β can be calculated. It is also called Cox estimator.

Cox estimator is consistent, i.e. the estimator $\to \beta$ when the sample size $n \to \infty$.

Sometimes, the covariates are time-dependent. Some counting process models with time-dependent covariates are given next.

7.5. *Multiple renewal process model*

A multiple renewal process with time-dependent covariates is used as a model for acute respiratory infections (ARI) in the paper of Fang.[13] Consider a marked renewal process with g marks $D_1, \ldots, D_g$ corresponding g classes of diseases.

Let $\mu_{si}(t)$, $s = 1, 2, \ldots, g$, $i = 1, 2, \ldots, n$ denote the intensity that disease D_s for individual i happens at time t. Let $Z_i(t) = (z_i'(t), (t - t_{ri}), (t - t_{ri})^2, \ldots, (t - t_{ri})^q)'$. It contains a set of p time dependent covariates $z_i(t)$ and quasi-covariates, $(t - t_{ri})$, $(t - t_{ri})^2, \ldots, (t - t_{ri})^q$, representing the effect of time, where t_{ri} is the latest renewal time of individual i before time t. And let

$$\mu_{si}(t) = \exp(C_s' Z_i(t) + \theta_s)\,,$$

where $C_s = (c_{s1}, c_{s2}, \ldots, c_{s,p+q})'$ is a $p+q$ dimensional column vector and θ_s are parameters related to the occurrence of D_s and expected to be estimated from the data.

For individual i, the records in the data include the beginning and the end of observed time, t_{0i} and t_{ei}, the occurrence times $t_{1i}, t_{2i}, \ldots, t_{k_i i}$, and the corresponding states $d_{1i}, d_{2i}, \ldots, d_{k_i i}$, where k_i is the number of transitions happening. The full log-likelihood function for n individuals can be written as

$$\ln L = \sum_{\substack{i=1 \\ k_i \neq 0}}^{n} \sum_{j=1}^{k_i} \ln \mu_{d_j i}(t_{ji}) - \sum_{i=1}^{n} \int_{t_{0i}}^{t_{ei}} \left[\sum_{s=1}^{g} \mu_{si}(t) \right] dt\,.$$

As a especial case, if the parameter vectors C_s are assumed to be equal to C for all $s = 1, 2, \ldots, g$, the parameters are fewer. To estimate the vector C, a numerical method such as the Newton–Raphson algorithm is used.

The child survey data on ARI was analyzed. Eighteen covariates are dealt with, of which nine are indices of health, seven are weather indices, and the last two are $(t - t_r)$ which is length of time since the latest illness (TEF), and $(t - t_r)^2$ which is square of TEF (TEF^2), respectively, serving to explore the effect of time. The nine indices of health include hemochrome (HEM), history of tracheitis (HTR), rickets (RIC), age (AGE), history of tuberculosis (HTB), dental caries (CAR), sex (SEX), ratio of height and weight (RHW), family history of tracheitis (FHT). The seven indices of

weather include low temperature for days (LTM), range of temperature for days (RTM), relative humidity for days (RHU), difference of minimal temperatures for two days (DLT), maximal wind velocity for days (MWV), atmospheric pressure for days (ATP).

By means of the likelihood ratio test, we find that one index of health HEM, three weather indices LTM, RTM, and RHU, and the effect of time, TEF and TEF^2 are significant on the risk of occurrence of disease D. The estimates of parameter θ_s corresponding to the six types of ARI are given as -2.0759, -3.9243, -7.6897, -4.4043, -2.0402 and -6.4517. The regression coeffients are 0.5115, -0.2971, 0.3055, 0.2176, 2.2937, -4.9689, respectively.

7.6. *Markov counting process model*

In the paper of Fang *et al.*,[4] a Markov counting process with time-dependent covariates is used as a model. Let N_{ij} denote the numbers of transitions $i \to j (i \neq j)$ that occur in $(0, t]$. Let $\mu_{ijh}(t)$ denote the intensity that the transition $i \to j$ for individual h happens at time t. We assume

$$\mu_{ijh}(t, z) = \mu_{ij0}(t) \exp(\beta'_{ij} Z_{ijh}(t)) .$$

Then the partial likelihood function is

$$L = \prod_t \prod_{i,j,h} \left(\frac{\exp(\beta'_{ij} Z_{ijh}(t))}{\sum_{h=1}^{n} \exp(\beta'_{ij} Z_{si}(t)) Y_{ih}(t)} \right)^{\Delta N_{ijh}(t)} ,$$

where $N_{ijh}(t)$ is the number of transitions $i \to j (i \neq j)$ that occur in $(0, t]$ for the individual h, $\Delta N_{ijh}(t) = N_{si}(t) - N_{si}(t-)$ and Y_{ih} is indicator of the individual h at risk, corresponding state i. i.e.

$$Y_{ih}(t) = \begin{cases} 1 & \text{if individual } h \text{ at risk at time } t, \text{ corresponding state } i , \\ 0 & \text{else} . \end{cases}$$

Using this model, Fang[4] analyzed a set of 12-year and 6-run screening data of cervical cancer in Jingan county, Jiangxi Province, China. The covariates are sex disorder, sex health, age, age-square and cervicitis. There are four states and five covariates are dealt with. To obtain maximal partial likelihood estimator of parameter vector β, Marquardt modification algorithm was used. The estimators are solved as

$$\beta'_{01} = (0.1348^{8},\ -0.6567,\ 0.1088^{*},\ -0.0012^{*},\ 0.1838^{*}) ,$$

$$\beta'_{10} = (0.0450,\ -0.8642,\ 0.0542^{*},\ -0.0010^{*},\ -0.0173) ,$$

$$\beta'_{12} = (-0.2087,\ -0.7219,\ 0.0691^*,\ -0.0008^*,\ 0.5787^*)\,,$$

$$\beta'_{23} = (1.7469^*,\ 0.0000,\ 0.0544,\ 0.0014^*,\ 0.6952)\,,$$

where * means the covariate is significant.

7.7. *General multiple counting process model*

Andersen *et al.*[10] proposed a statistical model based on multiple counting process. Now consider k types of event. Let $\mu_{si}(t,z)$ denote the occurrence intensity of type s event for individual i with time-dependent covariate vector z. We assume

$$\mu_{si}(t,z) = \mu_{s0}(t,\theta) f(\beta' Z_{si}(t))\,, \quad s = 1,2,\ldots,k\,,$$

where $Z_{si}(t) = (z_{si1}(t), z_{si2}(t), \ldots, z_{sip}(t))'$ is covariate vector, β is a regression coefficient vector and θ is a parameter.

The partial likelihood function is given as

$$L = \prod_t \prod_{s,i} \left(\frac{f(\beta' Z_{si}(t))}{\sum_{i=1}^n f(\beta' Z_{si}(t)) Y_{si}(t)} \right)^{\Delta N_{si}(t)}\,,$$

where $N_{si}(t)$ is the occurrence number of type s event by time t for the individual i, $\Delta N_{si}(t) = N_{si}(t) - N_{si}(t-)$, and Y_{si} is indicator of the individual i at risk, corresponding type s event.

References

1. Cheng, J. Z., Zhou, S. Y. and Xu, H. Y. (1994). Application of Markov process in structural dynamic forecasting of forest resources with tree species structure in Nanping region of Fujian province as an example. *Chinese Journal of Applied Ecology* **5**(3): 232–236.
2. Deng, Y. L. (1994). *Stochastic Models and Their Applications*, Advanced Education Press.
3. Fang, J. Q., Mao, J. H., Zhou, W. Q. *et al.* (1995). Two-stage models, nonhomogeneous Markov chains, for cancer latent period. *Chinese Journal of Applied Probability and Statistics* **11**(2).
4. Fang, J. Q. and Wu, C. B. (1995). Two-stage model, counting process, and Bootstrap for cancer latent period. *Chinese Journal of Applied Probability and Statistics* **11**(2).
5. Fudan University (1981). *Stochastic Processes*, Advanced Education Press.
6. Gao, G. M. (2000). Probability models and Bayesian analysis of capture-recapture data via Markov chain Monte Carlo simulation. *The PhD dissertation in Sun Yat-sen University.*

7. Huang, Z. N. (1995). *Multiple Analysis in Medical*, Hunan Science and Technology Press.
8. Mao, S. S., Wang, J. L. and Pu, X. L. (1998). *Advanced Statistics*, Advanced Education Press.
9. Wang, Q. and Jin, P. H. (2000). Applications in economic hygiene for Markov model. *Chinese Journal of Health Statistics* **17**(2): 86–88.
10. Anderson, P. K., Borgan, Ø., Gill, R. D. *et al.* (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
11. Chiang, C. L. (1980). *An Introduction to Stochastic Processes and Their Application*, Robert E. Krieger Publishing Company, New York. (The Chinese version is translated by Fang, J. Q., Shanghai translation press, 1986.)
12. Chiang, C. L. (1983). *The Life Table and Its Application.* (The Chinese version is translated by Fang, J. Q., Shanghai translation press, 1984).
13. Faddy, M. J. and Fenlon, J. S. (1999). Stochastic modeling of the invasion process of nematodes in fly larvae. *Applied Statistics* **48**, Part 1: 31–37.
14. Fang, J.-Q., Shi, Z. L., Wang, Y. *et al.* (1990). Parametric inference in a renewal process with time-dependentCovariates. *Biometrics* **46**(3): 849–854.
15. Fang, J. Q., Ni, T. Y., Fan, Q. *et al.* (1996). Two-state stochastic models for memory in ion channels. *Acta Pharmacologica Sinica* **17**(1): 13–18.
16. Green, P. J. (1995). Reversible jump Markov chain Monto Carlo computation and Bayesian model determination. *Biometrika* **82**(4): 711–732.
17. Lucas, W. F. (1983). *Modules in Applied Mathematics*, Vol. 4: Life Science Models. Springer-Verlag, New York.
18. Parzen, E. (1962). *Stochastic Processes*, Holden-Day, San Francisco. (The Chinese version is translated by Deng, Y. L. and Yang, Z. M., 1987.)
19. Richard, J. B., Daniel, A. H. and Darren, J. (2000). Wilkinson, Detecting homogeneous segments in DNA sequences by hidden Markov models, *Applied Statistics* **49**, Part 2: 269–285.
20. Ross, S. M. (1983). *Stochastic Processes*, John Wiley and Sons Inc., 1983.
21. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pat. Anal. Mach. Intel.* **6**: 721–741.
22. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. *et al.* (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087–1091.
23. Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

About the Author

Jiqian Fang born in Shanghai 1939, BS of Mathematics in 1961 from the Fudan University and PhD of Biostatistics in 1985 from the University of California at Berkeley. 1985 to 1990, Professor and Director, the Department of Biostatistics and Biomathematics, Beijing Medical

University; Since 1991, Director and Chair Professor, Department of Medical Statistics, School of Public Health, Sun Yat-Sen University. His research projects covers widely various fields, including "Stochastic Models of Life Phenomena", "Gating Dynamics of Ion Channels", "Biostatistical Theory and Methods for Research on Cancer Prevention", "Bootstrap Studies on Multi-state Models", "Statistical Methods for Data on Quality of Life", "Health and Air Pollution", "Analysis of DNA Finger Printing", and "Linkage Analyses between Complex Trait and Multiple Genes", etc. Some of them have received awards from the Beijing Municipal Government or Ministry of Public Health of China for their significant advances in the biostatistics fields.

CHAPTER 27

TREE-BASED METHODS

HEPING ZHANG

Department of Epidemiology and Public Health,
Yale University School of Medicine, New Haven, CT 06520-8034
heping.zhang@yale.edu

1. Introduction

In this chapter, I describe the development and applications of tree-based methods. The thrust of these methods is the *recursive partitioning* technique that facilitates a process to divide an initially heterogeneous sample of observations into smaller subgroups within each of which the outcome of interest is relatively homogeneous.

The book by Breiman *et al.*[2] on classification and regression trees (CART^{TM}) is the milestone of the tree-based methodology. It provides much historical background and describes the methods and applications systematically. The associated CART program has become a commercial software. Since 1984, there has been a great deal of methodological developments as well as applications of tree-based methods, particularly in the area of survival trees. As in CART, the idea of recursive partitioning is still the heart and soul of the more recently developed methods. Please refer to Zhang and Singer[28] for a detailed introduction to those methods.

The rest of this chapter is organized as follows. First, the basic ingredients in CART is introduced. Then, survival trees is described. Finally, tree-based methods for analyzing multiple correlated responses is discussed.

2. The Basics of CART

One of the important and original applications of CART was to develop expert systems that can assist physicians in diagnosing patients potentially suffering heart attacks. Traditionally, the physicians made diagnoses in a

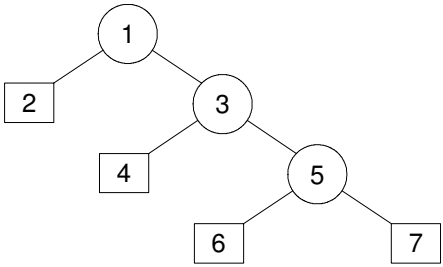


Fig. 1. A sample tree structure.

subjective, intuitive, idiosyncratic manner. A data-driven classification tree would enable the physician to interpret a patient's conditions by taking advantage of the empirical information from a large number of patients with similar conditions.[20] Now, the applications of the tree-based methods are far reaching.[1,3,4,8,11,12,14,24,27]

CART is arguably the most popular method among classification trees. All of tree-based approaches have in common the successive partitioning of a "feature space" of predictors into subsets. The partitioning is done on the basis of a *learning sample*, and it is sometimes validated by a *test sample*. Some of classification trees make use of a multi-level partition of a non-terminal node (a sub-group of the learning sample that is subject to a further division). However, only binary trees will be presented, i.e. a non-terminal node has exactly two daughter nodes. It is noteworthy that a tree that is constructed in a binary manner is not confined to be presented in the same manner as illustrated in Fig. 1 of Zhang and Bracken[27] for the sake of an easier interpretation. In other words, a multi-level partition can be derived in principle by repeated binary partitions on the same variable.

Suppose that we have collected p covariates $\mathbf{x}$ and a response y from n subjects. For the ith subject, the measurements are

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \quad \text{and} \quad y_i, i = 1, \ldots, n\,.$$

The objective is to model the probability distribution of $P(y|\mathbf{x})$ or sometimes a function of this probability. The covariates $\mathbf{x}$ can be an array of mixed categorical (nominal or ordinal) and continuous variables, and they may have missing values for some subjects. In this section, we consider a single response y of either a categorical or continuous form. Later sections will deal with censored response or multiple responses. The characteristics of y mandate the choice of methodology.

Let us begin with an arbitrary tree as depicted in Fig. 1. This tree has four layers of nodes. In general, the number of layers varies from case to case. At the top is the unique root node. Including the root node, there are three non-terminal, or internal nodes. They are represented by the circles and labeled as 1, 3, and 5. The tree has four terminal nodes, represented by boxes and labeled as 2, 4, 6, and 7. The root and the internal nodes are connected to two nodes in the next layer that are called left and right daughter nodes, whereas the terminal nodes do not have "offspring." Moreover, the tree is not necessarily symmetric in that one of the two daughter nodes can be an internal node and the other a terminal one; for instance, nodes 2 and 3 are both the daughter nodes of node 1, and node 2 is terminal whereas node 3 is internal. The connection from a parent node to the two daughter nodes is determined by a splitting rule that I will elaborate in detail shortly.

Although the details and the implementation are complex, the nutshell of tree construction is really a few key questions: (a) How are the nodes determined from the data? (b) How do we split a node? (c) When does a node become terminal? I divide the answers to these questions in two steps: tree growing and tree pruning. After a tree is constructed, we need to interpret the tree structure and make statistical inferences to reveal the relationships among the predictors and the response. This issue is important and may determine the final tree, but it belongs to the use and interpretation of trees and does not have a clear-cut answer.

2.1. *Tree growing through node splitting*

Node is the most basic element of a tree. A node is simply a collection of observations. For example, the root node contains the learning sample, namely, all of the observations that are used during a tree construction. All nodes except the root node are subsets of the learning sample. When an internal node is divided into its daughter nodes, it means that a subset of the sample is further divided into sub-groups. Because the node division is exclusive, the terminal nodes are disjoint subgroups of the learning sample and the union of all terminal nodes is the root node.

The tree growing procedure begins with the split of the root node into its two daughter nodes. Once this is done, the resulting two daughters can be split recursively in the same way. This is why the procedure is called recursive partitioning. Obviously, the fundamental step is to partition one parent node (e.g. the root node) to the two daughter nodes. How do we split a node? The division of the root node is carried out through a predictor.

The purpose of splitting is to generate two daughter nodes within which the distributions of the response are more homogeneous than that in the parent node. Every predictor in $\mathbf{x}$ competes against another to achieve a combined maximum homogeneities in the two daughters. If x_j is an ordered covariate such as age, two subgroups result from the question of the form "Is $x_j > c$?" Here the cutoff point c is in the range of the observed values of x_j. The ith subject goes to the right or left node according to whether or not $x_{ij} > c$. The number of such distinct questions that can be asked upon an ordered covariate is one fewer than the number of the distinctly observed value of x_j. On the other hand, if x_j is nominal such as nationality, we can send a subject to the left or right node by asking questions such as "is the subject an Asian?" and "is the subject a Hispanic or an African?" If x_j has k levels, we can ask $2^{k-1} - 1$ meaningfully different questions, considering that the designation of left and right daughter nodes is arbitrary. For example, if x_j has four levels, A, B, C, and D, we can make seven distinct cut as follows: $\{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{A, C\}$, and $\{A, D\}$. We do not list $\{B, C\}$ and others, because its compliment $\{A, D\}$ is listed and asking "$x_{ij} \in \{B, C\}$?" or "$x_{ij} \in \{A, D\}$?" has the same effect. Considering p covariates and the number of possible cutoff points from each of them, we see that there are usually many possibilities to split a parent node into two daughter nodes. Therefore, we need a criterion to decide which split is preferable over the rest. This leads to the concept of impurity.

Let us use age as a predictor and cancer status as the response to explain how to evaluate the splits for a node (t) based on this age predictor. Suppose that we consider an age cutoff at c, e.g. 35. As a result of the question "Is x_j (age) $> c$ (35)?", we have the following table:

		Normal	Cancer	
Left Node (t_L)	$x_j \le c$	n_{11}	n_{12}	$n_{1\cdot}$
Right Node (t_R)	$x_j > c$	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	

What do we like to see? As stated earlier, we want a split such that the distributions of y in the daughter nodes are homogeneous. In other words, we would like to push as many observations as possible either along the diagonal n_{11}, n_{22} or along the off-diagonal n_{12}, n_{21}. A perfect example is $n_{11} = n_{22} = 0$. In this case, the two nodes are pure (or completely homogeneous) because each of them contains either the cancer patients only or the normal subjects only. In contrast, their parent node includes

a mixture of n_{21} normal subjects and n_{12} cancer patients. Thus, the two daughter nodes are "more desirable" than their parent node. However, in most applications, whether the daughter nodes are more desirable than their parent node is not so clear cut and it generally requires a mathematical criterion to make the comparison.

One commonly used measure of node impurity for a categorical response is defined through the entropy function as follows:

$$i(t_L) = -\frac{n_{11}}{n_{1\cdot}} \log\left(\frac{n_{11}}{n_{1\cdot}}\right) - \frac{n_{12}}{n_{1\cdot}} \log\left(\frac{n_{12}}{n_{1\cdot}}\right). \tag{1}$$

Likewise, $i(t_R)$ and $i(t)$ can be defined. Then, we select a split that minimizes the weighted node impurity:

$$\frac{n_{\cdot 1}}{n} i(t_L) + \frac{n_{\cdot 2}}{n} i(t_R), \tag{2}$$

which can be regarded as the node splitting criterion.

For the later discussions, it is useful to note that $-i(t_L)$ is simply the maximum log likelihood of y by assuming that it follows a binomial distribution in node t_L. Minimizing criterion (2) amounts to maximizing the likelihood or homogeneity in this case.

When y is a continuous response, a node is pure when the responses within the node equal to the same constant. However, when the within-node responses are not constant, commonly used node impurity measures are the within-node variance and the absolute distance toward the median.

So far, I have described the splitting procedure based on completely observed ages from all subjects. In the presence of missing ages for some subjects, two strategies are available to deal with the splitting. One makes use of surrogate splits. The idea is this. If we cannot use age to decide how to send a subject to the left or right daughter node due to missing information, we try to find a split based on another predictor that hopefully resembles the age split sufficiently. The other strategy is much easier. We simply create another level for missing values. Then, all subjects with missing information will be sent to the same daughter node.

2.2. *Tree pruning by determining terminal nodes*

Applying the node splitting procedure described above to the root node, then to the resulting daughter nodes, and so on, we usually end up with a tree of excessive nodes. We do not need to worry about when to stop the recursive partitioning process because it stops by itself when further splitting is not possible or meaningless. In an extreme case, for example,

we cannot split a node with one observation. And, the number of study subjects is always finite. What we need to be concerned with is how to deal with a tree of excessive size. In usual practice, such a tree is too large to be useful and trustworthy. This is why we need the tree pruning step to trim off some over-fitting nodes. Let us pretend that the tree in Fig. 1 is large and subject to pruning. We need to address the question: "can we prune away some of the nodes?" If we have a general answer for this question, then we can prune any tree.

Tree pruning starts at the bottom of a tree, and the pruned tree is a subtree of the original one. Thus, pruning the tree in Fig. 1 is equivalent to selecting one of its subtrees. The latter requires a measure of tree quality, reflecting our objective of extracting homogeneous subgroups of the study sample. Whether we construct trees for classification or prediction purpose, we make our decision based on the distributions of the response in the terminal nodes. All internal nodes play an intermediate role ultimately to lead to relatively homogeneous terminal nodes. Therefore, the quality of a tree depends on the quality of its terminal nodes. Let $Q(T)$ denote a certain quality measure of tree T, and we have

$$Q(T) = \sum_{t \in \tilde{T}} p(t) q(t) \,, \tag{3}$$

where $\tilde{T}$ is the set of terminal nodes of tree T, $q(t)$ summarizes the quality of node t, and $p(t)$ is the proportion of subjects falling into node t.

For binary outcomes, $q(t)$ is usually replaced with the within-node misclassification cost $r(t)$, and $Q(T)$ with a tree misclassification cost $R(T)$. In other words, a tree is assessed by

$$R(T) = \sum_{t \in \tilde{T}} p(t) r(t) \,.$$

There are two types of misclassifications, each of which is associated with a certain misclassification cost. The misclassification cost should reflect the severity of the error, for instance, when a cancer patient is classified to be cancer free or vice versa. Let $C(i|j)$ be the misclassification cost that a class j patient is classified as a class i patient. Here, there are two classes of subjects: 0 for normal and 1 for cancer patients. For medical reasons, it is natural to choose $C(0|1) > C(1|0)$ because the consequence is potentially more severe when a patient with disease is wrongly diagnosed than when a normal person is classified to have the disease. Without loss of generality, we can set $C(1|0) = 1$ as the cost unit and let $C(0|1) = c$, which means that the

a false positive diagnosis costs as many as c false negative ones. In addition, there is no cost when the classification is correct, namely, $C(i|i) = 0$. Unless a node is pure, we makes mistakes one way or another. The within-node misclassification cost is the minimum of the two possible misclassification costs.

Although defining the measure in Eq. (3) is easy, using it is not as straightforward unless there is an independent set of sample–the so-called test sample. When a test sample is available, we can estimate $p(t)$ and $r(t)$ from it, leading to an estimate of $R(T)$. Then, we can select a subtree that has the lowest estimated cost $\hat{R}(T)$. However, in many applications, such a second sample is not feasible or is too costly. Sample re-use methods are used as an alternative. For these methods, the size of a tree is another important aspect, indicating the tree complexity. Note that the total number of nodes in a tree, T, is $2|\tilde{T}| - 1$, where $|\tilde{T}|$ is the number of the terminal nodes of T. Hence, the complexity of T can be defined directly as $|\tilde{T}|$. Usually, a unit penalty, called a complexity parameter, is assigned to each terminal node, and the sum of these penalties becomes the penalty for the tree complexity. Therefore, the final quality measure of a tree is the following cost-complexity:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|\,, \tag{4}$$

where $\alpha(> 0)$ is the complexity parameter.

For a given complexity parameter and an initial tree such as the one in Fig. 1, there is a unique smallest subtree of the initial tree that minimizes the cost-complexity measure Eq. (4). Importantly, if $\alpha_1 > \alpha_2$ the optimally pruned subtree corresponding to α_1 turns out to be a subtree of the one corresponding to α_2. So, as we increase the complexity parameter, we have a sequence of nested optimally pruned subtrees. The fact that the successive optimally pruned subtrees are nested can entail important savings in computation.[2] This nested sequence of subtrees has a finite length, because the number of subtrees is finite, and the last one is the root node. On the other hand, the complexity parameter takes a continuous value, which implies that an interval of the complexity parameter must correspond to the same subtree. Let T_0 be the initial tree. To prune off some nodes from T_0, we need to find the smallest α, denote by α_1, to allow some of the terminal nodes to be removed such that $R_{\alpha_1}(T_1)$ for the pruned T_1 is better than $R_{\alpha_1}(T_0)$ of the initial tree. It turns out

$$\alpha_1 = \min_{t \in \tilde{T}_0} \frac{r(t)p(t) - R(T(t))}{|\tilde{T}(t)| - 1}\,,$$

where $T(t)$ represents a tree rooted at node t. Likewise, we proceed to find the next smallest α, denoted by α_2, to allow some of the terminal nodes to be removed from T_1 such that $R_{\alpha_2}(T_2)$ for the pruned T_2 is better than $R_{\alpha_2}(T_1)$. As we continue this process until we reach the tree with the single root node, we end up with a sequence of increasing complexity parameters $\{\alpha_i\}_0^m$ (here $\alpha_0 = 0$) and a sequence of nested and shrinking subtrees $\{T_i\}_0^m$ (here T_m is the single root node tree).

The next step is to select a subtree from the nested sequence, and a cross-validation procedure is usually recommended. For example, we can randomly divide the study sample into several, say 5, sub-samples of about the same size. We use 4 of the 5 sub-samples to grow a large tree and prune it using the sequence $\{\alpha_i\}_0^m$ that leads to a new sequence of subtrees. Then, we compute $R(T)$ for each of those subtrees based on the left-over subsample, giving us one set of estimates for $\{R(T_i)\}_0^m$. We can do this 5 times and the average will be the final estimates for $\{R(T_i)\}_0^m$. With this sequence of estimates, we can select the subtree with the smallest or near the smallest $\hat{R}(T)$. Please refer to Breiman *et al.*[2] and Zhang and Singer[28] for details. Once the subtree is selected, the pruning step is accomplished.

3. Survival Trees

Although CART is a well-known brand name, the most frequently used tree-based method in biomedical research is survival trees, partly because survival analysis per se is a major topic in the health sciences. In this section, how to adapt the ideas expressed above for censored survival data will be explained. We face the same basic issues. One is to define a splitting criterion to divide a node into two, and the other is to choose a "right-sized" tree for subsequent use. Many criteria have been proposed in the literature, but they differ primarily in the way of declaring which daughter nodes are desirable. A few major ideas have emerged. First, as in CART, we can split a node to achieve better impurities in its daughter nodes. The concept of impurity is very intuitive in CART; however, for survival trees, we have to decide what we mean by node impurity. The second idea is to maximize the distributional difference between the two daughter nodes. In classical ANOVA, reducing within-group variances increases the between-group variances. But, for survival trees, it is not clear that reducing node impurity increases the distributional difference between the two daughter nodes. Finally, as hinted earlier, there is a connection between node impurity and

maximum likelihood in CART. Although this connection may not hold in survival trees, we can nonetheless base our splitting decision on likelihood. In the following, the focus will be on presenting the approaches that have been implemented in the author's STREE program. Also see Zhang and Singer[28] and Zhang *et al.*[25]

3.1. *Use of the difference*

I begin the introduction of splitting rules with the use of Wasserstein metrics to measure the between-nodes distributional difference as proposed by Gordon and Olshen.[13] The within-node survival distribution is estimated by the Kaplan–Meier curve.[17] A desirable split can be characterized as one that results in two very different survival functions in the daughter nodes. Gordon and Olshen[13] used the so-called L^p Wasserstein metrics, $d_p(\cdot\,,\cdot)$, as the measure of discrepancy between the two survival functions. Specifically, for $p = 1$, the Wasserstein distance, $d_1(S_L, S_R)$, between two Kaplan–Meier curves, S_L and S_R, is the area sandwiched by the two Kaplan–Meier curves. Suppose that S_L and S_R are respectively the Kaplan–Meier curves for the left and right daughter nodes. We choose the split that maximizes the distance, $d_1(S_L, S_R)$. As before, we employ the recursive partitioning process to produce an initially large tree that will be pruned later.

A standard approach for comparing the survival times of two groups is the log-rank test. Thus, it is no surprise in the literature that the log-rank test is also used to separate the left and right daughter nodes. Indeed, Ciampi *et al.*[6] and Segal[18] adopted the log-rank test statistic as the splitting criterion.

3.2. *Use of likelihood functions*

One very flexible way of forming a splitting criterion is to use likelihood functions. Not only is this true for survival trees, but it is also the case for analyzing more complex responses. This approach is useful and convenient because we can assume a simple within-node distribution when we assess a split or node. In fact, a few likelihood based splitting and pruning criteria have been proposed. Davis and Anderson[9] assume that the survival function within any given node is an exponential function with a constant hazard. LeBlanc and Crowley[15] and Ciampi *et al.*[7] assume the proportional hazard models in two daughter nodes, but the hazard functions are unknown, but they respectively used the full and partial likelihoods for maximization.

3.3. *Use of impurity*

Note that we observe a binary death indicator and the (failure or censored) time. If we take these two outcomes separately for the moment, we can compute the within-node impurity, i_δ, of the death indicator and the within-node variation, i_y, of the time toward the median. By considering both of them together, we have the within-node impurity for both the death indicator and the time using a weighted combination, $w_\delta i_\delta + w_y i_y$. Zhang[21] examined a similar approach and recommended some choices of weights w_δ and w_y. Even though this approach does not fully incorporate the relationship between the censoring and observed time variables, existing evidence suggests that this simple extension outperforms the more sophisticated ones in discovering the underlying structures of data.

3.4. *Pruning survival trees*

I explained various ways of growing a survival tree. There is also a variety of options to prune a survival tree. We need the same recipes as before: a quality measure and a cost-complexity of a tree. They enable us to use the cross validation procedure again to finish the pruning step.

Gordon and Olshen[13] suggested using the deviation of survival times toward their median as a measure of node quality $q(t)$ for a node t and model (4) as the cost-complexity where $R(T)$ is taken as $Q(T)$.

In addition, a variety of tree cost-complexities has also been proposed by using the likelihood ratio statistic that compares the survival times in a parent node with those in its daughter nodes. A related method due to Therneau *et al.*[19] makes use of what are termed martingale residuals from the Cox model as the input to a cost-complexity scheme using least squares as the cost.

LeBlanc and Crowley[16] introduced the notion of "goodness-of-split" complexity as a substitute for cost-complexity in pruning the tree. Now, let $q(t)$ be the value of the log-rank test at node t. Then the split-complexity measure is

$$Q(T) = \sum_{t \in \tilde{T}} q(t) - \alpha(|\tilde{T}| - 1)\,,$$

where the summation above is over the set of internal (non-terminal) nodes rather than the terminal nodes as in Eq. (4). The negative sign in front of the complexity part is a reflection of the fact that Q is to be maximized here, whereas the cost-complexity R is minimized there. In CART, cross

validation is used to determine an optimal complexity parameter value, while here LeBlanc and Crowley[16] recommend choosing α between 2 and 4. I refer to their work for the justification of this choice.

Although the tree pruning procedures as described above are statistically elegant, they are rather sophisticated for researchers outside of the statistical society to comprehend. Even within the statistical community, we do not necessarily agree upon the correct use of these procedures. Thus, a practical and intuitive approach is appealing. Segal[18] recommended the following alternative for pruning survival trees. For each internal node (including the root node) of an initial tree, we assign it a value that equals the maximum of the log-rank statistics over all splits starting from the internal node of interest. Then, plot the values for all internal nodes in an increasing order and decide a threshold from the graph. If an internal node corresponds to a smaller value than the threshold, we prune all of its offspring. Although this usually works out fine, the choice of threshold could be arbitrary. In the author's RTEE program, pruning a tree (not necessarily a survival tree) at a different significance level is chosen. Analyzing genetic data, Zhang and Bonney[23] demonstrated how to decide a final tree based on both the scientific implication and computer output.

It is clear that there are plenty of choices to construct survival trees, which is good and bad. More choices give the data analysts the opportunity to select the ones that make a better scientific sense. Sometimes, however, too many choices can also lead the data analysts to wonder what to do. The state of the art is still to construct survival trees using a number of approaches and discuss them with experts to come up one or more trees that are as simple, informative, and interpretable as possible.

4. Classification Trees for Multiple Binary Responses

In this section, I introduce a tree-based approach for analyzing multiple correlated binary outcomes. Such outcomes are sometimes referred to as clustered outcomes. Correlated discrete responses can be generated from a single endpoint by repeatedly measuring it on individuals in a temporal or spatial domain. They are called longitudinal discrete responses. The correlated responses may also consist of distinct endpoints, which are actually the focus of this section.

Suppose that $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iq_i})'$ is a vector of binary responses for subject i, $i = 1, \ldots, n$. The index length q_i may vary from individual to individual. This is particularly useful when some responses are missing for

some individuals. Multiple correlated outcomes arise from many applications. For example, (Y_1, Y_2) may indicate the blindness of the left and right eyes or cancer status for a sib pair.

We have seen in both classification and survival trees that some of the simple parametric distributions form the foundations for tree constructions. For multiple correlated responses, the following distribution from an exponential family is proven useful:

$$P\{\mathbf{Y}_i = \mathbf{y}_i\} = \exp\left[\sum_{j=1}^{q_i} \theta_{ij} y_{ij} + \sum_{j_1<j_2} \theta_{ij_1j_2} y_{ij_1} y_{ij_2} + \cdots + \theta_{i1\cdots q_i} y_{i1} \cdots y_{iq_i} + A_i(\theta_i)\right], \tag{5}$$

where

$$\theta_i = (\theta_{i1}, \ldots, \theta_{iq_i}, \theta_{i12} \cdots \theta_{i,q_i-1,q_i}, \ldots, \theta_{1\cdots q_i})$$

is the $(2^{q_i-1} - 1)$-vector of canonical parameters and $\exp[A_i(\theta_i)]$ is the normalizing constant. The above model is commonly referred to as log-linear model.

In lieu of extensive search of node splitting, it is important that the distribution used to form the splitting criterion is as simple as possible. Even in more traditional, parametric data analyses, a much simplified version of model (5) is generally used by setting the canonical parameters with respect to the terms with the third- or higher-orders to zero.[10,29] Thus, Zhang[22] considered the following quadratic exponential model:

$$P\{\mathbf{Y} = \mathbf{y}\} = \exp\left[\sum_{j=1}^{q} \theta_j y_j + \sum_{j<k} \theta_{jk} y_j y_k + A(\Psi, \theta)\right], \tag{6}$$

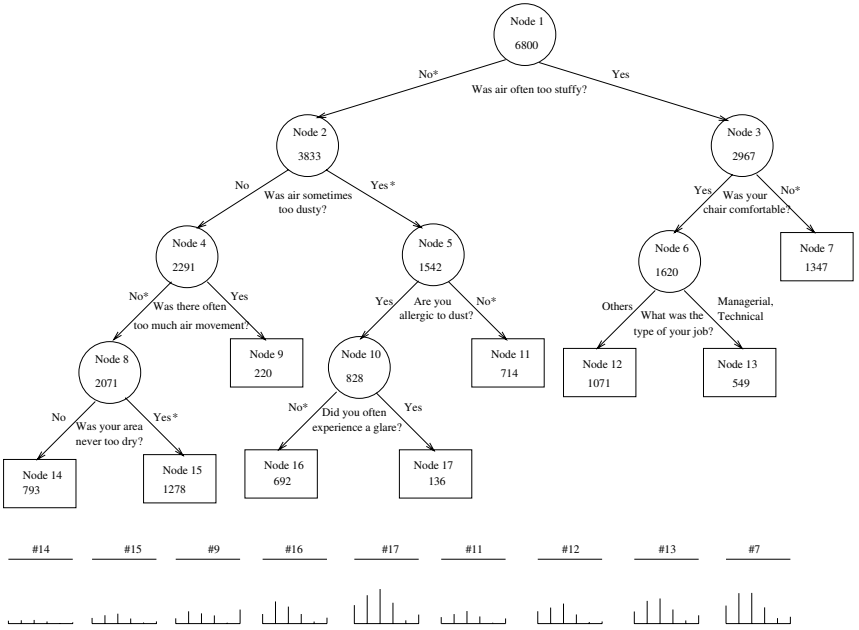
where

$$\Psi = (\theta_1, \ldots, \theta_q)', \quad \theta = (\theta_{12} \cdots \theta_{q-1,q}).$$

He defined the generalized entropy criterion of node t as the maximum of the log-likelihood derived from this distribution, which equals

$$h(t) = \sum_{\{\text{subject } i \in t\}} (\hat{\Psi}' \mathbf{y}_i + \hat{\theta} w_i - A(\hat{\Psi}, \hat{\theta})), \tag{7}$$

where $\hat{\Psi}$ and $\hat{\theta}$ are the maximum likelihood estimates of Ψ and θ, respectively. Then, he chose a split to maximize $h(t_L) + h(t_R)$.



Terminal	Node	Cluster of symptoms					
node No.	Size	CNS	UA	Pain	Flu	Eyes	LA
7	1347	377	637	642	340	114	143
9	220	18	42	35	28	3	48
11	714	72	106	139	79	10	57
12	1071	206	267	333	152	27	35
13	549	103	194	214	120	27	71
14	793	36	41	45	26	2	16
15	1278	113	166	197	101	22	43
16	692	103	238	182	103	22	64
17	136	39	60	73	44	7	19

Fig. 2. Tree Structure for the Risk of Symptoms.[22] Inside each node (◯ or a □) are the node number and the number of subjects. The splitting question is given under the node. The asterisks indicate where the subjects with missing information are assigned. The nine diagrams under the tree show the incidence rates of the six clusters (CNS, upper airway, pain, flu-like, eyes, and lower airway) in the nine terminal nodes. The top and bottom lines in each diagram define the unit of 1.

To determine the terminal nodes, Zhang[22] defined the tree quality measure as:

$$R(T) = -\sum_{t \in \tilde{T}} h(t)\,. \tag{8}$$

To illustrate the use of this classification tree approach, Zhang[22] analyzed a subset of the data from a 1989 survey of 6,800 employees of the Library of Congress and the headquarters of the Environmental Protection Agency in the United States. The response variables are six cluster indicators for building-related occupant complaint syndrome, which is a nonspecific set of related symptoms of discomfort reported by occupants of buildings. The six clusters are: central nerve system, upper airway, pain, flu-like, eyes, and lower airway. Zhang[22] considered 22 predictors including those used in Fig. 2 as the risk factors. Please refer to Zhang[22] for more details.

Figure 2 reveals that terminal nodes 7 and 17 have the highest incidences of symptoms. The table below the figure gives the number of symptoms reported in each terminal node and for each cluster. The respondents in terminal nodes 7 and 17 experienced more problems in nearly all clusters than others. The figure shows that it is because the air quality in their working area was poor, namely, often too stuffy or sometimes dusty. On the other hand, respondents in terminal node 14 had the least discomfort because they had the best air quality.

The subjects in terminal nodes 16 and 17 were allergic to dust whereas those in terminal node 11 were not. Due to this personal difference, many more symptoms in the central nervous system, upper airway, pain, and flu-like were reported among the allergic subgroups than among the non-allergic ones. Overall, the incidence rate of the eye symptoms is very low, and it appears to be mostly related to air stuffiness as shown in Fig. 2. In fact, this figure reveals a lot more information about the symptom incidents than what is mentioned here. Please refer to Zhang[22] for a detailed analysis.

5. Final Remarks

The tree-based methods have become increasingly popular in medical research.[1,3–5,8,14,26,27] In addition, they have proven to be very useful in machine learning, marketing, finance, etc. The tree-based methods may become one of the standard analytic choices, but they likely complement rather than replace the classic statistical methods such as logistic regression models and Cox proportional hazard models. The tree-based methods

enable us to produce intuitive and interpretable tree structures without making restrictive parametric assumptions as in the classic models. For the same reason, however, it is more difficult to make statistical inference based on tree structures.

References

1. Bacchetti, P. and Segal, M. R. (1995). Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of aids. *Lifetime Data Anal.* **1**: 35–47.
2. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, California. Since 1993 this book has been published by Chapman and Hall, New York.
3. Carmelli, D., Halpern, J., Swan, G. E. *et al.* (1991). 27-year mortality in the western collaborative group study: Construction of risk groups by recursive partitioning. *Journal of Clinical Epidemiology* **44**: 1341–1351.
4. Carmelli, D., Zhang, H. P. and Swan, G. E. (1997). Obesity and 33 years of coronary heart disease and cancer mortality in the western collaborative group study. *Epidemiology.*
5. Chou, P. A., Lookabaugh, T. and Gray, R. M. (1989). Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transaction Information Theory* **35**: 299–315.
6. Ciampi, A., Thiffault, J., Nakache J.-P. and Asselain B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics and Data Analysis* **4**: 185–204.
7. Ciampi, A., Hogg, S., McKinney, S. and Thiffault, J. (1988). A computer program for recursive partition and amalgamation for censored survival data. *Computer Methods and Programs in Biomedicine* **26**: 239–256.
8. Curran, W. J. Jr., Scott, C. B., Horton, J. *et al.* (1993). Recursive partitioning analysis of prognostic factors in three radiation therapy oncology group malignant glioma trials. *Journal of the National Cancer Institute* **85**: 704–710.
9. Davis, R. and Anderson, J. *et al.* (1989). Exponential survival trees. *Statistics in Medicine* **8**: 947–962.
10. Fitzmaurice, G. and Laird, N. M. (1995). Regression Models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**: 845–852.
11. Goldman, L., Cook, F., Johnson, P., Brand, D., Rouan, G. and Lee, T. (1996). Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain. *NEJM* **334**: 1498–504.
12. Goldman, L., Weinberg, M., Olshen, R. A., Cook, F., Sargent, R. *et al.* (1982). A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *NEJM* **307**: 588–597.
13. Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports* **69**: 1065–1069.

14. Kwak, L. W., Halpern, J., Olshen, R. A. and Horning, S. J. (1990). Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: Results of a tree-structured survival analysis. *Journal of Clinical Oncology* **8**: 963–977.
15. LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics* **48**: 411–425.
16. LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness-of-split. *Journal of the American Statistical Association* **88**: 457–467.
17. Miller, R. G. (1981). *Survival Analysis*, Wiley, New York.
18. Segal, M. R. (1988). Regression trees for censored data. *Biometrics* **44**: 35–48.
19. Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**: 147–160.
20. Wasson, J. H., Sox, H. C., Neff, R. K., Goldman, L. (1985). Clinical prediction rules: Applications and methodologic standards. *The New England Journal of Medicine* **313**: 793–799.
21. Zhang, H. P. (1995). Splitting criteria in survival trees. In *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling*, Innsbruck, Austria, July 1995, Springer-Verlag, pp. 305–314.
22. Zhang, H. P. (1988). Classification trees for multiple binary responses. *Journal of the American Statistical Association* **93**: 180–193.
23. Zhang, H. P. and Bonney, G. (2000). Use of classification trees for association studies. *Genetic Epidemiology.*
24. Zhang, H. P. and Bracken, M. B. (1995). Tree-based risk factor analysis of preterm delivery and small-for-gestational-age birth. *American Journal of Epidemiology* **141**: 70–78, 1995.
25. Zhang, H. P., Crowley, J., Sox, H. and Olshen, R. A. (1998). Tree structured statistical methods. *Encyclopedia of Biostatistics* **6**: 4561–4573, Wiley, Chichester, England.
26. Zhang, H. P., Holford, T. and Bracken, M. B. (1996). A tree-based methods of analysis for prospective studies. *Statistics in Medicine* **15**: 37–49.
27. Zhang, H. P. and Bracken, M. B. (1996). Tree-based, two-stage risk factor analysis for spontaneous abortion. *American Journal of Epidemiology* **144**: 989–996.
28. Zhang, H. P. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences*, Springer, New York.
29. Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**: 642–648.

About the Author

Heping Zhang is a tenured Associate Professor of Biostatistics, Statistics, and Child Study at Yale University. He received his BS in Mathematics from Jiangxi Normal University in 1982 and his PhD in Statistics from Stanford University in 1991. He is a fellow of the American Statistical Association, an elected member of the International Statistical Institute, and a member

of the board of director of the International Chinese Statistical Association. He currently serves on the editorial board of *Biometrics* and *Current Index to Statistics*, and a study section of the National Institutes of Health. His research interests include nonparametric methods, longitudinal data, statistical genetics and bioinformatics, statistical modeling of epidemiological data, brain imaging analysis, statistical computation, and statistical methods in behavioral sciences. He is an author of a book on *Recursive Partitioning* in the Health Sciences published by Springer, and has also published extensively in statistical, genetic, epidemiological, and psychiatric journals from the *Annals of Statistics* and the *American Journal of Statistical Association* to *Science*.

This page intentionally left blank

CHAPTER 28

MAXIMUM LIKELIHOOD ESTIMATION FROM INCOMPLETE DATA VIA EM-TYPE ALGORITHMS

CHUANHAI LIU

Rm 2C-262, Bell Laboratories, Lucent Technologies,
700 Mountain Av., Murray Hill, NJ 07974-0636
Tel: (908) 582-3986; liu@research.bell-labs.com

This article reviews EM-type algorithms, including the Expectation-Maximization (EM), Expectation-Conditional-Maximization (ECM), Expectation-Conditional-Maximization-Either (ECME), and Parameter-Expanded-Expectation-Maximization (PX-EM) algorithms, which are popular tools for modal estimation from incomplete data. These algorithms are presented along with maximum likelihood estimation of the t-distribution, which has played an important role in the development of EM-type algorithms and robust estimation. Existing algorithms are reviewed and new algorithms are proposed for maximum likelihood estimation of the general linear mixed-effects models, which has become a popular tool for analyzing repeated measures and longitudinal data in many fields such as biology and medicine.

1. Introduction

Incomplete data are pervasive in scientific investigations for many reasons. Unexpected interruptions of scheduled experiments create fully missing values; limitations of measurement methods when values are below detection limits produce censored values; and the use of latent variables in modeling data with complex structures introduces unobservable values.

Missing data make it difficult to analyze incomplete data using complete-data methods. Numerous statistical methods have been developed in the last twenty five years for dealing with missing data problems. Rubin[33–35] and Little and Rubin[10] provided fundamental principles in analyzing incomplete data. Dempster, Laird, and Rubin[1] formulized the Expectation-Maximization (EM) algorithm for maximum likelihood estimation from incomplete data. The simplicity and stability of EM may well

explain both the unprecedented explosion of applications of EM in last two decades[26] and the comprehensive expansion and extension of EM in last decade.[29]

The EM algorithm has been applied to fitting a wide range of statistical models, including multivariate normal distributions with incomplete data;[1,10,15] factor analysis;[1,20,21,37] general linear mixed-effects models;[1,5,6,21,29] loglinear models and general location models;[1,20] Poisson imaging models and their extensions;[7,16,39] mixture models;[1,22,41] and (multivariate) t-distributions.[8,12,17,19,21,29,36] While not all inclusive, these references are useful starting points for understanding the power of the EM algorithm.

The expansions and extensions of EM in last decade include the ECM algorithm,[28] the ECME algorithm,[18] the AECM algorithm,[29] and the PX-EM algorithm.[21] The Data-Augmentation (DA) algorithm[40] and its extensions, such as the Gibbs sampler,[3] can also be viewed as stochastic versions of EM-type algorithms. These are useful for fitting Bayesian models.

The rest of the article is arranged as follows. Section 2 discusses the t-distribution, which is used throughout to explain the EM, ECM, ECME, ACEM, and PX-EM algorithms. Section 3 reviews EM-type algorithms and illustrates their application with the t-distribution. Section 4 considers the general linear mixed effects models. Existing algorithms are reviewed. New algorithms are proposed. Finally, Sec. 5 concludes with a brief discussion.

2. A Gamma-Normal Hierarchical Model and the t Distribution

The t-distribution is a useful model for data analysis, especially for robust estimation.[8,11,17,36] The development of likelihood-based methods for estimation of the multivariate t distribution has also stimulated many methods that more efficient, such as the ECME algorithm of Liu and Rubin,[18] the efficient data augmentation algorithm of Meng and van Dyk,[29] and the PX-EM algorithm of Liu, Rubin, and Wu.[21] This section describes a simple model with the univariate t-distribution, which can be obtained from a Gamma-normal hierarchical model.

2.1. *A Gamma-normal hierarchical model*

Suppose that (i) the random variable τ follows the χ^2_ν-distribution, or more generally, the Gamma distribution

$$\tau \sim \hat{\mathrm{G}}(\nu/2, \nu/2)\,; \tag{1}$$

with $\nu(\nu > 0)$ number of degrees of freedom and density function

$$h(\tau|\nu) = \frac{1}{\Gamma(\nu/2)} \left(\frac{\nu}{2}\right)^{\nu/2} \tau^{\nu/2-1} \exp\left\{-\frac{\nu\tau}{2}\right\}, \quad \tau \in (0, \infty),$$

where $\Gamma(.)$ denotes the Gamma function. Also suppose that (ii), given τ, the conditional distribution of z follows the normal distribution with mean μ and variance σ^2/τ; that is,

$$y|(\tau, \mu, \sigma^2, \nu) \sim \mathrm{N}(\mu, \sigma^2/\tau).$$

It follows that the joint distribution of (τ, y) has the density function

$$f(\tau, y|\mu, \sigma^2, \nu) = \frac{1}{\Gamma(\nu/2)} \left(\frac{\nu}{2}\right)^{\nu/2} \tau^{\nu/2-1} \exp\left\{-\frac{\nu\tau}{2}\right\}$$
$$\times \frac{1}{(2\pi\sigma^2/\tau)^{1/2}} \exp\left\{-\frac{\tau(y-\mu)^2}{\sigma^2}\right\}.$$

Maximum likelihood estimation of the parameters μ and σ in the joint distribution of (τ, y) with known ν from a sample of n observations $X = \{(\tau_i, y_i) : i = 1, \ldots, n\}$ is straightforward. The loglikelihood function of the parameter (μ, σ^2), ignoring constants, is

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \tau_i y_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} \tau_i y_i - \frac{\mu^2}{2\sigma^2} \sum_{i=1}^{n} \tau_i.$$

Thus, the joint distribution of X belongs to the exponential family. The loglikelihood is linear in the sufficient statistics

$$S_{\tau y} = \sum_{i=1}^{n} \tau_i y_i, \quad S_{\tau y^2} = \sum_{i=1}^{n} \tau_i y_i^2, \quad \text{and} \quad S_\tau = \sum_{i=1}^{n} \tau_i \tag{2}$$

for the unknown parameter (μ, σ^2). The maximum likelihood estimate $(\hat{\mu}, \hat{\sigma}^2)$ of (μ, σ^2) can be written, in terms of the sufficient statistics as follows,

$$\hat{\mu} = \frac{S_{\tau y}}{S_\tau} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\left[S_{\tau y^2} - \frac{S_{\tau y}^2}{S_\tau}\right]. \tag{3}$$

The solution is known as weighted least squares.

2.2. *The t-distribution*

Suppose that $\tau_1, \ldots,$ and τ_n in the sample X are missing, and thereby the observed data Y consist of $y_1, \ldots,$ and y_n, that is, $Y = \{y_i : i = 1, \ldots, n\}$.

The marginal distribution of y is $t(\mu, \sigma^2, \nu)$, which has the density function

$$\begin{aligned} g(y|\mu,\sigma^2,\nu) &= \int_0^\infty f(\tau, y|\mu,\sigma^2,\nu)d\tau \\ &= \int_0^\infty \frac{1}{\Gamma(\nu/2)}\left(\frac{\nu}{2}\right)^{\nu/2}\tau^{\nu/2-1} \\ &\quad \times \exp\left\{-\frac{\nu\tau}{2}\right\}\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{\tau(y-\mu)^2}{\sigma^2}\right\}d\tau \\ &= \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)(\nu\pi\sigma^2)^{1/2}}\left[1+\frac{(y-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2}, \\ & y \in (-\infty, \infty), \end{aligned} \tag{4}$$

where μ and σ are the center and scale parameters, respectively, and ν is the number of degrees of freedom. Thus, with known degrees of freedom ν, the loglikelihood function of the parameter (μ, σ^2), ignoring constants, can be written as

$$L(\mu,\sigma^2) = -\frac{n}{2}\ln(\sigma^2) - \frac{\nu+1}{2}\ln[1+(y_i-\mu)^2/(\nu\sigma^2)].$$

It is difficult to find the maximum likelihood estimate of (μ, σ^2), $(\hat{\mu}, \hat{\sigma}^2)$ that maximizes $L(\mu, \sigma^2)$ over (μ, σ^2). Making use of the complete data X, the EM algorithm provides an iterative procedure to find the maximum likelihood estimates of μ and σ^2.

3. The EM Algorithm and its Extensions

3.1. *The EM algorithm*

Let X be the complete-data with the density $f(X|\theta)$ and the sample space $\mathcal{X}$, where the parameter θ lies in the parameter space $\Theta \subset R^d$, the d-dimensional Euclidean space; and let Y be the observed incomplete-data, which is obtained by a many-to-one mapping $Y = Y(X)$ from $\mathcal{X}$ to $\mathcal{Y}$, the sample space of Y. Also let $g(Y|\theta)$ denote the density of Y. Then

$$g(Y|\theta) = \int_{\mathcal{X}(Y)} f(X|\theta)dX,$$

where $\mathcal{X}(Y) = \{X : X \in \mathcal{X}, Y(X) = Y\}$. The objective is to find the maximum likelihood estimate $\hat{\theta}$ of θ, which maximizes the loglikelihood

$$L(\theta) \equiv \ln[g(Y|\theta)].$$

Denote by $k(X|Y,\theta)$ the conditional density of X given Y. Then

$$f(X|\theta) = g(Y|\theta) \cdot k(X|Y,\theta)\,.$$

For any $\theta' \in \Theta$, the loglikelihood can be written as

$$L(\theta) = Q(\theta|\theta') - H(\theta|\theta')\,,$$

where

$$\begin{aligned} Q(\theta|\theta') &= E\{\ln f(X|\theta)|Y,\theta'\} \\ &= \int_{\mathcal{X}(Y)} \ln[f(X|\theta)]k(X|Y,\theta')dX \end{aligned}$$

is the expected complete-data loglikelihood, and

$$\begin{aligned} H(\theta|\theta') &= E\{\ln k(X|Y,\theta)|Y,\theta'\} \\ &= \int_{\mathcal{X}(Y)} \ln[k(X|Y,\theta)]k(X|Y,\theta')dX \end{aligned}$$

is the expected missing-data loglikelihood.

The EM algorithm maximizes $L(\theta)$ by iteratively maximizing $Q(\theta|\theta')$ over θ with θ' replaced with the current estimate of θ. More precisely, starting with $\theta^{(0)} \in \Theta$, the tth ($t \geq 1$) iteration of the EM algorithm consists of two steps: An Expectation (E) step and a Maximization (M) step.

3.1.1. *The EM algorithm*

E step: Compute the expected complete-data loglikelihood $Q(\theta|\theta^{(t-1)})$.
M step: Find $\theta^{(t)}$ that maximizes $Q(\theta|\theta^{(t-1)})$ over $\theta \in \Theta$.

Dempster *et al.*[1] showed that (i) each iteration of EM increases $L(\theta)$, which implies that EM is stable, and (ii) if EM converges to θ^*, then θ^* is a (local) maximum of $L(\theta)$ (see Wu (1983) for more discussion).

When the complete-data model for X belongs to the exponential family, the complete-data loglikelihood of the parameter θ is linear in a set of sufficient statistics $T(X)$ for θ. Denote by $\hat{\theta}(T(X))$ the complete-data maximum likelihood estimate of θ. Then the EM algorithm can be written as follows:

3.1.2. *The EM algorithm (for exponential family)*

E step: Compute the expected complete-data sufficient statistics $T^{(t)}(X) = \mathrm{E}\{T(X)|\theta^{(t-1)}, Y\}$.

M step: Calculate $\theta^{(t)} = \hat{\theta}(T^{(t)}(X))$.

For the t-distribution with the sample $Y = \{y_i : i = 1, \ldots, n\}$, the sufficient statistics are linear functions of the missing data $\{\tau_i : i = 1, \ldots, n\}$. Given the observed data Y, the missing data $\tau_1, \ldots,$ and τ_n are independent, and for each $i = 1, \ldots, n$,

$$\tau_i | (Y, \theta = (\mu^{(t-1)}, (\sigma^2)^{(t-1)})) \sim \hat{\mathrm{G}}\left(\frac{\nu+1}{2}, \frac{\nu + (y_i - \mu^{(t-1)})^2/(\sigma^2)^{(t-1)}}{2}\right),$$

which leads to

$$\tau_i^{(t)} \equiv \mathrm{E}\{\tau_i | Y, \theta = (\mu^{(t-1)}, (\sigma^2)^{(t-1)})\} = \frac{\nu + (y_i - \mu^{(t-1)})^2/(\sigma^2)^{(t-1)}}{\nu+1}.$$

Then, the EM algorithm for the t-distribution with known degrees of freedom is given as follows.

3.1.3. *The EM algorithm (for the t-distribution with known degrees of freedom)*

E step: Compute the expected complete-data sufficient statistics $S_{\tau y}^{(t)} = \sum_{i=1}^n \tau_i^{(t)} y_i$, $S_{\tau y^2}^{(t)} = \sum_{i=1}^n \tau_i^{(t)} y_i^2$, and $S_\tau^{(t)} = \sum_{i=1}^n \tau_i^{(t)}$.

M step: Calculate

$$\mu^{(t)} = \frac{S_{\tau y}^{(t)}}{S_\tau^{(t)}} = \frac{\sum_{i=1}^n \tau_i^{(t)} y_i}{\sum_{i=1}^n \tau_i^{(t)}},$$

and

$$(\sigma^2)^{(t)} = \frac{1}{n}\left[S_{\tau y^2}^{(t)} - \frac{(S_{\tau y}^{(t)})^2}{S_\tau^{(t)}}\right] = \frac{\sum_{i=1}^n \tau_i^{(t)} (y_i - \mu^{(t)})^2}{n}.$$

This iterative procedure is known as re-weighted least squares.[36]

3.2. *The ECM, ECME, and ACEM algorithms*

When the M-step of EM is difficult, it can be replaced with a sequence of constrained (on some functions of parameters) maximizations of the Q function, called CM-steps. This extension of the EM algorithm is called the ECM algorithm by Meng and Rubin.[28] Liu and Rubin[18] realized that an algorithm that converges faster can be obtained by replacing some of

CM-steps of ECM with CM-steps that maximize the corresponding constrained actual loglikelihood. For the sake of convenience, we call a step that maximizes a constrained expected complete-data log-likelihood a CMQ-step and a step maximizing a constrained actual loglikelihood a CML-step. This extension of EM and ECM is called the ECME algorithm by Liu and Rubin,[18] with "E" for "either". As noticed by Meng and van Dyk,[29] an E-step is generally required after a sequence of CML-steps and before a call to a sequence of CMQ-steps to guarantee the monotone convergence of the likelihood. Starting with the CML algorithm that iteratively maximizes constrained actual loglikelihood functions, Fessler and Hero[2] considered an EM-step, i.e. an iteration of the EM algorithm, for each CML-step and proposed the SAGE algorithm (for Space-Alternating Generalized EM). Meng and van Dyk[29] extended EM, ECM, ECME, and SAGE further to allow data-augmentation schemes as well as the constraining functions for the CM-steps to vary from a CM-step to another CM-step. They call this algorithm AECM (for Alternating ECM).

Suppose that the number of degrees of freedom ν of the t-distribution is also unknown. The complete-data sufficient statistics for $\theta = (\mu, \sigma^2, \nu)$ are $T(X) = (S_\tau, S_{\tau y}, S_{\tau y^2}, S_{\ln\tau - \tau})$, where S_τ, $S_{\tau y}$, and $S_{\tau y^2}$ are given in Condition (2) and

$$S_{\ln\tau-\tau} = \sum_{i=1}^{n} [\ln(\tau_i) - \tau_i] \,.$$

The complete-data maximum likelihood estimates of μ and σ^2 are the same as those with known number of degrees of freedom given in Eq. (3). The complete-data maximum likelihood estimate of ν is obtained by maximizing

$$Q(\nu) = -n \ln(\Gamma(\nu/2)) + \frac{n\nu}{2} \ln(\nu/2) + \frac{\nu}{2} S_{\ln\tau-\tau}$$

by a one-dimensional search, e.g. the Newton–Raphson method. If the parameter space of $\theta = (\mu, \sigma^2, \nu)$ is partitioned as $\theta_1 = (\mu, \sigma^2)$ and $\theta_2 = \nu$, then the EM algorithm can be used to find the maximum likelihood estimate of $\theta = (\mu, \sigma^2, \nu)$. The extra computation involved in the E-step is the evaluation of

$$\begin{aligned} S^{(t)}_{\ln\tau-\tau} &\equiv \mathrm{E}\{S_{\ln\tau-\tau} | Y, \theta^{(t-1)}\} \\ &= \sum_{i=1}^{n} [\phi((\nu^{(t-1)}+1)/2) - \ln((\nu^{(t-1)}+1)/2)] \\ &\quad + \sum_{i=1}^{n} [\ln(\tau_i^{(t)}) - \tau_i^{(t)}] \,, \end{aligned} \tag{5}$$

where $\phi(.)$ is the Digamma function. In this situation, the ECM algorithm is the same as the EM algorithm.

3.2.1. *The EM algorithm and the ECM algorithm (for the t-distribution with unknown degrees of freedom)*

E step: This is the same as the E-step of the EM algorithm for the t-distribution with known ν. In addition, compute $S^{(t)}_{\ln\tau-\tau}$ in Eq. (5).

CM step 1: This is the same as the M-step of the EM algorithm for the t-distribution with known ν.

CM step 2: Find $\nu^{(t)}$ that maximizes $Q^{(t)}(\nu) = -n\ln(\Gamma(\nu/2)) + \frac{n\nu}{2}\ln\frac{\nu}{2} + \frac{\nu}{2}S^{(t)}_{\ln\tau-\tau}$ over $\nu(>0)$.

Liu and Rubin[18] proposed a version of the EMCE algorithm with a CMQ-step that is the same as the above CM-step 1 and a CML-step that maximizes the actual constrained likelihood function of ν

$$L(\mu,\sigma^2,\nu) = n\ln[\Gamma((\nu+1)/2)] - n\ln[\Gamma(\nu/2)] - \frac{n}{2}\ln(\nu\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\nu+1}{2}\ln[1+(y_i-\mu)^2/(\nu\sigma^2)]\,,$$

instead of $Q(\nu)$, with μ and σ^2 fixed at their current estimates. As with the CM-step 2 of ECM (or EM) for updating ν, the CML step needs a one-dimensional search algorithm. They showed that this version of ECME converges dramatically faster than both EM and ECM.

3.3. *Accelerating EM via Parameter Expansion: The PX-EM algorithm*

Notice that the scale parameter $2/\nu$ of the γ distribution (1) for the missing τ_i is constrained to the inverse of the shape parameter $\nu/2$. The scale parameter could be considered as an unknown parameter to be estimated from the complete data $X = \{(\tau_i, y_i) : i = 1,\ldots,n\}$. For the sake of clarity, denote by

$$\theta_x = (\mu_x, \sigma_x^2, \nu_x, \lambda_x)$$

the corresponding parameters for X, where μ_x, σ_x^2, and ν_x correspond to μ, σ^2, and ν, respectively, and $\lambda_x(>0)$ is the extra scale parameter for τ_i.

More precisely, the model for the complete data is written as follows

$$\tau_i|(\nu_x, \lambda_x) \sim \mathrm{G}\left(\frac{\nu_x}{2}, \frac{\nu_x}{2\lambda_x}\right), \quad \text{and} \quad y_i|(\tau, \nu_x, \mu_x, \sigma_x^2) \sim \mathrm{N}(\mu_x, \sigma_x^2/\tau_i).$$

Integrating out the τ_i from the joint distribution of (τ_i, y_i) gives the marginal distribution of y_i

$$\frac{\Gamma((\nu_x+1)/2)}{\Gamma(\nu_x/2)(\nu_x\pi)^{1/2}(\sigma_x^2/\lambda_x)^{1/2}}\left[1+\frac{(y-\mu_x)^2}{\nu(\sigma_x^2/\lambda_x)}\right]^{-(\nu+1)/2}, \quad y \in (-\infty, \infty). \tag{6}$$

The parameters σ_x^2 and λ_x in the marginal distribution (6) of y is unidentifiable from the observed data $Y = \{y_i : i = 1, \ldots, n\}$ when $\tau_1, \ldots,$ and τ_n are missing. It is thus necessary to constrain (σ_x^2, λ_x). For example, fix λ_x at $\lambda_x = 1$ as in Sec. 2.2. However, the EM algorithm converges faster if it is applied to the expanded model with the extra parameter λ_x. This extra parameter can be used to capture the information of "imputed" S_τ (i.e. $S_\tau^{(t)}$) which goes to n as $t \to \infty$, and then the information is used to adjust the current estimate of σ^2. This leads to the parameter-expanded EM algorithm.[21]

PX-EM expands the complete-data model $f(X|\theta)$ $(\theta \in \Theta)$ used in EM to a larger model $f_x(X|\theta_x, \lambda_x)$ $((\theta_x, \lambda_x) \in \Theta \times \Lambda)$ with the expansion satisfying two conditions: (i) the observed-data model is preserved in the sense that there is a many-to-one reduction function

$$\theta = R(\theta_x, \lambda_x), \quad \theta_x \in \Theta;\ \lambda_x \in \Lambda;\ \theta \in \Theta \tag{7}$$

from $\Theta \times \Lambda$ onto Θ such that $Y|\theta_x \sim g(Y|\theta = R(\theta_x, \lambda_x))$; and (ii) there is a (fixed) null value λ_x, λ, in the sense that, for all θ, $f_x(X|\theta_x = \theta, \lambda_x = \lambda) = f(X|\theta)$. PX-EM extends EM by replacing the complete-data model $f(X|\theta)$ with the expanded complete-data model $f_x(X|\theta_x, \lambda_x)$. More specifically, starting with $(\theta_x^{(0)} = \theta^{(0)}, \lambda_x^{(0)} = \lambda)$, the tth iteration of PX-EM consists of a parameter-expanded E-step and a parameter-expanded M-step.

3.3.1. *The PX-EM algorithm*[21]

PX-E step: Compute $Q_x(\theta_x, \lambda_x|\theta_x^{(t-1)}, \lambda_x^{(t-1)}) = \mathrm{E}\{\log f_x(X|\theta_x, \lambda_x)|Y, \theta_x^{(t-1)}, \lambda_x^{(t-1)}\}$.

PX-M step: Find $(\theta_x^{(t-)}, \lambda_x^{(t-)})$ that maximizes $Q_x(\theta_x, \lambda_x|\theta_x^{(t-1)}, \lambda_x^{(t-1)})$, then apply the reduction function $R(\theta_x, \lambda_x)$ to obtain $\theta^{(t)} = R(\theta_x^{(t-)}, \lambda_x^{(t-)})$ and set $\theta_x^{(t)} = \theta^{(t)}$ and $\lambda_x^{(t)} = \lambda$.

For the t-distribution with known number of degrees of freedom $\nu_x = \nu$, the reduction function is given by the many-to-one mapping

$$\mu = \mu_x \quad \text{and} \quad \sigma^2 = \sigma_x^2/\lambda_x$$

with the null value of $\lambda_x : \lambda = 1$. The sufficient statistics for the expanded parameter $(\mu_x, \sigma_x^2, \lambda_x)$ are the same as those for (μ, σ^2) given in Condition (2). The complete-data maximum likelihood estimates of μ_x and σ_x^2 are the same as those of μ and σ^2 given in Eq. (3), respectively, that is,

$$\hat{\mu}_x = \hat{\mu} = \frac{S_{\tau y}}{S_\tau} \quad \text{and} \quad \hat{\sigma}_x^2 = \hat{\sigma}^2 = \frac{1}{n}\left[S_{\tau y^2} - \frac{S_{\tau y}^2}{S_\tau}\right].$$

The complete-data maximum likelihood estimate of λ_x is

$$\hat{\lambda}_x = \frac{S_\tau}{n}.$$

The null value of λ_x is $\lambda = 1$. Thus, PX-EM t-distribution with the sample $Y = \{y_i : i = 1, \ldots, n\}$, can be written as follows.

3.3.2. *The PX-EM algorithm (for the t-distribution with a known number of degrees of freedom)*

PX-E step: This is the same as the E-step of EM

PX-M step: This is the same as the M-step of EM except that $(\sigma^2)^{(t)}$ is replaced by

$$(\sigma^2)^{(t)} = \frac{1}{S_\tau^{(t)}}\left[S_{\tau y^2}^{(t)} - \frac{(S_{\tau y}^{(t)})^2}{S_\tau^{(t)}}\right] = \frac{\sum_{i=1}^n \tau_i^{(t)}(y_i - \mu^{(t)})^2}{\sum_{i=1}^n \tau_i^{(t)}}.$$

This algorithm for the t-distribution is first proposed by Kent *et al.*[4] Meng and van Dyk[29] showed that this algorithm is an EM with a different complete-data model. Liu *et al.*[21] provided this PX-EM version. The ECME algorithm can also be applied to the expanded model with unknown degrees of freedom.[12]

4. General Linear Mixed Models

4.1. *The model*

Mixed-effects models are among the most important applications of EM. van Dyk[42] and Pinherio *et al.*[32] provided recent such examples. The most

commonly used linear mixed-effects model was proposed by Laird and Ware,[5] which, without loss of generality, is simplified as follows

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i\,, \quad i = 1, \ldots, m\,, \tag{8}$$

where i is the subject index, $\boldsymbol{y}_i$ is an n_i-dimensional vector of observed responses, $\boldsymbol{X}_i$ is a known $n_i \times p$ design matrix corresponding to the p-dimensional fixed effects vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{Z}_i$ is a known $n_i \times q$ design matrix corresponding to the q-dimensional random effects vector $\boldsymbol{b}_i = (r_{i,1}, \ldots, r_{i,q})'$, and $\boldsymbol{e}_i = (e_{i,1}, \ldots, e_{i,n_i})'$ is an n_i-dimensional vector of within-subject errors.

The random effects $\boldsymbol{b}_1, \ldots,$ and $\boldsymbol{b}_m$ are independent of each other. For each i, $\boldsymbol{b}_i$ follows the q-dimensional normal distribution

$$\boldsymbol{b}_i \sim \mathrm{N}_q(\boldsymbol{0}, \boldsymbol{\Psi})\,, \quad i = 1, \ldots, m\,,$$

where $\boldsymbol{0}$, a vector of q zeros, is the mean vector and $\boldsymbol{\Psi}$ is the $(q \times q)$ variance-covariance matrix. The errors $\boldsymbol{e}_1, \ldots,$ and $\boldsymbol{e}_m$ are independent of each other and independent of $\boldsymbol{b}_1, \ldots,$ and $\boldsymbol{b}_m$. For each i, $\boldsymbol{e}_i$ follows the n_i-dimensional normal distribution

$$\boldsymbol{e}_i \sim \mathrm{N}_{n_i}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$$

with a common unknown variance parameter σ^2, where $\boldsymbol{I}_{n_i}$ denotes the $(n_i \times n_i)$ identity matrix. For more discussion of the structure of the variance-covariance matrices for the random effects and errors, see Pinherio *et al.*[32] and the references therein.

Integrating out the unobservable (missing) random effects $\boldsymbol{b}_1, \ldots,$ and $\boldsymbol{b}_m$ leads to the observed data model

$$\boldsymbol{y}_i|(\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2) \sim \mathrm{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}_{n_i})\,, \quad i = 1, \ldots, m\,. \tag{9}$$

It is difficult to find the maximum likelihood estimates of the parameter $\theta = (\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2)$ from the observed data $Y = \{\boldsymbol{y}_i : i = 1, \ldots, m\}$, especially when q, the dimension of $\boldsymbol{\Psi}$, is large.

4.2. *The complete-data maximum likelihood estimates*

Consider the complete data that consist of both observed data $Y = \{\boldsymbol{y}_i : i = 1, \ldots, m\}$ and missing data $\{\boldsymbol{b}_i : i = 1, \ldots, m\}$. The complete-data

model can also be written as

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{y}_i \end{bmatrix} \Bigg| (\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2) \sim \mathrm{N}_{n_i+q} \left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{X}_i\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{\Psi}\boldsymbol{Z}_i' \\ \boldsymbol{Z}_i\boldsymbol{\Psi} & \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i' + \sigma^2\boldsymbol{I}_{n_i} \end{bmatrix} \right). \tag{10}$$

The complete-data model belongs to the exponential family with the sufficient statistics

$$S_{bb'} = \sum_{i=1}^{m} \boldsymbol{b}_i\boldsymbol{b}_i',$$

$$S_{(y-zb)'(y-zb)} = \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{b}_i), \quad \text{and}$$

$$S_{x'(y-zb)} = \sum_{i=1}^{m} \boldsymbol{X}_i'(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{b}_i) \tag{11}$$

for the parameter $\theta = (\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2)$. Let $S_{x'x} = \sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{X}_i$ and let $S_n = \sum_{i=1}^{m} n_i$. The complete-data maximum likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, and σ^2 are given as follows

$$\hat{\boldsymbol{\beta}} = S_{x'x}^{-1} S_{x'(y-zb)}, \quad \hat{\boldsymbol{\Psi}} = \frac{1}{m} S_{bb'},$$

and

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{S_n}(S_{(y-zb)'(y-zb)} - S'_{x'(y-zb)} S_{x'x}^{-1} S_{x'(y-zb)}) \\ &= \frac{1}{S_n} \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\boldsymbol{b}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\boldsymbol{b}_i). \end{aligned}$$

4.3. *EM-type algorithms*

4.3.1. *The EM algorithm*

Given the observed data and the current estimate $\theta^{(t-1)} = (\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\Psi}^{(t-1)}, (\sigma^2)^{(t-1)})$, under the complete-data model (10), the $\boldsymbol{b}_i$ are independent of each other and

$$\boldsymbol{b}_i | (\boldsymbol{y}_i, \theta^{(t-1)}) \sim \mathrm{N}_q(\hat{\boldsymbol{b}}_i, \hat{\boldsymbol{V}}_i),$$

where the mean vector and variance-covariance matrix are

$$\begin{aligned}\hat{\boldsymbol{b}}_i &\equiv \mathrm{E}\{\boldsymbol{b}_i|Y,\theta^{(t-1)}\}\\ &= \boldsymbol{\Psi}^{(t-1)}\boldsymbol{Z}_i'(\boldsymbol{Z}_i\boldsymbol{\Psi}^{(t-1)}\boldsymbol{Z}_i'+(\sigma^2)^{(t-1)}\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}^{(t-1)})\\ &= [(\sigma^2)^{(t-1)}(\boldsymbol{\Psi}^{(t-1)})^{-1}+\boldsymbol{Z}_i'\boldsymbol{Z}_i]^{-1}\boldsymbol{Z}_i'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}^{(t-1)}) \qquad (12)\end{aligned}$$

and

$$\begin{aligned}\hat{\boldsymbol{V}}_i &\equiv \mathrm{cov}\{\boldsymbol{b}_i|Y,\theta^{(t-1)}\}\\ &= \boldsymbol{\Psi}^{(t-1)}-\boldsymbol{\Psi}^{(t-1)}\boldsymbol{Z}_i'(\boldsymbol{Z}_i\boldsymbol{\Psi}^{(t-1)}\boldsymbol{Z}_i'+(\sigma^2)^{(t-1)}\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\boldsymbol{\Psi}^{(t-1)}\\ &= \left[(\boldsymbol{\Psi}^{(t-1)})^{-1}+\frac{1}{(\sigma^2)^{(t-1)}}\boldsymbol{Z}_i'\boldsymbol{Z}_i\right]^{-1} \qquad (13)\end{aligned}$$

respectively. Thus, the (standard) EM algorithm[18] is given as follows.

E step: Compute $\hat{\boldsymbol{b}}_i$ and $\hat{\boldsymbol{V}}_i$ for $i=1,\ldots,m$, and then

$$\hat{S}_{bb'}=\sum_{i=1}^{m}\hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i'+\sum_{i=1}^{m}\hat{\boldsymbol{V}}_i\,,\quad \hat{S}_{x'(y-zb)}=\sum_{i=1}^{m}\boldsymbol{X}_i'(\boldsymbol{y}_i-\boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)\,,$$

and

$$\hat{S}_{(y-zb)'(y-zb)}=\sum_{i=1}^{m}(\boldsymbol{y}_i-\boldsymbol{Z}_i\hat{b}_i)'(\boldsymbol{y}_i-\boldsymbol{Z}_i\hat{b}_i)+\sum_{i=1}^{m}\mathrm{tr}(\boldsymbol{Z}_i\hat{\boldsymbol{V}}_i\boldsymbol{Z}_i')\,,$$

where the trace function $\mathrm{tr}(\boldsymbol{Z}_i\hat{\boldsymbol{V}}_i\boldsymbol{Z}_i)$ denotes the sum of the diagonal elements of the $(q\times q)$ matrix $\boldsymbol{Z}_i\hat{\boldsymbol{V}}_i\boldsymbol{Z}_i$.

M step: Compute

$$\boldsymbol{\beta}^{(t)}=S_{x'x}^{-1}\hat{S}_{x'(y-zb)}\,,\quad \boldsymbol{\Psi}^{(t)}=\frac{1}{m}\hat{S}_{bb'}\,,$$

and

$$\begin{aligned}(\sigma^2)^{(t)} &= \frac{1}{\sum_{i=1}^{m}S_n}[\hat{S}_{(y-zb)'(y-zb)}-\hat{S}'_{x'(y-zb)}S_{x'x}^{-1}\hat{S}_{x'(y-zb)}]\\ &= \frac{1}{S_n}\left[\sum_{i=1}^{m}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}^{(t)}-\boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)'(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}^{(t)}-\boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)\right.\\ &\qquad \left.+\mathrm{tr}(\boldsymbol{Z}_i\hat{\boldsymbol{V}}_i\boldsymbol{Z}_i')\right].\end{aligned}$$

4.3.2. *The ECME algorithm*

It is well known that EM for general linear mixed-effects models can converge very slowly. Many methods for accelerating EM have been proposed in the literature.[6] A modified version of the above EM was proposed as an EM implementation by Laird and Ware.[5] This algorithm can be represented as the ECME algorithm with a CML step that updates $\boldsymbol{\beta}$ and a CMQ step that updates $(\boldsymbol{\Psi}, \sigma^2)$.

4.3.2.1. The ECME algorithm (version 1)

E step: This is the same as the E-step of EM

CMQ step: Update the estimates of $\boldsymbol{\Psi}$ and σ^2 : $\boldsymbol{\Psi}^{(t)} = \hat{S}_{bb'}/m$ and

$$(\sigma^2)^{(t)} = \frac{1}{S_n}\Bigg[\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t-1)} - \boldsymbol{Z}_i\boldsymbol{b}_i)' \times (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t-1)} - \boldsymbol{Z}_i\boldsymbol{b}_i) + \sum_{i=1}^{m}\operatorname{tr}(\boldsymbol{Z}_i\boldsymbol{V}_i\boldsymbol{Z}_i')\Bigg].$$

CML step: Updates the estimate of $\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ fixed at their current estimates:

$$\boldsymbol{\beta}^{(t)} = \left[\sum_{i=1}^{m}\boldsymbol{X}_i'(\boldsymbol{Z}_i\boldsymbol{\Psi}^{(t)}\boldsymbol{Z}_i' + (\sigma^2)^{(t)}\boldsymbol{I}_{n_i})^{-1}\boldsymbol{X}_i\right]^{-1} \times \left[\sum_{i=1}^{m}\boldsymbol{X}_i'\Big(\boldsymbol{Z}_i\boldsymbol{\Psi}^{(t)}\boldsymbol{Z}_i' + (\sigma^2)^{(t)}\boldsymbol{I}_{n_i}\Big)^{-1}\boldsymbol{y}_i\right].$$

Liu and Rubin[18] considered another version of ECME with three CM steps: A CMQ step for updating the estimate of $\boldsymbol{\Psi}$, a CML step for updating the estimate of $\boldsymbol{\beta}$, and a CML step for updating the estimate of σ^2. The CML step for updating σ^2 does not have a closed-form solution. A modified version (see Schafer[38] and Lindstrom and Bates[9]) has a closed-form solution for σ^2.

4.3.2.2. The ECME algorithm (version 2)

E step: This is the same as the E-step of EM

CMQ step: Update the estimate of $\boldsymbol{\Psi}$: $\boldsymbol{\Psi}^{(t-)} = \hat{S}_{bb'}/m$.

CML step: Update the estimates of σ^2 and $\boldsymbol{\beta}$ with $\boldsymbol{\Phi} \equiv \boldsymbol{\Psi}/\sigma^2$ fixed at its current estimate $\boldsymbol{\Phi}^{(t)} = \boldsymbol{\Psi}^{(t-)}/(\sigma^2)^{(t-1)}$:

$$\boldsymbol{\beta}^{(t)} = \left[\sum_{i=1}^{m} \boldsymbol{X}_i'(\boldsymbol{Z}_i\boldsymbol{\Phi}^{(t)}\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})^{-1}\boldsymbol{X}_i\right]^{-1} \times \left[\sum_{i=1}^{m} \boldsymbol{X}_i'(\boldsymbol{Z}_i\boldsymbol{\Phi}^{(t)}\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})^{-1}\boldsymbol{y}_i\right].$$

and

$$(\sigma^2)^{(t)} = \frac{1}{S_n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)})'(\boldsymbol{Z}_i\boldsymbol{\Phi}^{(t)}\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})^{-1} \times (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)})\,;$$

and then adjust the current estimate of Ψ:

$$\boldsymbol{\Psi}^{(t)} = \boldsymbol{\Phi}^{(t)}(\sigma^2)^{(t)} = \frac{(\sigma^2)^{(t)}}{(\sigma^2)^{(t-1)}}\boldsymbol{\Psi}^{(t-)}\,.$$

As was noted by van Dyk,[42] this ECME version can also be obtained using the more convenient parameterization that replaces $\boldsymbol{\Psi}$ by $\sigma^2\boldsymbol{\Phi}$.

4.3.3. *The PX-EM algorithm*

Meng and van Dyk[30] considered efficient data augmentation for deriving fast implementations of EM for general linear mixed model. The basic idea is to augment less missing information, as was described in Meng and van Dyk.[29] As shown by Liu *et al.*,[21] PX-EM is easier to implement and converges faster than the efficient implementation of EM by Meng and van Dyk.[30] The idea of PX-EM is to expand the complete-data model (10) to capture the information on (*i*) the covariance matrix between missing data $\boldsymbol{b}_i$ and the observed responses $\boldsymbol{y}_i$, which is implicitly constrained by the variance-covariance matrix $\boldsymbol{\Psi}$ of $\boldsymbol{b}_i$ in the original model; and (*ii*) the mean vector of $\boldsymbol{b}_i$, which is fixed at $\boldsymbol{0}$ in the original model. First, we describe the PX-EM version of Liu *et al.*[21] (see also Liu[13]), who considered the case (*i*) and implemented the model with $n_i = 1$ for all $i = 1, \ldots, m$, is presented. Second, we discuss a new version that takes both (*i*) and (*ii*) into account.

The expanded model that activates the covariance matrix between $\boldsymbol{b}_i$ and $\boldsymbol{y}_i$ can be written as

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{y}_i \end{bmatrix} \Bigg| (\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\beta}_*, \boldsymbol{\Psi}_*, \sigma_*^2, \boldsymbol{C}_*)$$

$$\sim \mathrm{N}_{n_i+q} \left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{X}_i \boldsymbol{\beta}_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi}_* & \boldsymbol{\Psi}_* \boldsymbol{C}_*' \boldsymbol{Z}_i' \\ \boldsymbol{Z}_i \boldsymbol{C}_* \boldsymbol{\Psi}_* & \boldsymbol{Z}_i \boldsymbol{C}_* \boldsymbol{\Psi}_* \boldsymbol{C}_*' \boldsymbol{Z}_i' + \sigma_*^2 \boldsymbol{I}_{n_i} \end{bmatrix} \right), \quad (14)$$

where $\boldsymbol{C}_*$ is a $(q \times q)$ matrix. The observed-data model is thus given by the marginal distribution of $\boldsymbol{y}_i$ in model (14). The reduction function is

$$\boldsymbol{\beta} = \boldsymbol{\beta}_*\,, \quad \sigma^2 = \sigma_*^2\,, \quad \text{and} \quad \boldsymbol{\Psi} = \boldsymbol{C}_* \boldsymbol{\Psi}_* \boldsymbol{C}_*'$$

with the null values of $\boldsymbol{C}_*$: $\boldsymbol{C} = \boldsymbol{I}_q$, the q-dimensional identity matrix. The complete-data maximum likelihood estimates of $\theta_* = (\boldsymbol{\beta}_*, \boldsymbol{\Psi}_*, \sigma_*^2, \boldsymbol{C}_*)$ is obtained from the following linear model

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_* + (\boldsymbol{b}_i' \otimes \boldsymbol{Z}_i) \vec{\boldsymbol{C}} + \boldsymbol{e}_i\,,$$

where $\otimes$ denotes the Kronecker operator, $\vec{\boldsymbol{C}}$ denotes the vector obtained by stacking the columns of $\boldsymbol{C}$ (i.e. $\vec{\boldsymbol{C}} = (\boldsymbol{C}_1', \ldots, \boldsymbol{C}_q')'$, which $\boldsymbol{C}_j$ is the jth column of $\boldsymbol{C}$), and $\boldsymbol{e}_i \sim \mathrm{N}_{n_i}(\boldsymbol{0}, \sigma_*^2 \boldsymbol{I}_{n_i})$ for $i = 1, \ldots, m$. This leads to the following PX-EM algorithm.

4.3.3.1. The PX-EM algorithm (version 1)

PX-E step: This is the same as the E-step of EM

PX-M step: Compute $\boldsymbol{\Psi}^{(t)}$, which is the same as the M-step of EM,

$$\begin{bmatrix} \boldsymbol{\beta}_*^{(t)} \\ \vec{\boldsymbol{C}}_*^{(t)} \end{bmatrix} = \left[\sum_{i=1}^m \begin{pmatrix} \boldsymbol{X}_i' \boldsymbol{X}_i & \boldsymbol{X}_i' (\hat{\boldsymbol{b}}_i' \otimes \boldsymbol{Z}_i) \\ (\hat{\boldsymbol{b}}_i \otimes \boldsymbol{Z}_i') \boldsymbol{X}_i & (\hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i' + \hat{\boldsymbol{V}}_i) \otimes (\boldsymbol{Z}_i' \boldsymbol{Z}_i) \end{pmatrix} \right]^{-1}$$

$$\times \left[\sum_{i=1}^m \begin{pmatrix} \boldsymbol{X}_i' \\ \hat{\boldsymbol{b}}_i \otimes \boldsymbol{Z}_i' \end{pmatrix} \boldsymbol{y}_i \right],$$

and

$$(\sigma_*^2)^{(t)} = \frac{1}{S_n} \left[\sum_{i=1}^n \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i{}' + \sum_{i=1}^n (\vec{\boldsymbol{C}}_*^{(t)})' (\hat{\boldsymbol{V}}_i \otimes (\boldsymbol{Z}_i' \boldsymbol{Z}_i)) \vec{\boldsymbol{C}}_*^{(t)} \right],$$

where $\hat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_*^{(t)} - \boldsymbol{Z}_i \boldsymbol{C}_*^{(t)} \hat{\boldsymbol{b}}_i$, and then apply the reduction function to obtain

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}_*^{(t)}\,, \quad (\sigma^2)^{(t)} = (\sigma_*^2)^{(t)}\,, \quad \text{and} \quad \boldsymbol{\Psi}^{(t)} = \boldsymbol{C}_*^{(t)} \boldsymbol{\Psi}_*^{(t)} (\boldsymbol{C}_*^{(t)})'\,.$$

In order to activate the mean vector of the random effects, suppose that there is a known $(p \times q)$ matrix $\boldsymbol{K}$ such that $\boldsymbol{Z}_i = \boldsymbol{X}_i \boldsymbol{K}$ for all $i = 1, \ldots, m$. For example, the design matrix $\boldsymbol{Z}_i$ consists of the columns of $\boldsymbol{X}_i$, as is often the case in practice. In this situation, PX-EM can be used to accelerate EM further by activating the mean vector of the random effects. The complete-data model can be written as

$$\begin{aligned} \boldsymbol{b}_i | (\boldsymbol{\beta}_*, \boldsymbol{\Psi}_*, \sigma_*^2, \boldsymbol{C}_*, \boldsymbol{\mu}_*) &\sim \mathrm{N}_q(\boldsymbol{\mu}_*, \boldsymbol{\Psi}_*) \\ \boldsymbol{y}_i | (\boldsymbol{b}_i, \boldsymbol{\beta}_*, \boldsymbol{\Psi}_*, \sigma_*^2, \boldsymbol{C}_*, \boldsymbol{\mu}_*) &\sim \mathrm{N}_{n_i}(\boldsymbol{X}_i \boldsymbol{\beta}_* + \boldsymbol{Z}_i \boldsymbol{C}_* \boldsymbol{b}_i, \sigma_*^2 \boldsymbol{I}_{n_i}) \,, \end{aligned}$$

The corresponding complete-data model is

$$\begin{aligned} \boldsymbol{y}_i | (\boldsymbol{\beta}_*, \boldsymbol{\Psi}_*, \sigma_*^2, \boldsymbol{C}_*, \boldsymbol{\mu}_*) \\ \sim \mathrm{N}_{n_i}(\boldsymbol{X}_i(\boldsymbol{\beta}_* + \boldsymbol{K}\boldsymbol{C}_* \boldsymbol{\mu}_*),\, \boldsymbol{Z}_i \boldsymbol{C}_* \boldsymbol{\Psi}_* \boldsymbol{C}_*' \boldsymbol{Z}_i' + \sigma_*^2 \boldsymbol{I}_{n_i}) \,. \end{aligned}$$

The reduction function is then given by

$$\boldsymbol{\beta} = \boldsymbol{\beta}_* + \boldsymbol{K}\boldsymbol{C}_* \boldsymbol{\mu}_* \,, \quad \sigma^2 = \sigma_*^2 \,, \quad \text{and} \quad \boldsymbol{\Psi} = \boldsymbol{C}_* \boldsymbol{\Psi}_* \boldsymbol{C}_*' \,,$$

with the null values of the extra parameters $\boldsymbol{C} = \boldsymbol{I}_q$ and $\boldsymbol{\mu} = \boldsymbol{0}$. Thus, we have the following new version of PX-EM.

4.3.3.2. The PX-EM algorithm (version 2)

PX-E step: This is the same as the E step of EM.

PX-M step: Compute

$$\mu_*^{(t-)} = \frac{1}{m} \sum_{i=1}^{m} \hat{\boldsymbol{b}}_i \,,$$

$$\boldsymbol{\Psi}_*^{(t-)} = \frac{1}{m} \sum_{i=1}^{m} [(\hat{\boldsymbol{b}}_i - \mu_*^{(t-)})(\hat{\boldsymbol{b}}_i - \mu_*^{(t-)})' + \hat{\boldsymbol{V}}_i] \,,$$

and $(\boldsymbol{\beta}_*^{(t-)}, \vec{\boldsymbol{C}}_*^{(t-)}, (\sigma_*^2)^{(t-)})$, that is the same as that in M step of PX-EM version 1; and then apply the reduction function to obtain

$$\begin{aligned} \boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}_*^{(t)} &= \boldsymbol{\beta}_*^{(t-)} + \boldsymbol{K}\boldsymbol{C}_*^{(t-)} \boldsymbol{\mu}_*^{(t-)} \,, \\ (\sigma^2)^{(t)} = (\sigma_*^2)^{(t)} &= (\sigma_*^2)^{(t-)} \,, \\ \boldsymbol{\Psi}^{(t)} &= \boldsymbol{C}_*^{(t-)} \boldsymbol{\Psi}_*^{(t-)} (\boldsymbol{C}_*^{(t-)})' \,, \\ \mu_*^{(t-)} &= \boldsymbol{0} \,, \quad \text{and} \quad \boldsymbol{C}_*^{(t-)} = \boldsymbol{I}_q \,. \end{aligned}$$

Another PX-EM version can be obtained by activating only the mean vector of the random effects. Parameter expansion can also be used to accelerate the ECM and ECME algorithms. van Dyk[42] provides such examples with the expanded model that includes the extra parameter $\boldsymbol{C}_*$.

5. Discussion

There are many other important issues about EM that this brief review has not touched on. These include how to compute asymptotic variance-covariance matrix,[14,25,27] the relationship between EM and Markov chain Monte Carlo methods, e.g. EM and the Data-Augmentation algorithm;[40] ECM and the Gibbs sampler;[3] ECME and collapsed Gibbs sampler;[18] and PX-EM and the Parameter-Expanded-Data-Augmentation algorithm,[24,31] and Monte Carlo EM algorithms.[43,44]

Acknowledgement

The author thanks Dr. Diane Lambert for her helpful comments.

References

1. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* **B39**: 1–38.
2. Fessler, J. A. and Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing* **42**: 2664–2677.
3. Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 141–151.
4. Kent, J. T., Tyler, D. E. and Vardi, Y. (1994). A curious likelihood identity for the multivariate t distribution. *Communication Statistical Simulation Computing* **23**: 441–453.
5. Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**: 963–974.
6. Laird, N., Lange, N. and Stram, D. (1987). Maximizing likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association* **82**: 97–105.
7. Lange, K. and Carson, R. (1984). EM reconstruction for emission and transmission tomography. *Journal of Computing Assistant Tomograph* **8**: 306–312.
8. Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989), Robust statistical modeling using the t distribution, *Journal of the American Statistical Association* **84**: 881–896.

9. Lindstrom, M. J. and Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**: 1014–1022
10. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data*, Wiley, New York.
11. Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data, *Journal of the American Statistical Association* **91**: 1219–1227.
12. Liu, C. (1997a). ML estimation of the multivariate t distribution and the EM algorithm, *Journal of Multivariate Analysis* **63**: 296–312.
13. Liu, C. (1997b). Comment on "The EM algorithm — An old folk song sung to fast new tune" by Meng and van Dyk, *Journal of the Royal Statistical Society* **B59**: 557–558.
14. Liu, C. (1998). Information matrix computation from conditional information. *Biometrika* **85**: 973–979.
15. Liu, C. (1999). Efficient ML estimation of the multivariate normal distribution from incomplete data, *Journal of Multivariate Analysis* **69**: 206–217.
16. Liu, C. (2000a). Estimation of discrete distributions with a class of simplex constraints, *Journal of the American Statistical Association*.
17. Liu, C. (2000b). Robit regression: A simple robust alternative to logistic and probit regression, *Technical report*, Bell Labs.
18. Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**: 633–648.
19. Liu, C. and Rubin, D. B. (1995). ML estimation of the multivariate t distribution with unknown degrees of freedom. *Statistica Sinica* **5**: 19–39.
20. Liu, C. and Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistical Sinica* **8**: 729–747.
21. Liu, C., Rubin, D. B. and Wu, Y. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**: 755–770.
22. Liu, C. and Sun, D. (1997). Acceleration of the EM algorithm for mixture models using ECME. *American Statistical Association Proceedings of the Statistical Computing Section*, 109–114.
23. Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* **89**: 958–966.
24. Liu, J. S. and Wu, Y. (1999). Parameter expansion scheme for data augmentation. *Journal of the American Statistical Association* **94**: 1264–1274.
25. Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society* **B44**: 226–233.
26. Meng, X. L. and Pedlow, S. (1992). EM: A bibliographic review with missing articles. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 24–27.
27. Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**: 899–909.

28. Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**: 267–278.
29. Meng, X. L. and van Dyk, D. (1997). The EM algorithm — An old folk song sung to a fast new tune (with Discussion). *Journal of the Royal Statistical Society* **B59**: 511–67.
30. Meng, X. L. and van Dyk, D.(1998). Fast EM implementations for mixed-effects models. *Journal of the Royal Statistical Society* **B60**: 559–578.
31. Meng, X. L. and van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**: 301–320.
32. Pinheiro, J. C., Liu, C. and Wu, Y. (2000). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution, revised for *Journal of Computational and Graphical Statistics*.
33. Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journla of the American Statistical Association* **69**: 467–474.
34. Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**: 581–590.
35. Rubin, D. B. (1978). Multiple imputation in sample surveys — A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34. Also in *Imputation and Editing of Faulty or Missing Survey Data*, US Dept. of Commerce, Bureau of the Census, 1–23.
36. Rubin, D. B. (1983). Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences*.
37. Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47**: 69–76.
38. Schafer, J. L. (1999). Some improved procedures for linear mixed effects. Unpublished technical report.
39. Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Image Processing* **2**: 113-122.
40. Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**: 528–550.
41. Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, New York.
42. van Dyk, D. A. (2000a). Fitting mixed-effects models using efficient EM-type algorithms. *Journal of Computational and Graphical Statistics*.
43. van Dyk, D. A. (2000b). Nesting EM algorithms for computational efficiency. *Statistica Sinica*.
44. Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association* **85**: 699–704.
45. Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**: 95–103.

About the Author

Chuanhai Liu is a technical staff member of Bell Laboratories, Lucent Technologies, (http://www.stat.bell-labs.com/stat/liu/). He obtained MS in Probability and Statistics in 1987 from Wuhan University, MA in Statistics (1990) from Harvard University, and PhD in Statistics (1994) also from Harvard University. His research areas include Bayesian statistics, missing data problems, multiple imputation, robust statistics, time series, EM algorithms, and Markov chain Monte Carlo methods.

This page intentionally left blank

CHAPTER 29

INTRODUCTION TO ARTIFICIAL NEURAL NETWORKS

XIA JIELAI and JIANG HONGWEI

Department of Health Statistics,
The Fourth Military Medical University, Xi'an, Shanxi 710033, PR China
Tel: (86) 29-3376979; medstat@fmmu.edu.cn

TANG QIYI

Department of Plant Protection, Zhejing University,
268 Kaixuan Road, Hangzhou, 310029, PR China
Tel: (86) 571-86971621; tqy@mail.hz.zj.cn

1. Introduction

While recent progressions of neurology lead the rapid development of artificial neural networks (ANN), the growing requirement of digital computer and artificial intelligence (AI) also promotes ANN. Today, in all problems that involve AI, human intelligence is still performed over AI. To develop new generation of intelligent computers, we must fully understand the human intelligent processes; in particular, the mechanisms of dealing with information by the neural network systems in human brains. On the other hand, although the initial intention of ANN was merely to explore and simulate informational processing of human, its superior capability has been demonstrated in problems that traditional digital computer systems and artificial intelligence encountered. Indeed, ANNs can be viewed as a major new break-through to various fields such as computational methodology and AI, etc.

Artificial neural networks (ANN) is an engineering method that simulates the structures and operating principles in the information processing systems possessed by human brain. It was a milestone that psychologist McCulloch and mathematician Pitts had originally proposed the first

mathematical model of ANNs in 1940s. Since then ANN has made rapid progresses, and various perceptron models have been brought forth subsequently by many researchers such as F. Rosenblatt, Widrow, Hopf and J. J. Hopfield, etc.

In a magnitude of ANN studies, simulated annealing (SA)[1,2,4] and Genetic Algorithm (GA)[3,4] are two popular stochastic optimization algorithms. The former was proposed by Metropolis to simulate the annealing process of metal heating, and the latter was proposed by Holland to simulate the natural evolutional process of living beings. Although the stimulated objectives are at all different, both algorithms are extremely similar each other in formulation of algebraic structures. SA holds for the ergodicity of state spaces by generation functions and ensures the directions of iteration processes by acceptation operator. GA holds for the ergodicity of state spaces by crossover operator and mutation operator, and ensures the directions of iteration processes by selection operator.

The traditional statistics, especially parametric statistics, usually assume a population distribution with unknown parameter. It is the mostly perplexing to assure the validity of the assumptions that samples indeed come from the population specified before using statistical techniques such as t test, ANOVA, regression and so on. However, in accordance to directly learning from data sets, ANN dynamically modulates the "weight" of neurons, and sequentially be able to perceive newly resembled data. Because of its favorable resilience against distortions, ANN has unique advantages to processing imperfect data sets and to problems of complex nonlinear systems. Statistically ANN can be described as the nonparametric nonlinear models. Its applications include predictions, cluster analysis, pattern recognition engines, time series analysis and wick relationship gauge among complex systems. Depending on the nonlinear linkage of numerous simple rule sets (neurons), ANN, especially multilayer perceptron networks, is different in essence from normal expert systems, which is some enumerative procedures based on comprehensive rule systems. As knowledge of experts is collected and represented using some traditional measurements, the establishment of expert systems is more difficult than ANN.

2. Back Propagation (BP) Neural Networks

There are many different types of ANN, including the popular Hopfield model,[5] the connection networks by Feldmann,[6] the Baltzmann machine model by Hinton,[7] the multilayer perceptron model by Rumelhart[8] and

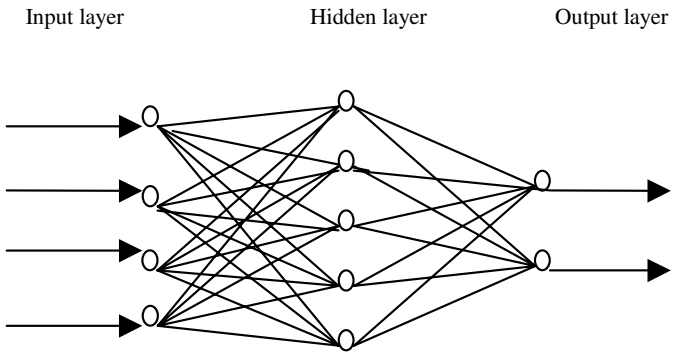


Fig. 1. A BP neural network model.

the self-organization networks models by Kohonen,[9] etc. Multilayer perceptron model is the most general among these ANN models. Although ANN had been around since the late 1940's, no major progress was made until the mid-1980's when the multilayer forward-propagation perceptron model was proposed by Minsky[10] and became sophisticated enough for general applications through combination with back-propagation (BP) algorithm by Rumelhart and simulated annealing (SA) algorithms. A manifold of three-layer BP neural network is shown in Fig. 1.

The BP neural networks (BPNN) systems with the hierarchical structure, including one input layer, several hidden layers and one output layer. Each layer consists of various neutrons taking on two phases: Activity and inactivity. Figure 1 illustrates a typical network with one input layer, one hidden layer and one output layer.

Typically in BPNN, after having processed the signals received from the input layer, the neutrons of the hidden layers propagate it forward to the output layer that completes the finial procedure to export the results. A conventional stimuli function of every neutron usually is a S-shaped curve function such as the logistic function.

$$f(x) = \frac{1}{1 + e^{-x/Q}} \,.$$

Here, Q is the threshold parameter to adjust the formulation of stimuli function. The learning procedure of this algorithm is made up of forward-propagation and backward-propagation. The special characteristic of this type of network is its simple dynamics: when a signal is inputted into the BPNN, it is propagated to the next layer by the interconnections between

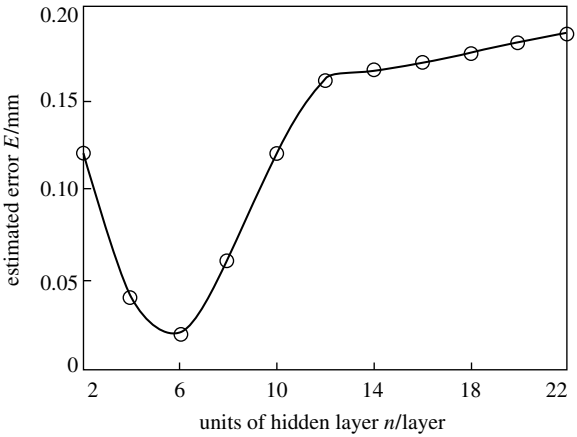


Fig. 2. Relationships between units of the hidden layer and estimated error.

the neurons. The sign is processed by the neurons of one layer and then be propagated onto the next layer. It means that the state of each layer locally influence the next layer. This procedure will not stop until the signal reaches the output layer sending out the processed signal. In order to upgrade the precision of the system, signal errors feedback across the same pathways. Through modifications of the weights for all units of each layer, the differences between the expected and observed outcome are minimized.

At present, there is no matured theory on how to select the number of units and hidden layers. In general, the more units of hidden layers neural networks possess, the more complexity they reflect and higher precision of learning. Nevertheless, with the increment of units in hidden layers, overfitting to the learning data comes into being easily. If an ANN model is trained on a learning data set very well, its ability to predict subsequently future data set will be enhanced.

Figure 2 shows a special example in which the lowest error achieved when the system has 6 units in the hidden layer.

For simplicity, we assume that there are n sigmoid type units in a neural network which possesses only one unit x in the input layer and one unit y in the output layer. Let (x_k, y_k) $(k = 1, 2, 3, \ldots, N)$ be observations in which x_k is the input signal and y_k is the output signal for the kth sample. Also, let the output of any unit i as O_{ik} and the input of unit j is

$$\mathrm{net}_{jk} = \sum_i W_{ij} O_{ik} \,.$$

And the error function is

$$E = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 .$$

In this function $\hat{y}_k$ is the predicted value of the network output. If $E_k = (y_k - \hat{y}_k)^2$, $\delta_{jk} = \frac{\partial E_k}{\partial \text{net}_{jk}}$ and $O_{jk} = f(\text{net}_{jk})$, then

$$\frac{\partial E_k}{\partial W_{ij}} = \frac{\partial E_k}{\partial \text{net}_{jk}}\frac{\partial \text{net}_{jk}}{\partial W_{ij}} = \frac{\partial E_k}{\partial \text{net}_{jk}}O_{ik} = \delta_{jk}O_{ik} .$$

If the unit j is in the output layer, $O_{jk} = \hat{y}_k$

$$\delta_{jk} = \frac{\partial E_k}{\partial \hat{y}_k}\frac{\partial \hat{y}_k}{\partial \text{net}_{jk}} = -(y_k - \hat{y}_k)f'(\text{net}_{jk}) . \tag{1}$$

Else if unit j is not in the output layer, then

$$\delta_{jk} = \frac{\partial E_k}{\partial \text{net}_{jk}} = \frac{\partial E_k}{\partial O_{jk}}\frac{\partial O_{jk}}{\partial \text{net}_{jk}} = \frac{\partial E_k}{\partial O_{jk}}f'(\text{net}_{jk})$$

$$\begin{aligned}\frac{\partial E_k}{\partial O_{jk}} &= \sum_m \frac{\partial E_k}{\partial \text{net}_{mk}}\frac{\partial \text{net}_{mk}}{\partial O_{jk}} \\ &= \sum_m \frac{\partial E_k}{\partial \text{net}_{mk}}\frac{\partial}{\partial O_{jk}}\sum_i W_{mi}O_{ik} \\ &= \sum_m \frac{\partial E_k}{\partial \text{net}_{mk}}\sum_i W_{mj} = \sum_m \delta_{mk}W_{mj} .\end{aligned}$$

Thus,

$$\begin{cases} \delta_{jk} = f'(\text{net}_{jk})\sum_m \delta_{mk}W_{mj} \\ \dfrac{\partial E_k}{\partial W_{ij}} = \delta_{mk}O_{ik} . \end{cases} \tag{2}$$

If a neural network has M layers in which the Mth only owns the output units and the first layer only possesses the input units, then BP algorithms are

(i) Select the initial weights W.
(ii) Repeat following procedures until converging:
  a. For K from 1 to N
    (a) Calculate O_{ik}, net_{jk} and $\hat{y}_k$ (in the procedure of forward-propagation)

(b) Implement the reversed calculation of layers from M to 2 (in the procedure of backward-propagation)

b. For the same unit $j \in M$, calculated δ_{jk} by (1) and (2).

(iii) Modulate weights, $W_{ij} = W_{ij} - \delta \frac{\partial E}{\partial W_{ij}}$, $\delta > 0$, for $\frac{\partial E}{\partial W_{ij}} = \sum_k^N \frac{\partial E_k}{\partial W_{ij}}$

From BP algorithms, it concludes that BP models transform input-to-output patterns of sampling data sets into the optimization of nonlinear models. Its optimization is different at all from the traditional gradient descend method. So neural networks are the absolutely nonlinear mapping projects between input and output.

The focal design of a neural network lies in how to estimate the structure of models and the selection of learning algorithms. To establish appropriate learning algorithms and model structure, we must rely on current theoretical developments of ANN and train these systems with enormous datasets. By dynamically adjusting the parameters of networks in the continuous procedures of learning, ANN can reach the precision required.

3. Introduction to Operation of DPS Data Process System

ANN packages have been embedded in statistical software packages, such as SPSS, MATLAB and so forth. They can be browsed in those statistical software websites. DPS, Data Processing System, programmed by Qiyi Tang, will be showed below in this section. The basic data structure is that each row is a single case (observation), each column is a single variable and the left is the data of input units (independent variables), the right is the data of output units (dependent variables). And all values of cases are entered one by one. Do not need to enter the outputs (dependent variables) for the individuals to be recognized (predicted).

After the data-entering step has finished, press CTRL and right button of mouse to define the predicted data as the second block.

Before the learning procedure of neural networks, an optional dialogue, showed in Fig. 3 below, will appear to require some parameters of neural network. The principles of setting parameters are:

(1) *Number of units*: The number of units of the input layer equals to the number of characteristic factors (independent variables), and the units of the output layer just amount to the number of system targets. Generally the number of units in hidden layers, greatly varying according to individual experiences, is 75% units of input layer. For

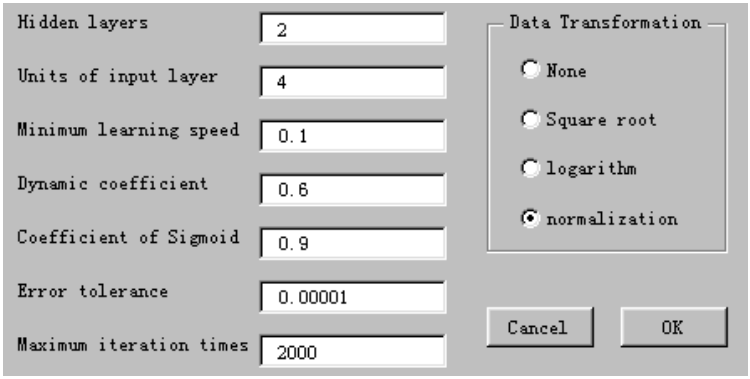


Fig. 3. Optional dialogue of parameters in neural network.

example, if there are 7 units in an input layer and 1 unit in an output layer, the number of units in a hidden layer will be 5 to constitute a 7-5-1 model of neural network. In practice, due to comparing the output consequences of various units in a hidden layer, the most reasonable structure is established conclusively after the learning procedure of neural network.

(2) *Initial weights*: All of initial weights must not be exactly equal to each other. For the fact that it has been verified that once the initial weights are identical, even if there exists a set of diverse weights so as to the minimum error of neural network, the weights of units will remain to be equal. Thus, in our software, a random generator is programmed to yield a set of random numbers ranged from -0.5 to $+0.5$ as the initial weight of neural network.

(3) *Optimum learning speed*: As a typical BP algorithm, the larger the learning speed is, the greater the change of the weights is, and the faster the convergence is. However when learning speed is beyond a certain limitation, the neural network will oscillate. Consequently learning speed is larger with the guarantee against system oscillation. So, in DPS, learning speed is optimized automatically, though user can specify a certain value, say 0.9.

(4) *Dynamic coefficient*: It is chosen empirically too, just as the range from 0.6 to 0.8.

(5) *Error tolerance*: Generally ranges from 0.001 to 0.00001. If the error between the results of two successive iterations is below the tolerance, computing stops systematically to provide the results.

(6) *Times of iteration*: The default value is 1000. Due to the possible divergence of neural network computing, the maximum iteration times is given beforehand.
(7) *Coefficient of Sigmoid function*: The value, regulating the stimuli formulas of neutron, ranges from 0.9 to 1.0 generally.
(8) *Data transformation*: DPS has advantage of allowing data transformations in several functions, such as logarithm, square root and normalization.

4. Application Examples

Example 1 is an illustration to use our software. Physicians under randomization collect the dataset. The influencing factors set of body surface area consists of 4 physical factors: Sex, age, weight and height. Figure 4 shows the nonlinear relationships between predictor variables and response variable.

Before establishing BP neural network, we split the data set into two segments: From No. 1 to No. 70 as learning sample and from No. 71 to No. 90 severally as predicted sample. The data structure is defined as the following blocks in Table 1.

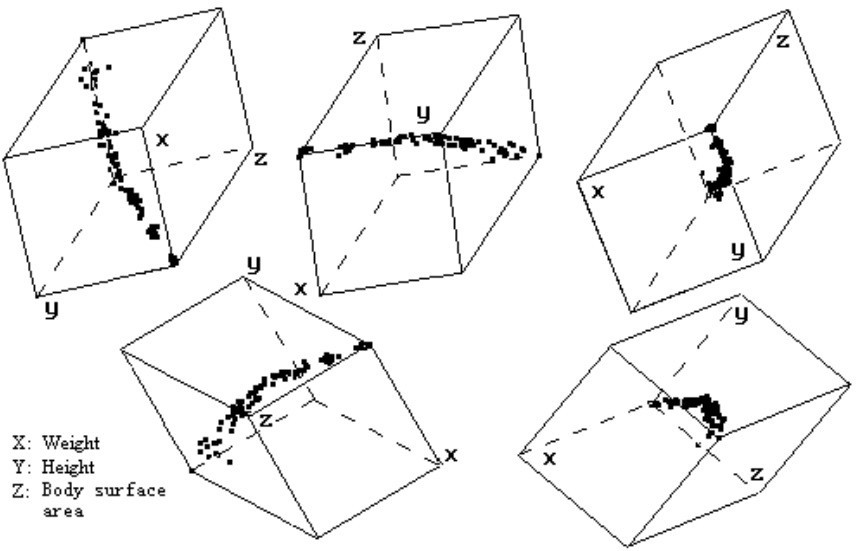


Fig. 4. 3-D Scatter of weight, height and body surface area.

Table 1. Random allotment of 90 persons' physical measurements.

No.	Sex	Age (year)	Weight (kg)	Height (cm)	Body Surface Area (cm^2)	No.	Sex	Age (year)	Weight (kg)	Height (cm)	Body Surface Area (cm^2)
1	1	13	30.5	138.5	10072.9	46	0	15	43	152	12998.7
2	0	5	15	101	6189	47	0	13	27.5	139	9569.1
3	0	0	2.5	51.5	1906.2	48	0	3	12	91	5358.4
4	1	11	30	141	10290.6	49	0	15	40.5	153	12627.4
5	1	15	40.5	154	13221.6	50	1	5	15	100	6364.5
6	0	11	27	136	9654.5	51	1	1	9	80	4380.8
7	0	5	15	106	6768.2	52	1	5	16.5	112	7256.4
8	1	5	15	103	6194.1	53	0	3	12.5	91	5291.5
9	1	3	13.5	96	5830.2	54	1	0	3.5	56.5	2506.7
10	0	13	36	150	11759	55	0	1	10	77	4180.4
11	0	3	12	92	5299.4	56	1	9	25	126	8813.7
12	1	0	2.5	51	2094.5	57	1	9	33	138	11055.4
13	0	7	19	121	7490.8	58	1	5	16	108	6988
14	1	13	28	130.5	9521.7	59	0	11	29	127	9969.8
15	1	0	3	54	2446.2	60	0	7	20	114	7432.8
16	0	0	3	51	1632.5	61	0	1	7.5	77	3934
17	0	7	21	123	7958.8	62	1	11	29.5	134.5	9970.5
18	1	11	31	139	10580.8	63	0	5	15	101	6225.7
19	1	7	24.5	122.5	8756.1	64	0	3	13	91	5601.7
20	1	11	26	133	9573	65	0	5	15	98	6163.7
21	0	9	24.5	130	9028	66	1	15	45	157	13426.7
22	1	9	25	124	8854.5	67	1	7	21	120	8249.2
23	1	0	2.25	50.5	1928.4	68	0	9	23	127	8875.8
24	0	11	27	129	9203.1	69	0	7	17	104	6873.5
25	0	0	2.25	53	2200.2	70	1	15	43.5	150	13082.8
26	0	5	16	105	6785.1	71*	1	15	50	168	14832
27	0	9	30	133	10120.8	72*	0	7	18	114	7071.8
28	0	13	34	148	11397.3	73*	1	3	14	97	6013.6
29	1	3	16	99	6410.6	74*	1	7	20	119	7876.4
30	1	3	11	92	5283.3	75*	0	0	3	54	2117.3
31	0	9	23	126	8693.5	76*	1	1	9.5	74	4314.2
32	1	13	30	138	9626.1	77*	0	15	44	163	13480.9
33	1	9	29	138	10178.7	78*	0	11	32	140	10583.8
34	1	1	8	76	4134.5	79*	1	0	3	52	2121
35	0	15	42	165	13019.5	80*	0	11	29	141	10135.3
36	1	15	40	151	12297.1	81*	0	3	15	94	6074.9
37	1	1	9	80	4078.4	82*	0	13	44	140	13020.3
38	1	7	22	123	8651.1	83*	1	5	15.5	105	6406.5
39	0	1	9.5	77	4246.1	84*	1	9	22	126	8267
40	1	7	25	125	8754.4	85*	0	15	40	159.5	12769.7

Table 1. (*Continued*).

No.	Sex	Age (year)	Weight (kg)	Height (cm)	Body Surface Area (cm^2)	No.	Sex	Age (year)	Weight (kg)	Height (cm)	Body Surface Area (cm^2)
41	1	13	36	143	11282.4	86*	1	1	9.5	76	3845.9
42	1	3	15	94	6101.6	87*	0	13	32	144	10822.1
43	0	0	3	51	1850.3	88*	1	13	40	151	12519.9
44	0	1	9	74	3358.5	89*	0	9	22	124	8586.1
45	0	1	7.5	73	3809.7	90*	1	11	31	135	10120.6

Note: The sign * denote predicted sample.

	A	B	C	D	E	F
1	No.	Age (year)	Weight (kg)	Height (cm)	Square of Body Surface (cm^2)	
2						
3	1	0	3.00	54.0	2117.3	
4	2	0	2.25	53.0	2200.2	
5	3	0	2.50	51.5	1906.2	
6	4	0	3.00	51.0	1850.3	
7	5	0	3.00	51.0	1632.5	
8	6	1	7.50	77.0	3934.0	
9	7	1	10.00	77.0	4180.4	
10	8	1	9.50	77.0	4246.1	
11	9	1	9.00	74.0	3358.5	
12	10	1	7.50	73.0	3809.7	
13	11	3	15.00	94.0	6074.9	
14	12	3	13.00	91.0	5601.7	
15	13	3	12.00	92.0	5299.4	
16	14	3	12.00	91.0	5358.4	
17	15	3	12.50	91.0	5291.5	

Page 1 / Page 2 / Page 3

Fig. 5. Diagram of the data editor window for BP neural network.

Figure 5 is an editor window of DPS system. The format of data is inputted as the following.

After launching the learning procedure of neural network, a window similar to Fig. 5 will be displayed. And then assign the parameter of network: Units in input layer is 4, hidden layer has 2 layers, optimum learning speed is 0.1, dynamic coefficient is 0.6, the coefficient of Sigmoid function is 0.9, error tolerance is 0.00001, maximum times of iteration are 2000, and the selection of data transformation is normalization.

Then press the "OK" button. Then we set 5 to the units of the first hidden layer and set 3 to the units of second hidden layer. After 2000

iterations, the error is 0.00000170. The weights of neutron in output layer is shown below:

Weights matrix of units in the first hidden layer

0.597230	−0.824710	0.566580	−1.065810	0.051900
0.750700	−0.151260	0.172180	−0.369140	−0.139280
−0.715220	2.044800	0.194420	−0.869060	−2.243410
−2.33773	−0.07406	1.46518	−0.12269	1.168

Weights matrix of units in the second hidden layer

−0.507940	−4.616450	3.675080
−0.928980	1.937520	−1.824980
−0.268910	0.21253	−3.046950
−0.708560	−4.81552	2.276470
−0.08934	−4.92864	−0.5283

Weights matrix of units in the output layer

1.12130
7.20956
−5.5911

Table 2 compared the predicted values with the observed values of the body surface area. And these predicted values from No. 71 to No. 90, used as predicted sample, are very close to the observed values. So the facts illustrate that the neural network has favorable abilities in model fitting and predicting.

Table 2. The predicted values and the observed values of neural network.

No.	Predicted Value	Observed Value	No.	Predicted Value	Observed Value	No.	Predicted Value	Observed Value
1	10226.87	10072.9	31	8511.77	8693.5	61	3793.63	3934
2	6370.82	6189	32	10085.21	9626.1	62	10024.22	9970.5
3	2076.073	1906.2	33	10138.25	10178.7	63	6370.82	6225.7
4	10320.32	10290.6	34	3896.33	4134.5	64	5528.03	5601.7
5	12589.19	13221.6	35	12830.50	13019.5	65	6183.77	6163.7
6	9544.59	9654.5	36	12484.89	12297.1	66	13052.96	13426.7
7	6651.57	6768.2	37	4353.68	4078.4	67	8072.70	8249.2
8	6510.22	6194.1	38	8353.89	8651.1	68	8536.07	8875.8
9	6034.13	5830.2	39	4053.10	4246.1	69	6733.68	6873.5
10	11881.13	11759	40	8993.87	8754.4	70	12893.16	13082.8
11	5461.66	5299.4	41	11713.91	11282.4	71*	13282.39	14832
12	2166.992	2094.5	42	6090.36	6101.6	72*	7358.28	7071.8
13	7748.67	7490.8	43	2082.186	1850.3	73*	6173.54	6013.6
14	9342.00	9521.7	44	3746.90	3358.5	74*	7875.99	7876.4
15	2275.284	2446.2	45	3485.76	3809.7	75*	2163.63	2117.3
16	2082.186	1632.5	46	12914.47	12998.7	76*	3909.322	4314.2
17	8109.89	7958.8	47	9703.54	9569.1	77*	13002.22	13480.9
18	10528.90	10580.8	48	5387.18	5358.4	78*	10854.05	10583.8
19	8803.51	8756.1	49	12658.19	12627.4	79*	2213.59	2121
20	9157.74	9573	50	6306.22	6364.5	80*	10094.41	10135.3
21	8894.74	9028	51	4353.68	4380.8	81*	6026.73	6074.9
22	8797.66	8854.5	52	7228.11	7256.4	82*	12950.38	13020.3
23	2145.221	1928.4	53	5457.31	5291.5	83*	6703.66	6406.5
24	9414.36	9203.1	54	2386.589	2506.7	84*	8315.04	8267
25	2104.641	2200.2	55	4120.19	4180.4	85*	12601.92	12769.7
26	6732.92	6785.1	56	8871.79	8813.7	86*	4075.91	3845.9
27	10243.62	10120.8	57	11159.89	11055.4	87*	10903.51	10822.1
28	11424.97	11397.3	58	6947.78	6988	88*	12564.37	12519.9
29	6584.85	6410.6	59	9869.52	9969.8	89*	8277.76	8586.1
30	5395.49	5283.3	60	7658.39	7432.8	90*	10423.74	10120.6

Note: The sign * denote predicted sample.

5. ANNs Based on Genetic Algorithm

Genetic Algorithm (GA), firstly proposed in 1975 by Holland in Michigan University, USA, is inspired by natural selection of Darwinism and the genetics machine. As a brand-new global optimization technique, the algorithm uses the population evolution principles to continuously optimize the prediction weights and eventually finds the optimal or nearly optimal solutions. Because it is simple, universal, robust and applicable to parallel computing, this method is effectively used in a wide variety of fields, such as computer, dispatch optimization, transport problems and constitution optimizations etc.

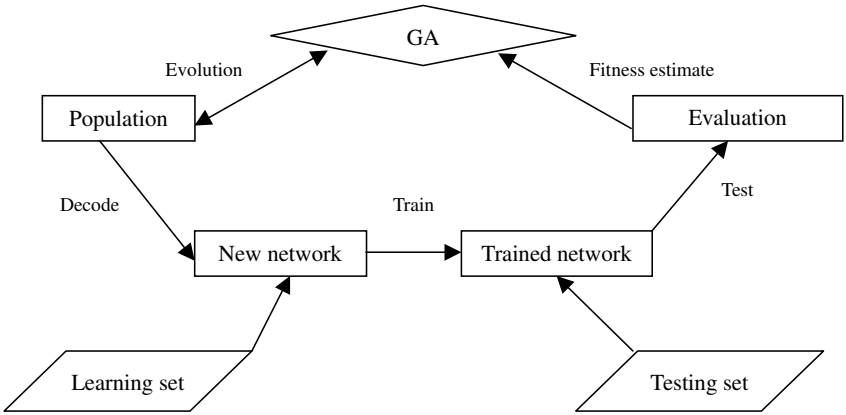


Fig. 6. The conjunction of GA and ANN.

GA, instead of many traditional methods, has been increasingly applied in ANNs design in the learning steps. The conjunction of GA and ANNs, showed below in Fig. 6, will be applied to the evaluation of crops in current research.

5.1. *Value encoding GA oriented to learning of weights in ANN*

5.1.1. *Encoding*

The procedure of weight learning in neural network is a continuously complicated optimization problem of parameters. Binary encoding gives too many possible chromosomes even with a small numbers of alleles. On the other hand, this encoding is often not natural for many problems and sometimes corrections must be made after crossover and/or mutation. By this method, the change of weights will step forward so as to influence the precision of learning in neural network. Therefore value encoding is adopted in current study.

5.1.2. *Fitness function*

The weights of chromosomes are allocated to ANN, and the learning data sets are served as input/output. Then inverse mean squared error, coming out after ANN computing, is defined as fitness function:

$$f = 1 \Big/ \sum_{i=1}^{n} e_i^2 \,.$$

5.1.3. *Weights initialization*

Against that in ordinary BP algorithms performing with the uniform distribution from 0.0 to 1.0, the initial weights of neural network are obtained in accordance with the distribution of $e^{-|\gamma|}$, supported by enormous known trials, because all possible solution can be run around. As a result, the absolute values of weights are relatively small after neural network converges.

5.1.4. *Genetic operator*

Although genetic operators are different from various application circumstances, the weight crossover and weight mutation are two most important operators.

5.1.5. *Selection*

The selection probability of each individual is determined not by means of proportion, but by an elite ratio S, a measure of the surviving ratio of offspring. This can be written as:

$$P_2 = P_1 \cdot S$$
$$P_3 = P_2 \cdot S$$
$$\vdots$$

where $P_1,\ P_2,\ P_3,\ \ldots$ represent the individual probabilities of various fitness functions: First-rate, second-rate, third-rate and so on.

5.2. *The conjunction of GA and ANN*

There is a 3-layer BP neural network with one input layer, one output layer and one hidden layer. Now first, using training sample A_k and expected output $C_k(k = 1, 2, \ldots, m)$, we calculate the stimuli values from the input layer to the hidden layer by formula

$$b_i = f\left(\sum_{h=1}^{n} a_h V_{hi} + \theta_i\right),$$

where $i = 1, 2, \ldots, p$, the units in output layer are a_h, the connection weights from input layer to hidden layer are V_{hi}, the thresholds of units in hidden

layer are θ_i. The values of a_h, V_{hi}, θ_i are determined by the probability distributions, $e^{-|\gamma|}$. The stimuli algorithm is a logistic function:

$$f(x) = 1/(1 + e^{-x}) . \tag{3}$$

And then using formula (3) and (4), we compute the stimuli values of units in the output layer:

$$C_j = f\left(\sum_{i=1}^{p} W_{ij} b_i + \gamma_j\right).$$

where $j = 1, 2, \ldots, q$, the connection weights from hidden layer to output layer are W_{ij}, the thresholds of units in the output layer are γ_j. And also the values of W_{ij}, γ_j are determined by the probability distribution, $e^{-|\gamma|}$. The method for calculating normalized error of output layer is given in (5):

$$d_j C_j (1 - C_j)(C_j^k - C_j) , \tag{4}$$

where the expected value of unit j in output layer is C_j^k.

Finally, compute the error of units in hidden layer compared with each d_j

$$e_i = b_i(1 - b_i) \sum_{j=1}^{q} W_{ij} d_j .$$

On the basis of abiding on the above-mentioned steps and recombining crossover and mutation, we modulate the hidden-layer-to-output-layer connection weights and the thresholds of units in the output layer following the adjustment of the input-layer-to-hidden-layer connection weights and the thresholds of units in hidden layer. When the error between expected and observed outcomes converge to the pre-determined error tolerance, the learning of neural network stops.

6. Future Research Trends

Because the intelligent computing techniques, such as ANN, succeed in many applications, they obtain considerate attention and play important role in magnitude of research fields. As a popular utility, should become more matured in the future. But in terms of intelligent computing itself and statistics, several outstanding obstacles, which be urgently solved in this field are:

(1) It is known that AI still does not possess multitude inherent characteristics of brain, such as tolerance and robustness. Although ANNs have

solved some problems in AI, their theories are not perfect in and are still in their infancy. Moreover, AI and ANN may be completely different realizations of the natural principles on which the brain is based. So a number of mathematical principles should been developed right now. But this is ignored in lots of standard textbooks and reviews. Researchers should pay more attention to studying existing theories and to establishing mathematical foundations. To avoid losing in "forest" of biological mechanisms, the nature of AI and ANN theories should be figured out. As there have been a multitude materials and experiences, it is time to replace many seemly faultless algorithms, which are full with analogues and metaphors, with the specifically objective methods and theories of quantification.

(2) It will take longer time to reveal and understand the intelligent mechanisms of human being. These biological discoveries are increasingly considered as a way to open up the scopes of AI and ANN. Once biology, neurology, genetics make break-through, AI and ANN can simulate high-level intelligent mechanisms to solve some unsolvable problem encountered today. Furthermore, they also urge biologic researches to unveil more sealed puzzle in intelligence. Although AI and ANN are gray-box algorithms and approximately correspond to intelligence of human being, they can provide some useful clues and foundations for further research. While verifying rationality of previous models, the methodology of AI and ANN, how to construct more sophisticated model, become more individualistic and explicit.

(3) Although AI and ANN have developed for near 50 years, their terminology is not standardized by any cases. Especially as to random system, most networks completely conceal or rigidly utilize the statistics. In the words of Anderson, Pellionisz and Rosenfeld[22]: Neural networks are statistics for amateurs. Most statisticians still soberly stand by the development of ANNs and will not to accept it in a short time because ANNs are quite imperfect compared with statistics. They, especially in statistical applications, are reluctant to waste valuable data and time in automatic processing of computer. However, along with maturity of MCMC theory and Gibbs sampling and the increasing Interactions between Frequency School and Bayesian School, the ANN based on Bayesian theory is growing rapidly. All the advantages of ANN are evaluated empirically based on practical applications rather than in theoretical comparisons with statistic methods. The impact of ANNs on the theory and application of statistics is rather obscure at

this stage. At present, as a system method between gray-box and black-box, we cannot evaluate ANNs advantages in generalized range. So how to combine ANN with statistics may be a feasible approach to get out of current dilemma. Statistician should pay more attention to the aspects of AI and ANN.

References

1. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1092.
2. Rumelhart D. E., Hinton G. E. and Williams R. J. (1986). Learning representation by back propagation errors. *Nature* **323**(6188): 533–536
3. Holland J. H. (1992). *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, Massachusetss.
4. Dong, C., Li, Z. N., Xia, R. W. and He, Q. Z. (1995). Advance and some problems on multilayer perceptron neural network. *Mechanics advance* **25**(2): 186–196.
5. Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of National Academic Sciences USA* **79**: 2445–2458.
6. Feldmann, R., Monien, B. and Mysliwietz, P. (1990). Distributed game tree search. In *Parallel Algorithms for Machine Intelligence and Vision*, eds. Kumar, Kanal and Gopalakrishan, Springer-Verlag, New York.
7. Hinton, G. E., Van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory*, Santa Cruz, 5–13.
8. Rumbelhart, D. E., Hinton, G. E., Willians, R. J. (1986a) Learning representations by back-propagating errors. *Nature* **323**: 533–536.
9. Kohonen, T. (1984). *Associative Memory and Self-Organization*, Springer-Verlag, New York.
10. Minsky, M. L. and Papert, S. A. (1988). *Perceptrons. Expanded Edition*, MIT Press, Cambridge, Massachusetss. (First edn 1969.)
11. Lai, Y. X. and Lu, Y. S. (1999). BP neural network application in the distribution study of fluid diversity. *Journal of Beijing Chemistry University* **26**(2).
12. Radford, M. J. (1996). *Bayeian Learning for Neural Networks*, Springer, New York.
13. Watt, R. (1991). *Understanding Vision*, Academic Press, London.
14. Barndorff-Nielsen, O. E. and Jensen, J. L. (1993). *Networks and Chaos Statistical and Probabilistic Aspects*, Chapman and Hall, London.
15. Zhang, S. Q., Chen, C. and Wan, E. P. (1998). Gray system application in production evaluation. *Geography* **18**(6): 581–585.
16. Hechi Nielsen, R. (1989). Theory of the back propagation neural network. *International Journal of Conference Neural Network* **1**: 593–605.

17. Bornholdt, A. (1992). General asymmetric neural network and structure design by genetic algorithms. *Neural Network* **5**(2): 327–334.
18. Li, M. Q., Xu, B. Y. and Kou, J. S. (1999). The combination of GA and ANN. *Theory and Practice of System Engineering* **21**(2): 65–69.
19. Jin, L., Luo, Y., Mou, Q. L. *et al.* (1998). Study on ANN prediction model of the humidity in the soil of cornfield. *Agrology* **35**(1): 25–35.
20. Li, Z. and Zhang, J. T. (2000). The study on evaluation of mealie based on the combination of GA and ANN. *Natural Resource Journal* **15**(3).
21. Deng J. N. (1992). *The Basic Methods of Gray System*, Center China Sci. Tech. Univ. Press, 304–312.
22. Anderson, J. A., Pellionisz, A. and Rosenfeld, E. (1990). *Neuro-computing 2*: *Directions for Research*, MIT Press, Cambridge, Massachusetts.

About the Author

Xia Jielai, professor of the Fourth Military Medical University, earned his bachelor in Applied Mathematics from the Anhui University and master and PhD in Healthy Statistics from the Fourth Military Medical University. He teaches in bio-statistics department of the university as an assistant professor (1983–1988), lecturer (1988–1995), associate professor (1995–1998) and professor (1988-present). He visited clinical and epidemiological research center of Prince Welsh Hospital, the Chinese University of Hongkong (January–February 1994), department of genetics and biometry, Louisiana state university medical center (March–October 1994). His research fields are biostatistics and data processing, including theory and methods of statistical modeling and soft development of NoSA (Non-typical data statistical analysis system). He has published more than 50 articles in various scientific journals.

Index