Making Sense of Data

Making Sense of Data

A Self-Instruction Manual on the Interpretation of Epidemiological Data

Third Edition

J. H. ABRAMSON
Emeritus Professor of Social Medicine
The Hebrew University-Hadassah School
of Public Health and Community Medicine
Jerusalem

Z. H. ABRAMSON, M.D., M.P.H.
Department of Family Medicine
The Hebrew University-Hadassah Medical School
Jerusalem

OXFORD
UNIVERSITY PRESS
2001

OXFORD
UNIVERSITY PRESS

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw

and associated companies in
Berlin Ibadan

Published by Oxford University Press, Inc.,
198 Madison Avenue, New York, New York, 10016
http://www.oup-usa.org

Library of Congress Cataloging-in-Publication Data
Abramson, J.H. (Joseph Herbert), 1924–
Making sense of data : a self-instruction manual on the interpretation of epidemiological data /
J.H. Abramson, Z.H. Abramson.—3rd ed.
p. : cm.
Companion v. to: Survey methods in community medicine /
J.H. Abramson, Z.H. Abramson, 5th ed. 1999.
Includes bibliographical references and index.
ISBN 0-19-514524-0 (cloth) — ISBN 0-19-514525-9 (pbk)
1. Epidemiology—Statistical methods. 2. Epidemiology—Statistical methods—Problems, exercises, etc.
I. Abramson, Z.H. II. Abramson, J.H. (Joseph Herbert), 1924–
Survey methods in community medicine. III. Title.
[DNLM: 1. Epidemiologic Methods—Programmed Instruction.
2. Data Interpretation, Statistical—Programmed Instruction.
WA 18.2 A161m 2001] RA652.4.A27 2001 614.4'2'0727—dc21 2001021134

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

For Jonathan, Dafna,
Roni, Efrat,
Boaz, Tamar,
Dan, and Michal

Preface

The purpose of this manual is to provide readers with basic epidemiological concepts and skills that will help them to appraise published reports as well as their own findings. Consideration is given to applications in clinical medicine, public health and community medicine, and research. The book should thus be useful to a wide range of students and practitioners.

The aim is to produce competence in the ABCs of data interpretation. The manual is not a textbook of statistics, nor does it cover data-processing techniques or advanced epidemiological methods. It is, in a sense, a companion volume to our book *Survey Methods in Community Medicine,* which deals with the planning of studies and the gathering of data.

This edition includes a new section on the practical application of epidemiological findings, and other new topics have been added (Cox proportional hazards regression, qualitative studies, ROC curves). Numerous minor changes have been made, including the addition of new examples, updating of examples based on fictional data, and updating of references. Examples based on official statistical reports have also been updated, but we have not tried to replace all examples with more recent ones ("If it ain't broke, why fix it?").

The book can be used for independent study. In the framework of organized courses, experience indicates that many students prefer to work on the exercises together, in small groups; formal or informal discussions with instructors are helpful.

We are grateful to the many students who participated as involuntary guinea pigs in the testing of the exercises, and to a number of colleagues for their criticism and suggestions.

J.H.A.
Z.H.A.

Jerusalem
December 2000

Contents

B. Rates and Other Measures, 69

C. How Good Are the Measures?, 129

D. Making Sense of Associations, 165

E. Causes and Effects, 233

F. Meta-Analysis: Putting It All Together, 277

G. Putting Study Findings to Use, 335

Introduction

"Why" said the Dodo, "the best way to explain it is to do it."
(Carroll, 1865)

The Aim of This Book

The purpose of this book is to help you to interpret and use data concerning health and disease, health care, and their determinants in populations, population groups, or groups of patients. The book aims to equip you with basic concepts and skills that will enable you to appraise your own data or data collected or published by others, and apply the findings in clinical practice, community medicine and public health, or research.

The book has seven sections. Section A, which deals with basic concepts and procedures, presents a basic step-by-step procedure for the appraisal of data, starting with the assessment of single tables and diagrams. It introduces fundamental terms and directs attention to the variety of uses to which epidemiological data may be put. Section B deals with rates and other simple measures used in epidemiology; and Section C, with their accuracy, the appraisal of accuracy, and the ways in which inaccurate measures can bias results. The appraisal of associations between variables is given detailed consideration in Section D, and Section E deals with the appraisal of cause–effect relationships and ways of measuring the impact of causal factors. Section F focuses on meta-analysis (the critical review and integration of the findings of separate studies of the same topic), and Section G formulates the questions that should be asked before deciding to apply study results in practice.

By the time you reach the end, you should be competent in the use of basic epidemiological tools and capable of exercising critical judgment when assessing results reported by others. When you read a paper, you should be able to identify shortcomings in the study methods or inferences, and make due allowance for them when drawing your conclusions, but without succumbing, it is to be hoped, to the "I am an epidemiologist" bias (Owen, 1982) that leads to the complete repudiation of any study with a flaw.

This book does *not* aim to make you an epidemiological expert; it is an introductory manual that tries to deal in a simple way with fundamental epidemiological approaches and procedures for use in data interpretation. It does not pre-

tend to be a comprehensive textbook of epidemiology. It does *not* deal with techniques of data processing. And, it is *not* a textbook of survey methods or statistics.

How to Use This Book

This is a workbook. There is no point in just sitting down and reading it, skipping the exercises. You will reap little benefit unless you systematically do the exercises.

Each of the book's seven sections is made up of numbered units. These contain short exercises, comments on the exercises in the previous unit, and other explanatory text. Preferably, work on the sections in sequence (but this is not essential). Within each section, go through the units in order; each exercise leads to the next one. Most of the exercises are easy; few require much calculation (but have a pocket calculator handy). To derive the most benefit from the exercises, *write down* your answer to each one. And don't peek! Only when you have written down your answer should you read the detailed comments in the next unit. When you are sure that you have learned all there is to learn from one unit, go on to the next.

At the end of each section there is a self-test. This is a checklist of "what you should now be able to do." Test yourself on each item; if you have any doubts, refer back to the respective unit before proceeding to the next section.

The book is intended to be reasonably self-contained, and sufficient explanations, notes, and definitions are included to minimize the need to refer to other texts. You are encouraged, however, to consult textbooks and other sources for in-depth explanations.

The book may be used for independent study, but if there is an opportunity to work on the exercises in collaboration with others, you may find this an advantage.

Making Sense of Data

Section A

Basic Concepts and Procedures

The White Rabbit put on his spectacles. "Where shall I begin, please your majesty?" he asked.

"Begin at the beginning," the King said gravely, "and go on till you come to the end; then stop."

(Carroll, 1865)

Unit A1

Introduction

This initial series of exercises has three main purposes. First, it introduces a basic approach to the appraisal of data. Step by step, what is the procedure we should use when we look at a table or graph? What are the basic questions to be asked, and in what order? What kinds of explanation should be considered, and how should they be tested?

Second, a number of fundamental terms and concepts that are relevant to the interpretation of epidemiological data are introduced. These include incidence rates; associations; confounding; effect modification; absolute and relative differences; epidemiological models, and many others.

Third, attention is directed to the variety of uses that may be made of epidemiological data. Clinicians, practitioners of public health and community medicine, researchers, and others have different interests, so that though their basic approach to data is the same, they may be interested in asking different questions and reaching conclusions of different kinds.

Exercise A1

Table A1 provides information on the occurrence of cases of acute gastroenteritis (diarrhea and vomiting) in Epiville, an imaginary town in a developing region.

When a table or graph is examined, the first steps are to determine what facts are shown, and then to summarize the facts (unless, of course, they are so simple they do not need to be summarized).

Table A1. Number of Cases of Acute Gastroenteritis Occurring in Epiville in Selected Years, 1970–2000*

Year	No. of Cases
1970	400
1975	600
1980	800
1985	900
1990	1,000
1995	1,100
2000	1,200

**Note:* The above imaginary data are the same as in the previous edition of this book, except that 15 years have been added to each date.

Question A1–1

State the facts shown in Table A1.

Question A1–2

Summarize these facts.

■■■■■■■■■■ Unit A2

Determining What the Facts Are

To be sure of what facts are shown by a table, always read the words as well as the figures. If you read the title of the table, the column and row captions, the footnotes (if there are any), and any explanatory material in the accompanying text, this should enable you to understand *what the numbers represent* and *how they were obtained or calculated.*

The detailed facts shown in Table A1 are easily stated: In 1970, there were 400 cases of acute gastroenteritis in Epiville; in 1975, there were 600; in 1980, 800; in 1985, 900; in 1990, 1,000; in 1995, 1,100; and in 2000, 1,200.

Stating the facts in such detail is, of course, seldom necessary. But what is important is that one should always be sufficiently certain of what the numbers represent to be *able* to spell out the facts in detail. This may not be easy if the table is complicated, badly constructed, or poorly labeled, or if the requisite information is not available.

Unfortunately, Table A1 gives no information on the manner in which the data were obtained. The data are admittedly imaginary, but we are not told from what imaginary source (interviews, a survey of medical records, a case notification system, etc.) they are derived. This uncertainty will have to be taken into account when we later go on to consider possible explanations for the findings. In extreme instances, such serious doubts about the accuracy of the data may arise at this point that further consideration of the findings may be deemed superfluous.

Also, we are unfortunately not told whether the "cases" in Table A1 are *individuals* who had gastroenteritis, or are *episodes* (spells) of illness. If the same child had the disease twice in one year, did he or she count as one case or as two? (In answer to an SOS, the honorary official epidemiologist in Epiville tells us that the table actually refers to spells of illness.)

Summarizing the Facts

Obviously, there was a rise in the number of cases per year between 1970 and 2000. A full summary of the facts in Table A1 would mention at least three features of this increase:

1. The continuing, or "monotonic" (see Note A2–1), nature of the increase—that is, the occurrence of a rise between each observation and the next.
2. The overall extent of the increase. This may be expressed in absolute or relative terms. The absolute difference is 800 cases per year (1,200 minus 400). The relative difference can be expressed as a simple ratio: 1,200/400 (i.e., 1,200 divided by 400)—a threefold increase. Alternatively, it can be stated as a percentage change: $[(1{,}200 - 400)/400] \times 100$—a rise of 200%.
3. The variation in the rate of change. The trend is not uniform: the increase is steeper in earlier than in later periods. This variation is apparent whether we look at the absolute or relative changes in the numbers of cases. The absolute differences between successive observations are 200 for each of the first two intervals, and only 100 for each of the subsequent intervals. If you have not already done so, examine the relative changes as well, by calculating the ratio of each observation to the preceding one, and/or the percentage change between each pair of successive observations. (For answers, see Note A2–2.)

When you listed or summarized the facts, you may have included such items as "sanitary conditions got worse," "the population grew in size," or "the number of deaths from gastroenteritis increased." These are *not* empirically observed facts; they are *inferences*. They may or may not be true, and they should not be regarded or reported as facts. It is usually important to consider possible explanations for the observations, but only *after* the facts themselves have been determined. (Sometimes, of course, there is no need to go beyond determining the facts. These may be all we want, and there may be no interest in drawing inferences or finding explanations.)

Table A2–1. Number of Cases of Influenza

	Wuntown	Nuthertown
1998	500	5,000
2000	200	4,000

Exercise A2

In Table A1, we saw that initially there was a steep increase in the annual number of cases of gastroenteritis in Epiville, and later the rise became less steep. This change in trend was obvious whether we looked at the absolute changes or the relative ones. Sometimes, however, absolute and relative differences may give us conflicting messages, and we may have to decide which mean more to us.

Question A2–1

Table A2–1 shows the numbers of cases of influenza in two imaginary towns in 1998 and 2000. Health programs for preventing influenza were introduced in both towns in 1999. Calculate the absolute and relative changes in each town. In which town is there stronger evidence that the program was effective in reducing the occurrence of influenza?

Question A2–2

You are a health administrator concerned with the provision of facilities for health care. Table A2–2 shows the numbers of new patients with end-stage renal disease who required renal dialysis (a life-saving but elaborate and expensive form of treatment) in two regions in 1998 and 2000. Calculate the absolute and relative changes. Looking forward to 2001, in which region would you be more concerned about the increase?

Question A2–3

Table A2–3 shows the numbers of infant deaths in the same two regions in 1998 and 2000; the numbers of births did not change. Programs aimed at reducing infant mortality were started in both regions in 1999.

Table A2–2. Number of Patients Requiring Dialysis

	Pepi	Quepi
1998	30	2,000
2000	90	3,000

Table A2–3. Number of Infant Deaths

	Pepi	Quepi
1998	300	5,000
2000	60	4,000

1. In which region is there more convincing evidence that the reduction in mortality was caused by the program?
2. If the program can be continued in only one region, which would you choose? (Assume that the reductions are caused by the programs.)

Question A2–4

Can you suggest a rule of thumb for deciding when to use the relative difference and when to use the absolute difference?

Notes

A2–1. *Monotonic sequence.* A sequence is monotonically increasing if each value is more than or equal to the previous one, and monotonically decreasing if each value is less than or equal to the previous one. If each value is more than the preceding one, or if each value is less than the preceding one, the sequence is strictly monotonic (increasing or decreasing).

A2–2. The successive ratios were 1.50, 1.33, 1.12, 1.11, 1.10, and 1.09. The percentage changes were 50%, 33%, 12.5%, 11%, 10%, and 9%.

■■■■■■■■■■ Unit A3

Absolute and Relative Differences

In some circumstances we may be more interested in absolute differences; and in others, in relative differences.

In answer to *Question A2–1,* Table A2–1 shows a larger relative decrease in influenza in Wuntown (60%) than in Nuthertown (20%), and a larger absolute decrease in Nuthertown (1,000) than in Wuntown (300). The evidence that the program was effective is stronger in Wuntown, where over half the cases were apparently prevented. In this context, the relative difference is more meaningful.

In answer to *Question A2–2,* Table A2–2 shows a larger absolute increase in patients needing renal dialysis in Quepi (1,000) than in Pepi (60), and a larger relative increase in Pepi (200%) than in Quepi (50%). The administrator would

probably be more concerned with the change in Quepi, where the personnel, equipment and other facilities needed to treat a very large additional number of patients must be found. In this context, the absolute difference is more meaningful.

In answer to *Question A2–3,* the evidence that the program was effective is more convincing in Pepi, where the number of infant deaths decreased by 80%, than in Quepi, where the relative decrease was only 20%. But the program apparently prevented 1,000 deaths in Quepi in 1999, and only 240 in Pepi. If we had to choose, we would probably decide to continue the program in Quepi, where more lives are saved.

In answer to *Question A2–4,* a general rule of thumb is that when we are concerned with the magnitude of a public health problem—how many lives, how many facilities, how much cost—it may be appropriate to give emphasis to absolute rather than relative differences. Relative differences, on the other hand, are of more interest when we wish to study processes of causation—for example, to examine the effect of health care or of a supposed risk factor or protective factor, on the occurrence of diseases or deaths. It is not always easy to choose between the use of relative and absolute differences, and sometimes both are important.

Exercise A3

Diagrams are often used to summarize and clarify findings. They provide a useful way of showing trends and differences at a glance.

In this exercise you are asked to draw diagrams by hand and interpret them, although in real life you might use one of the many computer programs that draw diagrams.

Question A3–1

Draw a graph showing the data of Table A1. Put the scale for numbers of cases (i.e., the dependent variable—see Note A3–1) along the *Y* (vertical) axis, and put the scale for time (the independent variable) along the *X* (horizontal) axis. It is customary to use the *Y* axis for dependent variables and the *X* axis for independent variables. Use ordinary (arithmetic) scales along both axes.

Question A3–2

Draw another graph showing the data of Table A1. Again use an ordinary scale for time, but this time use a logarithmic scale for numbers of cases. This is easy if you have semilogarithmic graph paper (see Note A3–2). If you have only ordinary graph paper, plot the logarithms of the numbers of cases instead of the actual numbers (see Note A3–3). If you have forgotten what logarithms are, see Note A3–4.

Table A3. Occurrence of Cases of Acute Gastroenteritis in Epiville in 1998

Period	No. of Cases
January–March	60
April–June	150
July	280
August–September	300
October–December	210
Total	1,000

Question A3–3

Which scale—ordinary or logarithmic—is more appropriate for showing absolute differences, and which one gives a better representation of relative differences? If the answer is not obvious to you, examine the absolute and relative changes displayed by the following two sequences of values, and then plot them against both kinds of scale. In each instance, use 1, 2, 3, 4, 5, 6, and 7 on the horizontal axis.

Sequence A: 1, 3, 5, 7, 9, 11, 13.
Sequence B: 1, 2, 4, 8, 16, 32, 64.

Question A3–4

Draw a diagram to summarize the data provided in Table A3 on the distribution of gastroenteritis during the year.

Question A3–5

Figure A3–1 shows the change in mortality from ischemic heart disease of males and females in the Philippines between 1964 and 1976. (At last! Real data!) In which sex was there more change? The actual figures (rates per 100,000) were: males, 33.3 (1964), 40.3 (1968), 55.8 (1972), and 78.0 (1976); females: 15.4, 18.4, 25.2, and 34.5, respectively (Note A3–5).

Question A3–6

Figure A3–2 (more real data!) shows the change in the rate of suicide among unemployed men and women in Italy between 1982 and 1991 (Note A3–6). Notice the use of a logarithmic scale. The relative increase over time is greater in women than in men. Might the absolute increase be larger in men? How could you find out?

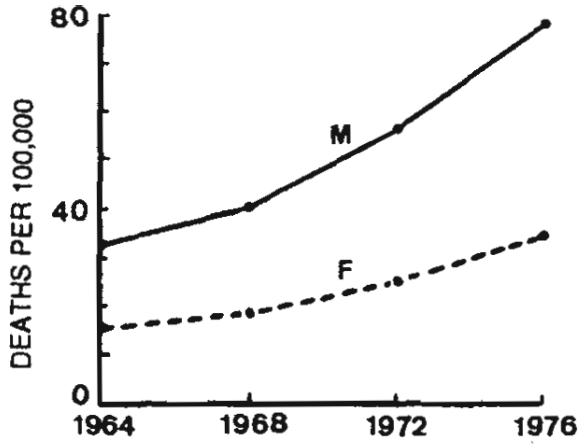

Figure A3–1. Mortality from ischemic heart disease, Philippines, 1964–1976. M = males; F = females. (Data from Ruomilehto et al., 1984.)

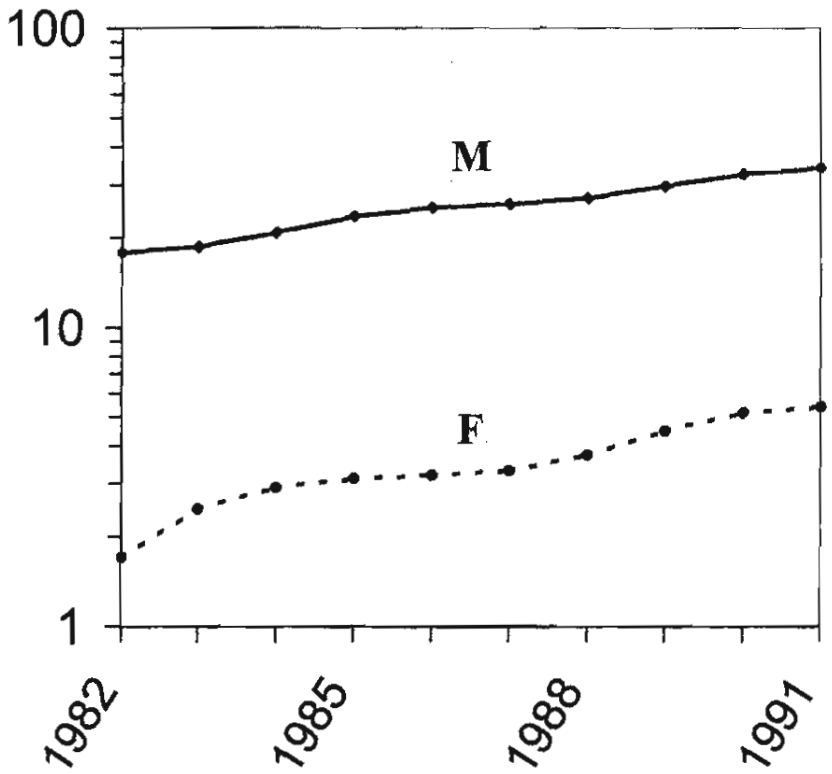

Figure A3–2. Suicide rates among unemployed in Italy, 1982–1991. Logarithmic scale. M = males; F = females. (Data from Preti and Miotto, 1999.)

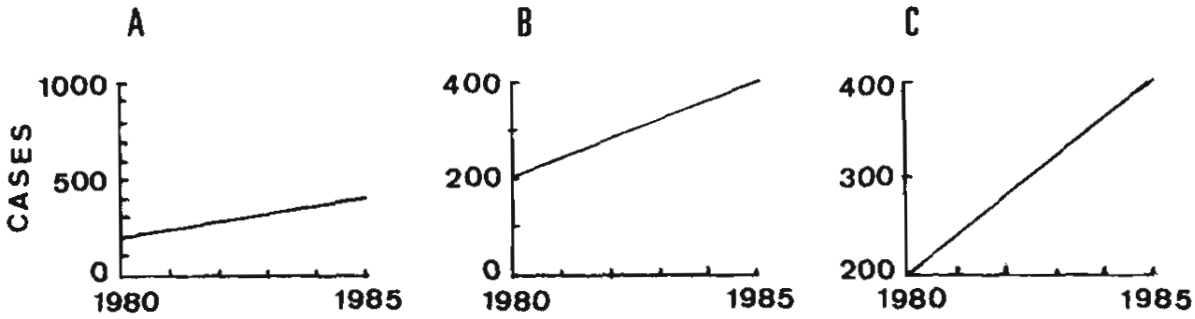

Figure A3–3. Cases of diseases A, B, and C, 1980–1985.

Question A3–7

The three graphs in Figure A3–3 show the changes in the annual number of cases of diseases A, B, and C between 1980 and 1985. Which disease showed the biggest change, and which the smallest?

Notes

A3–1. A *dependent variable* is "a variable the value of which is dependent on the effect of other variable(s)—independent variables—in the relationship under study. A manifestation or outcome whose variation we seek to explain or account for by the influence of independent variables"—*A Dictionary of Epidemiology* (Last, 2001).

A3–2. Semilogarithmic paper has a logarithmic scale along the Y (vertical) axis, and an ordinary (arithmetic) scale along the other. You need not look up logs; just plot the numbers against the scale. The paper probably has figures from 1 to 10 printed along the Y axis (starting at the bottom), and then another set of figures from 2 to 10; take the second set to represent 20, 30, 40, and so on—up to 100; if there is a third set, it will represent 200, 300, and so on—up to 1,000. If you had smaller values to plot, you could designate the first set of figures as (say) 0.1 to 1 and the second as 2 to 10. A logarithmic scale has no zero.

A3–3. If you have ordinary graph paper, use a table of logs or a pocket calculator to obtain the logarithms of the numbers of cases, and then plot these logs on an ordinary (arithmetic) scale. Instead of 400, plot its log, which is 2.60; instead of 600, plot 2.78, and so on.

A3–4. To refresh your memory about logarithms, the log of 100 is 2, because common logs use 10 as their base, and 100 is 10^2. The antilog (or exponential) of the log 2 is 100. Every positive number has a log, and the logs and antilogs can be obtained from tables, calculators, or computers. Adding two logs and then taking the antilog of their sum is equivalent to multiplying the numbers they represent: if the logs are 2 and 3 (representing 100 and 1,000) their sum is 5, the antilog of which is 100,000. Similarly, if the absolute difference between two logs is x, this means that one of the numbers they represent is antilog (x) times as large as the other; the difference between the logs of 1,000 and 100 is 1, which is the antilog of 10; this tells us that the ratio of 1,000 to 100 is 10. The ratio is also 10 for any other two numbers whose logs differ by 1. On a logarithmic graph, the distance between two points (which represents the absolute difference between the logs) thus expresses the ratio or relative difference between the numbers they represent. Use is often made of *natural logs*, which have a mysterious number called e, the value of which is about 2.71828, as their base.

A3–5. Data from Tuomilehto et al. (1984). The rates are age-standardized rates for the 35–64 age group.

A3–6. Data from Preti and Miotto (1999). The curves were smoothed by the running-medians procedure, using SMOOTH, a computer program in the PEPI package (see Note A3–7). Smoothing by eye can produce misleading curves, and

it is wise to be suspicious of smoothed curves if the method of smoothing is not specified.

A3–7. Most of the statistical procedures mentioned in this book can be performed by programs in the PEPI package, a set of over 40 statistical programs for epidemiologists (Abramson and Gahlinger, 2001). The package can be downloaded free; to find a convenient source, contact www.shareware.com and search for "pepi" in the "DOS" category. The programs are in DOS format, but can be run in Windows. For installation programs (not essential) and a manual, contact www.sagebrushpress.com. Some PEPI programs have been rewritten in a Windows format and can be downloaded free from www.myatt.demon.co.uk/index.htm.

For other free statistical software, try

www.vetmed.wsu.edu/courses-jmgay/EpiLinks.htm
www.undp.org/popin/softproj/software/software.htm or
www.softseek.com/Education_and_Science/Math/Statistics/

(but these links may be out-of-date: the Internet keeps changing).
Epidemiological software packages are reviewed by Goldstein (2000).

■■■■■■■■■■ Unit A4

Diagrams

The graphs requested in *Questions A3–1* and *A3–2* should have a general resemblance to those shown in Figure A4–1. In graphs (line diagrams) like these, the slope represents the rate of change: the steeper the slope, the more the change. Rates of change can be compared by comparing different segments of a line, or by comparing different graphs (*but only if they are plotted against the same scales*).

In answer to *Question A3–3,* the slope of a graph plotted against an ordinary (arithmetic) scale represents the rate of absolute change, whereas the slope of a graph plotted against a logarithmic scale represents the rate of relative change. Sequence A (1, 3, 5, 7, etc.) displays a constant rate of absolute change (an increase of 2 between each pair of numbers) and a decreasing rate of relative change (the percentage increase between successive numbers decreases from 200% to 18%). When an arithmetic scale is used, the graph is a straight line, showing that the rate of absolute change is constant; but a logarithmic scale provides a curve that rises steeply at first, and then progressively rises less steeply (Fig. A4–2). Sequence B (1, 2, 4, 8, etc.), conversely, displays a constant rate of relative change (each number is double the previous one), and a logarithmic

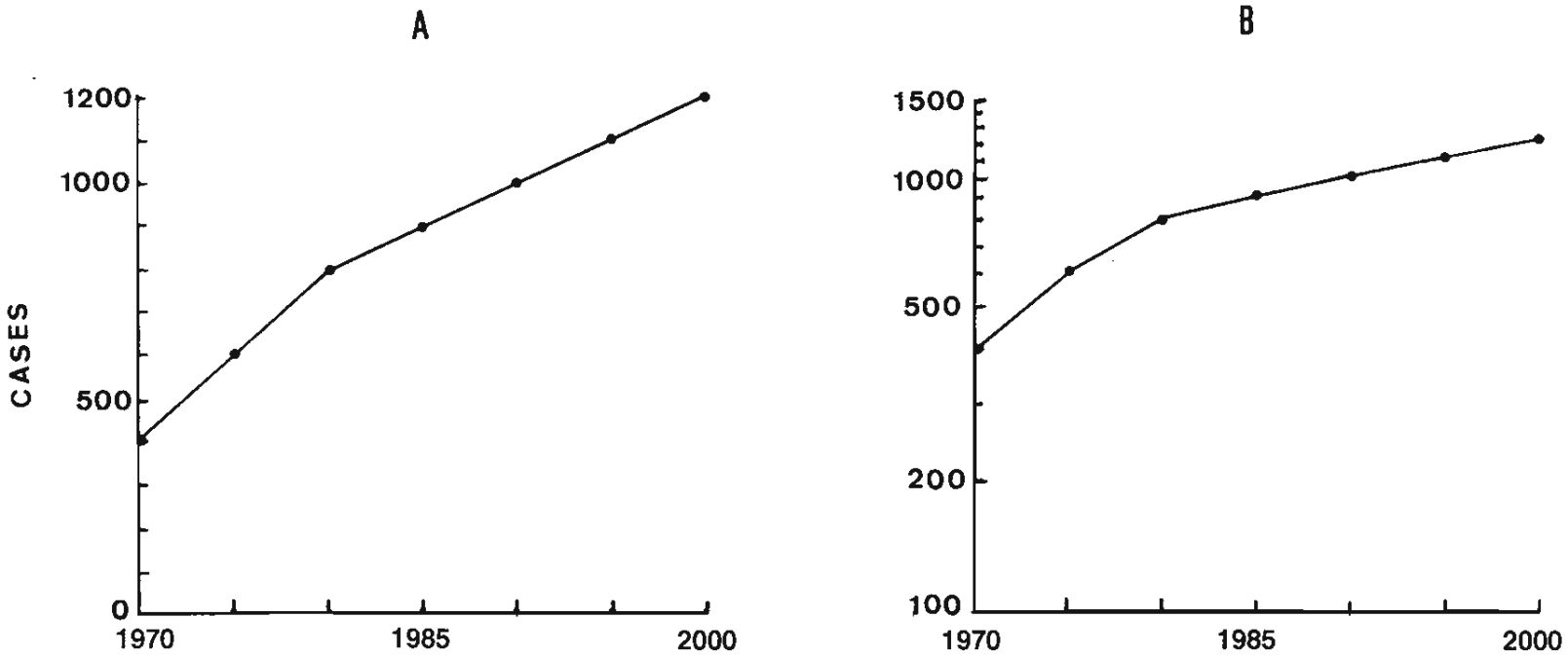

Figure A4–1. Cases of acute gastroenteritis, 1970–2000. (A) Arithmetic scale; (B) logarithmic scale.

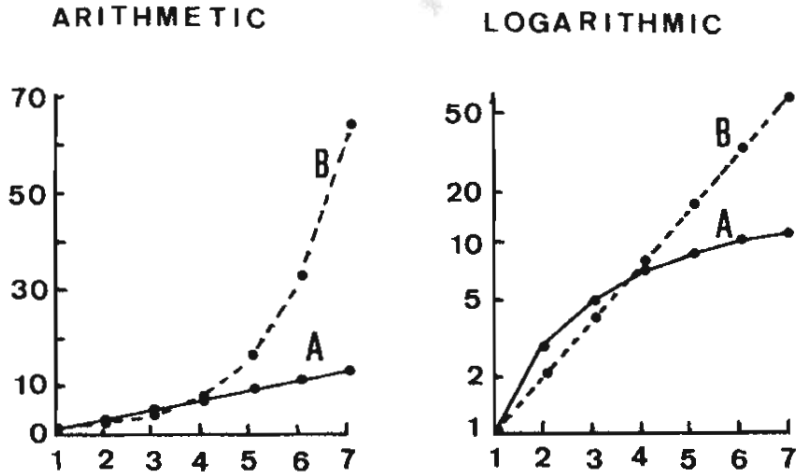

Figure A4–2. Comparison of arithmetic and logarithmic scales. Sequence A: 1, 3, 5, 7 9, 11, 13. Sequence B: 1, 2, 4, 8, 16, 32, 64.

scale therefore provides a straight-line graph. There is an increasing rate of absolute change (the successive changes increase from 1 to 32), and an arithmetic scale shows a progressively steeper rise.

Both of the graphs based on Table A1 (Fig. A4–1) show a slowing in the tempo of change, providing a pictorial summary of our previous observation that the increase in cases of gastroenteritis was steeper in earlier than in later years, whether we looked at absolute or relative changes.

Various kinds of diagrams are shown in Figure A4–3. You may have used one of these in answering *Question A3–4.* Different diagrams are appropriate in different circumstances.

In this instance, where the data (Table A3) refer to periods of different lengths, the diagrams in the top row of the figure may be misleading. These are a *bar diagram,* in which the height of the bar portrays the number of cases in each period, a *line graph* (or *curve*) in which each period is represented by a single point, and a *pie chart* showing the proportion of cases in each period. (To draw a pie chart, calculate the degrees for each segment by multiplying the percentage in the segment by 360/100, i.e., 3.6.) Better solutions are shown in the bottom row of Figure A4–3. The best diagram when successive values represent ranges that differ in width, as in this instance, is probably a *histogram.* This comprises adjacent blocks whose widths are proportional to the class interval (the number of months), and whose areas are proportional to the number of cases. The height of the blocks portrays, not the number of cases but the number of cases divided by the class interval (e.g., 20 instead of 60 for the 3-month January–March period). Note how the bar diagram and histogram give quite different impressions. Use may also be made of a *frequency polygon,* which is a line diagram constructed from a histogram; it is the dotted line in Figure A4–3. The same rules for choosing an appropriate kind of diagram and for correctly presenting the data apply both to computer-drawn and hand-drawn diagrams.

In answer to *Question A3–5,* Figure A3–1 clearly shows a steeper increase in mortality from ischemic heart disease among men. But an arithmetic scale was

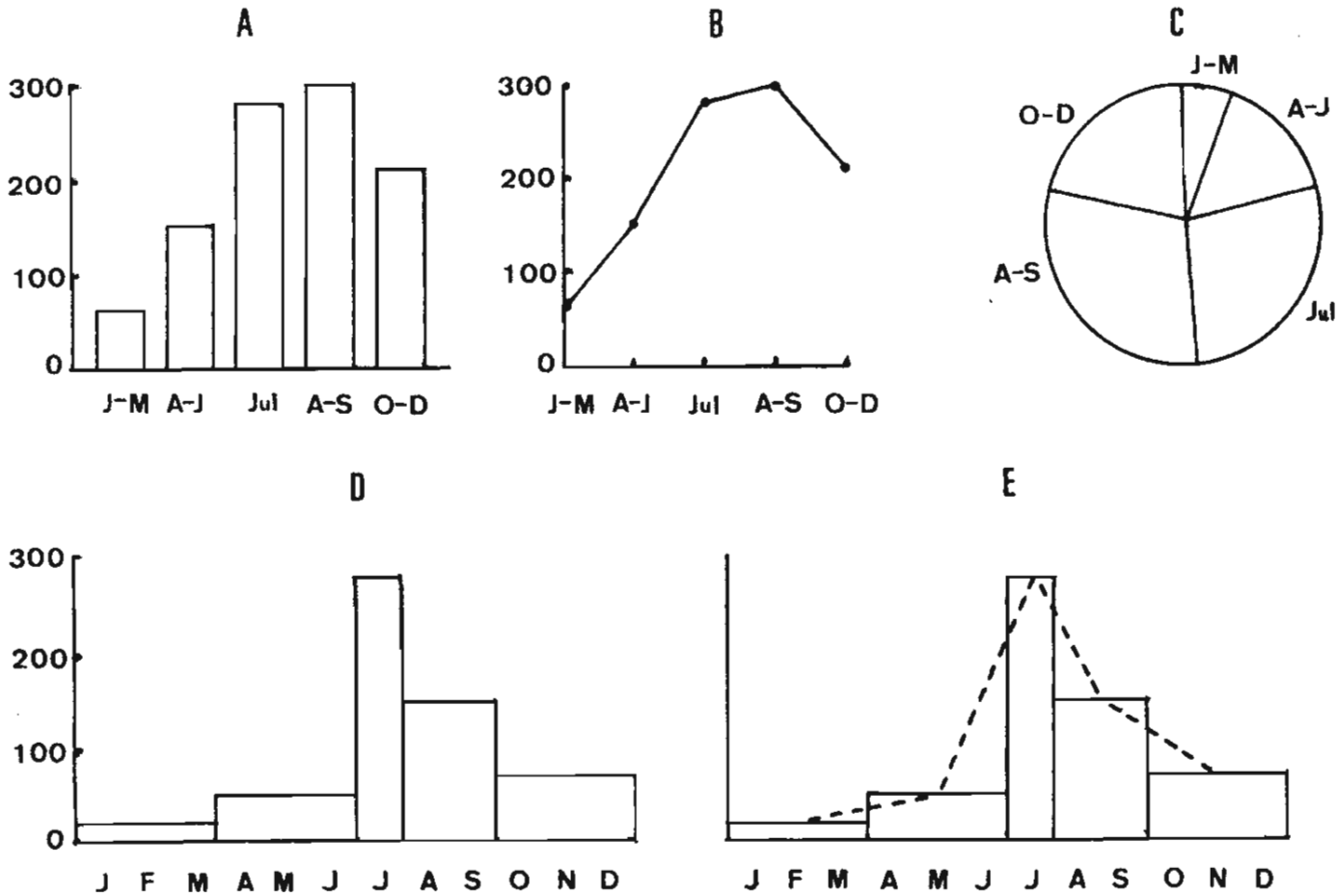

Figure A4–3. (A) Bar diagram; (B) line diagram; (C) pie chart; (D) histogram; and (E) frequency polygon. [J–M = January to March, etc.]

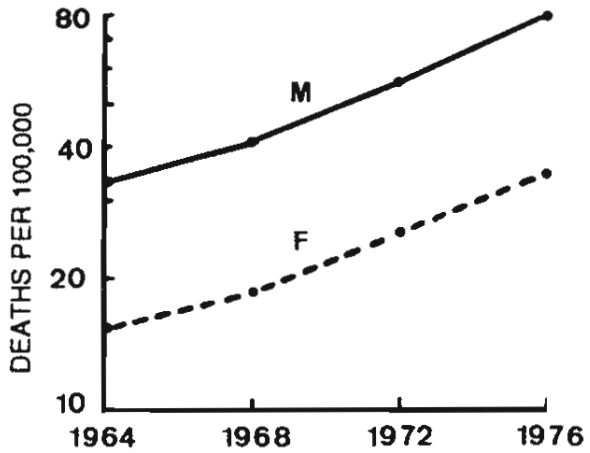

Figure A4–4. Mortality from ischemic heart disease, Philippines, 1964–1976. Logarithmic scale.

used, and it is only the *absolute* change that is greater. If we plot the same data against a logarithmic curve (Fig. A4–4) we see that the *relative* change—which may be of more interest to us—is about the same in the two sexes.

The absolute increases in suicide rates in unemployed men and women (*Question A3–6*) could be compared by using an arithmetic scale. This is done in Figure A4–5, which shows that the absolute increase in suicide rates is much larger in men. One could also appraise the absolute and relative changes non-graphically, by calculating them from the rates at the beginning and end of the period.

Question A3–7 shows how easily graphs can mislead. The three graphs in Figure A3–3 present identical data—a steady rise from 200 in 1980 to 400 in 1985.

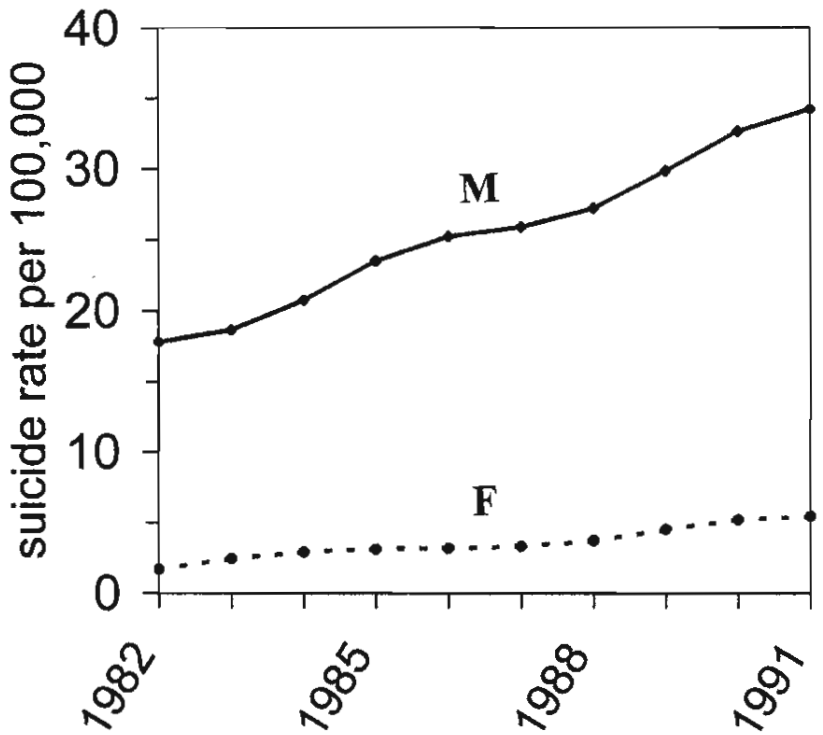

Figure A4–5. Suicide rates among unemployed in Italy, 1982–1991.

The first graph looks flat because the vertical scale is compressed, whereas the third one looks steep because the vertical scale is spread out and because it does not begin at zero. (This is the easiest way to give a deceptive impression of the facts.) Care should be taken when presenting and reading diagrams.

Exercise A4

Question A4–1

Let us come back to Epiville. Both in words and in pictures, we have summarized the facts about the rise in cases of gastroenteritis between 1970 and 2000 (Table A1). Now let us consider possible explanations. What explanations can you suggest?

Question A4–2

There is an important principle of economy in scientific thinking, often called *Occam's razor.* William of Occam (c.1285–c.1349) was an English philosopher who formulated the maxim, *Entia non sunt multiplicanda praeter necessitatem*—that is, "assumptions to explain a phenomenon must not be multiplied beyond necessity." In 1853, Sir William Hamilton termed this the "law of parsimony" and expressed it as follows: "Neither more, nor more onerous, causes are to be assumed, than are necessary to account for the phenomena." Karl Pearson (1892), in *The Grammar of Science,* calls this canon of economy "the most important in the whole field of logical thought."

Which of the explanations you listed in your answer to Question A4–1 would you test first? What additional information do you need to test it? If you can, formulate a specific hypothesis for testing.

■■■■■■■■■■ Unit A5

Seeking Explanations for the Facts

Your list of possible explanations for the findings in Epiville (*Question A4–1*) may include a wide variety of factors that could have led to an increase in the number of cases of gastroenteritis—a worsening of sanitary conditions, changes in infant feeding practices, an increase in population size, and so on. However long or short the list of possible causes, it is important, nevertheless, that "non-causal" explanations also be considered.

First, it is possible that the occurrence of the disease did *not* actually increase; the rise may be not a fact but an *artifact,* attributable to a flaw in study methods. The increase may, for example, have been only in the number of cases that were *identified,* rather than in the number that *occurred.* This might be due to an im-

provement in the completeness of clinical records, to an increase in the public's readiness to use medical services, and so forth.

Second, consideration should also be given to the possibility that the apparent upward trend is due solely to *chance*. We possess data for 7 of the 31 years in the period from 1970 to 2000. It is possible that the number of cases varied randomly from year to year during this period, with no upward trend, but that merely by chance the particular seven observations that were selected show a rise. Most other sets of seven observations might have shown no rise. We cannot completely exclude this possibility. But common sense suggests that it is extremely unlikely, and we would probably decide that we can safely ignore it. If we are in doubt, we can do a test of *statistical significance* to help us make a decision. Actually, an appropriate significance test reveals that if in fact there is no increase in the number of cases with time, the probability that a sample of seven observations would display a monotonic increase is only 2 in 10,000 ("P = .0002"). This probability is so low that we would certainly decide to regard the finding as *nonfortuitous* (i.e., not due to chance).

These two questions—*Is the finding actual or artifactual?* and *Can the finding safely be regarded as nonfortuitous?*—should always be asked, and are often the first ones asked.

Keeping Occam's razor in mind, the first explanation chosen for testing (*Question A4–2*) should be one that, if confirmed, would go a long way toward explaining the findings. The explanation should also be a testable one; there is little point in selecting it for testing—however cogent the reasons may be—if the requisite data cannot be obtained. Use these two criteria in appraising your choice of an explanation for testing.

In this instance, most epidemiologists would probably agree that the chief possible explanation for the increase in cases of gastroenteritis in this town in a developing region is that the population increased between 1970 and 2000, so that there was a rise in the number of individuals who were at risk of incurring the disease. This possibility should probably be explored before serious consideration is given to *any* other explanation.

This requires data on the size of the population in the period under consideration. We will examine such data in the following exercise. The method usually used is to calculate and compare gastroenteritis rates per (say) 1,000 population. We will do this in a subsequent exercise.

Testing Explanations

To test an explanation we usually require additional information, drawn from the same study or from another one. We can then see whether the new facts are consistent with the explanation. If they are, our explanation may be (but is not necessarily) correct; if they are not, we can rule out the explanation.

When we seek new information, we should know *why* we want it and *how* we will use it. This enables us to be selective both in seeking and in appraising in-

formation. In the present instance, if we can pinpoint the population findings that would explain the increase in cases, we will know exactly what to look for. Our hypothesis is that the population has grown in the same way as has the number of cases. The specific hypotheses are therefore that

1. There was a monotonic increase in the size of the population.
2. There was a threefold increase between 1970 and 2000. (We specify a *relative* increase, because we can assume that a tripling of the number of cases would be associated with a tripling of population size.)
3. The trend in population size changed in the same way as did the trend in the number of cases; that is, there was a rapid increase in earlier years and a slow increase in later years.

If these specific hypotheses are not confirmed, population growth cannot be the sole explanation for the increase in cases.

To appraise your formulation of a specific hypothesis (in your answer to *Question A4–2*), ask whether it is testable and whether, having obtained the new information you requested, you would be able to come to a clear decision as to whether your explanation is tenable. Can the new information refute the hypothesis?

Exercise A5

Table A5–1 provides information about population size. You may assume that the figures are accurate. The table shows the average population of Epiville in the given year—that is, the mean of the population at the beginning and end of the year.

Question A5–1

Summarize the facts in Table A5–1.

Table A5–1. Population of Epiville in Selected Years, 1970–2000

Year	Population
1970	20,000
1975	30,000
1980	40,000
1985	45,000
1990	50,000
1995	55,000
2000	60,000

Table A5–2. Deaths From Choking on Food* in Infants Aged Under One Year, England and Wales, 1974–1984

Year	No. of Deaths
1974	126
1975	93
1976	97
1977	97
1978	90
1979	110
1980	74
1981	62
1982	41
1983	29
1984	30

*"Inhalation and ingestion of food causing obstruction or suffocation," code E911 in the International Classification of Diseases.

Question A5–2

Can the increase in cases of gastroenteritis in Epiville be completely explained by the change in population size?

Question A5–3

Choking on food is an important cause of accidental deaths in infancy. Information about deaths from this cause in England and Wales is shown in Table A5–2 (data from Roper and David, 1987).

Summarize the facts, list the possible explanations for the decrease between 1979 and 1984, select one explanation for testing, and state how you would test it.

■■■■■■■■■■ Unit A6

The Basic Scientific Process

The sequence we are following is the one we should adopt whenever we examine a table or graph: first, determine and summarize the facts; then, formulate possible explanations; and then, decide what additional information is needed to test the explanation (or for other reasons). There is often a temptation to start

by saying "These data tell me nothing, because I don't have information on such-and-such" (e.g., "because I don't have information about population size"). It is generally more helpful, however, if we first see precisely what the data *do* tell us and only then decide what extra information to seek.

It may be helpful to look at this procedure in the context of the process of scientific inquiry as it is used in epidemiology (Note A6–1). There are two basic approaches. The *inductive* approach, which moves from the particular to the general, starts with observed facts, which form the basis for inferences; whereas the *deductive* approach, which moves from the general to the specific, starts with a theory or hypothesis that can be proved false by observed facts. In practice (and despite philosophical objections), consistent failure to find facts that falsify a hypothesis may be taken as support for its validity—that is, as verification.

Combining these two approaches, the basic scientific process is:

- If there is no hypothesis:

  Observe and consider the facts.
  Formulate hypotheses that explain them.

- If there is a hypothesis (which may be derived from the facts):

  Observe and consider the new facts.
  See whether they refute or conform with the hypothesis.

- If the hypothesis is refuted, or if there are new ideas (which may be derived from the new facts):

  Formulate new or modified hypotheses.
  Seek information that can refute them.
  Observe and consider the new facts.
  See whether they refute or conform with the hypotheses.

and so on.

The procedure we have been following (determine the facts, then formulate possible explanations, and then decide what additional information is needed) is the one to be used whenever we "observe and consider the facts."

To test whether the increase in cases of gastroenteritis in Epiville is explained by a change in population size, we formulated three specific hypotheses, or refutable predictions, and obtained new facts to test them (*Question A5–1*). The new facts show that the changes in population size paralleled the changes in the occurrence of cases. The increase was monotonic, the overall increase was threefold, and the relative changes in successive 5-year periods were identical with those observed for gastroenteritis (percentage changes of 50%, 33%, 12.5%,

11%, 10%, and 9%, respectively). You may have drawn a graph to show the change in population size. If you used the same logarithmic scale as you used for cases of gastroenteritis (*Question A3–2*), you obtained a curve parallel to the previous one, showing that the trends of relative change were identical.

In answer to *Question A5–2,* therefore, the change in population size can completely explain the increase in cases. The explanation is not refuted.

The data on infant deaths in Table A5–2 are real, and do not display the smooth trends that characterize fictional data. Your summary (*Question A5–3*) should include the fact that the annual number of deaths from choking on food declined monotonically between 1979 and 1983, and remained low in 1984. The annual numbers in 1980–1984 were lower than in previous years, and in 1983 and 1984 they were one-third or less than those in any year between 1974 and 1979. You may also have mentioned the stability of the annual number between 1975 and 1978, and the sharp peaks in 1974 and 1979.

Possible explanations for the decline after 1979 include

1. A decrease in the annual number of births. This explanation can be tested by seeing whether there was a decline in births, paralleling the change in deaths from choking. Alternatively, we could examine *rates,* rather than numbers, of deaths from choking. The specific hypothesis (or refutable prediction) would be that the rate did not decline during this period; if it did, the decrease in deaths cannot be attributed solely to this cause.
2. A change in doctors' habits of death certification. During this period there was a rise in reported deaths due to sudden infant death syndrome (SIDS), and possibly deaths were assigned to SIDS that would previously have been attributed to choking. We might examine the annual numbers of deaths from both these causes (combined), to see whether the overall number decreased.
3. Chance variation. This seems an unlikely explanation, but if we wish we can do a test of statistical significance.
4. Changes in infant feeding practices. This is the most important possibility, as it might point the way to preventive measures; but "noncausal" explanations require rebuttal first.

In a discussion of these ratings, Roper and David (1987) concluded that the fall in deaths was not merely a reflection of the decline in births, as the infant mortality rate attributable to choking fell in this time from 0.23 to 0.05 per 1,000 live births in boys, and from 0.16 to 0.05 in girls. They pointed out that the pattern of change of SIDS deaths was different, reaching a peak in 1982 and declining slightly in 1983 and 1984. The explanation they favored was a change in infant feeding practices; they pointed out that since the early 1970s, when it was recommended that the early introduction of solid food should be avoided, there had been a decrease in the proportion of infants receiving solid foods before the age of 3 months. According to surveys in England and Wales, this proportion was 85% in 1975 and 55% in 1980.

Rates

Information about the frequency of an event in a group or population is commonly summarized by dividing the number of events (the *numerator*) by a suitable *denominator* (e.g., the number of people in the group or population). This controls for the effect of the size of the denominator on the number of events. The result is generally multiplied by 100, 1,000, or another convenient figure. For simplicity, we will refer to all measures of this kind as *rates,* although (as we will see in Unit B1) this term is often defined more strictly.

Incidence rates can be computed in different ways, as we will see later (Unit B5). In the following examples, they refer to the occurrence of events in a given population during a specified period. An *incidence rate (spells)* is based on the number of spells (episodes) of disease, and an *incidence rate (persons)* on the number of people who incur the disease (each person can appear in the numerator only once). Death rates (*mortality rates*) are incidence rates that measure the frequency of deaths. By convention, the *infant mortality rate* is the number of infant deaths (under the age of 1 year) divided by the number of live births during the same period.

Exercise A6

Question A6–1

You will be asked to calculate the annual rates of gastroenteritis per 1,000 population in Epiville between 1970 and 2000. Before you do so, can you say what findings you would expect if the increase in cases of gastroenteritis is completely explained by the incidence in population? In other words, formulate a specific hypothesis for testing.

Question A6–2

Calculate the annual incidence rates of gastroenteritis per 1,000 population in Epiville between 1970 and 2000, using the numbers of episodes (Table A1) as numerators and the average population figures (Table A5–1) as denominators. The formula is

$$\frac{\text{Number of episodes}}{\text{Average population}} \times 1{,}000$$

Question A6–3

Can you draw an inference about the risk of acute gastroenteritis for an individual in Epiville during this time? (If you want definitions of "risk," see Note A6–2.)

Question A6–4

Is there any possibility that the risk of incurring acute gastroenteritis for an individual in Epiville actually *decreased* between 1970 and 2000? Is there any way

in which this kind of confusion could occur? (In answering this question, you may assume that the information on incidence and population size is accurate.)

Question A6–5

In a given year the incidence rate (persons) of acute gastroenteritis was 10 cases per 100 population in region A, and 5 per 100 population in region B. The population size was 10,000 in region A and 5,000 in region B. Which (if any) of the following statements are true?

1. There were the same numbers of cases in both regions.
2. There were twice as many cases in region A as in region B.
3. There were four times as many cases in region A as in region B.
4. The risk of incurring the disease during the year was about the same for individuals in the two regions.
5. The risk of incurring the disease during the year was twice as high for individuals in region A as for those in region B.
6. The risk of incurring the disease during the year was four times as high for individuals in region A as for those in region B.
7. The incidence rate in the total area (A and B combined) was 7.5 per 100 population.
8. The incidence rate in the total area (A and B combined) was 15 per 100 population.

Notes

A6–1. If you wish to embark on the deep waters of the philosophy of epidemiologic research and plumb the acceptability of *inductive reasoning* (i.e., inferring a general law or principle from the observation of particular instances) as opposed to *deductive reasoning* (which leads to the use of observations to test hypotheses), see Greenland (1998a) and the diverse views expressed in collections edited by Greenland (1987) and Rothman (1988). For a simple commonsense approach, see Susser (1973, 1987). The central question is: "Besides refuting the hypothesis that the Earth is flat, can we not affirm that it is spherical? To naysayers we may retort, did Magellan circumnavigate the world, or did he and his shipmates cook the results of the voyage of 1519 to 1522? And what of the thousands who have followed under sail or steam or on the wing?" (Susser, 1988). "All of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature. . . . The tentativeness of our knowledge does not prevent practical applications, but it should keep us skeptical and critical" (Rothman and Greenland, 1998, p. 22).

A6–2. "*Risk.* The probability that an event will occur (e.g., that an individual will become ill or die within a stated period of time or age). Also, a nontechnical term encompassing a variety of measures of the probability of a (generally) unfavorable outcome"—*A Dictionary of Epidemiology* (Last, 2001). "Risk is defined as the probability of a disease-free individual's developing a given disease

over a specified period, conditional on that individual's not dying from any other disease during the period" (Kleinbaum et al., 1982).

■■■■■■■■■■ Unit A7

Rates (Continued)

In answer to *Question A6–1,* if the increase in gastroenteritis is completely explained by the increase in population, we would expect the incidence rate to be the same each year. The specific hypothesis for testing is that there was no change in the annual incidence rate between 1970 and 2000. When you calculated the rates (*Question A6–2*), you found that each year the rate was 20 per 1,000, in accordance with this hypothesis. The rate could also be expressed as 2 per 100, 200 per 10,000, etc., or as 0.02 (per 1).

The rate of incidence of an event in a population is an estimate of the risk (on average) for its individual members. (As we will see later, the accuracy of this estimate depends on how the rate was calculated.) As the rate was 20 episodes of gastroenteritis per 1,000 population per year, individuals in Epiville had a 20 in 1,000 (or 2%) risk of having an episode in each of the years for which data were available (*Question A6–3*).

We will return to *Question A6–4* at a later stage.

To answer *Question A6–5,* the numbers of cases in the two regions must be calculated. This is easily done:

$$\text{Rate per } 100 = \frac{\text{Number of cases}}{\text{Population}} \times 100$$

Hence,

$$\text{Number of cases} = \frac{\text{Rate per } 100}{100} \times \text{population}$$

Thus,

$$\text{Number of cases in region A} = (10/100) \times 10{,}000 = 1{,}000$$

$$\text{Number of cases in region B} = (5/100) \times 5{,}000 = 250$$

Statements (1) and (2) are therefore false; statement (3) is true.

As the annual incidence rate in region A was double that in region B, the risk for individuals was twice as high in region A. Statement (5) is therefore true, and statements (4) and (6) are untrue.

In the total area (regions A and B combined), the number of cases was (1,000

+ 250) = 1,250. The total population was (10,000 + 5,000) = 15,000. The overall rate was therefore (1,250/15,000) × 100, or 8.33 per 100. Statements (7) and (8) are thus both untrue. Statement (7) uses the simple average (mean) of the two rates, and statement (8) uses the sum of the rates. The overall rate is actually the *weighted mean* (see Note A7) of the two separate rates, using the population sizes as weights. *The contribution of a subpopulation to the findings in a total population depends on the relative size of the subpopulation.* This may be a truism, but as we will see later, it has important implications.

Inspecting a Two-Dimensional Table

Age is a variable whose role should be considered in all epidemiological studies; this is because health status is probably more strongly related to age than to any other personal characteristic. In the next exercise, we will therefore look at the age composition of the population of Epiville and examine its changes over the years. To do this we require a two-dimensional table (or cross-tabulation), in which population figures are shown both by age and by calendar year (Table A7–1).

When inspecting a table of this sort in order to determine and summarize the facts, it is generally advisable to do at least the following (not necessarily in this order):

- Examine each row (horizontal line) of figures.
- Compare the rows (look for similarities and differences).
- Examine each column.
- Compare the columns.

Here, each column represents the time trend in a specific age category. When examining the columns, you may use the same procedures that you used previously to examine the time trends in the population as a whole.

Table A7–1. Population* by Age in Selected Years (1970–2000)

	Age (Years)				
Year	0–4	5–14	15–44	≥45	Total
1970	1,400	3,000	8,000	7,600	20,000
1975	2,700	5,000	12,000	10,300	30,000
1980	4,600	9,000	15,000	11,400	40,000
1985	6,000	11,000	16,500	11,500	45,000
1990	8,000	12,000	18,000	12,000	50,000
1995	10,000	13,500	19,000	12,500	55,000
2000	11,500	15,000	20,500	13,000	60,000

*The average population in the given year is shown—that is, the mean of the population in the specific age group at the beginning and end of the year.

Table A7–2. Percentage Distribution of Population of Epiville by Age in Selected Years (1970–2000)

	Age (Years)				
Year	0–4	5–14	15–44	≥45	Total
1970	7.0	15.0	40.0	38.0	100.0
1975	9.0	16.7	40.0	34.3	100.0
1980	11.5	22.5	37.5	28.5	100.0
1985	13.3	24.4	36.7	25.6	100.0
1990	16.0	24.0	36.0	24.0	100.0
1995	18.2	24.5	34.5	22.7	100.0
2000	19.2	25.0	34.1	21.7	100.0

Each row shows the frequency distribution, by age, of the population in a given year. When examining frequency distributions it is generally helpful to calculate percentages, using the total (the row total) as the denominator. In the first row, for example, 1,400 is 7% of 20,000, 3,000 is 15%, and so on. These percentage distributions are shown in Table A7–2. In such a table it is helpful if "100%" is indicated in the appropriate places, in order to show what totals were used as denominators. Note that in one instance the percentages do not add up to precisely 100%; this discrepancy is caused by rounding-off, and is acceptable.

When we compare the columns in Table A7–2, we are examining time trends with respect to the *percentage* of the population in each age category. This overcomes the effect of the changes in the total size of the population.

Exercise A7

Question A7–1

Summarize the facts shown in Tables A7–1 and A7–2.

Question A7–2

What is the most plausible explanation for these changes in the age composition of the population? You may assume that the information is accurate.

Question A7–3

Could the changes in the age composition of the population of Epiville have influenced the incidence rate of acute gastroenteritis in the town?

Note

A7. The formula for the weighted mean M of a set of values x_i, where x_i is the value for group i, the size of which is N_i, is

$$M = \frac{\Sigma\,(x_i \cdot N_i)}{\Sigma\, N_i}$$

The symbol Σ (the Greek capital letter "sigma") means "the sum of the values of." In the present instance,

$$M = \frac{(10 \times 10{,}000) + (5 \times 5{,}000)}{10{,}000 + 5{,}000} = 8.33.$$

■■■■■■■■■■ Unit A8

Inspecting a Two-Dimensional Table (Continued)

In answer to *Question A7–1*, we want to examine both the age composition of the population in different years (the rows), and the time trends in population size in different age groups (the columns). Examining the rows, we see that both the absolute numbers (in Table A7–1) and the percentage distribution (in Table A7–2) changed from year to year. The only consistent features seen in Table A7–2 are that the 0–4 age group was the smallest category each year, and the 15–44 age group was the largest.

When we inspect the columns in Table A7–1, we see that in each age group there was a monotonic increase between 1970 and 2000. The relative increase during this period varied with age, being largest in children aged 0–4 years and smallest in the oldest group. The ratios of the 2000 figures to the 1970 ones in Table A7–1 were: 0–4 years, 8.2; 5–14 years, 5.0; 15–44 years, 2.6; and ≥45 years, 1.7. You may have summarized these findings by drawing a graph, using a logarithmic scale. Such a graph would clearly show the difference between the time trends in different age groups. It would also show that in each age group the trend of relative increase was steeper in 1970–1980 than in subsequent years.

Inspection of the columns in Table A7–2 shows very different time trends in the different age groups. The percentages in the 0–4 and 5–14 age groups tended to increase, whereas the percentages in the older groups decreased monotonically.

Note that the columns in Tables A7–1 and A7–2 show different relative changes. For the 0–4 age group, for example, the ratio of the 2000 figure to the 1970 one was 11,500/1,400 = 8.2 in Table A7–1, and only 19.2%/7.0% = 2.7 in Table A7–2. For the ≥45 year age group, the corresponding ratios were 1.7 and 0.6. Can you suggest a reason for these discrepancies? (For answer, see Note A8.)

Changes in age composition may be due to aging, inward or outward migration, births, and deaths. The most plausible explanation for the extreme change observed in this growing community is selective immigration (*Question A7–2*).

A high proportion of the added population apparently consisted of families with young children, born before or after entry into the town.

In answer to *Question* A7–3, we have previously seen that the overall rate in a population is a weighted mean of the rates in its constituent subpopulations, and that the relative size of each subpopulation determines its contribution to the findings in the total population (see *Question A6–5*). We now know that the age composition of Epiville changed with time. This may well have influenced the incidence of gastroenteritis in the town. If, for example, the incidence of the disease was especially high in young children, the rise in the percentage of young children must have increased the overall incidence. The next exercise will make this clear.

At this stage, you may like to reconsider your answer to *Question A6–4.*

Exercise A8

The incidence rates we have been using are based on the occurrence of gastroenteritis in the total population; such rates are termed *crude* rates. We can clarify matters by using the gastroenteritis rates in different age groups—that is, *age-specific* rates. A specific rate is one whose numerator and denominator refer to the same defined category: for example, children aged 0–4 (an age-specific rate), or males (a sex-specific rate), or boys aged 0–4 (an age- and sex-specific rate).

We can calculate age-specific rates if we know the age distribution both of the population (Table A7–1) and of the cases of gastroenteritis. If we know that in 1970, for example, there were 350 episodes in 1,400 children aged 0–4 years, the specific rate for this group in 1970 was $(350/1{,}400) \times 100 = 25$ per 100.

The age distribution of the cases is shown in Table A8–1, and the age-specific rates are listed in Table A8–2. Check the calculation of some of the rates, to be sure you know how they were obtained.

Question A8–1

Summarize the facts shown in Table A8–2.

Table A8–1. Numbers of Cases of Acute Gastroenteritis in Epiville in Selected Years (1970–2000) by Age

	Age (Years)				
Year	0–4	5–14	15–44	≥45	Total
1970	350	50	0	0	400
1975	540	60	0	0	600
1980	690	110	0	0	800
1985	780	120	0	0	900
1990	880	120	0	0	1,000
1995	970	130	0	0	1,100
2000	1,060	140	0	0	1,200

Table A8–2. Incidence Rates of Acute Gastroenteritis in Epiville in Selected Years (1970–2000) by Age (Episodes per 100 Population of Specified Age)

	Age (Years)				
Year	0–4	5–14	15–44	≥45	Total
1970	25.0	1.7	0	0	2.0
1975	20.0	1.2	0	0	2.0
1980	15.0	1.2	0	0	2.0
1985	13.0	1.1	0	0	2.0
1990	11.0	1.0	0	0	2.0
1995	9.7	1.0	0	0	2.0
2000	9.2	0.9	0	0	2.0

Question A8–2

Did the risk of incurring acute gastroenteritis in Epiville change between 1970 and 2000? (In answering this question, you may assume that the data on incidence and population size are accurate.) Refer to your reply to *Question A6–4*.

Question A8–3

How can we reconcile the changing incidence rate observed in the children with the unchanging rate seen in the population as a whole?

Note

A8. There is no reason why the columns in Tables A7–1 and A7–2 should show identical trends. Each column in Table A7–1 shows the trends in the *number* of individuals in a given age group, whereas each column in Table A7–2 shows the trends in the *percentage* of the age group. The percentage depends not only on the number in the given age group, but also on the numbers in other groups. The reason for the decrease in the percentage of older people, for example (Table A7–2), despite the increase in their absolute number (Table A7–1), was the marked increase in the number of younger residents.

■■■■■■■■■■ Unit A9

Inspecting a Two-Dimensional Table (Continued)

Inspecting the rows in Table A8–2, we find that the rates were consistently much higher in the 0–4 than in the 5–14 age group. The differences (in absolute or

relative terms) between these age groups were larger in 1970 and 1975 than in subsequent years. The rates in the 15–44 and ≥45 age groups were consistently zero. This, incidentally, might be due to absence of the disease, failure of adults with the disease to request medical care, or a tendency to use other diagnostic labels (enteritis, dysentery, food poisoning) for adult patients; but in fact, it was due merely to a wish to simplify the exercise.

When we examine the columns, we find that in both the 0–4 and 5–14 age groups there was a monotonic decrease between 1970 and 2000. In the older groups, the rate was consistently zero, and we already know that in the total population it was consistently 2.0 per 100. The relative decrease was greater in the 0–4 than in the 5–14 age group, the ratios of the 2000 to the 1970 rates being, respectively, 0.37 and 0.53 (if you think these are misprints, see Note A9–1). In both age groups, the decline was steeper between 1970 and 1985 than between 1985 and 2000. (You may have shown this graphically. If you wish, calculate the relative changes in these two periods; for answers, see Note A9–2.) In both of the 15-year periods, the decrease was steeper in the 0–4 than in the 5–14 age group.

The salient facts then, in answer to *Question A8–1,* are that the rate was consistently higher in younger than in older children; that there were no adult cases; and that between 1970 and 2000 the rates in children fell steeply, especially in children under 5 years, and especially in the first half of this period.

We may infer that for children—who were the only ones to get the disease—the risk of incurring acute gastroenteritis declined markedly between 1970 and 2000 (*Question A8–2*). Our previous inference—based on the constancy of the crude rates—that the risk of incurring the illness did not change (*Question A6–3*) turns out to be misleading.

The disparity has an obvious explanation. As we have seen, the incidence rate varied markedly with age. In a previous exercise (see Unit A7), we saw that the crude (overall) rate of a disease in a population is a weighted mean of the specific rates in the population's subgroups, the weights being the sizes of the subgroups. In other words, a subgroup's contribution to the rate in a total population depends on the relative size of the subgroup. The relative size of the child population of Epiville increased with time (Table A7–2), and the contribution of this high-incidence age group to the overall incidence therefore also increased with time. This increased weight was just enough to cancel out the effect of the decreasing risk of gastroenteritis in children, so that the crude rates remained constant. The average risk for residents of Epiville remained constant, but only because of the increased chance that the resident was a child. If the child population had grown even more, the crude gastroenteritis rates would have shown a rising trend—and this despite the decline in the risk of the disease! (By hindsight, we now see that the correct answer to *Question A6–4* was yes, and the above circumstances explain why.)

What we have seen is an example of *confounding* of an association. Before looking at this important phenomenon in more detail, let us consider what is meant by an "association."

Associations

An association (or "statistical dependence") between two variables is said to be present if the probability that one variable will occur or be present, or the quantity of the variable, depends on the occurrence, presence, or quantity of the other variable.

If 30% of bald men are ugly and 30% of hairy men are ugly, being bald does not alter the probability of being ugly, and there is thus no association between baldness and ugliness. If the prevalence of ugliness differs in bald and hairy men, there is an association between alopecia and ugliness. The detection of associations is usually based on comparisons of this sort. A difference means there is an association.

The association between two variables is called positive if they "go together"—that is, if one event or characteristic, or high values of one variable, are associated with the presence or occurrence of another event or characteristic or with high values of a second variable. The association is negative if they "go in opposite directions"—for example, if the presence of one characteristic is associated with the absence of another. If we know that 30% of men are bald and 40% of men are ugly, and if being bald does not alter the probability of being ugly (no association), we would expect 40% of bald men to be ugly; that is, 30% $\times$ 40%, or 12%, of men would be both bald and ugly. If we find that the proportion of bald ugly men in the population is above or below 12%, we can say that these two attributes are associated. If the proportion is above 12%, they are positively associated; and if it is less than 12%, they are negatively (or inversely) associated—that is, they occur together *less* frequently than we would expect.

An association does not necessarily imply a causal relationship. Associations may be artifacts caused by flaws in study methods, or they may arise by chance, or they may be attributable to confounding effects.

Conditional associations are associations that are observed in defined conditions (e.g., in specific population groups). For example, a positive association between baldness and self-appraised ugliness—that is, bald people regarding themselves as ugly—might be found in one ethnic group and not in another, or in one sex and not the other. A negative association between these variables—that is, bald people regarding themselves as attractive—might be found in another ethnic group or in the other sex. An association may be present in one group and not another, or may be stronger in one group than in another, or may be opposite in direction in different groups. When we examined the columns in Table A8–2, we looked at the conditional associations between gastroenteritis incidence and time in the 0- to 4- and 5- to 14-year age groups.

Exercise A9

State whether the following statements are true or false.

1. If you find that 60% of students who develop infectious mononucleosis (the "kissing" disease) are habitual smokers, this shows the presence of an association between the disease and smoking.

2. If you find that 5% of students who smoke develop infectious mononucleosis during a 1-year follow-up period, this shows the presence of an association between the disease and smoking.
3. If 60% of a large sample of male students and 30% of a large sample of female students smoke, there is an association between sex and smoking.
4. If, in a class of five male and five female students, none of the males smoke and all of the females smoke, there is an association between sex and smoking.
5. If 75% of the smokers in a college are males and 25% are females, there is an association between sex and smoking.
6. If over half the adults in a neighborhood have sedentary occupations and over half the residents have recurrent low back pain, there is an association between sedentary work and low back pain.
7. If adults with low body weights tend to have lower blood pressures than adults with high body weights, there is an inverse association between body weight and blood pressure.
8. If during an influenza epidemic there is a higher incidence rate of the disease among smokers than among nonsmokers, there is an association between smoking and influenza.
9. If during an influenza epidemic there is a lower incidence of the disease among smokers than among nonsmokers, there is no association between smoking and influenza.
10. If during an influenza epidemic there is a lower incidence rate among people who had influenza shots than among people who did not have shots, there is a positive association between influenza shots and incidence of the disease.
11. If you compare children of four ethnic groups and find that they differ in their mean hemoglobin levels, there is an association between ethnic group and hemoglobin level. The association is neither positive nor negative.
12. If the incidence rate of gastroenteritis is higher in infants than in older children, there is a positive association between gastroenteritis and age.
13. If a follow-up study shows relatively high mortality rates among people with very low and very high blood cholesterol levels, and a relatively low mortality rate among people with intermediate cholesterol levels, there is no association between blood cholesterol and mortality.
14. If a comparison of countries shows that the more personal computers there are per 100 population, the higher the mortality rate from coronary heart disease, this shows an association between the prevalence of PCs and coronary mortality.

Notes

A9–1. Some readers have been surprised to encounter ratios that are less than 1. A ratio is the number of times that one number contains another, and is calculated by dividing the one number by the other. The ratio of 25 to 9.2 is 2.72 (or 2.72 to 1, or 2.72:1), and the ratio of 9.2 to 25 is 0.37 (which is the reciprocal of 2.72; i.e., it is 1 divided by 2.72). If the numbers are equal, the ratio is 1.

A9–2. According to Table A8–2, the percentage decrease in the 0–4-year group was $(25 - 13)/25 \times 100 = 48\%$ in 1970–1985, and 29% in 1985–2000. In the 5–14 age group, the corresponding figures were 35% and 18%. Or (using ratios): in the 0–4 age group the ratio of the 1985 rate to the 1970 rate was $13/25 = 0.52$, and the ratio of the 2000 rate to the 1985 rate was $9.2/13 = 0.71$; in the 5–14 age group the corresponding ratios were 0.65 and 0.82.

■■■■■■■■■■ Unit A10

Associations (Continued)

Here are the answers to the true–false questions (Exercise A9):

1. False. We must have a comparison before we can conclude that there is an association. It is not enough to know the smoking habits of students who develop the disease; we must also know the smoking habits of students who do not develop the disease. If we find a difference between the proportions who smoke, we have an association. This is called a "retrospective" approach, because we move from the postulated outcome to the postulated cause.
2. False. Without a comparison we cannot conclude that there is an association. It is not enough to know the incidence rate of the disease in smokers; we must also know the incidence rate in nonsmokers. If the incidence rates are different, there is an association. This is called a "prospective" approach, because we move from the putative cause to the putative outcome.
3. True. There is a difference; therefore, there is an association.
4. True. There is a difference; therefore, there is an association. In such a small sample, there is a high likelihood that the association is fortuitous, but it certainly exists.
5. False. We have no comparison and hence can draw no conclusion about an association. It is possible that among nonsmokers also, 75% are males.
6. False. We have no comparison, for example, between the proportion of sedentary workers who had back pain and the proportion of nonsedentary workers who had back pain. You may have thought of an association at a population (not necessarily individual) level, because sedentary work and low back pain seem to "go together" in the same neighborhood. But here too we have no comparison. What were the proportions with back pain in neighborhoods with fewer or more sedentary workers? We have no data for other neighborhoods, and cannot draw a conclusion about the presence of an association: the rate of low back pain may be the same in populations with more active occupations.

7. False. The association is a positive one. Low body weights hang together with low blood pressures; that is, the variables tend to go in the same direction.
8. True. There is a positive association between smoking and influenza. "Positive" does not necessarily mean "beneficial."
9. False. If smoking is linked with a low incidence of influenza, there is a negative association between these two variables.
10. False. If influenza shots are associated with a low incidence rate—that is, the presence of one characteristic is linked with low values of another—the association is a negative one. "Negative" does not necessarily mean "harmful."
11. True. There is a difference; therefore, there is an association. As ethnic categories do not fall into a natural order (there are no "high" or "low" values), we cannot call the association positive or negative.
12. False. The association is a negative one. Low age goes together with a high incidence of gastroenteritis.
13. False. There is an association, but it is not a simple "linear" (straight-line) one. If plotted on a graph, the mortality rates would form a U-shaped curve, or maybe a J-shaped or reverse J-shaped one.
14. True. But the association is, of course, not necessarily a causal one. The association exists at a group or population level (this is sometimes called an "ecological" association), but it does not necessarily exist at an individual level; individuals who possess or use personal computers do not necessarily have an increased risk of dying of coronary heart disease.

Confounding

Let us return to Epiville and the distorted picture we obtained of the time trend in the incidence of gastroenteritis.

We were interested in the association between two variables: time (A) and the occurrence of the disease (B) (Fig. A10–1). Specifically, we were interested in the effect of time (the independent variable) on the occurrence of the disease (the dependent variable). When we looked at the crude rates (in *Question A6–2*), we found no association between these variables. But when (in *Question A8–1*) we introduced a third variable, age, we found clear evidence of an association; the age-specific rates showed a strong downward trend in both the age groups in which the disease occurred.

This distortion occurred because the crude data reflected the mingled effects of time and age on incidence. Age was strongly associated with both time and the incidence of the disease; that is, the age composition of the population varied with time, and the incidence of gastroenteritis varied with age. This is shown schematically in Figure A10–2, where *A* is time, *B* is the occurrence of the disease, and *C* is age.

The essential elements are that C must affect B (hence the arrow in the diagram), and that A and C must be associated with each other. The association be-

Figure A10–1. Causal association between two variables.

tween A and C need not be causal (hence no arrow), but C can affect A. If, however, the association between A and C is solely due to the effect of A on C, then C cannot be a confounder (the marked change in age composition referred to in *Question A7–2* was mainly an effect of selective immigration, not of the mere passage of time). When this constellation exists, it may be difficult to separate whatever effect A may have on B from the effect of C on B; the interplay of the associations may distort the picture of the A–B relationship. C is a *potential confounder* of the association between A and B (from the Latin *confundere*, "to mix together"). If distortion of the A–B relationship actually happens, as in our Epiville example, C is a *confounder* (confounding factor, confounding variable).

It should be noted that only if the associations between the confounder and the other variables are strong ones can there be a confounding effect of any importance (Note A10–1). If distortion is slight, it can usually be ignored.

If confounding occurs, we can obtain an undistorted picture only if we control the effects of the confounding variable (C), as we did by looking at each age group separately.

In Epiville, age was a factor that distorted the relationship between time and gastroenteritis incidence. In this instance the confounder *masked* the association. In other instances a confounder may *diminish*, *reverse*, or *exaggerate* an association. Commonly, it *produces an apparent association* when none really exists.

If the relationships pictured in Figure A10–2 are present or suspected to be present, the variable denoted by "C" may be considered a potential confounder. This is a simple operational method for the selection of possible confounders, satisfactory in most situations, and the only one used by many epidemiologists. When this model is used, should age be considered a possible confounder in a study of the effect of smoking on stomach cancer? (See Note A10–2.)

This simple model is generally adequate, although it does not give full expression to the complexity of the requirements for confounding (see Note A10–

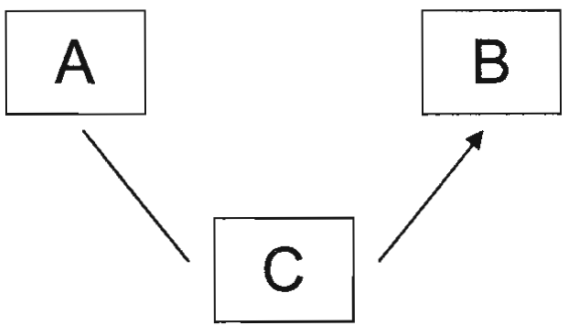

Figure A10–2. Conditions for confounding of A–B association by C.

3). A somewhat more complete formulation (which you may prefer to skip) is provided in Note A10–4.

Decisions about possible confounding effects cannot be made in an offhand way. They demand prior knowledge or assumptions about causal processes, may need examination of the data, and require the application of judgment.

When we try to explain an association between two variables, we should not seriously consider the possibility that it is a cause–effect relationship until we have asked three questions:

Is the association an artifact?
Can it be regarded as nonfortuitous?
Is it produced by confounding?

Exercise A10

The following questions refer to the sharp decline in the incidence rate of gastroenteritis in children aged 0–4 years in Epiville between 1970 and 2000 (Table A8–2).

Question A10–1

In trying to explain this decline, how would you decide whether sex should be considered as a possible confounder? You may assume that the time trend is not an artifact, and is not due to chance.

Question A10–2

How would you decide whether sex is *actually* (not *potentially*) a confounder?

Question A10–3

If sex should turn out to be a confounder, how could you control (i.e., neutralize or eliminate) its effect?

Question A10–4

What (if any) are the important confounders to be considered in trying to explain the decline in gastroenteritis incidence seen in children aged 0–4 years in Table A8–2?

Notes

A10–1. For numerical examples showing that the confounding effect is weak unless the associations with the confounder are strong, see Breslow and Day (1980, p. 96) and Bross (1966, 1967).

A10–2. The required association between C and B is present, since the risk of stomach cancer varies with age. The required association between C and A is

also present, since smoking habits vary with age, and this is not because smoking affects age. Age is therefore a potential confounder of the causal association between smoking and stomach cancer.

A10–3. For a fuller description of confounding, see Rothman and Greenland (1998, pp. 60–62, 120–125). Confounding has diverse definitions; our working definition is that the ratio of rates (or whatever measure of the association is used) is different when the confounder is ignored and when the effects of the confounder are held constant by stratification, standardization, or other methods.

A10–4. If we are interested in appraising the causal influence of variable *A* (independent) on variable *B* (dependent) and wish to identify possible confounders, the requirements are:

1. The potential confounder (C) must be causally related to B; it must be a variable that (according to prior knowledge or theory) influences B or is a stand-in for a variable that influences B. That is, it may itself be a cause of B or may cause a change in B, or it may be an indicator of a known or unknown correlated factor (or set of factors)—other than A—that affects B. Age, for example, may be considered a potential confounder because it is a surrogate for age-related causal factors. If B is a disease, C (or what it represents) may influence the likelihood of its diagnosis, not only the risk of its occurrence. If C is affected by B, or if it is a manifestation of B, it is not (in this context) a potential confounder. It is not essential to demonstrate an association between C and B in the data; but failure to find the expected association in the data may point to an inadequacy in prior knowledge.
2. C must be associated with A in the study population (or in the study sample, if it is representative). Prior knowledge may be a better guide than the data in some study designs, especially if numbers are small (Miettinen and Cook 1981), but it is seldom available. C is not a potential confounder (in this context) if the reason for the association is that (according to prior knowledge or theory) A influences C. We return to this important requirement in a later unit.
3. Although the selection of possible confounders is commonly based on the existence of simple associations between C and the other variables, it is actually the conditional associations (see Unit A9) with A and B that matter: The association with A must exist when B is held constant, and vice versa. If the independent variable (A) is exposure to a supposed etiologic factor, a criterion commonly used is that the association between C and B must occur even in the absence of exposure to the causal factor. If B is a disease, the association between C and A should be apparent in the source population from which the cases are derived, or in controls representative of that population.

■■■■■■■■■■Unit A11

Confounding (Continued)

In answer to *Question A10–1,* sex can have a confounding effect on the association between time and gastroenteritis incidence only if it is associated with both the latter variables. Gastroenteritis incidence may well have differed in the two sexes, as does the incidence of many other diseases; but there is no reason to believe that the sex composition of the child population changed appreciably during this period. We are therefore probably safe in concluding that sex need *not* be regarded as a possible confounder.

To determine whether a confounding effect actually exists (*Question A10–2*), we must compare what we see in the crude data with what we see when we neutralize or eliminate the effect of the suspected confounder. Is there an important difference in the findings? One way of doing this is to look separately at the data for each category (or "stratum") of the suspected confounder. It was by using this *stratification* procedure that we detected the confounding effect of age: we compared the time trend shown by the crude incidence rates with the time trends shown by age-specific incidence rates. We can now repeat this procedure, for sex. We can "control for sex" by calculating sex-specific rates (for children aged 0–4 years), and seeing whether the time trend shown by the crude data for these children is a satisfactory reflection of the time trends seen in the two sexes.

By using stratified data, such as age- or sex-specific rates, we eliminate the effects of the stratifying variable (age or sex) on the associations that interest us (*Question A10–3*). We could also control these effects in other ways—for example, by standardization (which we will deal with later). Whatever method is used, two birds can be killed with one stone: the same procedure can both demonstrate the existence of a confounding effect and neutralize it.

The variables that are candidates for inclusion in a list of possible confounders (*Question A10–4*) are those that are known or suspected to affect the dependent variable. Any of these that is known or believed to be associated with the independent variable as well, but is not affected by it, may be listed as a possible confounder. It must be remembered that there can be an important confounding effect only if the associations are strong. There are a number of variables that are so often of relevance in epidemiological studies that consideration should always be given to their inclusion. These "universal variables" include age, sex, parity, ethnic group, religion, marital status, social class, and its components (occupation, education, income), rural or urban residence, and geographical mobility.

In Epiville, where we know that the population has grown extensively because of immigration, and where we have found that a change in its age composition has distorted the time trends in gastroenteritis incidence, we should give serious consideration to the possibility that selective immigration has resulted in changes in other demographic characteristics as well, resulting in other con-

founding effects. For example, the composition of the population may also have changed with respect to ethnic group or social class. If we know or suspect that such changes occurred, and if we believe that these variables may influence gastroenteritis incidence, we should investigate the possibility that they are confounders.

Effect Modification

In a previous exercise we extended our understanding of the association between two variables (gastroenteritis incidence and time) by investigating the influence of another variable (age) on the association. This, a very common analytic procedure, may be termed *elaboration* of the association. Stratification according to the categories of the other variable is the simplest way of doing this.

When we compared the associations seen in Table A8–2, which showed incidence rates by year and age, we observed two kinds of discrepancy. First, there were differences between the associations shown by the specific and crude rates; this was our evidence for the confounding effect of age. Second, there were discrepancies between the findings in the various specific strata–a striking decline in children aged 0–4, a less marked decline in older children, and no change in adults. The conditional associations (see Unit A9) between gastroenteritis incidence and time were different. This phenomenon may be termed the *modifying effect* of age on the association between gastroenteritis incidence and time. Age turned out to be both a confounder (because the time trends shown by the crude and age-specific rates differed) and a modifier (because the time trends in the various age strata differed from one another). The same stratification procedure demonstrated both effects.

Exercise A11

Question A11–1

For your convenience, the decline in gastroenteritis incidence among children aged 0–4 in Epiville is again shown in Table A11. Do you think it might be advantageous to use narrower age categories, and if so, why?

Question A11–2

Suppose you suspect that social class has a confounding effect on the association seen in Table A11, as a result of selective immigration with regard to social class. You propose to examine this possibility by using stratification. Construct a skeleton table (a table with captions, but without figures) to accommodate the new data you require for this purpose; provide space both for the raw figures and for whatever summary statistics are needed. For simplicity, use two social classes ("high" and "low") in this exercise.

Table A11. Incidence Rate of Gastroenteritis Among Children Aged 0–4 Years in Epiville in Selected Years, 1970–2000

Year	Incidence Rate per 100
1970	25.0
1975	20.0
1980	15.0
1985	13.0
1990	11.0
1995	9.7
2000	9.2

Question A11–3

What else—with specific reference to associations between variables—might you learn from the new figures you hope to put in the skeleton table?

■■■■■■■■■■ Unit A12

Refinement

In answer to *Question A11–1,* it might be useful to use narrower age categories, in order to discover whether the incidence of gastroenteritis varies *within* the categories we have so far used. In the 0–4 age group in particular, are the rates higher in the first 6 months, in the second 6 months, in the second year, or in the third, fourth, or fifth year of life? This knowledge might help to pinpoint groups that are at high risk and need special preventive care, and might also provide useful clues to the causation of gastroenteritis in this community.

The use of finer instead of broad categories is an example of a procedure termed *refinement,* which is often used in order to throw added light on an association. This procedure also sometimes reveals associations that were not previously apparent. We may refine a crude scale of measurement, as in the instance of age, or we may refine the variable itself. For example, instead of regarding acute gastroenteritis as a single entity, we might calculate the rates of acute gastroenteritis of different severity or duration or those associated with various specific microorganisms.

Skeleton Tables

The drawing of a skeleton table to accommodate new information is often a challenge that serves to clarify one's thinking and translate a fuzzy idea of "what I would like to know" into a clear-cut need for well-defined facts.

A skeleton table may be meant for raw data, for summary measures (such as rates, percentages, and means), or for both. Designing the table may necessitate decisions about the selection of variables, of categories, and of summary measures, and about the arrangement of variables (e.g., in cross-tabulations) so as to provide information on the associations of interest. Sometimes the table serves to draw attention to practical difficulties that have been overlooked; only when the requirements for data are clearly stipulated may it be realized that they cannot be met.

A skeleton table need not be prepared with obsessive attention to detail, but it should meet the basic requirements for a well-constructed table. It should include column and row captions. If categorical scales are used, they should be comprehensive and their categories should be mutually exclusive. Allowance should be made for "unknowns"; if there are many cases with missing data, it may be difficult to draw useful conclusions from the findings. If the figures are to be provided by a computer with the use of a ready-made package of programs, the arrangement of the table should conform to one of the formats offered by these programs.

The skeleton table requested in *Question A11–2* should look something like Table A12–1. It should show year-by-year incidence rates for each social class, together with the raw data (population figures and numbers of cases) required for calculating these rates.

Elaborating an Association

In answer to *Question A11–3,* the figures inserted in the skeleton table would not only help us to detect and control for possible confounding by social class, it would also tell us about:

1. The association between social class and time. Did the social class distribution of the population change?
2. The association between social class and gastroenteritis incidence. Did the rates in the social classes differ?
3. The modifying effect of social class on the association between gastroenteritis incidence and time. Did the time trends differ in the social classes?
4. As a corollary to (3), we would also learn about the modifying effect of time on the association between gastroenteritis incidence and social class. (Did the social class differences in incidence vary at different times?) These two modifying effects—(3) and (4)—are different expressions of the same phenomenon; one cannot exist without the other.

Table A12–1. Incidence of Gastroenteritis in Children Aged 0–4 in Epiville in Selected Years (1970–2000) by Social Class

	Social Class											
	High			Low			Unknown			Total		
Year	Pop.	No. of Cases	Rate per 100	Pop.	No. of Cases	Rate per 100	Pop.	No. of Cases	Rate per 100	Pop.	No. of Cases	Rate per 100
1970												
1975												
1980												
1985												
1990												
1995												
2000												

Table A12–2. Incidence of Gastroenteritis in Children Aged 0–4 Years in Epiville in Selected Years (1970–2000) by Social Class: Rates per 100

	Social Class		
Year	High	Low	Total
1970	14.6	31.9	25.0
1975	13.0	24.7	20.0
1980	11.1	17.6	15.0
1985	10.1	14.9	13.0
1990	9.1	12.3	11.0
1995	8.4	10.6	9.7
2000	8.2	10.5	9.2

As we will see later, elaboration of an association can also help us to test the possibility that the added variable is an *intermediate cause*—that is, a link in the chain of causation between the independent and dependent variables.

Exercise A12

Let us assume that there were no children whose social class was unknown. The incidence rates in children aged 0–4 are shown in Table A12–2, separately for each social class and for the age group as a whole.

Question A12–1

Summarize the facts shown in Table A12–2. In your summary, state what associations are shown.

Question A12–2

Does social class have a modifying effect on the association between gastroenteritis incidence and time?

Question A12–3

Does social class have a confounding effect on this association?

Question A12–4

What would be the importance of finding a modifying effect?

Question A12–5

What would be the importance of finding a confounding effect?

■■■■■■■■■■Unit A13

Modifying and Confounding Effects

To answer *Question A12–1,* we should inspect the table's columns and rows. Each column shows a monotonic decrease in gastroenteritis incidence with time. The ratio of the 2000 to the 1970 rate was 0.37 in the 0–4 age group as a whole, 0.56 in the high social class, and 0.33 in the low social class. The absolute differences between the rates in 2000 and 1970 were 15.8, 6.4, and 21.4 per 100 in the total group and in the children of high and low social class, respectively. The decline with time was thus much steeper in the low social class.

In each row we see a negative association between social class and gastroenteritis incidence—the rate is consistently higher in the low than in the high social class. The difference was biggest in 1970, when the absolute difference was 17.3 per 100 and the ratio (low:high) was 2.2. The difference became progressively less, but was still apparent in 2000, when the absolute difference was 2.3 per 100 and the ratio was 1.3. In each year, the rate in the total group was intermediate between those in the two social classes.

In answer to *Question A12–2,* social class is clearly a modifier of the association between gastroenteritis incidence and time; the time trends in the social classes differ.

To determine whether social class has a confounding effect on the association between gastroenteritis incidence and time (*Question A12–3*), we may compare the trends seen in the total group with those in the specific strata (social classes). The comparison is difficult because of the difference between the trends in the social classes, and the answer is not clear-cut. There is no basic difference (comparing 1970 and 2000) between the trend in the total group, as expressed by either the rate ratio or the rate difference, and the trends observed in the separate social classes; the direction of change is the same in each instance, and the values of the separate social classes straddle those in the total group. The rate ratio in the total group, however (unlike the rate difference), is closer to the rate ratio in the low social class than to that in the high social class, possibly as a result of confounding. We might conclude that the picture provided by the crude data (without controlling for social class) is not distorted enough to matter, and that there is no confounding effect of any importance: controlling the effect of social class (by stratifying) does not alter the conclusion that over the years there has been an appreciable decrease in the incidence rate of gastroenteritis.

Effect modification is pictured in two different ways in Figures A13–1 and A13–2, where *C* modifies the association between *A* (an independent variable) and *B* (the dependent variable). This means that the effect of *A* (in our example, time) on *B* (incidence) varies, depending on *C* (social class). It also always means (as a corollary) that the effect of *C* (social class) on *B* (incidence) varies, depending on *A* (time). It is the specific combination of *A* and *C* that determines the value of *B*. This may also be referred to as *interaction* between the two independent variables, *A* and *C*, in their association with *B*. Figure A13–1 is ap-

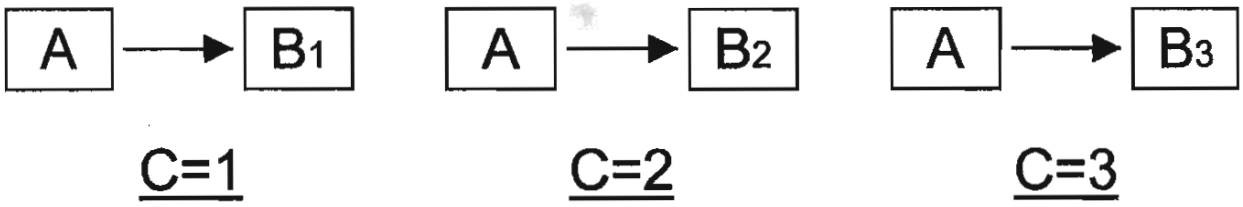

Figure A13–1. Effect modification; the modifier (C) has three categories in which the effect of the independent variable (A) on the dependent variable (B) differs.

propriate if *C* has two or more categories (e.g., different social classes); it highlights the fact that there is a different *A–B* association for each category of *C*. Figure A13–2 emphasizes the interaction between *A* and *C*, and is appropriate even if the modifier does not have separate categories (e.g., if *C* is weight in kilograms or height in inches).

When we detect a modifying effect (*Question A12–4*), we gain new information that may have important theoretical and practical implications. In Epiville, the fact that gastroenteritis declined more steeply in low-social-class children may help us in our search for the reasons for the decline. It is a clue that may help us to formulate appropriate hypotheses for testing. We may also use a different viewpoint: not only did the time trend in gastroenteritis incidence differ in the two social classes, but the difference between the rates in the social classes altered with time. This fact, too, may give us food for thought, and we may wish to explore it further. Third (at a simpler level), until we detected the modifying effect, we may not have known that social class was associated with incidence. As the diagram shows, a modifier is always associated with the dependent variable; in fact, it can usually be regarded as a cause or determinant. We may wish to go on to formulate and test possible explanations for the association between social class and gastroenteritis incidence.

The discovery of effect modification may also have practical implications. If *A* and *C* were sex and social class, for example, we would be able to identify children (say, boys in a low social class) who are especially likely to benefit from preventive intervention.

The importance of finding a confounding effect (*Question A12–5*) depends on whether it was previously known that the confounder influences the dependent variable. If this effect was already known (as is usually the case), discovery of the confounding effect leads only to a realization that the conclusions drawn

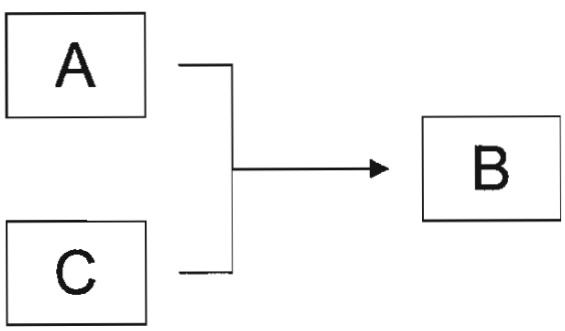

Figure A13–2. Effect modification (interaction).

from the crude data are misleading and require revision, by controlling for this "nuisance variable." Sometimes, however, a search for a confounder leads to new etiological insights—the fact that *C* affects *B* (or maybe both *A* and *B*) may be a new finding, and *C* may turn out to be a key factor in the causal processes.

A variable may be neither a modifier nor a confounder, or both, or a confounder and not a modifier, or a modifier with no material confounding effect. If the modifying effect is extremely strong, it is arguable that the confounding effect becomes irrelevant. Suppose, for example, that the incidence of gastroenteritis had risen sharply in one social class and had fallen steeply in the other. With such wide divergence, it would be so important to pay separate attention to the social classes that there might be little interest in the overall change in the town, confounded or not.

Exercise A13

In Table A12–2 we saw a strong association between gastroenteritis incidence and social class in children aged 0–4 in 1970. The rates in the two classes were 31.9 and 14.6 per 100. We now stratify the data in accordance with the mother's duration of residence in Epiville, and obtain the results shown in Table A13–1.

Question A13–1

Summarize the facts concerning the association between gastroenteritis incidence and social class. How would you explain the discrepancy between the associations shown by the crude and specific rates? Does mother's duration of residence in Epiville modify the association between gastroenteritis and social class?

Question A13–2

Let us suppose that when we stratify the data in accordance with the children's nutritional status (measured before the onset of gastroenteritis), we obtain the

Table A13–1. Incidence of Gastroenteritis Among Children Aged 0–4 Years in Epiville in 1970, by Social Class and Mother's Duration of Residence in Epiville

Mother's Duration of Residence in Epiville	Social Class					
	High			Low		
	Pop.	No. of Cases	Rate per 100	Pop.	No. of Cases	Rate per 100
Over 5 years	280	14	5.0	179	9	5.0
2–4 years	240	48	20.0	239	48	20.1
Under 2 years	40	20	50.0	422	211	50.0
Total	560	82	14.6	840	268	31.9

Table A13–2. Incidence of Gastroenteritis Among Children Aged 0–4 Years in Epiville in 1970, by Social Class and Nutritional Status

	Social Class					
	High			Low		
Nutritional Status	Pop.	No. of Cases	Rate per 100	Pop.	No. of Cases	Rate per 100
Well nourished	280	14	5.0	179	9	5.0
Slightly malnourished	240	48	20.0	239	48	20.1
Markedly malnourished	40	20	50.0	422	211	50.0
Total	560	82	14.6	840	268	31.9

results shown in Table A13–2. Summarize the facts concerning the association between gastroenteritis incidence and social class. How would you explain the discrepancy between the associations shown by the crude and specific rates?

■■■■■■■■■■ Unit A14

Elaborating an Association (Continued)

The association between gastroenteritis incidence and social class is elaborated in Table A13–1, where the data are stratified according to mother's duration of residence in Epiville. In answer to *Question A13–1*, the crude rates (in the bottom row of the table) show a strong association between gastroenteritis and social class. The ratio of the incidence rate in the low social class to that in the high social class is 31.9:14.6—or 2.2. But when mother's duration of residence is held constant, the association disappears; in each "duration of residence" category, the specific incidence rates in the two social classes are almost identical (the ratio of the rates is 1.0).

We can attribute this discrepancy between the associations shown by the crude and specific rates to the confounding effect of mother's duration of residence. The relationship with social class can be explained by the relationship with mother's duration of residence. As Table A13–1 shows, recency of immigration is strongly associated both with social class and with gastroenteritis incidence. (What is the evidence for these associations? For answer, see Note A14.) We may conclude that social class can be disregarded as a determinant of the occurrence of the disease.

Mother's duration of residence in Epiville does not modify the relationship

Figure A14–1. Intermediate cause.

between gastroenteritis incidence and social class. The ratio of rates is the same (1.0) in each "duration of residence" category.

The figures in Table A13–2 are identical with those in Table A13–1. Here too, the stratifying variable (nutritional status) is strongly associated both with gastroenteritis and with social class; and here too, the crude rates show an association with social class, whereas the specific rates do not. Yet the interpretation of the facts is different. We cannot conclude that social class has no role in the causation of gastroenteritis, since nutritional status may well be a link in the chain of causation between social class and gastroenteritis. We cannot regard nutritional status as just a confounder whose effect misled us to think that social class might play a causal role. Rather, we might infer that nutritional status is the intervening cause that accounts for the difference in incidence between the social classes: we could regard the association between social class and gastroenteritis as a meaningful one that might be explained by the effects on nutritional status of behavioral, economic, environmental, or other characteristics connected with social class.

This example carries an important message. The prerequisites for a confounding effect, as stated in Unit A10, were shown schematically in Figure A10–2. Both *A* and *C* must have an effect on *B*, and *A* and *C* must be associated with each other. The association between *A* and *C* may be noncausal. But if it is causal, with *A* affecting *C*, *C* is an intermediate link in the chain of causation between *A* and *B* (Fig. A14–1). It is then not a potential confounder, but an intermediate or intervening cause. Just as with a confounder, the associations seen in the crude data may differ from those seen when stratification or some other procedure is used to "hold *C* constant." However, although the statistical findings may be the same, their interpretation is different, as we have just seen in the Epiville example. If the association between *A* and *C* is a causal one with *C* affecting *A* (Fig. A14–2), *C* is a potential confounder and not an intermediate cause.

The above considerations apply not only if *C* is an intermediate cause, but also if it is a stand-in for an intermediate cause–for example, if it is a manifestation or result of some factor (known or unknown) that is affected by *A* and affects *B*. Variable *C* should then not be treated as a confounder. In the above instance, stratifying by skin dryness (an expression of nutritional status, but not a cause of

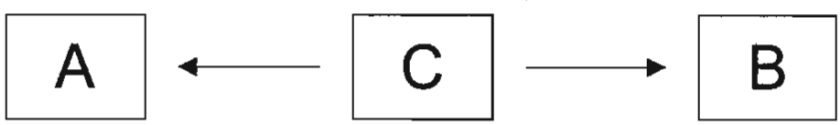

Figure A14–2. Confounding by a common cause.

gastroenteritis) might (like stratifying by nutritional status) have misled us to believe that social class has no role in the causation of gastroenteritis. It has been suggested that no variable that is determined, even in part, by *A* should be treated as a confounder (Weinberg 1993).

A dilemma arises if the causal processes are unclear, and it is not certain whether *C* is determined by *A* or not. There is also a dilemma if *C* is partly caused by *A*, and partly a marker for some quite different variable. In such circumstances if may be advisable to do parallel analyses, one treating *C* as a potential confounder, and one not, and reach alternative conclusions about the *A*–*B* association (if so-and-so then so-and-so, and if such-and-such then such-and-such). The two approaches may lead to similar conclusions.

Emphasis has been given in this unit and in Unit A10 to the conditions that must be met before a variable can be regarded as a potential confounder of a causal association and therefore held constant in the analysis. This does not mean, however, that a variable should not be held constant unless these conditions are met. There may be other reasons for doing so. We might, for example, have the notion that the social class differences in gastroenteritis are only partly accounted for by differences in nutritional status, and test this hypothesis by holding nutritional status constant, as in Table A13–2; if we then found an association between social class and gastroenteritis (which we did not find in Table 13–2), this would support our notion.

Exercise A14

Question A14–1

The effects of a confounding variable may be controlled by stratification and by other techniques that we have not yet discussed. What technique, other than stratification, have we used for this purpose in these exercises? This may be regarded as a trick question; the technique is a widely used one that is often applied in a routine manner, without specific thought as to its function in controlling for confounding.

Question A14–2

The incidence rate of gastroenteritis is twice as high in Epiville as in Shlepiville. Can the data shown in Table A14 explain this difference?

Table A14. Population Size and Incidence of Gastroenteritis in Two Towns, 1999

	Epiville	Shlepiville
Total population	60,000	30,000
Cases of gastroenteritis per 1,000 population	20	10

Question A14–3

This question deals with the formulation and testing of causal explanations. To avoid confusion, let us move to fresh pastures—the town of Zepiville, where there is a strong association between ethnic group (Easterners or Westerners) and the incidence of gastroenteritis in children aged 0–4. The incidence rate is much higher among Easterners than among Westerners in this town.

As far as we can tell, this association is not an artifact, and a test of statistical significance shows that we can safely regard it as nonfortuitous. We have looked for evidence of confounding and have found none. Of course, we cannot be sure (one never can) that there is no confounding by some variable that we have not measured, tested, or maybe even thought of; however, we have decided that for practical purposes we will reject the possibility that the association is caused by confounding. In the course of the analysis, we found no evidence that the association was modified by sex, social class, mother's age, or mother's duration of residence; the association of incidence with ethnic group was apparent in each category of these variables.

List all the possible causal explanations you can think of for the difference in incidence between Easterners and Westerners (forget Occam's razor).

Note

A14. In Table A13–1, a strong association between recency of immigration and social class is shown by the striking difference between the two frequency distributions of mother's duration of residence (280, 240, and 40 in the high social class and 179, 239, and 422 in the low social class). The differences between the gastroenteritis incidence rates in the "duration of residence" groups (5, 20, and 50 per 100) show an association between recency of immigration and the disease.

■■■■■■■■■■ Unit A15

The Use of Rates

At the outset of this series of exercises, we saw (in Table A1) that the annual number of cases of gastroenteritis in Epiville rose markedly between 1970 and 2000. We subsequently found that this rise could be attributed to the increase in population. The association between the number of cases and time was in fact due to the confounding effect of population size, a variable that was strongly associated with both the dependent variable (number of cases) and the independent variable (time). When we calculated incidence rates, we found no time trend: the rate was the same each year (20 per 1,000). The time trend disappeared because we used the rate—the number of cases per 1,000 population—as our de-

pendent variable, rather than just the number of cases. By using rates, we were able to hold the effect of population size constant in the comparison.

This, of course, is one reason why rates are used. In answer to *Question A14–1,* when we compare the occurrence of a disease in two populations we are aware that a difference in the numbers of cases may be due mainly to a difference in population size. We therefore use rates rather than numbers of cases. This controls for the confounding effect of population size. Percentages and other ratios are also used for this purpose. When we wished to see whether the age composition of the population of Epiville changed between 1970 and 2000, we used percentages (Table A7–2) so as to neutralize the effect of differences in population size.

The use of rates and proportions is probably the most widely used method of controlling for confounding. The basic principle is replacement of the dependent variable by another variable, which is defined in such a way that it incorporates, and neutralizes the effect of, the confounder—for example, "cases per 1,000 population" instead of "cases." This technique may be used to deal with confounders other than population size. When one compares body weights, for example, the confounding effect of height can be controlled by using a weight–height index, such as the ratio of weight to height or to the square of height; or a relative weight can be used, calculated by expressing the observed weight as a percentage of the "standard" weight of people of the same age, sex, height, and so forth, in order to neutralize the effects of these variables; or weight can be replaced by a weight percentile that expresses the child's position in relation to the weights of other children of the same age and sex. Another common example is the use of an intelligence quotient or developmental quotient that expresses a test score as a percentage of the average score of children of the same age.

In answer to *Question A14–2,* the data shown in Table A14 *cannot* explain the difference in gastroenteritis rates. The difference in population size cannot explain the difference in gastroenteritis rates, since its effect is neutralized by the use of rates. There were (20/1,000) × 60,000 = 1,200 cases in Epiville, and (10/1,000) × 30,000 = 300 in Shlepiville. There was a fourfold ratio of cases, which is reduced to a twofold ratio (20:10) when we control for population size by using rates.

Causal Explanations

Causes are always multiple; nothing has a single cause. Swallowing pathogenic microbes may cause gastroenteritis, but the disease is also caused by the person's susceptibility to these microbes, and by antecedent factors such as his or her attendance at the party where the microbes were ingested, and the dirty fingers that put them in the food. A metaphor commonly used by epidemiologists is the "web of causation" (MacMahon *et al.,* 1960), which in a diagram would consist of many events or attributes connected to one another by one-way or two-way arrows showing direction of influence. When we list possible causal explanations we are not generally trying to suggest a set of alternatives, one of which will turn out to be the sole cause. We are enumerating various factors that may each con-

tribute in some degree to the phenomenon we are studying, exerting its effect in a direct or indirect manner, separately or in combination with other factors.

Any factor whose modification may be expected to change the frequency or quality of another can be regarded as causal (Note A15–1). Most of the possible causes in our list will be neither necessary nor sufficient. A *necessary cause* is one without which the outcome cannot occur; infection by the tubercle bacillus, for example, is a necessary cause of tuberculosis; but most causes are not necessary. Single causes that are sufficient (e.g., beheading as a cause of death) are hard to find. A *sufficient cause* is generally a constellation of single causes (Note A15–2)—that is, a set of events and attributes—that inevitably produces the effect, such as a combination of exposure to an infective agent and a lack of immunity. Most of the possible causes in our list would be described (if we wished to use these terms) as "predisposing," "enabling," "precipitating," "reinforcing," "concomitant," or "intermediate." By this line of reasoning, the importance of any single cause—that is, the strength of its association with the effect—will be influenced by the prevalence of the other components of the various constellations in which it features.

In thinking of causal explanations for an association, it might be helpful to use an epidemiologic model, two of which are pictured here. The well-known host–agent–environment triangle is shown in Figure A15–1, and Figure A15–2 shows a model suggested by Kark (1974), which features the interrelationships among (a) the state of health of a population or group (in terms of diseases, disabilities, and deaths, and somatic and psychological characteristics); (b) the biological, social, and cultural attributes of the population or group; (c) the environment (natural, human, and manmade) and material resources of the population or group; and (d) the health care system.

The "Chinese-box" model is a challenging recent suggestion (Susser, 1996). It envisages a conjurer's nest of boxes, each containing a set of smaller ones. Each box represents a different level of organization, the levels ranging from the physical environment, through societies and broad populations, local communities, families, and individuals, to body systems, tissues and cells, and finally molecules. In each box there is a complex of causal associations, and there are intricate causal links between the boxes. This model encourages the study of determi-

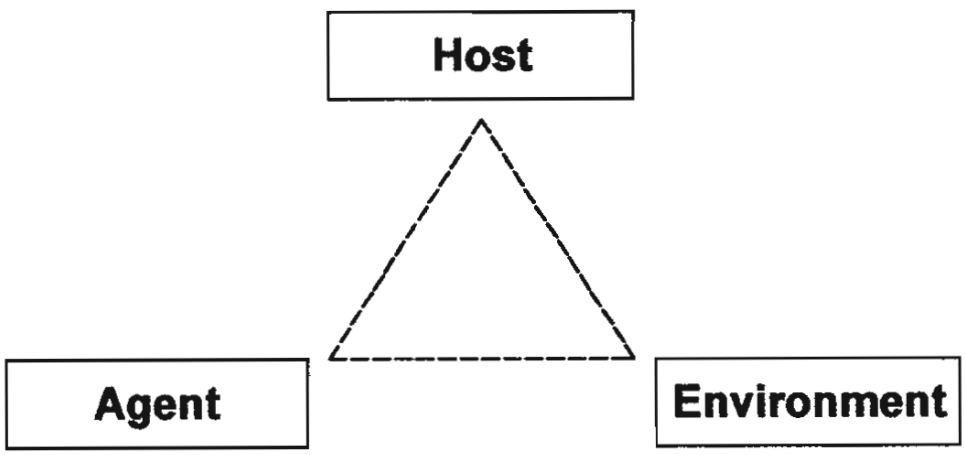

Figure A15–1. The epidemiological triangle.

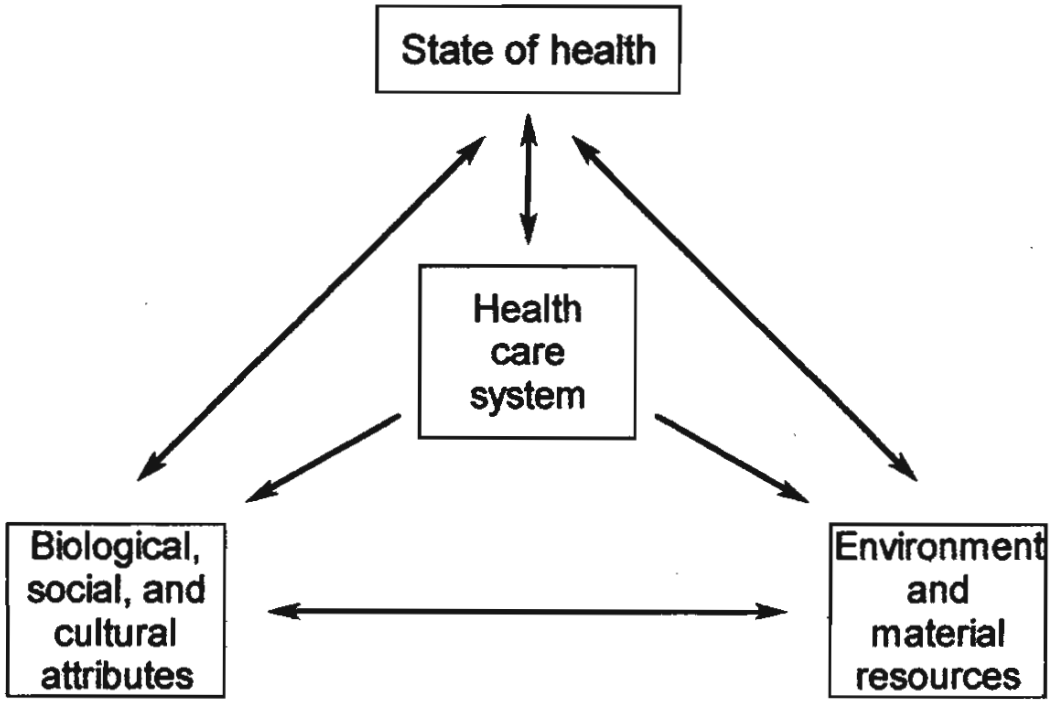

Figure A15–2. An epidemiological model of causal relationships.

nants and outcomes at different levels of organization, and it can accommodate the biological and social causal processes.

Associations with "universal variables" (see page 41), such as sex or ethnic group, usually have a variety of possible explanations. Members of different ethnic groups may differ not only in their culture (and hence in their habitual dietary, smoking, and other practices), but in their genetically determined characteristics, in their environmental exposures, in the availability of medical services, and in other respects.

There is, of course, no "correct answer" to *Question A14–3.* Your list of possible explanations for the ethnic difference in gastroenteritis may (inter alia) include differences in infant feeding practices, differences in nutritional status, differences in the hygiene of foodstuffs or food utensils, differences in handwashing practices, and genetic differences. You may have thought of more elaborate explanations, such as the possibility that differences in family size may lead to differences between ethnic groups in the amount of contact with other children, resulting in differences in the incidence of respiratory infections, and, as a consequence, differences in susceptibility to gastroenteritis. You may have also included factors (such as the way that mild diarrhea is treated at home) that may affect the severity rather than the occurrence of illness, leading to differences in the proportion of cases who have subclinical infections that do not meet the criteria required for definition as a "case."

Testing Causal Explanations

The basic way of testing a causal explanation is to seek new facts and see whether these fit in with what we might expect to find if the explanation was correct. If they do not, the explanation can be discarded; if they do, they provide support-

ive evidence for the explanation. This procedure may not really "prove" causality; but if enough new facts that could refute the explanation are sought and they persistently uphold a causal interpretation, they can constitute proof that is strong enough to provide a basis for decision and action.

Testing is best done by first formulating refutable predictions—statements of what findings may be expected if the causal explanation is correct. These statements are specific *"research hypotheses,"* which can then be tested by seeking the appropriate empirical facts. They are generally positive declarations, and not the "null hypotheses" required for tests of statistical significance (see Note A15–3).

To be useful, the hypothesis must be testable. It must be formulated in very specific terms, leaving no doubt as to what information is needed to test it; and obtaining this information must be feasible.

Exercise A15

Question A15–1

In the last exercise you suggested a number of possible explanations for the difference between Easterners and Westerners in the incidence of gastroenteritis in children in Zepiville. Now choose one of these explanations for testing (remember Occam's razor).

Question A15–2

Formulate an appropriate specific hypothesis (or hypotheses) that will test the explanation you have chosen.

Question A15–3

Construct a skeleton table (or tables) to accommodate the information you require for this purpose.

Notes

A15–1. "A causal association may be defined as an association between two categories of events in which a change in the frequency or quality of one is observed to follow alteration in the other. In certain instances the possibility of alteration may be presumed and a presumptive classification of an association as causal may be justified" (MacMahon et al., 1960).

"We can define a cause of a specific disease event as an antecedent event, condition or characteristics that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are found. . . . With this definition, it may be that no specific event, condition or characteristic is sufficient by itself to produce disease" (Rothman and Greenland, 1998, p. 8).

"In medicine and public health, it would appear reasonable to adopt a pragmatic concept of causality. A causal relationship would be recognized to exist whenever evidence indicates that the factors form part of a complex of circum-

stances that increases the probability of occurrence of a disease and that a diminution of one or more of these factors decreases the frequency of that disease" (Lilienfeld and Lilienfeld, 1980, p. 295).

A15–2. A *constellation of causes* (Rothman 1976, 1986, pp. 10–16; Rothman and Greenland 1998, pp. 7–16) is a set of minimal conditions and events that inevitably produce a given disease (or other effect) when an individual is exposed to them, "minimal" meaning that there are no superfluous factors in the set. Many alternative constellations of causes (known or unknown) may be involved in the etiological process in different individuals, and no single constellation is therefore a necessary cause. But in each constellation, every component is a necessary element, without which the combination of causes will not have their effect. When tackling causes in practice (see Unit G), prior consideration should be given to those that are always or frequently necessary—that is, those that feature in all constellations of causes, or in many of the frequently operating constellations that lead to the effect.

A15–3. Statistical testing requires a *null hypothesis*, which is a negative declaration such as: "There is no correlation between birth weight and the incidence of gastroenteritis," or "There is no positive correlation between birth weight and the incidence of gastroenteritis." The test indicates whether we can confidently reject this null hypothesis. What we have called the *research hypothesis* (e.g., "There *is* a correlation" or "There *is* a positive correlation") is generally what statisticians call "*the alternative to the null hypothesis.*" The precise formulation of the null hypothesis and its alternative depends on the kind of data available and the kind of statistical test used.

■■■■■■■■■■ Unit A16

Testing Causal Explanations (Continued)

In accordance with Occam's razor, the explanation chosen for examination should preferably be a likely one that, if true, would go a long way toward explaining the phenomenon we are studying (the ethnic difference in gastroenteritis incidence). It should also be a testable one. There is little point in selecting an explanation for testing—however cogent the reasons—if the information required for this purpose cannot be obtained. The explanation you chose in your answer to *Question A15–1* should meet these requirements.

Appraise your formulation of specific hypotheses (*Question A15–2*) by seeing whether the following criteria are satisfied:

- The hypothesis should be one that can meet its purpose; can observed facts refute the causal explanation?

- The hypothesis should be stated in clear, operational terms, so that there is no doubt as to what information is needed for testing it.
- Collection of the required information should be practicable.

As an illustration, suppose that the explanation selected for testing is that a difference in infant feeding practices caused the ethnic difference in gastroenteritis incidence. In phrasing a specific hypothesis for testing, we would start by eliminating the word "caused." Except perhaps in strictly experimental situations, it is not possible to test hypotheses containing such words as "produces," "causes," "results in," "influences," "reduces," "increases," or "affects." These are useful terms when we draw inferences or consider possible explanations for findings, but when we formulate specific hypotheses for testing, we should rather speak of associations (positive or negative), differences, and changes—for which empirical evidence may be available.

We might accordingly decide to test the hypotheses (a) that ethnic group is associated with infant feeding practices in this population, or (b) that infant feeding practices are associated with the occurrence of gastroenteritis. Alternatively, our hypothesis might be that if differences in infant feeding practices are controlled in the analysis, the difference between Easterners and Westerners in the incidence of gastroenteritis will be lessened. If any of these statements turns out to be untrue, we can reject our causal explanation.

These hypotheses are useful formulations but are not really specific enough to be operational: they do not tell us precisely what information we require. For example, what exactly is meant by "infant feeding practices"? Also, do we want information about all children, or about samples of children of different ethnic groups, or with different feeding histories or different experience of acute gastroenteritis? How do age and other variables enter into the hypotheses? and so forth. We might, for example, make the hypothesis more specific by postulating that differences in the mean duration of lactation and the mean age of introduction of fruit juices, cereals, eggs, and other specified food items will be found when Eastern children are compared with Western children; or our hypothesis might be that such differences will be found when children with two or more episodes of gastroenteritis in their third year of life are compared with age-matched controls with no episodes of the illness in their third year. We might sharpen these hypotheses by stating the direction of the expected differences.

If the hypotheses you drafted do not meet the criteria listed above, you may wish to try your hand again.

Skeleton tables can be properly constructed only if decisions have been made about the information to be collected. In answering *Question A15–3*, you may have found that constructing the skeleton tables helped you to clarify your thinking about the formulation of hypotheses. Appraise the table by asking whether the figures (when they are entered) will enable you to test your hypothesis, and whether the requirements for table construction (see Unit A12) are satisfied.

We will return to the topic of causality and its appraisal in Section E.

Basic Procedure for Appraisal of Data

As we may be in danger of losing sight of the wood for the trees, it will probably be helpful if we now review the basic procedure for the appraisal of data. This will bring together the highlights of what we have done and discussed so far. This review includes references to the units in which the topics were dealt with, so that you can refer to them if necessary.

When we examine a table, or graph, or a more substantial body of data, we should consider three questions:

- What are the facts?
- What are the possible explanations?
- What additional information is required, for its own sake or to test these explanations?

Usually all three of these questions are asked, but sometimes the second or third or both are omitted. We may need to know nothing but the facts themselves and be uninterested in explanations, or we may be able to draw simple inferences from the facts—for example, about the individual's risk (Unit A7)—that require no testing.

Figure A16 emphasizes the cyclic nature of the process of data appraisal.

1. What Are the Facts?

To answer this question, we must first ensure that we know what the numbers represent and how they were obtained or calculated (Unit A2). If the data are tabulated, we should carefully examine and compare the rows and columns of figures (Unit A7). We should not regard inferences as facts. We will generally need to summarize the findings; for this purpose we may have to calculate rates

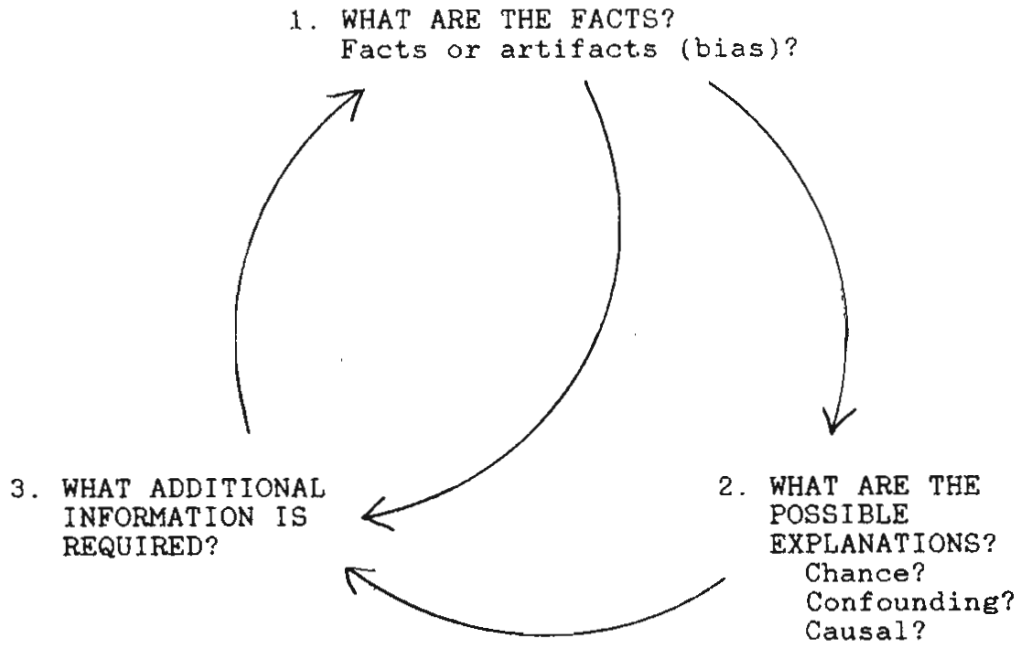

Figure A16. Appraisal of data.

(Unit A6), percentages, or other summary statistics, and it may be helpful to draw a diagram (Unit A4). We should see whether associations exist between variables (Units A9 and A10). If so, we should summarize the features of the associations not only in qualitative terms (direction, linearity, monotonicity), but in quantitative ones, using suitable measures of their strength (such as the difference between rates or proportions, or the ratio of rates or proportions). The data may tell us whether associations are consistent, or whether they differ in different strata.

Before or immediately after determining what the findings are, we should consider the possibility that shortcomings in the methods of gathering the data may have produced distortions. The findings may be biased (Note A16–1), and the ostensible facts may not be true ones (Unit A5). Apparent associations, or their absence, may be artifactual rather than actual. We may need to seek additional information that will enable us to decide whether these problems exist, and whether and how we can make allowance for them. The better our understanding of the basic techniques of study design and data collection (Note A16–2), the more likely we are to detect possible artifacts.

2. *What Are the Possible Explanations?*

Explanations of four kinds should be considered:

- artifactual effects (see above)
- chance occurrence
- confounding (Units A10 and A11)
- causal explanations (Unit A15)

We may be concerned with explaining the facts we have just observed, or facts observed previously as well. In considering possible explanations, we should take account of what we already know, as well as of the facts we have just observed.

A test of statistical significance may be needed to enable us to decide whether we can safely regard the finding as nonfortuitous (Unit A5). Sometimes simple inspection of the data (the "eye test") will enable us to make this decision.

We should list possible confounders that may have affected the associations that interest us. The variables to be considered as possible confounders are those that we know or suspect to be causally related to the disease or other dependent variable, and that are also related to (but not determined by) the other variable involved in the association (see Fig. A10–2). The confounding effect can be important only if the confounder is strongly associated with the other variables. The "universal variables" (age, sex, social class, etc.) are usually candidates for consideration as possible confounders (Unit A11).

Causal explanations can be given serious attention only when we have decided that we can safely ignore the possibility that the association is artifactual, due to chance, or distorted by confounding. We can then consider likely causal ex-

planations (using an appropriate epidemiological model to help us if we wish), select the one we want to test, and frame a hypothesis for testing.

3. *What Additional Information Is Required?*

If we suspect distortions due to flawed methods, we may need extra information about how the data were obtained and the accuracy of the methods used. (We will return to this topic in later exercises.)

If confounding is suspected, we may require new data that will enable us to detect its presence and control its effects, using stratification (Unit A11) or other procedures.

To appraise causal explanations, we will require whatever data are needed to test specific hypotheses.

We may also be interested in additional information for other purposes, not to test explanations for the facts we have, but to add in other ways to our understanding of the phenomenon we are studying. We may be interested in knowing whether an association is consistent in different categories of people or in different circumstances: is there effect modification (Units A11 and A13)? Or we may think that refinement of variables (Unit A12) may give us useful new knowledge; or we may be led by association of ideas to an interest in information about other variables.

New information can serve more than one purpose. Elaborating an association by stratification, for example, may reveal effect modification as well as testing the possibility that a variable is a confounder or an intervening cause (Units A12 and A14).

Whatever new information we require, we should be able to explain precisely why we want it.

Constructing a skeleton table (Unit A12) will often assist us to crystallize our thoughts about the additional data needed.

Exercise A16

This simple exercise, the last in this series, deals with the use of epidemiological data. (We will return to this topic in Section G.) Let us go to another town.

Question A16–1

We learn that the annual incidence rate (persons) of acute gastroenteritis in children aged 0–4 in this town is 60 per 100. What are the possible uses that can be made of this information?

Question A16–2

We also learn that the rate differs in the two ethnic groups. It is 90 per 100 in Easterners and 30 per 100 in Westerners. What are the possible uses of this additional information?

Question A16–3

If the ethnic difference disappears when social class is controlled in the analysis, how would this alter your answer (to *Question A16–2*) about the possible uses of the information that the rate differs in the two ethnic groups?

Question A16–4

Suppose that stratification reveals that the ethnic difference in incidence is not attributable to confounding by social class, but is *modified* by social class. How would this affect the use of the data?

Notes

A16–1. "*Bias.* Any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth"—*A Dictionary of Epidemiology* (Last, 2001). In this definition, "systematically" means "in a specific direction," for example, in the direction of a higher value than the true one. Detailed catalogues of the kinds of bias that may be encountered are provided by Sackett (1979) and Choi and Pak (1998).

A16–2. Methods of investigation are discussed by (*inter alia*) Abramson and Abramson (1999) and various authors in Detels et al. (2001).

■■■■■■■■■■ Unit A17

Uses of Epidemiological Data

Epidemiological data can be used for a variety of purposes (Note A17–1), depending on the interests of the user. Users fall into three main categories. First, in instances where the data relate to a defined community or population, there are users who have a practical concern with the specific community. These include practitioners of public health and community medicine, planners and administrators, physicians and other health professionals, community leaders, and citizens and others with a special interest in the health status or health care of the community. They may be interested in health and health care at the community level, or they may have a responsibility for the care of individuals who belong to the community; or they may be practitioners of community-oriented primary care (Note A17–2), who are concerned with health care at both the community and individual levels.

Second, there are other "pragmatic" users of epidemiological findings, who have no special concern with the community or sample that was studied, but wish to take what can be learned from the data and apply it in a practical way in their own work, wherever it is. They include practitioners of public health and

community medicine, administrators, and others who are interested in health care on a broad scale, as well as physicians and other professionals who provide care for individual patients.

Third, there are users whose basic interest is in "research," who seek knowledge of general interest, without reference to a special local situation or immediate practical applications. This may relate to etiological processes, the natural history of diseases, growth and development, and other topics.

To this list we may add people who use epidemiological data for teaching and learning purposes. The same user may of course fall into more than one category.

The information on the incidence of gastroenteritis in a specific town (*Question A16–1*) may thus have a variety of uses. It is of obvious interest to those who have a specific concern with the town. It becomes part of a community diagnosis that provides a factual basis for decisions on the planning and provision of health care. The incidence rate is a measure of the magnitude of the problem, and may help to determine what importance should be attached to the disease, and what priority it should be given in relation to other problems: does it warrant further investigation, and should intervention be undertaken? The rate indicates the extent of the need for primary and secondary preventive activities (Note A17–3). It may also be used as an indicator of the effectiveness (or ineffectiveness) with which the existing health services prevent the disease. If a decision is made to develop an active program, the present level of the rate may be used to determine a practical target for primary prevention: to what level is it hoped to reduce the rate within the first year or the first five years of the program? Knowledge of the incidence may help in the design of a detailed operational plan: what resources will be needed, in terms of time or manpower, oral rehydration salts, antibiotics, etc.? The incidence rate also provides a baseline for the measurement of change, and hence for evaluating the effectiveness of future efforts in primary prevention.

For physicians, workers in maternal and child health services, and others who provide care at the individual level in the town, the incidence rate provides an estimate of individual risk. Children aged 0–4 have a 60% risk of developing acute gastroenteritis each year. This knowledge may well influence the care and counseling that are given, both in health and in illness.

The incidence rate in this town is unlikely to be of practical interest to practitioners elsewhere, unless they have good reason to believe that their own population is so similar that the findings can validly be applied to it.

Finally, there is a slim possibility that the incidence rate in this town may be of interest to researchers who, by making comparisons with the rates in other populations, may develop interesting new hypotheses to explain the differences.

For users interested in this town, the information about the ethnic difference in incidence (*Question A16–2*) amplifies the community diagnosis. It identifies a population group at especially high risk, and may lead to decisions about the allocation of resources and concentration of attention on a high-risk target group. The ethnic difference may also provide clues to etiology, possibly leading

to a better understanding of the major causes of the disease in this town, so that suitable strategies and procedures can be selected, and the disease can be more effectively prevented.

For the clinician practicing in the town, the extra information provides a better way of identifying individual children who are at high risk, so that he or she can give them the preventive care they deserve.

For the researcher, there is a possibility (although maybe a slim one) that an exploration of the ethnic difference in incidence may yield new knowledge about etiology, not relevant to this town only.

Pending these discoveries, the only value the ethnic difference in this town is likely to have for practitioners elsewhere is that it may lead them to an interest in the possibility that similar differences may exist in their own populations.

The information that the ethnic difference is attributable to confounding by social class (*Question A16–3*) need *not* affect the use of ethnic group as an indicator of risk, at either the population or the individual level. Whatever the reason for the ethnic difference in incidence, this difference remains a fact. Easterners are at higher risk, even if this association is not due to ethnic factors themselves but rather to interrelationships with social class. When exploring the causes of gastroenteritis, however, we need no longer consider causes that are specifically connected with ethnicity. The ethnic difference provides no clues to etiology. The social class difference, however, may do so.

The information that the ethnic difference in incidence varies in different social classes (*Question A16–4*) brings two important benefits. First, it can sharpen the estimates of risk. The stratified data provide us with a specific incidence rate—and hence an estimate of individual risk—for each combination of ethnic group and social class. We now have a more effective way of identifying groups and individuals who are at special risk. Second, comparisons of the incidence rates for different combinations of these variables, and examination of the possible reasons, may lead us to a better understanding of causal factors.

Notes

A17–1. In his book *Uses of Epidemiology,* Morris (1975) described these uses under seven chapter headings: "Historical study," "Community diagnosis: community health," "Working of health services," "Individual chances and risks," "Identification of syndromes," "Completing the clinical picture," and "In search of causes." The uses of epidemiology in public health are listed by Detels (1997) as: "Describe the spectrum of the disease," "Describe the natural history of disease," "Identify factors that increase or decrease the risk of acquiring disease," "Predict disease trends," "Elucidate mechanisms of disease transmission," "Test the efficacy of intervention strategies," "Evaluate intervention programs," "Identify the health needs of a community," and "Evaluate public health programs." A textbook by Brownson and Petitti (1998) emphasizes the applications of epidemiology in public health and health care, rather than as its use in the study of disease etiology. Uses in clinical medicine are illustrated by Sackett et al. (1997).

A17–2. Community-oriented primary care (COPC) combines two elements, the health care of individuals in the community and the health care of the community as a whole, in a single integrated practice (Kark, 1981; Connor and Mullan, 1983; Nutting, 1987; Abramson, 1988; Kark et al., 1994; Gillam and Miller, 1997; Rhyne et al., 1998; Kark and Kark, 1999). Physicians and other clinical workers are responsible both for individual care and for programs that deal in a systematic way with the community's main health problems. Epidemiologic data provide the basis for the planning, implementation, monitoring, and evaluation of these programs (Abramson et al., 1983; Abramson, 1990).

A17–3. A distinction is usually made between different "levels" of prevention; these do not have universally agreed definitions, and their boundaries are not clear-cut. Primary, secondary, and tertiary prevention should not be confused with primary, secondary, and tertiary care. *Primary prevention* refers to the promotion of health (e.g., by improving nutritional status, physical fitness, and emotional well-being and by making the environment salubrious), and to the prevention of specific disorders (e.g., by immunization). *Secondary prevention* refers to the early detection of diseases and other departures from good health, and to prompt and effective intervention to correct them. *Tertiary prevention* refers to the avoidance or reduction of complications, impairments, disability, and suffering caused by existing (irremediable) disorders, and to the promotion of the patient's adjustment to such conditions (sometimes termed *quaternary prevention*).

■■■■■■■■■■ Unit A18

Test Yourself (A)

Now that you have completed this series of exercises, you should be able to do everything included in the following list. Go through the list carefully. If there is anything you think you may not be able to do, return to the relevant unit, which is indicated in parentheses.

You should be able to do the following:

- Describe, and use, the basic procedure for appraising data (A16).
- Determine and summarize the facts shown by a table (A2, A7).
- Determine the facts shown by line diagrams that use (a) arithmetic and (b) logarithmic scales (A4).
- State what condition must be met if graphs are to be used for comparing rates of change (A4).
- Explain the difference between a bar diagram and a histogram (A4).

- Draw
  a line diagram using an arithmetic scale (A3).
  a line diagram using a logarithmic scale (Note A3–2).
  a bar diagram (A4).
  a histogram (A4).
  a pie chart (A4).
  a frequency polygon (A4).
- Explain how graphs can deceive (A4).
- Formulate possible explanations for the facts shown in a table (A5, A11, A14, A16).
- State what criteria should be used in choosing an explanation for testing (A5, A16).
- Construct a skeleton table (A12).
- Explain what is meant by
  an association (A9, A10).
  a dependent variable (Note A3–1).
  positive and negative (inverse) associations (A9).
  an "ecological" association (A10).
  an artifactual association (A5).
- Calculate absolute and relative differences (A2).
- Compare the uses of absolute and relative differences (A3).
- Specify two ways of measuring the strength of an association (A16).
- Explain (in general terms)
  when and why statistical significance tests are done (A5).
  what is meant by a null hypothesis (Note A15–2).
  what is meant by "the alternative to the null hypothesis" (Note A15–2).
  the difference between inductive and deductive reasoning (A6).
- Explain what is meant by elaboration of an association (A11).
- Use stratification to elaborate an association (A11).
- State what new information may be provided by stratification (A13, A14).
- Explain (in general terms) what is meant by confounding (A10).
- State what effects confounding may have on an association (A10).
- Explain how to identify possible confounders (A10).
- Detect confounding (A11).
- Describe at least two methods of controlling for confounding (A11).
- Explain what is meant by effect modification (A11).
- Explain what is meant by interaction between variables (A13).
- Detect effect modification (A11, A13).
- Explain the value of detecting effect modification (A13, A17).
- Explain what is meant by
  a causal relationship (Note A15–1).
  an intermediate or intervening cause (A14).
- Describe three epidemiological models of causal relationships (A15).
- Test a causal explanation (A15, A16).
- Formulate a specific research hypothesis (A–17, A–18).

- State the criteria that should be met by a specific research hypothesis (A16).
- Explain
  what is meant by a rate (A6).
  why comparisons should be based on rates rather than on absolute numbers of cases (A15).
  the difference between crude and specific rates (A8).
  the difference between an incidence rate (spells) and an incidence rate (persons) (A6).
- Calculate
  an incidence rate (A6).
  an age-specific incidence rate (A8).
  a weighted mean (Note A7).
- Explain what is meant by
  risk (Note A6).
  bias (Note A16–1).
  "universal variables" (A11).
  statistical dependence (A9).
  refinement of variables (A12).
  Occam's razor (A4).
  monotonicity (Note A2–1).
  web of causation (A15)
  constellation of causes (Note A15–2)
  primary, secondary, and tertiary prevention (Note A17–3).
  community-oriented primary care (Note A17–3).
- State the main uses of epidemiological data (A17).
- State how epidemiological findings can be used for estimating individual risk (A7, A17).

When you feel that you have nothing more to learn from Section A take a (brief) rest, and proceed to Section B.

Section B

Rates and Other Measures

"Can you do Addition?" the White Queen asked. "What's one and one and one and one and one and one and one and one and one and one?"

"I don't know," said Alice. "I lost count."

(Carroll, 1872)

Unit B1

Introduction

Section B deals with rates and other simple summary measures that express the amount of a disease or other characteristic in a group or population. Its purposes are to ensure that you will be able to make sense of these measures when you encounter them, and use them to summarize your own data. The main topics are

- how rates of different kinds are calculated
- the questions to be asked if we want to know exactly what information a rate gives us
- sources of bias
- the uses to which rates, averages, and other measures may be put by practitioners of public health and community medicine, clinicians, and researchers

We will start with prevalence, and we will then deal with incidence rates, odds, odds ratios, averages, and other measures, and with standardized rates and the pros and cons of their use for detecting and controlling confounding effects.

Real data are used in these and most subsequent exercises. If imaginary numbers are used or the facts have been modified so as to simplify the exercise, you will be told so.

What Is a Rate?

The term "rate" is commonly used for a wide variety of measures of the frequency of a disease or other phenomenon, in relation to (for example) the size of a population. These may be measures of *prevalence*—that is, what *exists* (the

presence of a disease or other attribute in a group or population), or of *incidence*—that is, what *happens* (the occurrence of new cases of a disease, or other events).

All rates are *ratios,* calculated by dividing a *numerator* (e.g., the number of deaths in a given period) by a *denominator* (e.g., the average population during this period). The result is usually multiplied by 100, 1,000 or some other convenient figure, and then expressed per 100, per 1,000, etc. Some rates are *proportions* (i.e., the numerator is contained within the denominator).

The correct use of the term *rate* has unfortunately become controversial. To keep things simple we will use the term "rate" for all measures that are often called rates, even in instances where some epidemiologists regard this as incorrect; alternative terms will be mentioned, so that you can recognize them and use them if you prefer. Some authors restrict the use of "rate" to a ratio that reflects the relative changes (actual or potential) in two quantities, and others restrict it further, to a ratio that represents change over time; in this usage, a prevalence rate is not a "true" rate.

Prevalence Rates

A *prevalence rate* is the proportion of individuals in a group or population who have a given disease or other attribute at a given time, multiplied by 100, 1,000, etc. Sticklers for a strict usage of the term "rate" regard "prevalence rate" as a misnomer, and they prefer to call it just "prevalence" (a term that is also used for the number of people with the attribute, rather than for the ratio we have called a prevalence rate) or "prevalence proportion."

A *point prevalence rate* refers to a specific point of time. The number of people with the disease at that time is divided by the size of the group or population. The numerator contains people who developed the disease before the specified point of time and who were alive and in the population at that time. The rate depends on the incidence rate and the mean duration of the disease, until recovery or death.

A *period prevalence rate* is the proportion of the population with the disease at any time during a specified period (usually a year). The numerator comprises people who developed the disease before and during the period, including those who left, died, or recovered during the period.

A *lifetime prevalence rate* is the proportion of people who have had the disease at any time in their lives, generally until a specified age, sometimes until death.

When used without qualification, a prevalence rate usually refers to point prevalence.

Exercise B1

Question B1–1

A health center needs information for use in planning a home care program for people who are too disabled to leave their houses: for example, how many cases

can be expected to be under care at any given time, and what is the total number of cases that will be treated during a year? The following information is obtained from an agency that has a program in a similar neighborhood. At the beginning of 1999 the population size was 24,000, and at the end of the year it was 26,000. At the beginning of 1999 there were 96 house-bound patients; 20 of these died during 1999, and 4 moved elsewhere. Another 40 people became house-bound during 1999, and 8 of them died during the year.

Calculate the point prevalence rates at the beginning and end of 1999 and the period prevalence rate in 1999.

Question B1–2

A survey provides point prevalence rates of inguinal hernia in men of different ages. Are these lifetime prevalence rates?

Question B1–3

The prevalence of congenital anomalies was measured in a follow-up study of all the children born alive in a defined place and period. The numerator of the rate included children whose anomalies were detected at birth or only later in their lives, or (in some cases) only when they died. The denominator consisted of all the children studied. Is this a point or period prevalence rate?

Question B1–4

In a health survey in a city neighborhood (Note B1), 52 of 431 people aged 65 or more were found to have congestive heart failure, yielding a prevalence rate of 12.1 per 100. Each person was examined once, but the examinations were staggered over a period of 2 years. Is the rate a point or period prevalence rate? Is it a crude or age-specific rate?

Question B1–5

In recent years there has been a marked increase in the prevalence rate of pulmonary tuberculosis in the imaginary Hepi region, and a marked decrease in the equally imaginary Quepi region. Assuming that these are true changes (not artifacts due to changes in case-finding, migration, etc., and not caused by confounding), what are the main explanations you would consider, with special reference to changes in the effectiveness of health care?

Question B1–6

In the survey referred to in *Question B1–4*, the prevalence rate of congestive heart failure at 65–74 years was 6.6 per 100, and at ≥75 years of age it was 23.9 per 100. What is the probable explanation for this positive association with age?

Question B1–7

According to examinations of a representative sample of the total civilian noninstitutionalized population of the United States in 1988–1994, the prevalence rate of high serum cholesterol (240 mg/dl or more) among men rose until the age of 55–64 years, when it reached 28.0%, and then declined to 21.9% at age 65–74 and 20.6% at 75 years or more (National Center for Health Statistics, 2000). What are the possible explanations for this negative association with age?

Note

B1. The figures refer to the presence of "probable congestive heart failure," based on the presence of characteristic symptoms and physical signs (Kark et al., 1979; Gofin et al., 1981).

■■■■■■■■■■ Unit B2

Prevalence Rates (Continued)

In answer to *Question B1–1,* the point prevalence rate per 1,000 was (96/24,000) × 1,000, or 4, at the beginning of 1999 and [(96 + 40 − 20 − 4 − 8)/26,000] × 1,000, or 4, at the end of the year also. The denominator usually used for period prevalence is the average population during the period; the midyear population may be used, or the numbers at the beginning and end may be averaged. The average population was (24,000 + 26,000)/2, or 25,000. The period prevalence rate per 1,000 was therefore [(96 + 40)/25,000] × 1,000, or 5.44.

Point prevalence rates of inguinal hernia (*Question B1–2*) can be regarded as lifetime prevalence rates only in populations where hernias are never repaired. The numerator of a lifetime prevalence rate should include people who report hernia operations or (preferably) have herniorrhaphy scars.

In *Question B1–3,* the rate of congenital anomalies may be regarded as a point prevalence rate, the point of time being the individual's moment of birth—a single point of time for each individual, although the calendar time differs. The anomalies are present at birth but come to light only later. Fuller ascertainment of cases requires long-term follow-up.

A rate based on staggered examinations (*Question B1–4*) may also be regarded as a point prevalence rate—a single point of time for each individual, although the calendar time differs. A rate whose numerator and denominator refer to the same specific age group is, of course, an age-specific rate.

The prevalence of a disease depends on incidence and on the mean duration of the disease. The rise in the prevalence of tuberculosis in the Hepi region (*Question B1–5*) can therefore be attributed to a rise in incidence, an increase in mean duration, or both these factors. The increase in mean duration might be

due to a decrease in the chance of recovery or to a decrease in the risk of dying. Conversely, the declining prevalence in the Quepi region may be due to a drop in incidence, an improved chance of healing, or an increased risk of dying. Improved health care may reduce prevalence (fewer new cases, more cures) or may raise it (fewer deaths). A worsening of health care may raise prevalence (more new cases, fewer cures) or may reduce it (more deaths). Hence, no clear conclusion can be reached about the effectiveness of health care in the two regions.

The most obvious explanation for a rise with age in the prevalence of a disease like congestive heart failure (*Question B1–6*) is the continued accrual of new cases. If the incidence of new cases exceeds the loss of old ones by death or (less likely) recovery, cases accumulate and the prevalence rate rises.

Question B1–7 deals with the declining prevalence of high serum cholesterol among older men in the United States. There are a number of possible explanations for this negative association with age, apart from the very unlikely possibility that it happened by chance in this particular sample. First, perhaps this reflects metabolic changes related to the aging process. Second, the sample was drawn from men living at home; if men with a high serum cholesterol are more prone (because of associated disorders) to be in an institution, men living at home will have a relatively low prevalence rate; and this effect is likely to be most marked above the age of 65, when the risk of being in an institution is highest. Third—and this is the most obvious explanation—a raised serum cholesterol may reduce the chance of surviving to an advanced age. This *selective survival* will tend to reduce the rate in older people.

Fourth, there may be *confounding*. Especially in changing populations, people of different ages may differ in their ethnic group, social class or other characteristics, and these differences may confound associations with age. Fifth, it must be remembered that age groups represent the survivors of separate birth cohorts (people born at different times), who have had different lifestyles and have lived through different experiences. Age-related variation in the prevalence of high serum cholesterol may be expressions of this *birth cohort effect* (Note B2): older men in the United States may have been less exposed in their earlier lives to cultural and environmental influences that raise serum cholesterol, and these earlier influences may also find expression in their current lifestyle. This—rather than their more advanced age—might account for the decrease in prevalence.

Exercise B2

When we are presented with a prevalence rate, we must make sure we know exactly what the figure represents ("What are the facts?"), and appraise its accuracy, before making use of it.

In a paper entitled "Varicose veins and chronic venous insufficiency in Brazil: prevalence among 1755 inhabitants of a country town" (Maffei et al., 1986), we are told that the prevalence rate of varicose veins in adults was 47.6%.

List the questions that you would want answered in order to ensure that you know exactly what information you have been given.

Note

B2. A *cohort effect* or *generation effect* refers to "variation in health status that arises from the different causal factors to which each birth cohort in the population is exposed as the environment and society change. Each consecutive birth cohort is exposed to a unique environment that coincides with its life span"—*A Dictionary of Epidemiology* (Last, 2001).

■■■■■■■■■■ Unit B3

Questions About a Rate

To know what information a rate provides (Exercise B2), we need to ask four basic questions: What kind of rate is it? What is it a rate *of*? To what population or group does it refer? And, how was the information obtained? (These questions may be asked about any kind of rate, not only prevalence rates.)

1. What Kind of Rate Is It?

We might, for example, want to know whether it is a point or period prevalence rate.

2. Of What Is It a Rate?

How was the disease (or other attribute) defined? Was the same definition used in all instances? Most diseases exhibit a wide spectrum of abnormality, ranging from extremely mild to severe conditions, and different cutting-points might be used for deciding whether the disease is present or absent. Or, as often happens, does no one know what the diagnostic criteria were?

3. To What Population or Group Does the Rate Refer?

The denominator should be defined with respect to *place, time,* and sometimes *personal characteristics.* (Who? Where? When?) In the present instance we have some information about the place (a country town in Brazil), but we do not yet know when the study was done, or what precisely is meant by "adults." As we will see later (Unit B5), incidence rates of different kinds use denominators of different kinds.

4. How Was the Information Obtained?

Was the whole of the target population or group studied? If only part was studied, how was it selected? (Who were the 1,755 people who were studied?) Was

the sample a representative one, chosen by acceptable methods (see Note B3–1)? If not, the rate may be biased (see Note A16–1). Were many members of the population or sample excluded because they refused, could not be located, or for other reasons? If so, this may have biased the rate. (Is anything known about the characteristics of those who were excluded?) If a sample was studied, how big was it? The smaller it was, the greater the chance that the findings in the sample may differ from those in the population as a whole (sampling variation; see Note B3–2). How was the numerator information obtained? By observation (e.g., clinical or laboratory examinations), or by asking questions, or from documentary sources? If by observation, what methods were used (and were they standardized and tested)? If by asking questions, what was asked, who did the asking, and was a standard wording used? If from documentary sources, what records were used? Whatever methods were used, what is known about their accuracy? To understand what the rate of varicose veins tells us, we need answers to all these questions (and will probably find them if we carefully peruse the paper describing the survey). In some instances we may also need to ask how information was obtained about the size of the denominator.

Exercise B3

Question B3–1

This question asks you to consider possible sources of inaccuracy in prevalence studies. In each of the following instances, suggest one possible source of bias, and (if you can) specify the direction of the bias. ("Bias" was defined in Note A16–1).

1. What bias would you suspect in a survey of the prevalence of disability in the elderly population of a city, based on an investigation of members of old people's clubs?
2. What bias would you suspect in a household survey to determine the prevalence of senile dementia in a city?
3. What bias would you suspect in a survey of the prevalence of various electrocardiographic abnormalities after an acute myocardial infarction, conducted by examining all the patients treated for this condition in hospitals in the city?
4. What bias would you suspect in a questionnaire-based community survey of mental illness in which 30% of the study sample refused to be interviewed or examined?
5. What bias would you suspect in a survey of the prevalence of diabetes in a city, based on the use of the question "Has a doctor ever told you that you have diabetes?"
6. What bias would you expect in a survey of the prevalence of drug abuse?
7. What bias would you suspect in a survey of the prevalence of cigarette smoking, based on questions put to people who had been exposed to intensive antismoking education?

8. What bias would you expect in a survey of the prevalence of peptic ulcer, based on questions about the occurrence of typical ulcer pain?
9. What bias would you suspect in a survey of the prevalence of congestive heart failure based on one-time examinations?
10. What bias would you suspect in a survey of the prevalence of hypertension based on one-time measurements of blood pressure?
11. According to the U.S. National Health Interview Survey (Adams et al., 1999), the prevalence rate of diabetes in people aged 45–64 years was 58.2 per 1,000 in 1996, with a 95% confidence interval of 46.0 to 70.4. Can these findings be applied to the United Kingdom? Do you know what a confidence interval is?

Question B3–2

Although this question is also about bias, it is a digression, for it is based on a study with no pretension to the measurement of rates or other quantitative measures. An analysis of tape-recorded interviews with expectant mothers, in which they were permitted to speak in their own words, revealed that women who had seen relatives or friends breast-feeding successfully were more likely to intend to breast-feed and to be confident that they would be able to. Women who intended to breast-feed generally did so. The subjects were 21 White low-income London women expecting their first baby, recruited by doctors and nurses known to one of the researchers; an effort was made to ensure that the sample included some teenagers who intended to formula-feed. The women were interviewed early in pregnancy, and 19 of them again about 6 to 10 weeks after delivery. The main message of the study was that women hoping to breast-feed but with little exposure to breast-feeding might benefit from antenatal apprenticeship with a breast-feeding mother, preferably a relative or friend (Hoddinott and Pill, 1999). What are the possible biases? Has this study any value?

Notes

B3–1. A sample selected by *strictly random* methods—that is, by drawing lots or by using tabulated or computer-generated random numbers—can be regarded as a representative one. The population may first be divided into groups (strata), and a random sample selected from each group (*stratified random sampling*). The sampling units need not be individuals, but may be households, schools, or other aggregations whose members make up the sample (*cluster sampling*). A *systematic sample* (e.g., taking every third individual in a list) may often be regarded as equivalent to a random sample. Haphazard methods, not based on strictly random selection or a predetermined system, are sometimes misreported as "random," but do not guarantee representativeness.

B3–2. Chance differences may be expected between the findings in different random samples drawn from the same population, and the findings in any specific sample may differ from those in the whole population. This is called "random sampling variation or, more simply, "sampling variation" or "sampling error."

Unit B4

Sources of Bias

Question B3–1 illustrates two kinds of bias: selection bias and information bias. *Selection bias* occurs if the individuals for whom data are available are not representative of the target population (the population we wish to investigate). *Information bias* is caused by shortcomings in the way that information is obtained or handled. (See Note A16–1 if you want a more detailed inventory of kinds of bias.)

Questions (1) to (4) provide examples of possible selection bias. In (1), old people who are active enough to be members of clubs are not representative of the elderly population, and the prevalence of disability is therefore likely to be underestimated. In (2), people living at home (and not in institutions) are not representative of the elderly population of the city, and the prevalence of senile dementia is probably underestimated. In (3), patients treated in hospital for myocardial infarction are not representative of all patients with this condition, since those with very mild lesions or very severe ones (so serious that there is a strong chance of dying before reaching hospital) will tend to be excluded from the study; the direction of the bias with respect to electrocardiographic abnormalities is difficult to predict. In (4), the high nonresponse rate may well lead to a biased picture of the prevalence of mental illness, but it is difficult to guess the direction of the bias: mentally ill people may have been particularly eager, or particularly reluctant, to participate in the study.

Questions (5) to (10) provide simple illustrations of possible information bias. The use of questions is likely to yield underestimates of the prevalence of diabetes (many people with diabetes do not know they have it), and of drug abuse and smoking, because people tend to give answers they think are socially acceptable; a study in the Netherlands has shown that a question-based survey of alcoholism would miss over half the known problem drinkers (Mulder and Garretsen, 1983). On the other hand, the use of questions may overestimate the prevalence of peptic ulcer, since most people with typical symptoms do not have ulcers on gastroscopy (unless they are outnumbered by people who have ulcers without typical symptoms). If the definition of congestive heart failure includes patients who are temporarily in remission, one-time examinations may yield an underestimate of prevalence. On the other hand, if hypertension is defined as sustained hypertension, one-time measurements of blood pressure will provide an overestimate of prevalence.

In (11), the prevalence of diabetes was studied in a sample of the population of the United States, and we have no good reason to believe that we can apply the findings to the United Kingdom. The confidence interval (see below) does not help us in this respect.

Confidence Interval

Because of random sampling variation (see Note B3–2), the findings in a random sample may not accurately reflect the situation in the target population

from which the sample was drawn. The confidence interval expresses this uncertainty. It tells us within what range we can assume the true value in the target population to lie, with a specified degree of confidence. A narrower range (for a given degree of confidence) means a more precise estimate. The larger the sample, the more precise the estimate. The width of a confidence interval can be influenced by the size of the sample (the larger the sample, the more precise the estimate will be), the degree of confidence required (a 99% interval will be wider than a 95% interval), and the variability of whatever is being measured. Confidence intervals express uncertainty caused by random variation, not uncertainty caused by flaws in the study methods, and they may be misleading if such flaws are present. (See Note B4–1.)

In Exercise B3 (11) we are told that the true prevalence rate of diabetes in people aged 46–64 years in the United States, as measured by the methods used in the National Health Interview Survey, is probably between 46.0 per 1,000 (the lower confidence limit) and 70.4 per 1,000 (the upper confidence limit). This interval has a 95% probability of including the true value.

It can be calculated that if a sample four times bigger had been studied, the 95% confidence interval would have been 52.1–64.3 per 1,000. If a sample had been one-quarter the size, the confidence interval would have been 33.8–82.6 per 1,000.

Confidence intervals are sometimes used when it is wished to generalize the findings to a broad reference population, even though a random sample of this population was not studied. We are then estimating what findings might be expected in a hypothetical large population of which the study population was a random sample (see Note B4–2). This use of confidence intervals is open to question. In the present instance we have no reason to assume that the United States is representative of the world, and it would be wrong to use the confidence interval as an estimate of the probable prevalence rate in people (of this age) in general.

Validity

Exercise B3 can be used to illustrate the uses of the term "validity" (from the Latin *validus,* meaning "strong"). The term is used in three main ways.

First, it may be applied to a method for measuring a specific characteristic. The *validity of a measure* refers to the adequacy with which the method of measurement does its job; how well does it measure what we want to study? When we suspected information bias in Exercises B3 (5) to (10), we were expressing doubt about the validity of the measures.

Second, the term may be applied to a study as whole (*study validity*) or to the inferences drawn from a study. Inferences about causal associations, for example, are not well-founded if due attention has not been paid to possible artifacts, chance effects, and confounding; and a study is not valid if it cannot provide accurate information, or cannot enable well-founded inferences to be drawn concerning the target population that was studied. This is sometimes termed the *internal validity of the study.* A study's validity may be impaired by selection bias,

information bias, uncontrolled confounding, an unduly small sample, or other shortcomings.

Third, the term may be applied to generalizations to a broader reference population, beyond the target population that was studied. This is the *external validity of the study.* When we doubted that the findings of the U.S. Health Interview Survey could be applied to the United Kingdom or to people in general, we were questioning the study's external validity.

Qualitative Studies

Question B3–2 describes a study that uses *qualitative,* not quantitative, methods (see Note B4–3). The question may seem out of place, but qualitative studies might seem out of place anywhere in this book, for they are seldom mentioned in epidemiology texts. This place is as good as any to discuss them.

Qualitative studies do not measure quantities or frequencies, and their findings are described in words rather than numbers. They are useful in investigations of beliefs, perceptions, and practices regarding health; of the prevention and treatment of illness; and of the utilization of traditional and other health care. They provide "culture-specific maps [that] can help to improve the 'fit' of programmes to people"—maps that show the presence of beliefs and behaviors, but not their numerical prevalence in the population (Scrimshaw and Hurtado, 1987). A study of patients who had a heart attack, for example, pinpointed the misconceptions (about heart attack symptoms) that contribute to delay in calling for medical help (Ruston et al., 1998). These studies may be used in combination with quantitative ones—for example, by providing hypotheses for subsequent quantitative testing. Their methods include interviews and conversations in which key informants and other members of the community express their attitudes, perceptions, motivations, feelings, and behavior; *focus groups,* in which selected informants talk freely and spontaneously about themes chosen by the investigator; *field studies* (observations of social life in its natural setting, including observations in health care facilities); and *participant observation* (where the researcher is personally involved in the action being observed).

As in many qualitative studies, there is obvious selection bias in the breast-feeding study described in *Question B3–2;* it would be difficult to generalize from the findings, even if the sample size warranted this. But this does not alter the useful fact that in some women there was an association between previous witnessing of successful breast-feeding and the decision to breast-feed. A quantitative appraisal could possibly be performed subsequently, using appropriate sampling and the usual methods of epidemiologic research. The association may of course be a chance one, or attributable to the confounding effects of age or other variables. (How could these possibilities be explored? See Note B4–4.)

Information bias must also be considered, since different researchers analyzing the same transcripts might obviously reach different conclusions. In this instance, however (as in all good qualitative studies), pains were taken to minimize this type of bias. Two researchers were involved in the analysis of the transcripts; rigorous use was made of systematic methods of *content analysis* that have been

developed and well validated in the social sciences; and synopses were subsequently sent to the mothers for their confirmation.

The study is therefore of value, because it demonstrates an association that might not have been revealed by other methods and that (if it is not attributable to chance or confounding) may have practical implications in health care, for at least some expectant mothers.

Exercise B4

In this exercise you are asked to consider the uses of prevalence data. (You may wish to review Unit A17, which dealt with uses of incidence data.)

The prevalence of infection with *Schistosoma mansoni,* the parasite that causes intestinal bilharzia, was investigated in a rural district of Zambia (Sukwa et al., 1986). A sample of villages was selected (cluster sampling—see Note B3–1), and the parasite's eggs were sought in stool specimens from the residents of these villages. You may assume that there was no selection bias and that the methods of study were valid. The figures shown in Table B4 were calculated from the published findings.

Question B4–1

How would the facts shown in Table B4 help you if you were a doctor providing clinical care in this region of Zambia?

Question B4–2

What uses could you make of these facts if you were responsible for planning and organizing health services in this region? Give consideration to the possible use of prevalence data in evaluating the effectiveness of health services.

Question B4–3

Can facts like those shown in Table B4, or facts on the prevalence of infection in relation to characteristics other than age, be used to identify groups or individuals who have an especially high risk of becoming infected?

Question B4–4

Assuming that we knew very little about the causation of bilharzia, could the prevalence data provide clues to etiology? If we had a similar table for another region of Zambia, showing much lower rates, how would this help? What reservations might you have in making this type of use of prevalence data?

Question B4–5

The prevalence rate of untreated dental caries (of one or more teeth) in children aged 6–17 years in the United States in 1988–1994 was 23.1%, according to clinical examinations, conducted at mobile centers, of a representative sample of the

Table B4. Prevalence of *Schistosoma Mansoni* Infection, Zambian Villages, by Age

Age (yr)	Rate per 100*
5–9	66 (59–73)
10–14	80 (72–86)
15–19	75 (61–85)
20–39	69 (60–76)
≥40	62 (54–70)
Total (≥5)	69 (66–73)

*95% confidence intervals shown in parentheses.

population (National Center for Health Statistics, 2000). What other untreated dental caries prevalence rates would be useful as a guide to decisions on public health policy?

Notes

B4–1. More strictly, the 95% confidence interval is the interval calculated from a random sample by a procedure that, if applied to an infinite number of random samples of the same size, would, in 95% of instances, contain the true value in the population. To unravel this, consult a statistics textbook. Methods of estimating confidence intervals for a variety of measures are described by Altman et al. (2000). Appropriate computer programs include WHATIS and CONFINT in the PEPI package (Note A3–7) and the CIA program provided by Altman et al. (2000).

B4–2. Confidence intervals are sometimes estimated for total-population data (where there is no sampling error) on the grounds that "when the figures are used for analytical purposes such as the comparison of rates over a period, the number of events that actually occurred may be considered as one of a large series of possible results that could have arisen under the same circumstances" (National Center for Health Statistics, 2000, p. 372).

B4–3. *Qualitative studies* and their uses in studies of health and health care are described by (*inter alia*) Pope and Mays (2000), Greenhaigh and Taylor (1997), and Heggenhaugen and Pedersen (1997). They may be used in combination with quantitative studies (Black, 1994; Kroeger, 1983; Coreil et al., 1989).

B4–4. The possibilities that the association is fortuitous or caused by confounding might be explored in a larger study; confounding, for example, might be controlled by stratification. A successful controlled trial comparing the subsequent infant-feeding practices of mothers exposed and not exposed to a "breast-feeding apprenticeship" during their pregnancy would also answer this question. Experts on qualitative methods recommend the use of more than one qualitative method to see whether they lead to the same conclusions (*triangula-*

tion); this may offer a safeguard against artifactual, chance, and some confounding effects.

■■■■■■■■■■ Unit B5

Uses of Prevalence Data

In answer to *Question B4–1*, the prevalence rate of a disease tells a clinician what probability he or she can assign to the presence of the disease in an individual patient, before interviewing and examining that patient. This "pretest probability" can help the clinician decide what diagnoses to explore and what tests to perform. Doctors who are aware that the prevalence rate of *Schistosoma mansoni* is well above 50% (from the age of 5 years) would know that every one of their patients (from the age of 5 years) is more likely than not to have this infection. Thus a physician could decide to do specific diagnostic tests as a routine, or (if treatment is safe) to skip the tests and give specific treatment to all patients. The findings might also lead the clinician to undertake preventive activities.

Prevalence rates like those in Table B4 contribute to the community diagnosis that provides a factual basis for decisions on the planning and provision of health care (*Question B4–2*). They indicate the size of the problem and may help in determining priorities; how much effort should be put into investigating and controlling the problem? Prevalence rates may sometimes pinpoint groups requiring special care; but in our instance the rates in all age groups are so high that there seems little justification for giving special attention to older children, although their rate is especially high. The high rates might lead to a decision to undertake a mass treatment campaign, as well as intensive educational and environmental measures.

The prevalence of a condition that (like bilharzia) can be prevented or cured can be used to measure the effectiveness of health care. If an intervention program is in operation or contemplated, its effectiveness may be monitored by repeated measurements of the prevalence rate. It may be difficult to use prevalence data for the evaluation of recent preventive activities, since the prevalence of a long-term condition may be a reflection of what happened long before. In the present instance, however, the high rate (66%) among children aged 5–9 shows that recent preventive activities have not been effective. It is also obvious there is no effective program for the treatment of bilharzia in this region.

In answer to *Question B4–3*, prevalence is not determined solely by the incidence of new cases, and therefore a prevalence rate (unlike an incidence rate) cannot generally be used as an indicator of risk. Prevalence is determined both by incidence and by the mean duration of the condition. Table B4 shows a higher prevalence rate in older than in younger children, but this may not mean that

they are at higher risk of becoming infected. Their higher rate may be due solely to the cumulation of cases, and the lower rates in adults may be due to treatment or spontaneous disappearance of the infection. Prevalence rates can be used to indicate risk only if they reflect incidence, as they may do in short-term diseases. If we found a much higher prevalence of influenza in school *A* than in school *B*, we could certainly infer a difference in the risk of developing the disease. With respect to most long-term diseases, the prevalence rate of cases of recent onset may also be a useful indicator of risk.

Differences between prevalence rates can sometimes provide clues to etiology (*Question B4–4*), though they may reflect differences in the duration of the condition as well as the effect of etiological factors. The higher prevalence rate in older children may have no etiological significance. But if we knew that the infection was more prevalent in this region than in another, this might provide us with clues to etiology; but we would have to be certain that the difference was not due to a difference in the effectiveness of treatment.

The chance of finding clues to etiology in a prevalence study of a long-term condition may be limited because of the time that has passed since the initiation of the disease. The casual factors may no longer be present, or may be difficult to investigate. Even if interesting associations *are* found, it may be difficult to study time relationships: for example, did the postulated cause precede the postulated effect? It may be easy to find that the prevalence of diabetes is higher in obese people, but it is not so easy to know whether the obesity preceded the diabetes.

The prevalence of untreated dental caries (*Question B4–5*) is an obvious indicator of unsatisfactory dental care. As guides to decisions on public health policy, it might be useful to know the rates in narrower age groups, in age groups outside the 6–17-year range, and in population groups defined by income and other characteristics, as well as rates at different times (are they rising or falling?); rates of untreated caries of more than one tooth might also be helpful. The prevalence rate of untreated dental caries was in fact highest in poor children, especially those of Mexican origin (45.8%), but it was not negligible (14.5%) in families with incomes well above the poverty threshold. The rates in children aged 2–5 years were not much lower than those at 6–17 years. An auspicious finding was that the rates in 1988–1994 were less than half what they had been in 1971–1974.

Incidence Rates

Incidence rates describe the frequency of events. The events include the onset of a disease or disability, the occurrence of an episode, recurrence or complication of a disease, the occurrence of seroconversion or other evidence of infection, admissions to hospital, and visits to doctors. A *mortality rate* (death rate) is an incidence rate that measures the occurrence of deaths.

There are two types of incidence rate, with different types of denominator: *number-of-individuals* denominators (or "count" denominators) and *person-*

time denominators (or, for veterinary epidemiologists, cow-time, sheep-time, etc.). Both types may be measured in total populations or in restricted groups—we might, for example, want to measure the incidence of recurrences or deaths in people who have had a myocardial infarction.

If all members of a cohort (group) are followed up for a specified period, the number of individuals in the cohort at the outset can be used as the denominator (the *candidate population* or *population at risk*). The incidence rate of a disease is the number of disease onsets divided by the number of initially disease-free people. If we do a follow-up study of a cohort that contains 2,000 people and find 100 new cases during a year, the rate in 1 year is 100/2,000, or 50 per 1,000. It is a measure of the average individual's risk of incurring the disease during the specified period. It can be called a *cumulative incidence rate,* because the numerator is the number of new cases that accumulate during a defined period; it is sometimes called an *attack rate.* If the event cannot recur (onset of chronic disease, seroconversion, etc.), the rate is a proportion, multiplied by 100, 1,000, etc. If deaths are measured, the rate is a *cumulative mortality rate,* unless the cohort is confined to people with a specific disease, when the rate is a *case fatality rate.* Some epidemiologists refuse to use the word "rate" for incidence measures that are based on number-of-individuals denominators, and prefer terms like *risk, average risk, cumulative incidence, incidence proportion,* and *incidence probability.* In this book we will not apply this restriction to the use of "rate."

A different denominator—person-time at risk—is required if individuals differ in the length of their "at risk" periods. This may happen because individuals cease to be candidates for the event being studied—they may move away, refuse to cooperate, get lost, or die, or the period of risk may automatically end when the event occurs. "At risk" periods may also vary because individuals enter the study cohort at different times. In a study of the incidence of recurrences, complications, or death after a myocardial infarction, each subject might enter the study immediately after the infarction, but at different calendar times, and might be followed up for different periods.

In such instances, the incidence rate is calculated by dividing the number of events by the sum total of the individuals' periods at risk, measured in person-time units. Each individual's *period at risk* must be calculated—that is, the length of time from the start of follow-up until withdrawal from follow-up (including withdrawals because of occurrence of the endpoint event) or until the end of the study. In our 1-year follow-up study of 2,000 people, there were 1,900 who remained disease-free. Each one of these 1,900 was at risk of developing the chronic disease during an entire year, and each contributes one person-year to the denominator. The other 100 were at risk for various periods less than a year, from the onset of the study until the onset of the disease, and each contributes a part of a person-year. A subject who became diseased at midyear, for example, contributes 6 person-months, or 0.5 person-years. If the total number of person-years at risk was 1,950, the person-time incidence rate would be 100/1,950, or 5.13 per 100 person-years. This rate is not a proportion. (Why not? See

Note B5–1). There is here no disagreement about the use of the term "rate." Other terms you may encounter are *incidence density, average incidence rate,* and *interval incidence density.*

Incidence rates in cities, regions, nations, and other changing populations (i.e., in which there are births, deaths, and movements in and out) are generally calculated by dividing the number of events during a specified period by the average population size (then multiplying the result by 100, 1,000, etc.). To avoid confusion, we will refer to these as "ordinary" incidence rates. The total population (or, for a specific rate, the total population in a specific stratum; e.g., males or females) is used as the denominator, even when calculating the incidence rate of new cases of a chronic disease, although this denominator includes people who already have the disease and are not "at risk" of getting it. Can you suggest why a correction is not made? (See Note B5–2.) Into which of the two categories of incidence rate—rates with number-of-individuals denominators and rates with person-time denominators—would you put an "ordinary" incidence rate? (See Note B5–3.)

The two kinds of rate generally have very similar values, so that both can be used as indicators of average individual risk, although a rate with a person-time denominator is not a direct measure of risk. If the rate is very high or the follow-up period is very long, however, the cumulative incidence rate—the measure of risk—may be appreciably lower than the person-time rate. Even then, if a measure of risk is required and only a person-time incidence rate is available, a simple formula can generally be used to estimate risk (Note B5–4).

Although we refer to both these measures—incidence rates based on number-of-individuals and person-time denominators—as "rates," it is important to distinguish between them; this is easy if they are expressed, respectively, as (say) "per 1,000" or "per 1,000 person-years." The two types of rate often necessitate different formulae when they are used in statistical computations. We may not be able to recognize possible sources of bias unless we know with what kind of rate we are dealing.

Mortality rates are computed in the same way as other incidence rates—there are cumulative mortality rates (using number-of-individuals denominators), person-time mortality rates, and "ordinary" mortality rates.

Exercise B5

Question B5–1

Are the following statements acceptable, and why (or why not)?

1. The annual incidence rate of the disease was 1,200 per 1,000 persons at risk.
2. The incidence rate of the disease was 1,200 per 1,000 person-time units.

Question B5–2

The annual mortality rate from injuries among children aged 0–15 years in Finland decreased steadily between 1971 and 1995. The rate in boys decreased by

75%, falling from 36.7 to 9.5 per 100,000. A similar trend was seen in girls. The annual incidence rate of nonfatal injuries admitted to hospital showed no clear trend, in either gender (Parkkari et al., 2000). The data came from accurate official statistics on causes of death, hospitalizations, and population size. What kind of rate was used in this study? Can you suggest possible reasons for the disparity between the trends? What other death rates might help you to understand this disparity?

Question B5–3

For light relief, consider a highly imaginary army base, where there is a complete change of personnel every 3 months and the total strength is always 1,000. It is found that 2,000 soldiers incur syphilis each year. This gives an annual incidence rate (persons) of 200%. Is this a satisfactory measure of risk? If not, what measure do you suggest?

Question B5–4

You learn that the incidence rate of gonorrhea in the United States in 1997 was 122 per 100,000 population (National Center for Health Statistics, 1999). What questions would you ask to ensure that you know exactly what this figure represents ("What are the facts?")?

Notes

B5–1. A proportion is a ratio whose numerator is contained in its denominator. The numerator of a person-time incidence rate (the number of events) is not contained in the denominator (person-time).

B5–2. People who already have a chronic disease are not generally removed from the denominator when an "ordinary" incidence rate is calculated, for two reasons: the data are seldom available; and the correction makes a negligible difference, unless the prior prevalence is higher than it generally is. If the prevalence is 5 per 100, the correction will change the incidence rate by about 5% of its value.

B5–3. The "ordinary" incidence rate is an estimate of the person-time incidence rate, using the average size of the population at risk during a year as an estimate of the number of person-years of risk during that year. The estimate is a good one if the population did not change much in size or composition during the follow-up period—that is, if individuals who left were replaced by others who were similar to them in their chance of occurrence of disease, death, or whatever other event was measured.

B5–4. The cumulative incidence rate (risk) can easily be estimated from the person-time incidence rate, provided that the latter rate does not vary during the period we are interested in. The simplest formula is

$$\mathrm{CI} = \frac{\mathrm{PTI} \cdot t}{(\mathrm{PTI} \cdot t/2) + 1}$$

where CI is the cumulative incidence rate during t time units (e.g., years), and PTI is the rate per person-time unit. [Another formula is: $CI = 1 - \exp(-PTI \cdot t)$.] As an example, if PTI = 5.13 per 100 person-years, the estimated CI after one year is

$$\frac{0.0513 \times 1}{(0.0513 \times 1/2) + 1} = 0.05$$

that is, 5 per 100 persons. On the assumption that the PTI remains constant over a 5-year period, the estimated CI after 5 years ($t = 5$) is 22.7 per 100 persons. The reverse formula, for estimating the PTI per person-time unit from the CI after t time units, is

$$PTI = \frac{CI}{(1 - CI/2) \cdot t}$$

If the rate is low and it refers to a short period, and $PTI \cdot t$ therefore has a low value (say, less than 0.1), the denominator in the cumulative incidence formula is very close to 1, and the cumulative incidence rate during t time units is approximately equal to $PTI \cdot t$. The person-time incidence rate is then a good indicator of average risk. If individuals have equal follow-up periods and occurrence of the event does not remove them from the population at risk (e.g., when the incidence of headaches or spells of a recurrent disease is measured) the person-time and cumulative incidence rates are identical. Person-time incidence rates and cumulative incidence, and their mathematical relationships, are explained in detail by Rothman and Greenland (1998, pp. 30–42) and Kleinbaum et al. (1982, chap. 6).

■■■■■■■■■■ Unit B6

Incidence Rates (Continued)

If an incidence rate refers to an event that can happen to the same individual more than once, such as the occurrence of a new episode of an acute illness or an exacerbation of a chronic one, a rate of 1,200 per 1,000 persons is quite possible (*Question B5–1*). For example, if the average person contracts 1.2 colds a year, the incidence rate ("attack rate") would be 1,200 per 1,000 persons. A rate of 1,200 per 1,000 person-time units is possible even if it relates to an event that cannot recur, such as the onset of a lifelong disease. This is because the choice of the time component of a person-time unit is arbitrary. For example, if the sum total of the individuals' periods at risk is 3,650 days, and 12 events occur, we can express the incidence rate as 12/3,650 = 0.00329 per person-day, or 0.329 per 100 person-days, or 3.29 per 1,000 person-days. But if we measure the same pe-

riods of risk in years we have 10 years instead of 3,650 days, and the rate is 12/10 = 1.2 events per person-year, or 120 per 100 person-years, or 1,200 per 1,000 person-years. Both statements (1) and (2) in *Question B5–1* are therefore acceptable.

The rates used in *Question B5–2* can be presumed to be "ordinary" incidence rates. This is indeed so; their denominators were midyear population figures, used as estimates of the number of person-years of risk during the year (see Note B5–3). Because the rates are low and relate to short periods (single years), they are good indicators of individual risk (see Note B5–4). Possible reasons for the decreasing risk of fatal injuries, with no decrease in the risk of a serious nonfatal injury requiring hospitalization, are a decrease in the incidence of severe (life-threatening) injuries, and a decrease in the case fatality rate (i.e., in the risk of dying once an injury has been inflicted). A fall in case fatality could be due to prompter treatment at the site of the accident, better ambulance services, or improved medical care. We might understand the findings better if we knew the injury death rates in different parts of the child population (classified by age, region, or other variables), death rates for injuries from different causes (e.g., traffic accidents, drowning, poisoning), and injuries of different types (fractures, burns, etc.), as well as case fatality rates. The investigators supply some of these rates, and conclude that the most important single factors are probably improved traffic safety (including safety seats and belts) and better trauma care.

In *Question B5–3* a new cohort of 1,000 soldiers enters the army camp every 3 months and is followed up for 3 months. The simple and obvious way of measuring the risk of incurring syphilis is to calculate the cumulative incidence rate during a 3-month stay in the base.

This is easy to do. During a year there are 4,000 soldiers who are followed up for 3 months, and 2,000 of them contract syphilis. The cumulative incidence rate after three months in the base is therefore 2,000/4,000, or 50 cases per 100 soldiers. This rate, 50%, expresses the individual's risk of developing the disease during 3 months of service in the base. Our data do not enable us to estimate what the risk would be if soldiers remained in the base for a whole year. It might be anything from 50% to 100%.

The annual incidence rate of 200% is an "ordinary" incidence rate, with the average size of the population used as its denominator. It is therefore an estimate of the person-time incidence rate and may be expressed as 200 cases per 100 person-years. The person-time incidence rate is not a proportion (see Note B5–1) and may therefore exceed 100%; a rate of 200 per 100 person-years is quite acceptable. We can express this rate in terms of person-months: 200 cases per 100 person-years is the same as 200 cases per 1,200 person-months, or 16.7 cases per 100 person months, or 0.167 case per person-month or 0.5 cases per 3 person-months.

The person-time incidence rate is not a direct measure of risk. When incidence is high, as in the present instance, the person-time incidence and cumulative incidence rates may differ appreciably. If we wish, we can calculate the

estimated risk that corresponds to an incidence rate of 200 cases per 100 person-years (using the formula in Note B5–4). But we may hesitate to do this, on the grounds that in this instance the "ordinary" incidence rate is probably not a good estimate of the person-time incidence rate: there were many soldiers who contracted syphilis but remained in the denominator of the rate, although they stopped being at risk. This may have produced an appreciable downward bias of the rate, so that the rate underestimates the true risk. If we nevertheless calculate the risk from this rate (for the computation, see Note B6–1), we will find that the estimated risk of contracting the disease in 3 months is 40%; this is lower than the true value of 50%.

If you want practice in the calculation of a person-time incidence rate, assume that in each 3-monthly batch of 1,000 soldiers there were 250 who contracted the disease after precisely 1 month—on payday?—and another 250 who did so after precisely 2 months. Calculate the sum total of the soldiers' periods of exposure to risk, for use as a denominator, and calculate the person-time incidence rate. (For solution, see Note B6–2.)

In answer to *Question B5–4,* the same questions may be asked about an incidence rate as those we previously asked about a prevalence rate (Unit B3): What kind of rate is it? (It may not really be an incidence rate; not everyone knows the difference between incidence and prevalence.) What is it a rate *of?* To what population or group does it refer? And, how was the information obtained? In this instance, there seems no need to ask what kind of rate it is; it is obviously an "ordinary" incidence rate, based on spells of gonorrhea. When the incidence is as low as this, the difference between person-time and cumulative incidence rates is, in any case, negligible. The most important questions are about the numerator: How were the cases identified? How was gonorrhea defined? Were standard diagnostic criteria used? The data are in fact based on reporting of notifiable diseases to state health departments. We can be sure that the rate is an underestimate of the true incidence.

Exercise B6

In each of the following instances, state the main possible source of bias. If you can, specify the direction of the suspected bias. (The illustrations are fictional unless a reference is cited.)

1. In a study to determine the incidence of a chronic disease, 150 people were examined at the end of a defined follow-up period. Twelve cases were found, giving a cumulative incidence rate of 8%. Fifty other members of the initial cohort could not be examined, 20 of them because they had died.
2. In a study of a random sample of adults in Los Angeles County, the presence of depression was determined by asking a set of questions (which you may assume were satisfactory for this purpose). The sample included 809 people who were not depressed; the incidence of depression was measured

by interviewing them again after a defined period. Among 729 who were reinterviewed, 83 (11.4%) were found to be depressed; 80 others refused to be interviewed or could not be contacted (Clark et al., 1983).

3. Some children have convulsions when they are feverish. To determine what risk these children have of becoming epileptic, a series of children with febrile convulsions who had medical care at a university hospital were followed up for a period of many years. It was found that 40% became epileptic (Ellenberg and Nelson, 1980).
4. In a study of the incidence of headaches and other disorders for which medical care is usually sought only if they are severe, use was made of diaries in which the subjects recorded the symptoms they experienced, day by day for 2 months.
5. To determine the incidence of episodes of asthma in adults, detailed records of illnesses and reasons for absence from work were maintained by all the occupational health services in a city.
6. To study the incidence of impotence as a side effect of drug treatment for hypertension, patients were questioned after a year of treatment. They were not told the reason for asking the question.
7. In a similar study, the patients were told the reason for asking the question about impotence.
8. In a third study, in which the patients were not told why they were asked about impotence, two physicians reported very different rates of incidence of this symptom although they had very similar patients and used identical treatment schedules.
9. A two-stage case-finding procedure was used in a study of the incidence of pulmonary tuberculosis. All participants were subjected to mass miniature radiography, and all those with positive results were then given a complete diagnostic workup. What would you like to know in order to appraise the extent of the possible bias?
10. The annual incidence rate of pulmonary tuberculosis in a region was similar each year from 1985 to 1999. In 2000, it was five times as high.
11. The annual incidence rate of malaria in the United States decreased steeply between 1946 and 1949. The number of cases reported annually fell from 48,610 in 1946, through 17,317 and 9,797, to 4,239 in 1949 (Mainland, 1964).
12. According to death certificate data, the rate of mortality due to diabetes in the United States in 1999 was 13.6 per 100,000 (National Center for Health Statistics, 2000).
13. According to death certificate data, the death rate for motor vehicle accidents in the United States in 1998 was 15.6 per 100,000 (National Center for Health Statistics, 2000).
14. The incidence rate of road accident injuries in the Emirate of Sharjah was 810 per 100,000 in 1977, according to hospital records. Patients with these injuries have to be reported to the police, and are therefore specifically identified in the records (Weddell and McDougall, 1981).

15. The incidence rate of motor vehicle injuries in the United States in 1996 was 1.2 per 100 person-years, according to the National Health Interview Survey (Adams et al., 1999).

Notes

B6–1. By use of the formula in Note B5–4, the estimated cumulative incidence rate in 3 months ($t = 3$), calculated from the rate of 0.167 per person-month, is $(0.167 \times 3)/[(0.167 \times 3/2) + 1] = 0.4 = 40\%$.

B6–2. In each cohort of 1,000 soldiers, there are 250 who are at risk for 1 month (until they contract the disease), 250 who are at risk for 2 months, and 500 who are at risk for the full 3 months, without developing the disease. Each batch is therefore exposed to risk for $(250 \times 1) + (250 \times 2) + (500 \times 3) = 2{,}250$ person-months. This is the denominator. The numerator (the number of cases) is 500. The rate is therefore 500 per 2,250 person-months = 0.222 per person-month. This rate is based on a follow-up period of 3 months, and we have no information whatever about what would happen after a longer period in the base. If we wish to estimate individual risk, we can safely do so only for a 3-month period. We may say that the rate is 0.67 (i.e., 67%) per 3 person-months, and use this as a rough indication of a soldier's risk of incurring syphilis during 3 months at the base. Because the rate is high, it would be preferable, however, to calculate the corresponding cumulative incidence rate, which is a more direct measure of risk. The conversion formula (Note B5–4) gives us an estimated cumulative incidence rate of 0.50 (i.e., 50%).

■■■■■■■■■■ Unit B7

Bias in Incidence Studies

In Exercise B6, studies (1) to (5) provide examples of possible selection bias.

Losses to follow-up are a common source of bias. In (1), the incidence rate of 8% is likely to be an underestimate if having the disease increases the chance of dying. We can "play it safe" by calculating an extreme range: what would the rate have been if (a) none of or (b) all of the lost subjects had incurred disease? In the former instance the rate would have been $12/(150 + 50) = 6\%$, and in the latter $(12 + 50)/(150 + 50) = 31\%$; thus, the rate may be between 6% and 31%. This range is so wide (even without allowing for sampling variation) that we might well decide not to use the results. In (2), where the direction of the bias is hard to guess, the possible range is from 10.3% to 20.1% (83/809 to 163/809); on the basis of their knowledge of the nonrespondents' characteristics, the researchers estimated that the true incidence rate was 10.4%.

In (3), the results may have been biased by the fact that the children were a

selected group treated at a teaching hospital, which they may have reached because their convulsions were especially severe or frequent. Such children may be particularly likely to become epileptic. For physicians at this hospital, the finding may indeed be a useful prognostic indicator. But the external validity (see Unit B4) of the finding may be questioned; the rate may overestimate the risk of the average child with febrile convulsions. In fact, a literature search revealed 11 other studies of children treated at hospital clinics or speciality referral clinics, showing rates of subsequent epilepsy that ranged from 6% to 42%; whereas in five studies that tried to identify and follow up all children in a clearly defined population who experienced febrile seizures, the epilepsy rates ranged from 1.5% to 4.6%. Ellenberg and Nelson (1980) concluded that their findings are "probably generalizable to other common and frequently benign conditions. . . . Clinicians evaluating the need for therapeutic intervention should consider that studies from clinic-based populations may overestimate the frequency of unfavorable sequelae." This kind of bias has been called *referral filter bias* (Note B7–1).

In a study of symptoms based on diaries (4), there is a strong possibility of selection bias: people who are prepared to maintain diaries of this kind are not necessarily representative of the general population. Those who have symptoms and are concerned about their health may be more willing to cooperate. This is a kind of "volunteer bias." In some populations, literacy may also be a factor. There is also a possibility of information bias: there is likely to be underrecording, especially toward the end of the study period.

In study (5), the incidence of asthma episodes among workers may not be a valid reflection of their incidence in the total adult population; because people with troublesome asthma may be less likely to be in employment. This is sometimes called the "healthy worker effect."

In studies (6) to (15), there is possible information bias.

In (6), impotence is a symptom that people may prefer to keep to themselves, and underreporting may be suspected. In (7), where subjects were told that impotence was a possible side effect of the treatment they were getting, the direction of possible bias is difficult to guess. The patient's response to a question about impotence may be colored by his global attitude to his treatment. In (8), there is a possibility that the apparent variation in incidence is due to differences in the way the physicians questioned their patients: what phrasing they used, what their manner was, whether or not they suggested that an answer of a particular kind was expected, and how insistent they were. The results may reflect the physicians' prior opinions about the hazards of treatment.

When a screening test is used, as in (9), the possibility must be considered that the test may miss some cases. It would be helpful to know the validity of the test. In particular, what proportion of cases does it miss? What is its false-negative rate?

In (10) and (11), the sudden change in incidence strongly suggests that there were changes in case-finding methods or diagnostic criteria. The rise in tuberculosis incidence may have been due to an organized effort to detect cases. The

striking apparent decline in malaria incidence in the United States was largely due to a change in diagnostic methods; certain health authorities started to require demonstration of the malaria parasite in the blood before accepting a diagnosis of the disease (Mainland, 1964).

Statistics based on death certificates (study 12) usually grossly underestimate the incidence of deaths attributable to diabetes. The reason is that each death is assigned to a single underlying cause of death, and deaths are seldom assigned to diabetes if another disease appears in the certificate, even if the diabetes contributed to this other disease. Mortality rates are two to three times higher in diabetics, but only 10–20% of the death certificates of diabetics assign diabetes as the underlying cause of death. Despite the relatively low mortality rates (according to conventional statistics), diabetes is a leading cause of death in developed countries and many developing countries.

Each of the listed methods of studying the incidence of injuries caused by road accidents is likely to yield an underestimate. Death certificates (13) may have little bias as a source of information on fatal injuries; but if we are interested in all injuries caused by road accidents, they clearly provide only a partial picture. If reliance is placed on clinical records (14), only the injuries that received medical care will be ascertained, and then only if there are good records, including a statement of the cause of the injury. When information about accidental injuries is based on questions (15), there is a possibility that mild injuries will not be remembered or reported ("recall bias"); fatal injuries can obviously not be ascertained in this way. As with many other disorders, single sources of information are likely to yield incomplete data; the more sources that are used, the fuller the picture.

Exercise B7

This exercise deals with specific aspects of the use of incidence rates. The uses of incidence rates are covered in a more general way in Unit A17 (with reference to gastroenteritis in Epiville).

Question B7–1

It is sometimes said that incidence rates are used for acute (short-term) diseases and prevalence rates for chronic ones. Would you accept this as a recommendation? What use might be made of prevalence data for acute illnesses, or of incidence data for chronic ones?

Question B7–2

Incidence rates are often used for evaluating the effectiveness of health care, both in clinical trials of medical treatments and in evaluative studies of health programs directed at communities. What are the kinds of events whose incidence may tell us something about the effectiveness of care?

Question B7–3

A visit to a large (imaginary) hospital, during which a bed-by-bed survey is conducted, reveals that 10% of patients who have undergone surgical procedures have definite evidence of wound infection. Can you estimate the average risk of wound infection, for patients who underwent surgery in this hospital in the recent past?

Question B7–4

Follow-up studies of White women with breast cancer, based on data for 1989–1994 in the United States, show that 14% died in the first 5 years after the diagnosis of the disease (National Center for Health Statistics, 2000). Is this a cumulative mortality rate or a person-time mortality rate? Is it a case fatality rate? (For definition, see Note B7–2).) For patients with this neoplasm, what is the probability of surviving for at least 5 years after diagnosis? What is the probability of surviving for at least 1 year? What is the probability of surviving for at least 10 years? (The published results were computed by a method that controlled for the possible influence of other causes of death; ignore this complication.)

Question B7–5

A report on a series of 40 patients who were given a revolutionary new treatment for a previously incurable disease in a (make-believe) teaching hospital states that the cure rate (the cumulative incidence rate of complete recovery) was 50% in the first year, 50% in the second year, and 75% in the total 2-year period. Can these rates be correct?

Table B7. Number of Spells of Acute Gastroenteritis During a Year: Frequency Distribution

No. of Spells per Child	No. of Children
0	700
1	200
2	80
3	10
4	5
5	2
6	0
7	0
8	0
9	0
10	3
Total	1,000

Question B7–6

A hypothetical study of 1,000 children, all of whom were carefully followed up for a year, yielded the findings shown in Table B7. According to these data, what is the average child's risk of contracting gastroenteritis during a year? What is his or her risk of having two or more spells of the disease? How many spells may the average child be expected to have in a year?

Notes

B7–1. "*Referral filter bias.* As a group of ill are referred from primary to secondary to tertiary care, the concentration of rare causes, multiple diagnoses and 'hopeless cases' may increase." (Sackett, 1979)

B7–2. The *case fatality rate* is usually defined as the proportion of individuals with a specified disease who die of it during a stated period.

■■■■■■■■■■ Unit B8

Uses of Incidence Rates

In answer to *Question B7–1,* incidence and prevalence rates can be used for both acute and chronic diseases. For acute diseases, use is generally made of incidence rather than prevalence rates, for all purposes for which rates are employed. However, the prevalence of an acute disease is also sometimes of interest. During a cholera epidemic, for example, the health authorities may want to know not only how many new cases occur each day, but also how many cases are currently under treatment.

For chronic disorders, prevalence rates provide a basis for inferences about needs for curative and rehabilitative care and may provide clinicians with a useful guide to the probability of a diagnosis; they are less useful than incidence rates for other purposes. The rate of incidence of new cases of a chronic disease provides an indication of the present or recent activity of causal factors. Incidence rates may thus point to a need for primary prevention and may also identify the groups in which this need is most marked. A change in the incidence rate of new cases may be a measure of the effectiveness of primary prevention, and changes in the incidence of complications and other outcomes may be used to measure the effectiveness of curative and rehabilitative care. For the clinician, the incidence rate of new cases provides an estimate of individual risk, and the incidence rates of subsequent outcomes gives an indication of the prognosis. For the researcher, the incidence rates of various outcomes may provide an understanding of the natural history and clinical course of the disease, and comparisons of rates (of new cases or of outcomes) may throw light on etiological processes.

In answer to *Question B7–2*, the occurrence of any event that health care aims to prevent, or any desirable or undesirable effect of health care, may be used as an indication of the effectiveness of care. The goals of health care include the promotion, preservation, and restoration of health (see Note A17–3). Events whose incidence may be measured in clinical trials and other studies of the effectiveness of care thus include the occurrence of infection and other precursors of disease; the occurrence of the disease itself; and the occurrence of subsequent events, such as recovery, remission, complications, recurrences, various signs and symptoms, biochemical and immunological changes, return to work, incapacitation, and death. The occurrence of side effects of treatment may also be measured. In evaluative studies of health educational programs, the main events that are measured are changes in habitual practices, such as the commencement or cessation of cigarette smoking.

If we wish to know the risk of incurring a disease or the probabilities of various outcomes, it is essential to have incidence data. The prevalence data provided in *Question B7–3* cannot tell us the risk of wound infection. The point prevalence rate of such infections among postoperative patients, 10%, tells us nothing about risk. Like all prevalence rates, it is a reflection not only of incidence, but also of average duration; the longer the duration of the disorder, the higher the point prevalence. In this instance, the length of stay in hospital also plays a part: Are patients with wound infections kept in this hospital longer? Or, are they perhaps discharged especially early, to prevent their continued exposure to hospital pathogens or to reduce the hazard to other patients? All we can be sure of is that there *is* a risk of wound infection in this hospital, but we cannot say how big it is.

In *Question B7–4*, we are told that 14% of women died in the first 5 years after diagnosis of breast cancer. This is a cumulative mortality rate, not a person-time mortality rate; the denominator is the number of patients in the cohort at the beginning of the follow-up period, that is, at the time of diagnosis.

The probability of remaining alive for a given time can be calculated by subtracting the risk of dying during that time (the cumulative mortality rate, expressed as a percentage) from 100%. This is called the *cumulative survival rate*, or just the *survival rate*. These terms are sometimes used with reference not only to remaining alive, but to staying free of a particular disease, complication, or other endpoint event. A survival rate is thus the complement of (i.e., 100% minus) a cumulative incidence or mortality rate.

If the cumulative mortality rate for a 5-year period is 14% (*Question B7–4*), the individual patient's probability of surviving for 5 years is 86%. We can easily find the theoretical probability of surviving for 1 year after diagnosis, by computing the person-time mortality rate during the 5-year period, which is the average rate at which patients die, and using this to calculate the expected survival after 1 year (see Note B8). This procedure can be correct, however, only if the rate at which patients die during the 5-year period is a constant one. We have no certainty that this is so: all the patients who die within 5 years may do so in the first year, or all may die after the first year. We therefore cannot estimate the probability of surviving for 1 year. Similarly, we cannot estimate the 10-year sur-

vival rate; we have no reason to assume that the rate of dying in the second 5 years will be the same as in the first 5 years.

The rates cited in *Question B7–5* may look wrong, but they are correct. The follow-up study started with a cohort of 40 patients; 20 were cured in the first year (cure rate, 50%); of the 20 who were still ill at the end of the first year, 10 were cured in the second year (cure rate in the second year, 50%). In the total 2-year period, 30 of the 40 were cured (cure rate, 75%). The method used to combine cumulative incidence (or mortality) rates for separate periods, so as to obtain the rate for the total period, is simple: calculate the survival rates for each period, multiply them together to obtain the survival rate for the total period, and subtract this from 100%. In this study, the cure rate (the cumulative incidence rate of cures) was 50% each year; the survival rate ("freedom from cure") was therefore (100 − 50)%, that is, also 50%, each year. The survival rate in the 2-year period was 50% × 50%, that is 25%, and the cumulative incidence rate of cures in the 2-year period was (100 − 25)%, or 75%.

In the cohort study described in *Question B7–6,* there were 700 children who survived the year without contracting gastroenteritis, and 300 who had one or more spells during the year. The cumulative incidence rate (persons) was therefore 30%, and the risk for the average child was therefore 30%. There were 100 children who had two or more spells, and the risk of having two or more spells was therefore 10%. To know the number of spells a child can expect during a year, we must calculate the mean number of spells per child, by dividing the total number of spells by the total number of children. The total number of spells is (200 × 1) + (80 × 2) + (10 × 3) + (5 × 4) + (2 × 5) + (3 × 10) = 450, and the mean number of spells per child in the population is 450/1,000 = 0.45. This is also the annual incidence rate (spells).

Exercise B8

Incidence rates of fractures of the proximal femur ("fracture of neck of femur," "fractured hip") in women in Oxford, England, in 1983 are presented in Table B8 (Boyce and Vessey, 1985). The information, which came from hospital

Table B8. Annual Age-Specific Incidence of Fractured Neck of Femur in Women, Oxford, 1983

Age (yr)	Rate per 10,000
0–34	0
35–54	2
55–64	9
65–74	22
75–84	112
85–94	322

Data from Boyce and Vessey (1985).

records, refers to "nonpathological" fractures of the neck of the femur, not caused by tumors or other local bone diseases. Census figures were used as denominators. For the purpose of this exercise, you may assume that only patients with a first fracture were included, and that very few of these failed to reach the hospitals that were studied.

Question B8–1

Summarize the facts shown in Table B8. What kind of incidence rate was used?

Question B8–2

What are the possible explanations for the association with age?

Question B8–3

What risk does a woman aged 75 in Oxford have of sustaining a fracture of the neck of her femur within the next year? Do you have any reservations about your answer?

Question B8–4

What is the risk that she will have such a fracture during the next 10 years (if she lives that long)?

Question B8–5

Can you guess (or, if you are that way inclined, can you calculate) the probability that a woman in Oxford will sustain a fracture of the neck of the femur during her lifetime, if she lives to the age of 95. Is it about 1%, 2%, 3%, 4%, 5%, 20%, 40%, or more?

Question B8–6

Can the findings be generalized to men in Oxford?

Question B8–7

Can they be generalized to women who live elsewhere?

Note

B8. Using the last formula in Note B5–4, the person-time mortality rate that corresponds to a cumulative mortality rate of 0.14 after 5 years is 0.0301 per person-year. Using the first formula in Note B5–4, the estimated cumulative mortality rate after 1 year is 0.0297, or 2.97%. The expected survival after 1 year (on the unlikely assumption of a constant rate of dying during the 5-year period of observation) is therefore $(100 - 2.97)\% = 97.03\%$.

■■■■■■■■■■Unit B9

Estimating the Individual's Chances

The rates in Table B8 (*Question B8–1*) show a steep monotonic rise in incidence with increasing age. Looking at the differences between the rates, we see that the rise becomes steeper with increasing age. The rates are based on census figures; they are therefore "ordinary" incidence rates—that is, estimates of person-time incidence rates (see Unit B5). As they refer to patients with first fractures only, they are incidence rates (persons).

We have no reason to suspect that the association with age is an artifact, and it is very unlikely to be due to chance. It is also extremely unlikely that there can be any confounding factor strongly enough associated with both age and fractures of the femur to produce an age trend as strong as the one shown in Table B8. The main possibility, therefore (*Question B8–2*), is that the trend is caused by biological aging or some concomitant of aging, such as increased brittleness of the bones or a tendency to fall or to be involved in accidents of other sorts. We might tentatively suggest that a birth cohort effect (Note B2) might also play a part: older women may be particularly prone to this fracture because they belong to a generation whose bones are especially brittle in old age because of nutritional inadequacies at a younger age.

Incidence rates provide an indication of the average risk of an individual. Because the annual rate for women aged 75–84 was 112 per 10,000, we can infer that for a woman aged 75, the risk of having a first fracture within the next year (*Question B8–3*) is about 1.1%. The rates are not cumulative incidence rates, which would be direct measures of risk; however, they are so low that over short periods they are almost equivalent to the corresponding cumulative incidence rates. (If we use the first formula in Note B5–4, the highest annual rate in the table—322 per 10,000—is equivalent to a cumulative incidence of 317 per 10,000.) A more important reservation is that the rate we are using, 112 per 10,000, applies to a 10-year age group. In view of the steep rise in incidence with age, there is a strong possibility that for women aged 75, who are at the lower margin of the 75–84 age span, the annual incidence rate is lower than 1.1% (and for women aged 84, it is higher).

The risk that a woman aged 75 will have a fracture during the next 10 years (*Question B8–4*) is about 11%. The average annual rate at 75–84 years is 1.1%, so that if we follow up a cohort of women aged 75, we can expect about 1.1% to sustain a fracture each year, and ten times this proportion, or 11%, in 10 years.

The same approach can be used to obtain a rough idea of the lifetime probability of a fracture (*Question B8–5*). If we follow up a cohort from birth, we can expect few fractures below the age of 75; then about 1.1% of women will have a fracture in each of the next 10 years (11% in all), and another 3.2% will have a fracture each year in the next 10 years (another 32%), making the total lifetime probability about 43%.

This method is obviously not accurate, for women who sustain a fracture—

who (as we have just seen) are numerous—are not removed from the denominator. A better method is the one described in Unit B8 (see comment on *Question B7–5*): calculate the cumulative incidence rate for each year of life (using the first formula in Note B5–4), subtract it from 100% to obtain the corresponding survival rate (the rate of freedom from a fracture), multiply all the survival rates together to obtain the survival rate for the total period, and subtract this from 100%. If we do this, we obtain an estimated lifetime probability (to age 95) of 37%. This laborious but straightforward actuarial procedure is called *life table analysis*. Because it is based on "current" rates—that is, on incidence rates observed at a particular time (1983)—it is termed *current life table analysis*.

We must not forget that this estimate is a theoretical expectation, not derived from actual observations of a cohort. It is based on the assumption that the incidence rates observed in 1983 held good, and will continue to hold good, throughout the life-span of the women in question. This is not necessarily true. In fact, the age-specific incidence rates of fractures of the neck of the femur in Oxford were about twice as high in 1983 as they were 27 years earlier (Note B9–1), and we have no idea of what they will be 27 years later. For women who were old in 1983, the lifetime probability that we calculated is an overestimate of the risk they actually experienced during their lives. For women who were young in 1983, we do not yet know what their risk will be.

(Can you suggest a quite different way, conceptually simple although not necessarily feasible, of measuring the lifetime probability of incurring a fracture of the femur? A clue: it has something to do with information about people who die. For answer, see Note B9–2.)

In answer to *Question B8–6*, we should hesitate to apply the findings to men, unless we know from studies elsewhere that the incidence of fractures of the femur does not vary much with sex. In fact, men in Oxford had lower rates than did women, and their lifetime probability of a fracture by the age of 95 was 19%, as compared with 37% for women. (Can this difference be explained by the confounding effect of age? Above the age of 85, there were more than three times as many women as men in Oxford in 1983. For answer, see Note B9–3.)

We should also query the generalizability of the findings to women elsewhere (*Question B8–7*). As noted above, the rates for women in Oxford itself varied markedly over a 27-year period.

Time to Event (Survival Time)

In many follow-up studies there is interest not only in *whether* a specific event occurs, but in *when* it occurs (that is, after how long). The event may be death (the time lapse until its occurrence being the *survival time*), the occurrence of a disease or complication, recovery from a disease, return to work, becoming pregnant, etc. The methods of analysis are those developed for the study of survival times, and the terms "survival time" and "survival analysis" are often used irrespective of the nature of the event.

A *survival curve* is one way of summarizing the results of such a study. This

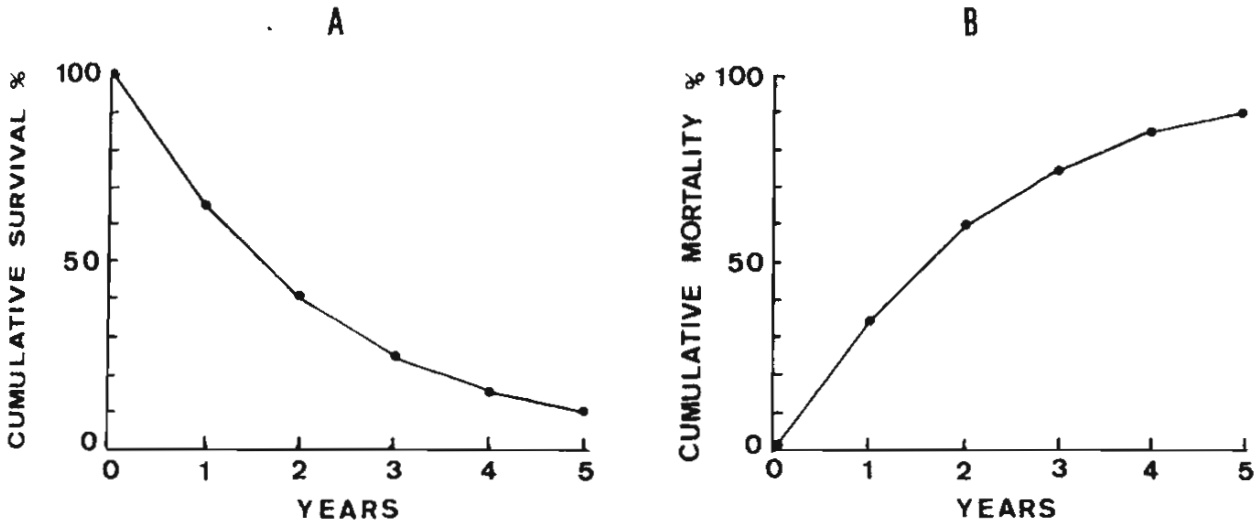

Figure B9–1. Survival curves: (A) cumulative survival rate; (B) cumulative mortality rate.

curve plots the survival experience against time. It may start at 100% and show the *cumulative survival rate*—that is, the proportion of people who have not yet experienced the event (curve *A* in Fig. B9–1). Or it may start at zero and show the *cumulative incidence rate* (the proportion who have experienced the event); if the event is death this is the *cumulative mortality rate* (curve *B* in Fig. B9–1); this is, of course, the complement of the survival rate. Figure B9–1 shows that 65% of patients were still alive 1 year after the onset of a particular disease and 10% were alive 5 years after the onset. Conversely, 35% died in the first year, and 90% in the first 5 years. Both the cumulative survival rate and the cumulative incidence (or mortality) rate have number-of-individuals denominators, and they express the average risk of surviving or not surviving for a specified period.

A survival curve can be drawn as a smooth line or in steps, each step representing a change due to the occurrence of one or more events. As an example, Figure B9–2 shows the cumulative incidence of hypertension at different times after the establishment of a diagnosis of borderline hypertension. Confidence intervals may be shown.

The information may be based on direct observation of a group of people who are all followed up for the same period. Usually, however, individuals are followed up for different periods, because of withdrawals or because they entered the study on different dates. Estimates of the cumulative survival and incidence rates (risks) can then be computed by the *Kaplan-Meier life table procedure* (Note B9–4). An individual might be withdrawn from observation for various reasons—for example, because of the occurrence of the event (so that he or she is no longer at risk), because of death or loss to follow-up, because the study has come to an end, or for other reasons. If the event has not occurred by the end of an individual's observation period, his or her observed survival time is called "censored" and requires special attention in the analysis.

In clinical trials and other follow-up studies, the survival experience of two groups is often compared. This generally requires statistical procedures that can

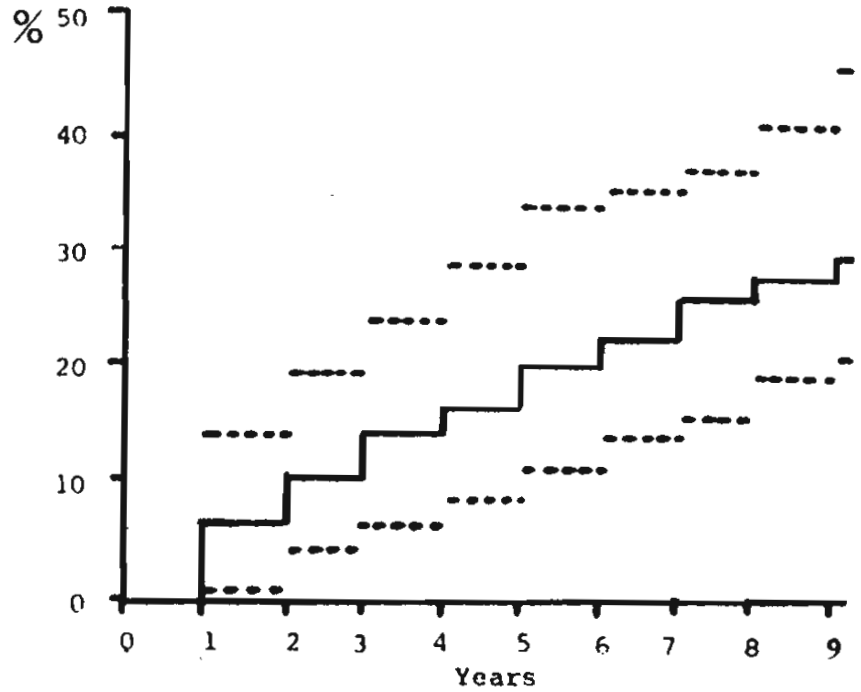

Figure B9–2. Cumulative probability of developing hypertension after establishment of diagnosis of borderline hypertention. Broken lines: 95% confidence limits. *Source* (with definitions): Abramson et al. (1983), data from Ban and Peritz (1982).

cope with censored survival times, such as the *log-rank test* for the difference between survival curves. A *hazard ratio* or relative risk (of the event) may be computed, expressing the ratio of the risks in the two groups during the period studied. (Can you suggest any other ways of comparing survival? See Note B9–5.)

Exercise B9

Question B9–1

The estimated average expectation of life at birth for females in South Africa was 57.6 years in 1970 and 64.5 years in 1996 (Udjo, 1998). These figures were calculated from estimates of the age-specific mortality rates at these times (current life table analysis; see Note B9–4). Does this mean that girls born in South Africa in 1996 can be expected, on average, to live to the age of 64.5 years?

Question B9–2

A survival curve based on a cohort study is portrayed in Figure B9–1. According to this curve, what is the 2-year survival rate? What is the average survival time?

Question B9–3

The median survival time of patients with a certain kind of cancer is 5 years (i.e., 50% of patients survive for 5 or more years). Several large-scale studies have shown that when special efforts are made to detect and treat patients early, the

median survival time is 7 years. What are the main possible explanations for this difference?

Question B9–4

What kind of incidence study will tell us what risk a child has of catching an infectious illness when another member of the family has it?

Notes

B9–1. The incidence of fracture of the neck of the femur in Oxford in 1983 was twice as high as in 1954–1958. The increase was observed in both sexes and at all ages. Boyce and Vessey (1985), who reported these findings, reexamined the data for 1954–1958 and found no evidence that the increase was an artifact.

B9–2. The simplest way of measuring the lifetime probability of a disease is to determine what proportion of people who die have had the disease during their lifetime, or (if the disease is irreversible) what proportion have it when they die. It may be possible to obtain this information for a sample of decedents by examining clinical records or death certificates, by autopsy, or by questioning relatives or medical attendants. Death certificates alone are not a very good source of information about the prevalence of most diseases at death, even if all the recorded causes of death (underlying and contributory) are taken into account (Abramson et al., 1971).

B9–3. The lifetime probability is calculated from age-specific rates, not crude ones, so they obviously control for effects connected with the number of people in each age group. If males and females have different age distributions in Oxford (as they do), this will not affect the age-specific rates in the two sexes, or the lifetime probabilities. In fact, the use of lifetime probabilities and other indices based on current life table analysis is an accepted method of controlling for the confounding effects of age when we are comparing mortality rates in different populations. If we find that life expectancy alters with time or varies in different countries, we can be sure that these findings are not due to age differences.

B9–4. The *Kaplan-Meier life table* procedure, which is based on a follow-up study of a cohort, provides estimates of the cumulative survival rate at different times. A survival probability is computed for each successive interval (the interval until the occurrence of the next event or events), using the experience of the subjects actually observed during this interval. At the end of each interval, the cumulative survival since the baseline time is computed by combining the survival probability in this period with the calculated survival probabilities in previous intervals. The complement of this survival rate is the *risk of the event.* For do-it-yourself explanations of the procedure, see Peto et al., 1977, Kahn and Sempos (1989, chap. 7), or Selvin (1996, pp. 367–371). *Current life table analysis* is similar, except that it uses predetermined time intervals (not those derived from the data) and "current" rates (e.g., those observed in the population in a given year), not those observed in a follow-up study.

B9–5. There are several commonly used methods of comparing the survival experience of two groups (besides use of a log-rank test and the hazard ratio). Comparisons often center on survival rates or cumulative incidence (or mortality) rates during a selected fixed period (e.g., 5-year survival rates or the probability of readmission to hospital within a year after discharge). Also, average (median or mean) survival times may be compared. It is often helpful to make a visual comparison of the survival curves, to see (for example) whether there is a difference throughout the period studied, and whether the difference increases or decreases with the passage of time.

■■■■■■■■■■ Unit B10

Estimating the Individual's Chances (Continued)

Average life expectancy at birth, calculated from the mortality rates at a given time cannot be used as a measure of the individual's chances. This would require the unwarranted assumption that these age-specific mortality rates were or will be valid throughout the individual's life-span. If they decrease, the average life-span will be longer. The average life-span of women born in South Africa in 1996 (*Question B9–1*) will depend largely on the course of the AIDS epidemic rampant in that country at the turn of the millennium. The value of life expectancy statistics is that they provide a way of controlling for the confounding effects of age when comparing populations (Note B9–3).

According to the survival curve (*Question B9–2*), the 2-year survival rate is 40%. There are two kinds of average survival time: the *median survival time,* and the *mean survival time.* The median survival time is the time at which the survival rate becomes 50%. This can be read from the curve; it is about 1.6 years after the onset of the disease. A survival curve does not tell us the mean survival time. To calculate this accurately, we need to know the survival times of all subjects so that we can add them and divide by the number of subjects. This is seldom feasible, as it can be done only after all subjects have incurred the event. An estimate of the mean survival time can be computed from censored data (Selvin, 1996, pp. 371–374).

The longer survival time of cancer patients who are detected early, as compared with those detected in the usual way (*Question B9–3*), may be explained in at least three different ways. First, early treatment may be beneficial. Second, the difference may be an artifact, as different starting points are used for measuring survival times in the two groups of patients. If we make a diagnosis earlier in the natural history of the disease, and measure survival from this earlier time, this alone will produce a spuriously longer survival. (This is referred to as *starting time bias* or *lead time bias.*) And third, there may be another kind of

bias. Cancers in the preclinical (i.e., asymptomatic, not clinically manifest) phase are a biased sample of all cancers, since slow-growing tumors remain in this phase longer than fast-growing ones, and therefore have a raised prevalence among preclinical cases. The cancers identified by early detection procedures therefore tend to have an overrepresentation of slow-growing tumors, which may continue to grow slowly after detection, resulting in a relatively long median survival time.

To determine a child's risk of catching an infectious disease introduced into his or her family (*Question B9–4*), we need to know the incidence rate of the disease in children exposed to the disease in this way. This can be measured by studying a series of families in which the disease has occurred. The required incidence rate is the *secondary attack rate.* This is a cumulative incidence rate whose denominator is the number of exposed contacts—that is, the total number of individuals (in this instance, children) in these families, excluding the first case (the *index case*) in each family. The numerator is the number of cases (excluding the index cases) that occur within a specified time period. If the disease is one to which some children are immune (as a result of prior disease or immunization), we may want to know the risk of susceptible children; for this purpose, we can restrict the denominator to the susceptible children in the families.

Other Rates

You may have to understand or use rates other than those we have so far employed here. *Question B10–1* will test you on some of the following rates. The base (100, 1,000, etc.) is arbitrary. "Per 1,000 population" usually means "per 1,000 in the average (midperiod) population"; for incidence rates the denominator can be person-time units or people, depending on how the information was obtained.

- *Crude birth rate:* live births in a specified period per 1,000 population
- *Fertility rate:* live births in a specified period per 1,000 women of childbearing age (usually defined as 15–44 years)
- *Proportional mortality ratio:* deaths assigned to a specific cause in a specified period per 100 total deaths in the period
- *Cause-specific death rate* (or cause-of-death rate): deaths assigned to a specific cause in a specified period per 1,000 population
- *Infant mortality rate:* deaths under the age of 1 year in a specified period per 1,000 live births in the same period
- *Neonatal mortality rate:* deaths in first 28 days of life in a specified period per 1,000 live births in the same period
- *Postneonatal mortality rate:* deaths in first year of life, excluding first 28 days, in a specified period per 1,000 live births in the same period
- *Fetal mortality rate:* fetal deaths (defined as ≥28 weeks' gestation, ≥20 weeks' gestation, or in some other way) in a specified period per 1,000 total births (live births plus fetal deaths) in the same period

- *Perinatal mortality rate:* fetal deaths plus deaths in the first 7 days of life in a specified period per 1,000 total births in the same period
- *Maternal mortality rate:* deaths from complications of pregnancy, childbirth, and the puerperium in a specified period per 100,000 live births in the same period
- *Admission rate:* hospital admissions in a specified period, per 1,000 population
- *Consultation rate:* consultations (usually with a doctor) in a specified period per 1,000 population

What Are the Odds?

Odds may be defined as the ratio of the probability that something is so or will occur, to the probability that it is not so or will not occur. If a follow-up study shows that 30 smokers develop chronic bronchitis and 20 do not, the odds (for smokers) in favor of developing chronic bronchitis are 30 to 20, or 60% to 40%, or 0.6 to 0.4, or 1.5 to 1, or—and this is the way they are usually expressed in epidemiology—simply 1.5. This is the odds in favor of future occurrence of the disease (also called the odds that the disease will occur, the odds of the disease, or the *disease odds*). Odds can also refer to the ratio of the probability that something is so in the present (or was so in the past), divided by the probability that it is (or was) not. If, for example, 40 people with chronic bronchitis are smokers and 10 are not, the odds (in these patients) in favor of being a smoker are 4 (to 1); these are the *exposure odds,* because they refer to exposure to a factor that affects health. The odds used in betting on a horse ("3 to 1") are the odds, in the bookmaker's view, *against* the horse's winning—the probability that it will lose, in relation to the probability that it will win.

An *odds ratio* is the ratio of one odds to another. It is a widely used tool in the appraisal of associations. By comparing the odds in favor of a disease in smokers with the corresponding odds in nonsmokers, we can see whether the disease is associated with smoking and measure how strong the association is.

Exercise B10

Question B10–1

Calculate the rates specified below, using the following information about the black population of the United States in 1997 (National Center for Health Statistics, 1999; numbers modified to simplify calculations). Average population 34 million, including 8.5 million women aged 15–44. Live births: 600,000. Fetal deaths (at 20 weeks of gestation, or more): 7,600. Deaths in first week of life: 4,600. Deaths in first 28 days of life (excluding first week): 1,000. Deaths in first year of life (excluding first 28 days): 2,900. Total deaths: 277,000. Deaths from heart disease: 77,000.

Calculate the following: crude birth rate, fertility rate, crude mortality rate, specific mortality rate for heart disease, proportional mortality ratio for heart disease, fetal mortality rate, infant mortality rate, neonatal mortality rate, postneonatal mortality rate, and perinatal mortality rate.

Question B10–2

Is the infant mortality rate that you calculated in the previous question a proportion? Is it a cumulative mortality rate (the number of events in a cohort during a given period, divided by the initial size of the cohort)? Is it a rate with a person-time denominator? All the above? None of the above? Who cares?

Question B10–3

If the annual incidence rate of stroke in Blacks aged 65–74 in Chicago was 3 per 100 (Ostfeld et al., 1974), what were the odds (in this population) in favor of having a stroke within a year? If 21 out of 75 swimmers who took part in a snorkel race in the Bristol City Docks developed gastrointestinal symptoms during the next week (Philipp et al., 1985), what were the odds that participants would develop these symptoms? Are the odds that an event will occur very different from the probability that it will occur?

Question B10–4

Table B10 shows the relationship between infant feeding and upper respiratory infections (URI) in American Indian children in Arizona. Use odds ratios to appraise this association. First calculate the disease odds (the odds in favor of having one or more episodes of URI) in bottle-fed babies, and the disease odds in breast-fed babies. Then divide the first odds by the second odds. (This ratio of two disease odds is the *disease odds ratio.*) Now calculate the odds in favor of being bottle-fed, first in the 241 infants with URI and then in the 310 without; divide the one odds by the other to obtain the *exposure odds ratio.* Do you know a short-cut way of calculating odds ratios?

Question B10–5

Now use probability ratios (rate ratios) to appraise the association between infant feeding and URI. First calculate the cumulative incidence rates (persons)

Table B10. Distribution of 551 Infants by Feeding Pattern in First 4 Months of Life, and Occurrence of Upper Respiratory Infections (URI) in First 4 Months of Life

	Episodes of URI*		
Feeding Pattern	One or More	None	Total
Bottle-fed (bottle only, or breast and bottle)	207	238	445
Breast-fed (breast only)	34	72	106
Total	241	310	551

*URI = upper respiratory infection (including otitis media) according to medical (including well-baby clinic) records.

Data from Forman et al. (1984).

of URI in bottle-fed and in breast-fed infants, and divide the first rate by the second. Then calculate the rates of bottle-feeding in children with URI and in children without, and divide the first rate by the second. Compare the rate ratios with the odds ratios.

Question B10–6

In *Question B10–3,* you calculated the odds ratio showing the association between URI and bottle-feeding. Now calculate the odds ratio showing the association between freedom from URI and breast-feeding—in other words, the ratio of the odds in favor of freedom from URI in breast-fed babies to the same odds in bottle-fed babies. In *Question B10–4,* you calculated the rate ratio showing the association between URI and bottle-feeding. Now calculate the rate ratio showing the association between freedom from URI and breast-feeding—that is, the ratio of the probabilities of being free from URI in breast-fed and bottle-fed infants. What do you conclude from the results?

Question B10–7

What are the possible explanations for the association between URI and bottle-feeding demonstrated in Table B10?

Question B10–8

What does an odds ratio of 1 mean?

Question B10–9

What does an odds ratio of 0 mean? If the ratio of odds *A* to odds *B* is 0, what is the ratio of odds *B* to odds *A*?

Question B10–10

The odds in favor of disease *A* are twice as high in vegetarians as in nonvegetarians (i.e., odds ratio = 2). The corresponding odds ratio for disease *B* is 0.5. Which disease is more strongly associated with eating habits?

■■■■■■■■■■ Unit B11

Other Rates (Continued)

The rates requested in *Question B10–1* are

1. Crude birth rate = 600,000/34,000,000 = 17.6 per 1,000 population.
2. Fertility rate = 600,000/8,500,000 = 70.6 per 1,000 women aged 15–44.

3. Crude mortality rate = 277,000/34,000,000 = 8.1 per 1,000 population.
4. Specific mortality rate for heart disease = 77,000/34,000,000 = 2.3 per 1,000 population.
5. Proportional mortality ratio for heart disease = 77,000/277,000 = 27.8%.
6. Fetal mortality rate = 7,600/(600,000 + 7,600) = 12.5 per 1,000 live births plus fetal deaths.
7. Infant mortality rate = (4,600 + 1,000 + 2,900)/600,000 = 14.2 per 1,000 live births.
8. Neonatal mortality rate = (4,600 + 1,000)/600,000 = 9.3 per 1,000 live births.
9. Postneonatal mortality rate = 2,900/600,000 = 4.8 per 1,000 live births.
10. Perinatal mortality rate = (7,600 + 4,600)/(600,000 + 7,600) = 20.1 per 1,000 live births plus fetal deaths.

The answer to *Question B10–2* is "None of the above." The babies who died in 1997 before reaching their first birthday (the numerator) were not necessarily drawn from the babies born in 1997 (the denominator); in fact, about half of them were born in 1996. The infant mortality rate is therefore not a proportion, because the numerator is not contained in the denominator. It is not a cumulative mortality rate, because it does not measure the events in a defined cohort. And it does not have a person-time denominator, because no allowance is made for the fact that infants who died were not at risk for a full year. It can be regarded as an estimate of either of the latter two rates, using the number of babies born in a given year as a substitute for the correct denominator. The estimate is obviously a very good one (and "Who cares?" is therefore an acceptable answer), except in populations with very rapid immigration or emigration or a suddenly changing birth rate, or (for a "person-time" rate) very high infant mortality.

Odds Ratio

In answer to *Question B10–3*, the odds in favor of having a stroke were 3% divided by 97%, or 0.031. The odds in favor of developing gastrointestinal symptoms were 21/54, or 0.39. The corresponding probabilities (expressed as decimal fractions) were 0.030 and 21/75, or 0.28. For stroke, the odds and probability are almost identical; but for gastrointestinal symptoms, they are rather different. The reason is that the probability of stroke was low, whereas the probability of tummy upsets was high. The formula is

$$\text{odds} = P/(1 - P)$$

where the probability P is expressed as a decimal fraction. If P is small the denominator is almost 1, so that odds $\approx P$. You may sometimes want to use the reverse formula, which is

$$P = \text{odds}/(1 + \text{odds})$$

Table B11. Odds Ratio*

	Disease	
Factor	Present	Absent
Present	*a*	*b*
Absent	*c*	*d*

*Odds ratio = *ad/bc*.

In *Question B10–4,* the disease odds are 207/238 = 0.870 in bottle-fed babies and 34/72 = 0.472 in breast-fed ones; the disease odds ratio is therefore 0.870/0.472 = 1.84. The exposure odds are 207/34 = 6.09 in infants with URI and 238/72 = 3.31 in infants without; the exposure odds ratio is 6.09/3.31, which is again 1.84. This is an important advantage of the odds ratio: the answer is the same, whichever way the calculation is done; thus, it becomes unnecessary to distinguish between disease and exposure odds ratios, and we can just refer to the "odds ratio" or "relative odds."

A short-cut formula for the odds ratio (without first calculating the separate odds) is *ad/bc* (see Table B11), where *a* represents the combined occurrence of the two factors (or categories) whose association we wish to appraise. The figures in the table can be frequencies (numbers of individuals), percentages or other proportions, or rates. The odds ratio is sometimes called the "cross-products" ratio.

If we wish to appraise the association between feeding and URI by comparing rates (*Question B10–5*), we can compare the rates of URI or the rates of bottle-feeding. The rates of URI are 207/445 = 46.5% in bottle-fed babies and 34/106 = 32.1% in breast-fed babies, so that the rate ratio is 46.5/32.1 = 1.45. This is the ratio of two risks, so we can call it a *risk ratio,* or *relative risk.* The rates of bottle-feeding are 207/241 = 85.9% in the infants with URI, and 238/310 = 76.8% in the infants without. The ratio of these two rates is 1.12. Note that the two rate ratios are different from each other, unlike the two odds ratios. Note also that the odds ratio is quite different from both the rate ratios.

Despite this example, the odds ratio is usually very close to the risk ratio. (Why is this? For answer, see Note B11). It is often called the "estimated relative risk."

Question B10–6 draws attention to another feature of the odds ratio. The odds ratio showing the association between URI and bottle-feeding is 1.84, and the odds ratio showing the association between freedom from URI and breast-feeding is (72/34)/(238/207), which is also 1.84. But the rate ratio for the association between URI and bottle-feeding is 1.45, whereas the rate ratio for the association between freedom from URI and breast-feeding is (72/106)/(238/445), which is only 1.27; thus if we look at the same data in another way, the association seems weaker! Fortunately, we seldom look at rates of freedom from disease, so perhaps this paradox should not worry us unduly.

In any case, it is clear that the odds ratio possesses desirable features that the rate ratio lacks: it has the same value whether the disease odds or the exposure odds are compared, and whether emphasis is placed on the presence or absence of the disease. As we will see later, it is sometimes possible to obtain an odds ratio but not a risk ratio. The odds ratio observed in a satisfactory sample is always an estimate of the odds ratio in the population, and, if the disease is rare, it is also an estimate of the relative risk. As will be pointed out in Unit D10, in some studies the odds ratio is also an estimator of the ratio of incidence rates based on person-time denominators. Conversely, a rate ratio has the advantage that it is easier to understand. Kahn and Sempos (1989) have summed up the situation:

> Since odds are not as much a part of ordinary usage as chance or probability or risk, many people find the concept of an odds ratio less meaningful than a relative risk. We think this is a matter of custom rather than of basic superiority of one method over the other and that odds and odds ratios will be increasingly used by epidemiologists in the future.

Whatever measure of association is used, Table B10 shows a clear positive association between bottle-feeding in the first 4 months of life and the occurrence of URI in this period. Possible explanations (*Question B10–7*) include (a) chance; (b) the effect of confounding factors (such as poor mothering, or the mother having URI, which may lead both to bottle-feeding and to an increased susceptibility to URI in the child); and (c) causal relationships in either direction: illness may affect the way a child is fed, and bottle-fed babies may be more susceptible to infection or (if infected) to illness—because of what the bottle contains, because of what it lacks, because of the posture in which babies are bottle-fed, or for other reasons. After considering data additional to that shown in Table B10, the authors concluded that their study showed that breast-feeding is beneficial, and reduces the risk of upper respiratory infections not only during the first 4 months, but up to 8 months of age (Forman et al., 1984).

An odds ratio of 1 (*Question B10–8*) means that there is no association; the two odds under comparison are identical. If an odds ratio is zero (*Question B10–9*) one of the odds being compared must be zero. The odds ratio thus indicates a strong (negative) association, unless the other odds is close to zero. If the ratio of odds *A* to odds *B* is zero, odds *A* must be zero, and the ratio of odds *B* to odds *A* (which requires division by zero) would be reported as infinity. In *Question B10–10,* the odds in favor of disease *A* are twice as high in vegetarians and the odds in favor of disease *B* are twice as high in nonvegetarians. The two diseases thus have equally strong associations with eating habits; only the directions differ. An odds ratio tells us both the strength and the direction of an association. If an odds ratio is under 1, it is often easier to understand its meaning if we convert it to its reciprocal (1 divided by the odds ratio).

Exercise B11

Rates, percentages and other proportions, and odds are measures of the frequency of an event or attribute. They are used for categorical variables. This ex-

ercise is concerned with measures used for noncategorical variables. You should consult a book on statistics if you do not know what standard deviations, standard errors, and percentiles and other quantiles are. You need not be a statistician to make sense of data, but you should know the elements of data summarization and understand the principles underlying basic statistical analyses.

Question B11–1

Name some measures that can be used to summarize the central tendency and the spread (dispersion, scatter) of a distribution.

Question B11–2

A study of elderly people with Alzheimer's disease in Finland showed that the concentration of HDL cholesterol in the blood serum was 1.26 ± 0.37 mmol/L (Lehtonen and Luutonen, 1986). What do the numbers mean?

Question B11–3

Examinations were performed of a sample of nonsmoking women living in homes where ten or more cigarettes, cigars, or pipes were smoked daily, and a sample of women not exposed to tobacco smoke in their homes (Brunekreef et al., 1985). The peak flow (a measure of lung function) was lower in the first sample (mean, 6.79 L/sec) than in the second (8.12 L/sec). Is such a difference likely to be due to random sampling variation? If you are not sure, what do you need to know or do in order to answer this question?

Question B11–4

The mean daily caffeine consumption of 2,724 Australian men was 240 mg, with a standard deviation of 145 mg and a standard error of 2.8 mg (Shirlow and Mathers, 1984). Can you calculate the 95% confidence interval (Unit B4)? Assume that the sample was representative, and that caffeine consumption is normally distributed.

Question B11–5

A report on antibodies to poliomyelitis in children in Barbados states that males had slightly higher geometric mean antibody titers than females (Evans et al., 1979). Why were geometric means used instead of ordinary means? (Skip this question if you do not know what titers are.)

Question B11–6

If a study of a large sample demonstrated a bimodal frequency distribution—yielding a curve with two humps, like a Bactrian camel—what explanation would you consider?

Note

B11. We have seen that if a probability is low, the odds are very close to the probability. The risk (incidence rate) of most diseases is—fortunately for humanity—low. The disease odds are therefore usually very close to the risk, and the ratio of two disease odds is very close to the risk ratio. This did not occur in Table B10, where the risks were high (46.5% and 32.1%).

■■■■■■■■■■ Unit B12

Other Measures

Measures commonly used to summarize the central tendency of a distribution (*Question B11–1*) are the *mean,* the *median* (which is the value of the middle observation when all the observations are arranged in ascending order), and the *mode* (which is the value that occurs most frequently). Measures of the spread of a distribution include the *range* and, for a normal distribution (one with an approximately bell-shaped curve), the *standard deviation* (see Note B12). The distribution may be described by stating at what points it can be divided into segments containing equal numbers of observations; these may be terciles, quartiles, quintiles, deciles, or percentiles (the 50th percentile is the median). The *interquartile range* between the upper and lower quartiles can be used as a measure of spread.

Question B11–2 tells us that the mean value was 1.26 mmol/L, but we do not know what the 0.37 represents. It may be the standard deviation of the distribution or the standard error (see Note B12) of the sample mean. (Actually it is the standard deviation.) The ± convention is best avoided.

Question B11–3 refers to the possibility of random sampling variation (Note B3–2). To know the probability that a difference of the observed size might be found between samples when there is no true difference (between the populations from which the samples were drawn), we must do a significance test. Most physiological attributes are normally distributed, and a *t* test would be appropriate. For this test we need the standard errors of the two sample means, or data from which we can calculate these standard errors—that is, the size of each sample and the standard deviation or variance of each distribution. If a *t* test is not appropriate, we can do a nonparametric significance test like the Mann–Whitney test, which makes no assumptions about the underlying distribution; for this we must know the detailed frequency distribution in each sample. If the difference is an artifact or attributable to confounding, there is, of course, little point in a test that appraises how likely it is to be due to random sampling variation.

The 95% confidence interval requested in *Question B11–4* is 234.5–245.5 mg. It is estimated by multiplying the standard error by 1.96 (or, roughly, 2), and then subtracting the result from the mean (to obtain the lower confidence limit), and adding it to the mean (to obtain the upper limit). The interval is from $[240 - (1.96 \times 2.8)]$ to $[240 + (1.96 \times 2.8)]$.

An ordinary (arithmetic) mean is the sum of the values, divided by N (the number of observations). The geometric mean (*Question B11–5*) is the Nth root of the product of the values. This is easily calculated by using logs. It is more useful than the ordinary mean for summarizing the central tendency of a series of titers. If we have five blood specimens, for example, with antibody titers of 1:2, 1:4, 1:8, 1:16, and 1:32, the median is 1:8; the arithmetic mean is $(0.5 + 0.25 + 0.125 + 0.0625 + 0.03125)/5$, that is, 0.194, or 1:5.2; and the geometric mean, the fifth root of $(0.5 \times 0.25 \times 0.125 \times 0.0625 \times (0.03125)$, is 0.125, or 1:8, like the median.

A bimodal curve (*Question B11–6*) may represent the combined findings in samples from two populations that have different but overlapping distributions.

Exercise B12

In this exercise we leave noncategorical variables and return to fractures of the femur. According to the study described in Exercise B8 (Boyce and Vessey 1985), the incidence of fractured neck of the femur in women aged 35 or more in Oxford in 1983 was 35.4 per 10,000. We now learn that in Epiville (which, you will remember, is an imaginary town in a developing region) the corresponding rate in 1983 was half this—18.0 per 10,000.

Following our basic procedure for appraising data (Unit A16), we must first consider the possibilities that this apparent difference may be an artifact, a chance finding, or caused by confounding. We are told that the methods of case identification were identical, and valid, in both localities, and that the difference between the rates is highly significant ($P = .0006$). We now wish to explore the possibility that the difference reflects the confounding effect of age.

Question B12–1

The age distributions of the populations of women aged ≥35 in Epiville and Oxford are shown in Table B12. Do these data support the possibility that age may be a confounder?

Question B12–2

One way of controlling for possible confounding is stratification: we could calculate age-specific incidence rates for Epiville and compare them with those for Oxford. What would be the advantage of using this method of controlling for age?

Table B12. Age Distribution of Women Aged ≥35 Years, Epiville and Oxford: Midyear Populations, 1983

	Epiville		Oxford	
Age (yr)	No.	%	No.	%
35–54	12,000	60.0	10,309	40.1
55–64	5,000	25.0	5,376	20.9
65–74	2,000	10.0	5,558	21.6
75–84	700	3.5	3,400	13.2
≥85	300	1.5	1,055	4.1
Total	20,000	100.0	25,698	100.0

Question B12–3

Unfortunately we cannot calculate age-specific rates, as we do not know the age distribution of the cases in Epiville. Instead, we will use an indirect way of compensating for the age difference between women in Epiville and Oxford.

We know the age distributions of both populations (Table B12), and we know the age-specific incidence rates in Oxford (Table B8). This enables us to calculate how many cases we would expect to find if the same age-specific rates occurred in Epiville as in Oxford. We can then compare the number of cases actually observed in Epiville (which was 36) with the number expected under this assumption. The observed and expected numbers are both determined by the actual age composition of the Epiville women, so that the effect of age is neutralized in this comparison. If there is a difference between the observed and expected numbers, this can be due only to differences between the unknown age-specific rates in Epiville and the known ones in Oxford.

Calculate the expected number of cases of fracture in Epiville by applying the Oxford age-specific rates (Table B8) to the women in Epiville, whose age distribution appears in Table B12. Compare the total expected number with the observed number (36). If there is a difference, how do you explain it?

Note

B12. The *standard deviation* (SD) describes the variability of individuals in a study sample; a large standard deviation means that the individual values are widely dispersed. By contrast, the *standard error* (SE) is a measure of the uncertainty of a statistic observed in a sample as an estimate of the value in the population from which the sample was drawn; the statistic may be a mean, median, proportion, rate, difference between rates, ratio of rates, odds ratio, etc. The larger the standard error, the less certain it is that the statistic derived from the sample (the point estimate) is a good estimate; the smaller the standard error,

the more *precise* the estimate. For some statistics, the estimated 95% confidence interval extends from 1.96 standard errors below to 1.96 standard errors above the point estimate; sometimes (particularly for ratio measures) the log of the point estimate, and its standard error, are used in this calculation.

■■■■■■■■■■ Unit B13

Indirect Standardization

In answer to *Question B12–1,* women in Epiville clearly tend to be younger than those in Oxford. The percentages in the younger groups are lower in Oxford than in Epiville, and the percentages in the older groups are higher in Oxford. This confirms the possibility of confounding, since age is strongly associated both with fracture of the femur (at least in Oxford; see Table B8) and with place of residence.

The confounding effect of age could be controlled by the use of age-specific incidence rates, which (in answer to *Question B12–2*) would serve additional purposes. They would tell us whether age is an effect modifier (Unit A11)—that is, whether there is a similar difference in incidence between Epiville and Oxford in every age group—and would also, of course, tell us the risks of women in different age groups in Epiville.

On the assumption that the Oxford age-specific rates hold good in Epiville, the expected numbers of cases to be expected in a year in Epiville (*Question B12–2*) are: 35–54 years, (2/10,000) × 12,000 = 2.40 cases; 55–64, (9/10,000) × 5,000 = 4.50 cases; 65–74, 4.40 cases; 75–84, 7.84 cases; and ≥85, 9.66 cases. The total expected number of cases is 28.8.

The observed number of cases in Epiville is 36, and the expected number (if the age-specific rates in Epiville were the same as those in Oxford) is 28.8. Both these numbers are determined by the actual age composition of the Epiville women. The observed number is a reflection of the age-specific incidence rates in Epiville, and the expected number is a reflection of the age-specific incidence rates in Oxford. The difference can mean only that, on balance, the age-specific rates in Epiville are higher than those in Oxford. Controlling for the confounding effect of age, the risk of fractures of the femur is higher in Epiville.

According to the crude rates, however, the incidence in Epiville was only half that in Oxford. We can conclude that this finding was a distortion caused by the confounding effect of age.

This simple method of controlling for a confounding effect is called *indirect standardization.* The ratio of the observed to the expected number of cases is called the *standardized morbidity ratio,* or *SMR.* It may be used for incidence or prevalence data, or for mortality data, when it is called the *standardized mortality ratio.* In this instance the SMR is 36/28.8, or 1.25.

To calculate the SMR (standardized for age), we require

- the age distribution of the group or population whose SMR is to be calculated
- the age-specific rates in a standard (reference) population; we used the rates of Oxford women for this purpose

The calculation itself is best left to a computer (see Note A3–7).

The SMR may be used in the same way to control for suspected confounders other than age, or for more than one confounder simultaneously. To control for age and ethnic group, for example, we would need to know the number of people in each age–ethnic category, and must have standard rates for such categories.

The SMR of the reference population is (of course) always 1, since the expected number of cases in this population (using its own specific rates) is the same as the observed number. In our example, the SMR was 1.25 for Epiville and 1 for Oxford.

The process is sometimes taken a step further, by multiplying the SMR by the overall (crude) rate in the standard population, to obtain what is called an *indirectly standardized rate.* (The rationale for this procedure is not simple; see Note B13.) This standardized (or "adjusted") rate is an indication of what the overall rate in the group or population would have been if it had been similar in composition (e.g., with respect to age) to the reference population. In our example, the crude rate in the standard population (Oxford women) was 35.4 per 10,000. If we multiply this by the SMR for Epiville, which is 1.25, we get an indirectly standardized rate of 44.2 per 10,000 for Epiville. The comparable rate for Oxford is, of course, 35.4 per 10,000. This comparison again shows that, controlling for age, the incidence rate was higher in Epiville.

Standardized rates and SMRs are used in the same way. We compare standardized rates or SMRs (based on a common standard) with one another to control for effects connected with the variable(s) we standardized for. Needless to say, SMRs or standardized rates based on different standards should not be compared.

The reference population should be one of the populations we wish to compare, as in the above example, or (less advisedly) some other population can be used as a standard.

Exercise B13

Question B13–1

If you want practice in indirect standardization, calculate SMRs and age-standardized rates for the incidence of fracture of the femur in women aged ≥35 in Epiville and Oxford, using data for men in Oxford in 1954–1958 (Boyce and Vessey, 1985) as the standard. You will find data on the age composition of the two female populations in Table B12, and the facts about the standard popula-

Table B13–1. Population Distribution by Age and Annual Age-Specific Incidence of Fractured Neck of Femur in Men, Oxford, 1954–1958

Age (yr)	Midperiod Population	Annual Rate per 10,000
35–54	14,217	1.1
55–64	4,303	6.5
65–74	2,695	6.7
75–84	1,100	21.8
85–94	164	48.8
Total	22,479	4.2

tion in Table B13–1. The numbers of observed cases in the women were 36 (Epiville) and 91 (Oxford). See if you get the figures shown in Table B13–2. Your results may differ slightly because of rounding-off.

Question B13–2

Table B13–2 shows the crude rates, SMRs, and indirectly age-standardized rates for fracture of the femur in women in Epiville and Oxford. What can we learn from this table?

Note

B13. An indirectly age-standardized rate is calculated by multiplying the observed crude rate by a standardizing factor. This factor is the ratio of the rate S in the standard population to the expected rate E in the population under study (calculated by applying the standard age-specific rates to the age distribution of

Table B13–2. Crude and Indirectly Age-Standardized Rates (per 10,000) and Standardized Morbidity Ratios (SMR) of Fractured Neck of Femur in Women, Epiville and Oxford, 1983

	Epiville (a)	Oxford (b)	Ratio (a:b)
Crude rate	18.0	35.4	0.5
SMR			
Using Oxford women (1983) as the standard	1.25	1.0	1.25
Using Oxford men (1954–58) as the standard	4.0	4.4	0.9
Indirectly age-standardized rate			
Using Oxford women (1983) as the standard	44.2	35.4	1.25
Using Oxford men (1954–58) as the standard	17.0	18.3	0.9

the latter population). *S/E* is an expression of the effect of the difference in age composition between the population under study and the standard population. The standardized rate in the study population is its crude rate *O* multiplied by *S/E*. This is the same as the SMR (i.e., *O/E*) multiplied by *S*.

■■■■■■■■■■ Unit B14

Indirect Standardization (Continued)

A basic way of detecting confounding is to compare the association shown by the crude data with the association seen after control of the suspected confounder. We have previously seen that this can be done by ascertaining whether crude and stratified data yield the same conclusions (Unit A11). Another way is to determine whether crude and standardized measures yield the same conclusions.

In this instance (*Question B13–2*), the crude rates clearly yield different conclusions from the SMRs and age-standardized rates; the ratios shown in Table B13–2 are very different. This shows that there was confounding by age.

Table B13–2 also shows that age-standardized morbidity ratios and indirectly age-standardized rates that use the same standard population yield the same conclusions; the ratios are the same (1.25 or 0.9) in each instance. This of course must be so, since standardized rates (using a given standard population) are calculated by multiplying the SMRs by a constant (the crude rate in the standard population). There is in fact no good reason for using indirectly standardized rates in these comparisons, rather than SMRs.

Table B13–2 also shows that the use of different standard populations may lead to different conclusions. If we use the women in Oxford as the standard, it appears that (controlling for age) the incidence was higher (ratio, 1.25) in Epiville than in Oxford; whereas when we use men in Oxford as the standard, the rates in the two localities become similar (ratio, 0.9). This is an unfortunate feature of indirect standardization. *The reference population should always be one of the populations we wish to compare.* If it is not, the results may be misleading (Note B14–1): the distortion may be negligible, but it can sometimes be substantial. When rates in different subgroups of a study sample are compared, the combined study sample—or the population from which it was drawn—is often used as a standard, but even then the findings may sometimes be distorted.

Table B13–2 also shows that the level of the standardized rate depends on the choice of a standard population: the two standardized rates for Epiville are 44.2 and 17.0! Indirectly standardized rates have no real-life meaning. Their only use is for comparison with the crude rate in the standard population, or with other age-standardized rates based on the same standard. We might as well use the SMRs.

Direct Standardization

Directly standardized rates are hypothetical rates based on the fiction that the groups or populations that are compared have a similar composition with respect to whatever confounder is under consideration. A standard population composition is used, not (as in indirect standardization) a standard set of specific rates.

To calculate an age-standardized rate by the direct method, we require

- the age-specific rates in the group whose standardized rate is to be calculated (The denominator in each age category must be large enough to give us a rate we can rely on.)
- the age distribution of a standard (reference) population

The standardized rate is a weighted mean of the stratum-specific rates in the study population, using the sizes of the strata in the standard population as weights (if this is not crystal-clear, see Note B14–2). Direct standardization can be used to control for confounders other than age, or for combinations of confounders. To control for age and sex together, for example, we would need to know the age- and sex-specific rates in the study population, and the size of the various age–sex categories in the standard population.

If two populations have the same age-specific rates, their directly age-standardized rates will always be identical, *whatever standard population is used.* (This is not true for indirectly standardized rates.)

There is a useful alternative way of standardizing rates for age, without using a standard population: this is to use the age intervals as weights (Note B14–3). The rate at 20–24 years, for example, gets a weight of 5 because it relates to a 5-year age period, and the rate at 25–34 years gets a weight of 10. The standardized rate is then the sum of the weighted age-specific rates. In effect, this simple method gives each single year of age the same weight. This procedure can be thought of as the use of an unrealistic hypothetical standard population with the same number of individuals at each year of age.

Exercise B14

Question B14–1

Unless you feel you do not need practice in direct standardization, calculate age-standardized rates for fractures of the femur in women in Epiville and Oxford, using the age distribution of men in Oxford in 1954–1958 as the standard. The age-specific rates you will need are in Table B14–1, and the facts about the standard population are in Table B13–1. See if you get the rates shown in Table B14–2.

Question B14–2

Table B14–2 shows the rates of fracture of the femur in women in Epiville and Oxford, standardized for age by the direct method. Five sets of rates, using dif-

Table B14–1. Annual Age-Specific Incidence of Fractured Neck of Femur in Women in Oxford and Epiville, 1983: Rates per 10,000

Age (yr)	Epiville (a)	Oxford (b)	Ratio (a : b)
35–54	1.7	1.9	0.9
55–64	12.0	9.3	1.3
65–74	30.0	21.6	1.4
75–84	142.9	111.8	1.2
85–94	400.0	322.3	1.2

ferent standards, are shown. Compare the findings with those shown in Tables B13–2 and B14–1. What are your conclusions about the use of standardized rates?

Question B14–3

Table B14–3 shows cerebrovascular disease mortality rates for Black and White men aged 45–84 in the United States in 1997. It displays age-specific rates, directly age-standardized rates using five different standard populations, age-standardized rates using age intervals as weights (with a footnote explaining the arithmetic), and the ratios of Black to White rates. When the U.S. population in 1977 is used as the standard population, the ratio of the rates is lower than when other standard populations are used. Can you suggest a reason for this? The ratio is even lower when age intervals are used as weights; can you suggest a reason? Can you think of any advantage to the use of age intervals as weights, apart from ease of computation?

Table B14–2. Age-Standardized Rates (per 10,000) of Fractured Neck of Femur in Women in Epiville and Oxford, 1983 (Standardized by the Direct Method, Using Five Different Standards)

Standard Population	Epiville (a)	Oxford (b)	Ratio (a : b)
Oxford women (1983)	45.0	35.4*	1.3
Oxford men (1954–58)	16.9	13.4	1.3
European standard population†	24.4	19.3	1.3
African standard population†	11.4	9.3	1.2
World standard population†	18.4	14.6	1.3

*This is the crude rate.

†See Note B14–4.

Table B14–3. Age-Specific and Age-Standardized Cerebrovascular Disease Mortality Rates for Black and White Men Aged 45–84 in the United States in 1997

Rate	Black (a)	White (b)	Ratio (a : b)
Age-specific, per 100,000			
45–54 yr	61.9	14.9	4.2
55–64 yr	135.7	43.4	3.1
65–74 yr	285.9	142.4	2.0
75–84 yr	650.3	494.2	1.3
Standardized by using standard population, per 100,000			
European standard population	180.3	90.4	2.0
African standard population	143.9	65.7	2.2
World standard population	163.6	77.0	2.1
U.S. population 1940	164.1	78.4	2.1
U.S. population 1997	209.4	115.2	1.8
Standardized by using age intervals as weights (%)*	11.3	6.9	1.6

Source: Center for Disease Control and Prevention, 1999.

*The age-specific rates in Black men are 0.000619, 0.001357, 0.002859, and 0.006503; each weight (age interval) is 10; the age-standardized rate is (10 × 0.000619) + (10 × 0.001357) + (10 × 0.002859) + (10 × 0.006503) = 0.11338 = 11.3%.

Notes

B14–1. "Indirect standardization is best used only for comparing two groups when one of these groups is the standard." For the mathematical basis for this conclusion, see Anderson et al. (1980). If several groups are being compared and one of them is used as the reference group, it is technically incorrect, although the error is usually negligible, to compare the SMRs of other groups with each other.

B14–2. A directly standardized rate is a weighted mean (Note A7) of the rates in specific strata. The formula is $\Sigma(w_i r_i)/\Sigma w_i$, where r_i is the rate in the stratum, and w_i is the weight given to it. If we apply this formula to the incidence rates (per 10,000) of fracture of the femur observed in Epiville (see Table B14–1), using the population figures in Epiville (Table B12) as weights (1.7 × 12,000, + 12.0 × 5,000, and so on, and then divide the total by 20,000) we will, of course, obtain the observed overall rate in Epiville women, which was 18.0 per 10,000 (as stated in Exercise B12). If we use different weights we will obtain a different (hypothetical) overall rate, and this is what is done in direct standardization, using the sizes of the strata in a standard population as the weights. Each weight w_i may be an absolute number or a proportion of the total standard population; in the latter instance $\Sigma w_i = 1$, which simplifies the calculation. Rates that are expressed as 11 per 10,000, 1 per 1,000, etc., can be taken as 11 and 1, respec-

tively, for the purposes of the calculation. Direct standardization can be applied to statistical measures other than rates, such as means.

B14–3. The use of age intervals as weights in direct standardization is described by Breslow and Day (1987, pp. 57–61), Abramson (1995), and Selvin (1996, pp. 360–362). See Note A3–7.

B14–4. The European, African, and world standard populations are hypothetical standard populations for use in direct age standardization. The European population is a relatively old one, with 11% aged ≥65 and 43% aged <30. The African population is a young one, with 3% aged ≥65 and 60% aged <30. For details, see Hill and Benhamou (1995) or Lilienfeld and Lilienfeld (1980, p. 81).

■■■■■■■■■■ Unit B15

The Use of Standardized Rates

In answer to *Question B14–2,* one obvious conclusion to be drawn from the tables is that a standardized rate based on a standard population has little meaning in itself. Table B14–2 shows that the level of a directly standardized rate depends on what standard is used, and Table B13–2 shows the same for indirectly standardized rates. These rates are useful only for comparison with other rates computed in the same way, using the same standard.

Table B14–2 also suggests that the ratio of two directly standardized rates is relatively little affected by the choice of a standard population. In this example, the ratio is consistently 1.2–1.3, which is similar to—and obviously reflects—the ratio of the specific rates in most age categories (Table B14–1). This is an advantage of directly standardized rates; the ratio of indirectly standardized rates or SMRs (Table B13–2) must be treated with circumspection, unless one of the groups compared is used as the standard.

Actually, the choice of a standard population can also affect the rate ratio when directly standardized rates are used. This is not shown by our example, because this distortion happens only when the confounder is also a strong effect modifier. In Canada, for example, where age had a strong modifying effect on time trends between 1971 and 1991 in asthma hospitalization, age-standardized rates showed different trends, depending on whether the 1971 or 1991 Canadian standard population was used (Choi et al., 1999). In such circumstances—where the associations in different strata are very different—it is arguable, however, that there is little interest in *any* summary measure (crude or standardized) that looks at all the strata together.

Both direct and (if an appropriate standard is used) indirect standardization are useful tools for detecting and controlling confounding effects. The ratio of

standardized rates provides a measure of the strength of the association when confounding is controlled. If this differs from the ratio of the crude rates, we know that confounding occurred.

A comparison of standardized rates is not as informative, however, as a comparison of specific ones. The standardized rates tell us that when age is controlled, the overall fracture rate is slightly higher in Epiville than in Oxford. But they cannot tell us that this difference does not occur among younger women (Table B14–1). There is an advantage in examining the specific rates if they are available. This is also demonstrated in Table B14–3, where comparisons of the age-specific rates show the modifying effect of age on the ratio of Black to White mortality rates.

However, there are at least two good reasons for using standardization. The first is its convenience. A single summary rate is much easier to use than an array of specific rates. This is an especial advantage if two or more confounders are controlled at the same time, so that the number of strata is large. Second, it often happens that specific rates are not available, or the denominators in separate strata may be so small that the specific rates are unreliable; indirect standardization may be used in these instances.

In answer to *Question B14–3,* the lower ratios of standardized rates when the U.S. population in 1977 is used as the standard are due to the fact that this is a relatively old population, and more weight is therefore given to the oldest age group, where (as the age-specific data show) the ratio is lowest. The low ratio when age-intervals are used as weights has a similar explanation, since the weights do not taper off with advancing age.

A useful feature of the "age intervals as weights" method of age standardization is that it provides a rate that is meaningful in itself, and not merely a reflection of the arbitrary choice of a standard population. The rate is the sum of the rates in successive years of age, and is hence a cumulative measure that may be regarded as the incidence or mortality rate during the total age-span covered. The rate is not a direct measure of risk, but it is easy to derive a cumulative incidence of mortality rate, or risk, from it (see Note B5–4). In this instance, the computed average risk of dying of cerebrovascular disease before the age of 85 is 10.7% for a Black man aged 40 and 6.7% for a White man aged 40. These estimates assume that the rate is approximately constant within the specific age periods; the narrower the intervals, the more valid the results; they take no account of the effect of mortality from other causes.

■■■■■■■■■■Unit B16

Test Yourself (B)

Now that you have completed Section B you should be able to do everything in the following list. If you have any doubt, return to the relevant unit.

- Calculate
  point and period prevalence rates (B1, B2).
  ordinary, cumulative, and person-time incidence rates (B5).
  cumulative survival rate (B8).
  crude birth rate and fertility rate (B10).
  cause-specific death rate (B10).
  infant mortality rate (B10).
  fetal and perinatal mortality rates (B10).
  neonatal and postneonatal mortality rates (B10).
  maternal mortality rate (B10).
  hospital admission and consultation rates (B10).
  a confidence interval from a standard error (Note B12).
  a standardized morbidity or mortality ratio (SMR) (B13).
  an indirectly standardized rate (B13).
  a directly standardized rate (B14, Note B14–2).
  a directly standardized rate, without a standard population (B14).
- Explain the difference between
  prevalence and incidence rates (B1, B5).
  point and period prevalence rates (B1).
  cumulative and person-time incidence rates (B5).
  direct and indirect standardization (B13, B14).
  standard deviation and standard error (Note B12).
- Explain what is meant by
  lifetime prevalence rate (B1).
  case fatality rate (Note B7–2).
  secondary attack rate (B10).
  median survival time (B10).
  an odds (B10).
  disease odds and exposure odds (B10).
  an odds ratio (B10).
  a risk ratio (relative risk) (B10).
  time to event (B9).
  censoring (B9).
- State what questions you would ask in order to understand what a rate tells you (B3).
- Appraise the possibility that a rate is biased (B3, B4, B7).
- State possible explanations for
  an increase with time in the prevalence of a disease (B2).

a decrease with time in the prevalence of a disease (B2).
an increase with age in the prevalence of a disease (B2).
a decrease with age in the prevalence of a disease (B2).

- Read a survival curve (B9).
- Use incidence rates to appraise the individual's risk (B9).
- Make sense of an odds ratio (B11).
- Compare the uses of prevalence and incidence rates in
  the clinical care of individual patients (B5, B8).
  the planning and provision of health services (B5, B8).
  the evaluation of health care (B5, B8).
  the investigation of etiology (B5, B8).
- State why and how standardized rates are used (B13, B15).
- Select an appropriate standard for calculating an indirectly standardized rate (B14).
- State what condition must be met if standardized rates are to be compared (B15).
- Explain the relative advantages of
  odds ratios and rate ratios as measures of association (B11).
  stratification and standardization as ways of detecting and controlling confounding (B15).
  direct and indirect standardization (B15).
- Give a list of
  measures of central tendency (B12).
  measures of dispersion (B12).
- Explain, in general terms, what is meant by
  a birth cohort effect (B2).
  a qualitative study (B4).
  triangulation (Note B4–4).
  selection bias (B4).
  information bias (B4).
  recall bias (B7).
  referral filter bias (Note B7–1).
  volunteer bias (B7).
  lead time (starting time) bias (B10).
  the "healthy worker effect" (B10).
  a confidence interval (B4).
  validity of a measure (B4).
  study validity (B4).
  external validity (B4).
  current life table analysis (Note B9–4).
  Kaplan-Meier life table analysis (Note B9–4).
  average life expectancy at birth (B10).
  random, stratified, cluster, and systematic samples (Note B3–1).
  sampling variation (sampling error) (Note B3–2).

Section C

How Good Are the Measures?

"Oh. I know!" exclaimed Alice, "It's a vegetable. It doesn't look like one, but it is."

"I quite agree with you," said the Duchess; "and the moral of that is—'Be what you would seem to be'—or if you'd like to put it more simply—'Never imagine yourself not to be otherwise than what it might appear to others that what you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise.'"

"I think I should understand that better," Alice said very politely, "if I had it written down."

(Carroll, 1865)

Unit C1

Introduction

Whether the results we wish to use are our own or those reported by others, we have to judge how accurate they are. The main topic of Section C is the validity of the measures used in the study. The more valid these are, the greater is the validity—both internal and external (Unit B4)—of the study as a whole.

We will consider methods of appraising the validity of measures, the ways in which poor validity can produce biased prevalence and incidence rates and erroneous conclusions about associations, and methods of making allowance for this bias. Other topics are reliability, its appraisal and its implications, and regression toward the mean. The series ends with exercises on the validity of screening and diagnostic tests.

Exercise C1

In this exercise you are asked to consider ways of appraising the validity of a measure. We will use a fictional example, to prevent you from being influenced by your prior knowledge about the measure.

TV dementia is an imaginary common disease caused by excessive exposure to television. It is characterized by a long symptom-free period, followed by progressive mental deterioration and culminating in inability to perform activities of daily living unaided. Assume that the diagnosis can be determined with certainty, before or after the development of symptoms, by accurate but costly and elaborate tests.

In a study using a new simple test, imaginatively named test A, the prevalence rate of the disease in a population was found to be 18.4 per 100.

How could you appraise the validity of the test? What kinds of evidence would be helpful? Mention as many possibilities as you can.

■■■■■■■■■■ Unit C2

Validity of a Measure

The *validity of a measure* refers to the degree to which it actually measures what it is designed to measure. The best and most obvious way of appraising validity is to find a criterion (or, in epidemiological jargon, a "gold standard") that we know or believe to be close to the truth, and to compare the results of our measure with this criterion. In this instance (Exercise C1) there is an elaborate but completely accurate diagnostic method that could be used for this purpose. This appraisal of *criterion validity* will tell us test A's sensitivity and specificity (see below).

In the absence of this kind of criterion, it would be helpful to know whether follow-up studies show an association between the results of the test and subsequent events (*predictive validity*). In the present instance, for example, are positive results associated with the subsequent development of complete incapacity? If the measure is to be used as an indicator of change in health status, an association might be sought between the change in its value and an external criterion of change in health, or with the provision of treatment (*responsiveness*).

Another possibility is to see whether there are associations with other variables—age, sex, social class, the amount of time spent watching TV—that there is reason to believe should be linked with the variable under study (*construct validity*—see note C2). These associations provide only weak evidence of validity, but their absence may be strong evidence against validity. Also, associations can be sought with other measures of the variable (*convergent validity*).

These associations—with a criterion, with an outcome, and with other variables or measures—may be examined in the study population itself, or in other samples.

There are other ways of appraising validity, not based on an examination of associations:

- The high or low validity of the measure may seem obvious (*face validity*). If the information is obtained by questioning, we can see whether the questions are clear and unambiguous; and common sense will tell us the likelihood of recall bias or other forms of bias. On the other hand, it may be obvious that the findings don't "make sense." In this instance, is a prevalence rate of 18% acceptable, in terms of what we know about the disease? If we are dealing with blood pressures, is there "zero preference" (an undue proportion of readings

ending in zero)? If so, the readings are obviously inaccurate. Are there very many "unknown" results? If so, the findings cannot tell us the true situation.

- If a set of questions is used, do they cover all the essential components of what they purport to measure (*content validity*)?
- We may also be influenced by the opinions of experts: Is there a consensus concerning the validity of the measure (*consensual validity*)?
- It may also be helpful to know whether the measure gives the same result when it is repeated. This is the *reliability* of the measure. If the results are consistent, they are not necessarily valid; but if they are very inconsistent, they can hardly be valid.

Sensitivity and Specificity

When a test is used to classify individuals as having or not having a specific attribute (say a disease), the *sensitivity* of the measure is the proportion of correct results among people who actually have the attribute, and the *specificity* of the measure is the proportion of correct results among people who are actually free of the attribute. The *false negative rate* is the proportion with incorrect results among people who actually have the disease, and the *false positive rate* is the proportion of incorrect results among people who are free of it.

Using the notation in Tables C2–1 and C2–2, which show the test results in diseased and disease-free people, respectively, the formulae are:

$$\begin{aligned} \text{Sensitivity} &= a/(a + b) \\ \text{False negative rate} &= b/(a + b) \\ \text{Specificity} &= d/(c + d) \\ \text{False positive rate} &= c/(c + d) \end{aligned}$$

These values are generally multiplied by 100 and expressed as percentages.

Exercise C2

Question C2–1

The validity of test A was measured by applying it to 100 patients known to have TV dementia and 400 people known to be free of this disease; there were 80 positive results in the first group, and eight in the second. What are the sensitivity and specificity of the test, and what are the false negative and false positive rates?

Table C2–1. Test Results in a Sample of Diseased People

Test Result	Number
Positive	a
Negative	b
Total	$a + b$

Table C2–2. Test Results in a Sample of Disease-Free People

Test Result	Number
Positive	c
Negative	d
Total	$c + d$

Question C2–2

Is there anything else you would like to know before using these findings?

Question C2–3

If a measure used for determining the prevalence of an attribute has a low sensitivity, how will this affect the prevalence rate?

Question C2–4

If the measure has a low specificity, how will this affect the prevalence rate?

Question C2–5

Can you calculate the prevalence rates that test A will yield in populations (Pepi and Quepi) where the true prevalence rates are 21% and 7%, respectively. If this is too complicated, just guess.

Question C2–6

According to the true prevalence rates in Pepi and Quepi, the rate ratio is 3. If we used the prevalence rates yielded by test A, do you think the rate ratio would be the same, lower, or higher?

Note

C2. *Construct validity:* "The extent to which the measurement corresponds to theoretical concepts (constructs) concerning the phenomenon under study. For example, if, on theoretical grounds, the phenomenon should change with age, a measurement with construct validity would reflect such a change" (Last, 2001).

■■■■■■■■■■ Unit C3

Misclassification

In answer to *Question C2–1*, the sensitivity of test A is 80/100 = 80%. The test's specificity is 392/400 = 98%. The false negative rate is the complement of sensitivity—that is, 100% minus 80%, or 20%—and the false positive rate is the complement of specificity—that is, 2%.

There are at least two things we might want to know before using these results (*Question C2–2*). First, how were the samples for testing validity selected? Many tests are more likely to be positive in full-blown cases of a disease, for example, than in early asymptomatic cases. Was the sensitivity of test A measured in hospital cases of TV dementia? If so, 80% may be an overestimate of its capacity to detect mild cases in the general population. Specificity, on the other hand, may be lower when the test is applied to hospital patients free of the disease under study (because such patients may have other disorders with similar manifestations) than when it is applied to disease-free people in the general population. Second, we might want to know the confidence intervals of the estimates of sensitivity and specificity.

When a measure is used to classify individuals (e.g., as diseased or disease-free), a low validity means that individuals will be misclassified. A low sensitivity (*Question C2–3*) means that people with the disease will be erroneously classified as free of it. This will result in an underestimate of prevalence or incidence. A low specificity, on the other hand (*Question C2–4*), means that there will be individuals who are erroneously classified as having the disease. This will result in an overestimate of prevalence or incidence. In both instances, there is *misclassification bias* (a kind of information bias).

The direction of the bias depends on whether there are more false positive or false negative results. The numbers of these false results are determined both by sensitivity and specificity and by the numbers of diseased and disease-free people in the population. The number of false positives is the false positive rate multiplied by the number free of the disease, and the number of false negatives is the false negative rate multiplied by the number with the disease.

To answer *Question C2–5*, let us construct Tables C3–1 and C3–2, showing the expected results in Pepi and Quepi. (We could also answer this question without constructing tables; see Note C3–1.) We will assume that the population of each locality is 10,000. First we enter the numbers of diseased and disease-free persons in the bottom lines—2,100 diseased people in Pepi, and 700 in Quepi. Then we calculate the expected numbers with positive tests; for example, in Pepi positive results can be expected in 158 (2%) of the 7,900 disease-free people and in 1,680 (80%) of the 2,100 diseased people. We can then easily complete the tables.

Looking at the right-hand columns in the two tables, we find that in Pepi, where the true prevalence rate is 21%, test A may be expected to yield a rate of

Table C3–1. Expected Results of Test A* in Relation to Presence of TV Dementia in Pepi (True Prevalence, 21%)

	Disease		
Test Result	Absent	Present	Total
Positive	158	1,680	1,838
Negative	7,742	420	8,162
Total	7,900	2,100	10,000

*Sensitivity 80%, specificity 98%.

only 1,838/10,000—that is, 18.4%; whereas in Quepi, where the true prevalence rate is 7%, the test will yield a rate of 7.5%.

When the rate of a disease is low (as is generally the case), even a very small rate of false positives can produce enough false positives to outweigh the false negatives, so that surveys that use tests of imperfect validity generally produce overestimates of the true incidence or prevalence rates.

We can use Tables C3–1 and C3–2 to answer *Question C2–6.* Test A may be expected to yield rates of 18.4% and 7.5%, so that the rate ratio will be 18.4/7.5 = 2.5, instead of the correct value of 3.

This is a typical example. When we compare two groups, using a measure whose sensitivity and specificity are the same in both groups, any misclassification that occurs will *always* reduce the difference between the groups (except in very exceptional circumstances, which we may ignore; see note C3–2). If we find a difference, we can therefore be sure that a difference exists, and is actually larger than it seems. The reverse, however, is not true: If we do *not* find a difference we cannot be sure that one does not exist. Misclassification may obscure a true association.

If a measure has the same sensitivity and specificity in both groups—that is, if its validity is *nondifferential*—the consequent misclassification is termed

Table C3–2. Expected Results of Test A* in Relation to Presence of TV Dementia in Quepi (True Prevalence, 7%)

	Disease		
Test Result	Absent	Present	Total
Positive	186	560	746
Negative	9,114	140	9,254
Total	9,300	700	10,000

*Sensitivity 80%, specificity 98%.

nondifferential. In the next exercise we look at *differential misclassification*—the effect of using a measure with a different validity (sensitivity, specificity, or both) in the groups under comparison.

Exercise C3

Question C3–1

Dissatisfied with test A, Dr. B has developed a new test for TV dementia. This test, named test B after its inventor, has a sensitivity of 99% and a specificity of 86%. Test B is now used to measure the prevalence of the disease in Quepi, and the result is compared with the rate (using test A) in Pepi; the latter rate, you will remember, was 18.4%, and the true prevalence rate in Pepi was three times that in Quepi.

Without doing any calculations, can you say whether the ratio of the rate in Pepi (using test A) to the rate in Quepi (using test B) will be more than 3, between 1 and 3, or less than 1?

Question C3–2

If you want to, construct a table (like Table C3–2) to show the expected results when Test B is used in Quepi. You can then supply the rate ratio requested in *Question C3–1.*

Notes

C3–1. The rate of positive test results in a population is the sum of the rates of true positives and false positives. The rate of true positives is the true prevalence rate multiplied by the test's sensitivity. The rate of false positives is the proportion of disease-free persons in the population, multiplied by the false positive rate. In Pepi, for example, the expected rate of positive test results is $(0.21 \cdot 0.80) + (0.79 \cdot 0.02) = 0.1838$.

C3–2. If two groups are compared, using a measure whose sensitivity and specificity are the same in both groups, misclassification will always reduce the difference between the groups, unless the measure is wrong more often than it is right, in which case the direction of the association may be reversed. The specific meaning of being "wrong more often than right" is that the false positive rate plus the false negative rate totals over 100%. Measures whose validity is as low as this are unlikely to be used at all, and this possibility can therefore safely be ignored. See Fleiss (1981), pp. 188–211, for full algebraic explanations of the effects of misclassification.

■■■■■■■■■■ Unit C4

Differential Misclassification

The correct answer to *Question C3–1* is no. It is *not* possible, without doing calculations, to say what the rate ratio will be. If misclassification differs in the groups under comparison—that is, if there is a difference in sensitivity, specificity, or both—*bias in any direction may occur.* A true difference may be artificially lessened, obscured, or increased, or its direction may change; a difference may be seen when really there is none. In the present instance, tests with a different validity were used. Misclassification may also differ when a single test is used, if for any reason its validity differs in the groups under comparison.

We happen to know what the true rate was in Quepi. We can therefore construct Table C4 to show the expected results when Test B is used in Quepi (as requested in *Question C3–2*). According to this table, test B can be expected to yield a prevalence rate of 1,995/10,000, or 19.9%. The ratio of the rate in Pepi (using test A) to the rate in Quepi (using test B) is 18.4/19.9, or 0.92. The disease appears to be more prevalent in Quepi!

Exercise C4

In which of the following studies would you suspect that an observed association might be an artifact (or spuriously strong) because of differential validity?

1. A comparison of the incidence of schizophrenia in two countries, based on the diagnoses recorded in clinical files by psychiatrists.
2. A study of the association of retinal disease with diabetes, based on the clinical records of people with and without diabetes.
3. A study of the efficacy of immunization against a specific disease, based on a comparison of the subsequent incidence of the disease in volunteers who were immunized and in people who were not immunized.
4. A study of the efficacy of a new treatment for painful menstruation, in which the proponents of this treatment questioned patients about the persistence

Table C4. Expected Results of Test B* in Relation to Presence of TV Dementia in Quepi (True Prevalence, 7%)

	Disease		
Test Result	Absent	Present	Total
Positive	1,302	693	1,995
Negative	7,998	7	8,005
Total	9,300	700	10,000

*Sensitivity 99%, specificity 86%.

of their symptoms, after randomly dividing them into two groups—one whose members received the new treatment (without their knowledge) and one whose members continued their usual treatment.

5. A study of the relationship between exposure to anesthetic gases and a specific immunodeficiency disorder, using a test (for the disorder) with a specificity of 100% but a sensitivity of only 60%.
6. A study of the association of senile dementia with educational level, using simple tests of cognitive functioning (general knowledge and intellectual capacity) to measure senile dementia.
7. A study of the association between fever in early pregnancy and congenital anomalies, in which mothers of deformed and normal babies were questioned about the illnesses they had had during their pregnancy.
8. A study of the effect of smoking on physical fitness, in which smokers were compared with people who had given up smoking.
9. A study of the effectiveness of an intensive educational program on hygienic practices, in which school children who had been exposed to the program were asked whether they washed their hands before eating, and their replies were compared with those of similar children who had not been exposed to this program.
10. A study to determine whether rheumatoid arthritis "runs in families," in which patients with this disease and controls who were free of it were asked whether their parents had arthritis.
11. A study of the association between respiratory disease and disease of the locomotor system (bones, joints and muscles), based on an analysis of the diagnoses recorded in hospital patients.
12. A study of international variations in the prevalence of gallstones, based on the crude findings of all autopsy studies published since 1890 (Brett and Barker, 1976).

■■■■■■■■■■ Unit C5

Effects of Misclassification

Spurious associations, or spuriously strong ones, could arise in all the studies listed in Exercise C4, except in (5), where the only problem is low sensitivity (nondifferential), which would reduce, and could not increase, the strength of any association. In studies (3), (8), and (11), and maybe in (12), however, the problem is not misclassification. In (3), there may be *volunteer bias:* volunteers may differ in many respects from other people, and these differences may be reflected in a different risk of contracting a given disease. In (8), people who give up smoking may differ from continuing smokers in many other ways—for ex-

ample, in their physical activity—and the effects of these differences may be confounded with the effects of ceasing to smoke. Study (11) provides an example of possible *Berksonian bias*—that is, bias due to selective admission to a study sample. Not all people with respiratory disease, nor all people with locomotor disease are hospitalized; however, people who have both types of disease may be especially likely to be hospitalized. Associations found in a highly selected sample, like hospital patients, may not exist in the general population. In this instance, a study in Ontario demonstrated that the rate of locomotor disease was 25.0% in hospital patients with respiratory disease and 7.6% in hospital patients without respiratory disease—giving a rate ratio of 3.3. There was no such association in the general population, where the corresponding rates were 7.6% and 7.2, with a rate ratio of 1.1 (Roberts et al., 1978). In (12), we cannot be sure that the methods of determining the presence of gallstones were uniform in all studies; but more obvious reasons for possible spurious differences in prevalence are selection bias (differences in the criteria for doing autopsies) and confounding (age differences).

In studies (1), (2), and (4) there is a possibility of differential validity because of the differences in the methods of measurement or the way they were used. In (1), it is very likely that different diagnostic criteria and techniques are used by psychiatrists in different countries, and these may produce apparent differences in the incidence of schizophrenia. The probability that a person with schizophrenia will receive psychiatric care and be blessed with a psychiatrist's diagnosis also varies from country to country. In (2), diabetics are probably more likely to have retinal examinations than other patients, because of the known hazard of diabetic retinopathy. In a study using clinical records, more retinal disease may therefore be missed in nondiabetics than in diabetics. In (4), there is a possibility that the findings may reflect the unconscious bias of the clinicians, who were proponents of the new treatment and knew which patients had which treatment. The questions they asked, the way they asked them, or the way they interpreted the responses may have differed in the two groups. This possibility of differential validity would not have existed if the appraisal of outcome had been "blind."

In (6), (7), (9), and (10), uniform methods of measurement were used, but their validity may have differed in the groups that were compared. In (6), the validity of the tests of cognitive functioning may well vary with educational status: for example, a low score may be due to lack of education rather than senile dementia. In (7), it is possible that mothers of deformed infants may, because of their concern or feelings of guilt, be especially likely to recall and report minor illnesses that occurred during early pregnancy. In (9), we may suspect that children who have been exposed to intensive brainwashing will tend to give the responses about hand-washing that they think are expected of them. And in (10), we may suspect that people who have a given disease will be especially likely to recall and report the occurrence of the same disease in their family members. In fact, in a study in which people with rheumatoid arthritis were questioned,

only 27% reported that their parents were free of arthritis. But when their unaffected siblings were questioned, 50% reported that the same parents were free of arthritis (Schull and Cobb, 1969).

The findings of a study can be taken at their face value only if the study methods are satisfactory. An appraisal of the validity of the measures and the possible effects of misclassification should never be overlooked. If we know what these effects may be, we can avoid unwarranted conclusions, and may be able to gauge the true situation by making allowance for the bias. Formulae are available for estimating the true situation from the observed findings, for both nondifferential (Note C5–1) and differential misclassification (Note C5–2).

Exercise C5

Question C5–1

In a study of the possible relationship of herpes to cancer of the lip, men with cancer of the lip and men with skin cancer elsewhere on the face (controls) were asked about the past occurrence of recurrent blisters on the lips or face. The results (Table C5–1) showed a positive association, with an odds ratio of 2.5 (Lindquist, 1979). Assume that men with lip cancer were more likely to remember and report their blisters. Without doing any calculations, can you say whether the observed association was stronger than the true one?

Question C5–2

A cohort study assessed the prognostic value of exercise electrocardiographic (ECG) testing in people with no symptoms of coronary disease. The subsequent incidence of coronary events (angina pectoris, myocardial infarction or sudden death) in individuals who initially had abnormal ECG findings was compared with the incidence of these events in those who initially had normal ECG findings (Giagnoni et al., 1983). The results (Table C5–2) showed a positive association, with a rate ratio of 4.5. However, there may have been bias, since the study was not "blind," and the physicians who made the appraisals may have had a greater tendency to diagnose coronary events in people whose previous exercise ECG was abnormal. Assume that this actually happened. Without any calculations, can you say whether the observed association was stronger than the true one?

Table C5–1. History of Herpetic Blisters in Patients With Lip Cancer and Controls

Herpetic Blisters	Cases	Controls
Yes	60	12
No	76	38

Table C5–2. Occurrence of Coronary Events in People With and Without Abnormal ECGs

	Exercise ECG	
Subsequent Coronary Event	Abnormal	Normal
Present	21	13
Absent	114	366

Notes

C5–1. The following formulae can be used to estimate the true situation if there is nondifferential misclassification with respect to one variable, and none with respect to the other. In a cohort study the true absolute difference between rates is the apparent difference (revealed by the survey) divided by (Se + Sp − 1), where Se and Sp are the sensitivity and specificity, expressed as decimal fractions (Fleiss, 1981). In the comparison of Pepi and Quepi (test A data, Tables C3–1 and C3–2), this formula gives a true difference of (18.38% − 7.46%)/(0.8 + 0.98 − 1), or 14%; the actual rates were 21% and 7%. If the disease has a low frequency, the true risk ratio can be estimated from the observed risk ratio R, provided that a definitive evaluation can be performed of unexposed people classified as diseased, to determine the proportion C of this group who are truly diseased. The true risk ratio is then approximately $(R + C - 1)/C$ (Green, 1983). In a case-control comparison where exposure to the factor under study has a low prevalence, the true odds ratio can be similarly estimated from the observed odds ratio (OR) by the formula $(OR + B - 1)/B$, where B is the proportion of controls classified as exposed who are truly exposed (Kelsey et al., 1986). The algebra of misclassification bias is described by Fleiss (1981, pp. 188–211) and Kleinbaum et al. (1982, chap. 12).

C5–2. The following formulae may be used if there is differential misclassification of one variable (Fleiss, 1981; Kleinbaum et al., 1982). If we use the symbols in Table B11 for the observed findings (after misclassification), the true number of exposed cases (in a case-control study) is $[a - (a + c)(1 - Se_X)]/(Sp_X + SE_X - 1)$, where Sp_X and Se_X are the specificity and sensitivity (with respect to the measure of exposure) in the cases, expressed as decimal fractions. To obtain the unexposed cases, subtract this number from $(a + c)$. The number of exposed controls is $[b - (b + d)(1 - Sp_Y)]/(Sp_Y + Se_Y - 1)$, where Sp_Y and Se_Y are specificity and sensitivity in the controls. Subtract this from $(b + d)$ to obtain the unexposed controls. In a cohort study the true number with the disease in the exposed group is $[a - (a + b)(1 - Sp_E)]/(Sp_E + Se_E - 1)$, where Sp_E and Se_E are the specificity and sensitivity (for detecting the disease) in those exposed; the true number with the disease in the unexposed group is $[c - (c + d)(1 - Sp_U)]/(Sp_U + Se_U - 1)$, where Sp_U and Se_U are the specificity and sensitivity in those unexposed to the factor under study.

■■■■■■■■■■Unit C6

Effects of Misclassification (Continued)

Differential validity can produce spurious associations, spuriously strong ones, or any other kind of distortion. But the correct answer to *Questions C5–1* and *C5–2* is no; it is not possible to guess the effect of differential misclassification. It is possible, however to calculate the true values from the observed results if assumptions are made about the sensitivities and specificities. This computation is easy if there is differential misclassification of only one variable (Note C5–2).

To see how the study described in *Question C5–1* might have been affected by misclassification, Sosenko and Gardner (1987) made the assumptions that sensitivity (with respect to prior herpes) was 98% in cases and 92% in controls, and that specificity was 95% in cases and 98% in controls—that is, that the cases had higher rates of both true and false positive responses. Using the first two formulae in Note C5–2, they calculated that the true odds ratio (OR) would then be 2.28—only very slightly less than the observed value of 2.50.

But when they made similar assumptions for the study described in *Question C5–2*, the results were different. They postulated that sensitivity (with respect to coronary events) was 98% in those with abnormal ECGs and 92% in those without, and that the respective specificities were 95% and 98%—that is, that people with prior ECG abnormalities had higher rates of both true and false positive diagnoses of coronary events. Under these conditions, the calculated true rate ratio was 7.0—higher than the observed value of 4.5. The direction of the bias is the opposite of what we might have expected, showing that one cannot guess the effect of differential misclassification. The bias depends on the balance between false positives and false negatives, which is not determined solely by sensitivity and specificity (as we saw in Unit C3).

In both these instances, simple computations demonstrated that (under the stated assumptions) the observed associations were not artifacts caused by differential misclassification. (If you are a martyr for punishment, check the calculations: apply the formulae in Note C5–2 to the data in Tables C5–1 and C5–2; to get the same answers, round off your results.)

When there is misclassification of both the independent and dependent variables, the kind of bias depends on whether the misclassification is differential or not (in the same way as when only one variable is misclassified). If there is no differential misclassification, a true association may be underestimated or obscured, but will not be increased or reversed. However, if there is differential misclassification of one variable or both, bias in *any* direction may be produced. Calculations to determine the true situation are complex if there is misclassification of both variables (see Note A3–7).

Exercise C6

Sensitivity and specificity can be used to gauge validity only in dichotomous (two-category) situations, where we have "yes–no" measures of "yes–no" enti-

ties (e.g., disease or no disease), and where a "gold standard" is available. This exercise presents other situations. Methods of appraising validity were reviewed in Unit C2.

Question C6–1

It is proposed to use ten questions about dyspeptic symptoms (belching, burning, nausea, pain, etc.) as a screening test for peptic ulcer, and to test their validity by a comparison with radiological findings. How could specificity and sensitivity be used as measures of validity? If validity is high, can the questions be used to study ethnic differences in the occurrence of peptic ulcer?

Question C6–2

In a survey of a population sample in Auckland, New Zealand, participants were asked their height and weight. People with a Quetelet's body mass index (weight in kilograms divided by the square of height in meters) of ≥ 30 were defined as obese (Stewart et al., 1987). How would you measure the validity of the self-reported measurements and the diagnosis of obesity, using actual measurements as the criteria?

Question C6–3

An epidemiological study of mental health in an Australian university was performed by asking students whether they had experienced any emotional or mental illness during the last year, and if so, whether it was serious, moderate, or minor (McMichael and Hetzel, 1974). How could these self-appraisals be validated?

Question C6–4

One of the variables measured in the Rand Health Insurance Study (a large-scale experiment designed to investigate the effects of different arrangements for financing health care) was "physical health in terms of functioning." A battery of questions about functional limitations was used ("Do you have trouble walking?" "Does your health keep you from working?" "Do you need help with dressing?" etc.). Each response was given a score, and the sum of the scores was used as a measure of physical health (Stewart et al., 1978). How could this measure be validated?

■■■■■■■■■■ Unit C7

Other Ways of Appraising Validity

To appraise the validity of the questions about indigestion (*Question C6–1*), sensitivity and specificity in relation to radiological evidence of peptic ulcer were measured for specific questions, for specific combinations of questions, and for the total number of symptoms reported. For the latter purpose, the range of responses was turned into a dichotomy, using alternative cutting-points: 3 or more, 4 or more, and so forth. Validity was best for a total score of 6 or more; sensitivity was then 80% and specificity 84% (Popiela et al., 1976). However high the validity of such questions, it would be unwise to use them to study ethnic differences, without first measuring their validity in different ethnic groups. Marked ethnic variation has been found in the validity of this kind of question (Epstein, 1969).

Sensitivity and specificity cannot be used for metric-scale variables like weight and height. (What is a metric scale? What kinds of scale of measurement do you know? See Note C7.) The criterion validity of measures of these variables (*Question C6–2*) can be appraised by comparing the findings with "true" ("gold standard") measurements, and using such indices as

1. the correlation between the observed and true measurements. (A correlation coefficient of 1 indicates perfect linear correlation; that is, a higher observed value always means a higher true value.)
2. the size of the discrepancies between the observed and true values (ignoring the direction of the differences), as an indication of the "precision" of the measurements.
3. the difference between the mean values, as an indication of the presence and direction of bias.

In this instance, the comparison showed that self-reported heights and weights had a high degree of accuracy in the population studied (Stewart et al., 1987). The coefficients of correlation between reported and measured values were .96 for height and .98 for weight. For 75% of participants the absolute discrepancy in height (i.e., ignoring its direction) did not exceed 3.5 cm and the discrepancy in weight did not exceed 2.4 kg. There was slight bias: the reported height tended to be more than the measured height (mean difference, 1.94 cm; 99% confidence interval, 1.78–2.10 cm), and the reported weight was lower than the measured weight (mean difference, 0.58 kg; 99% confidence interval, 0.41–0.75 kg).

The small biases in height and weight acted together to produce a larger bias in the diagnosis of obesity. The prevalence of obesity was 6.2% according to the reported measurements, and 9.3% according to the measured values. The sensitivity of the report-based diagnosis of obesity was 63%, and its specificity was 99.6%.

The self-assessments of mental illness used in the Australian study (*Question C6–3*) were validated in several ways (McMichael and Hetzel, 1974); you may have thought of other possibilities. Criterion validity was tested by a comparison with clinical records; among members of the study sample diagnosed as having an emotional illness during the previous year, the sensitivity of the self-assessment was 73%; the few students who were diagnosed as seriously ill all reported illness. Construct validity was demonstrated by correlations between the self-assessment and attributes that might be expected to go along with mental illness—namely a neuroticism score (the more serious the reported illness, the higher the score) and self-reported psychosomatic disorders. There was no correlation with the student's reported readiness to seek medical help when ill, a fact taken as evidence that the self-assessment of mental illness indicated the occurrence of illness rather than readiness to be labeled "ill." Also, 79% of students who reported mental illness one year reported it again the next year; and the more serious the illness reported the first year, the higher this proportion was. The authors regarded this as predictive validation.

It is not easy to find a "gold standard" for validating the questions used to measure physical health (*Question C6–4*). The investigators satisfied themselves that the questions had face validity (each question appeared to measure what it was supposed to) and content validity (the questions covered the areas included in measures of physical health found in the literature). Construct validity was appraised by seeking (and finding) the expected associations between the score and other questionnaire measures of functioning (physical abilities, role limitations, self-care limitation, performance of physical exercise, etc.), age, and income (Stewart et al., 1978).

The investigators also appraised the extent to which the separate questions "hung together"—how strongly the answers were correlated with each other and with the total score. This kind of *internal consistency* (also called *internal consistency-reliability*) is evidence that the items probably measure much the same thing. Alone, it is no guarantee of validity. But if face and content validity are satisfactory, internal consistency supports the probability that the measure is valid. In this instance, "coefficient alpha" (a measure of internal consistency you are very likely to encounter; possible values, 0–1) was .9; a value of $\geq .7$ is generally regarded as satisfactory.

Reliability

Reliability is defined as

> the degree of stability exhibited when a measurement is repeated under identical conditions. Reliability refers to the degree to which the results obtained by a measurement procedure can be replicated. Lack of reliability may arise from divergences between observers or instruments of measurement or instability of the attribute being measured. (Last, 2001)

Reliability is also called *reproducibility or repeatability.*

Reliability is no guarantee of validity: people of a certain age may give the same answer whenever they are asked how old they are, even over a period of years, but this may not be their true age. On the other hand, if a measure is unreliable this must detract from its validity. Especially in instances where criterion validity cannot be measured, it may therefore be useful to know how reliable the measure is.

Reliability is usually measured by performing two or more independent measurements and comparing the findings. The object may be to determine whether observers vary in their measurements (interobserver or interrater variation), whether differences exist between measurements made by the same observer at different times (intraobserver or intrarater variation), whether measuring instruments differ, or whether the attribute that is measured is itself labile.

Exercise C7

Cataract may be difficult to diagnose, especially in its early stages. A handbook on epidemiology for ophthalmologists states, "One observer may be more apt to diagnose cataracts . . . than another. One man's . . . cataract is not always another's" (Sommer, 1980).

In an imaginary study of the reliability of diagnoses, two ophthalmologists each examined the same 1,000 eyes, without knowing the other ophthalmologist's diagnoses.

Question C7–1

Suppose you are told that each ophthalmologist found 100 eyes with cataract. Does this mean that the diagnoses are reliable? Is there bias?

Question C7–2

Suppose you are told that the *percentage agreement* was 83%—that is, the ophthalmologists agreed with respect to 83% of the eyes they examined. Is this an adequate degree of reliability?

Question C7–3

You are now given the findings shown in Table C7–1. Is the reliability of the diagnoses satisfactory? (Can you see how the percentage agreement of 83% was calculated?)

Question C7–4

The full findings are shown in Table C7–2. Were the diagnoses more reliable for early or for advanced cataract?

Table C7–1. Presence of Cataract in 1,000 Eyes, According to Two Ophthalmologists

	Dr. Mackay		
Dr. McBee	Absent	Present	Total
Absent	815	85	900
Present	85	15	100
Total	900	100	1,000

Question C7–5

Using the data in Table C7–1, can you calculate the sensitivity and specificity of the diagnoses?

Note

C7. *Scales of measurement.* A *dichotomy* has two mutually exclusive categories (e.g., disease present, disease absent). A *nominal scale* has any number of mutually exclusive categories that do not fall into a natural order (e.g., Easterners, Westerners, Northerners). An *ordinal scale* has mutually exclusive categories that represent relative positions between which a natural order is assumed (e.g., social classes 1, 2, 3, 4, and 5; or no disease and mild, moderate, and severe disease). An *interval scale* is one in which any given difference between two numerical values has the same meaning, whatever the level of the values; the difference between the values reflects the magnitude of the difference in the attribute (e.g., age). The term *ratio scale* is sometimes used for interval scales whose zero values mean absence of the attribute (most interval scales used in epidemiology are ratio scales). Interval and ratio scales may be referred to as *metric.* These scales are *continuous* if an infinite number of values are possible along a continuum—for example, in measurements of height. They are *discrete* if only certain values are possible; for example, a woman's parity cannot be 2.3.

Table C7–2. Presence and Stage of Cataract in 1,000 Eyes, According to Two Ophthalmologists

	Dr. Mackay			
Dr. McBee	Absent	Early Cataract	Advanced Cataract	Total
Absent	815	85	0	900
Early cataract	85	9	1	95
Advanced cataract	0	0	5	5
Total	900	94	6	1,000

■■■■■■■■■■Unit C8

Appraisal of Reliability

The fact that the ophthalmologists detected the same numbers of cases of cataract (*Question C7–1*) does not ensure reliability, because they may not have decided that the same eyes had cataracts. Reliability may be very low. The fact that both ophthalmologists diagnosed the same number of cases does not necessarily mean there is no bias; they may have an equal tendency to overdiagnose or underdiagnose cataract.

The percentage agreement (*Questions C7–2* and *C7–3*) is 83%; this is because there were 830 agreements in 1,000 eyes (815, no cataract; 15, cataract). This high percentage suggests a high degree of reliability. However, this is misleading: as Table C7–1 shows, the ophthalmologists agreed on the presence of cataract in only 15 eyes, but in 170 others one said there was cataract and the other said there was not.

The percentage agreement is a widely used but obviously unsatisfactory measure of reliability. It does not allow for the fact that chance alone will lead to a large number of agreements; this is illustrated in hypothetical Table C8–1, where there is no association whatsoever between the diagnoses made by two physicians: Dr. Maxcy diagnoses trachoma in 10% of the eyes Dr. MacDee finds diseased, and in 10% of those Dr. MacDee finds free of trachoma. Yet the percentage agreement is 82%!

A better measure is *kappa* (Note C8–1), which is a measure of agreement "beyond chance." To calculate this for Table C7–1, we first estimate the number of agreements to be expected by chance, on the basis of the totals in the right-hand column and bottom row (the "marginal totals") of Table C7–1. Dr. Mackay found trachoma in 100/1,000 (10%) of the eyes he examined, and if the diagnoses were unrelated, he could therefore be expected to find trachoma in 10% of the 100 cases found by Dr. McBee, so that there would be ten agreements on a positive diagnosis. Similarly, Dr. Mackay reached a negative diagnosis in 900/1,000 (90%) of the eyes he examined, so that if the diagnoses were unrelated he could be expected to make a negative diagnosis in 90%, or 810, of the 900 eyes given negative diagnoses by Dr. McBee. In all, 820 agreements might be expected by

Table C8–1. Presence of Trachoma According to Two Physicians (No Association)

	Dr. Maxcy		
Dr. MacDee	Absent	Present	Total
Absent	810	90	900
Present	90	10	100
Total	900	100	1,000

chance (as in Table C8–1). We then subtract these chance agreements from the observed agreements (830), leaving ten agreements beyond chance. We also subtract the chance agreements (820) from the total number of comparisons (1,000), leaving 180 potential agreements beyond chance. Kappa is then 10/180 = 5.6%; that is, if chance agreements are excluded, the two eye doctors agreed in only 5.6% of instances. In Table C8–1, kappa is 0%.

A kappa value of 75% or more may be taken to represent excellent agreement, and values of 40–74% indicate fair to good agreement. Below 40% indicates poor agreement.

Agreement was closer for advanced than for early cataract (Question C7–4): Table C7–2 shows only one disagreement about the presence of advanced cataract. Kappa can be calculated for this diagnosis only, or for overall agreement (concerning both the presence and the stage of the disease). If you wish, calculate these kappas (solutions in Note C8–2).

In answer to Question C7–5, sensitivity and specificity of course cannot be calculated from the data in Table C7–1. We cannot regard either physician as providing us with the "true facts," for use as a criterion in appraising the other physician's diagnoses.

Exercise C8

Question C8–1

A medical group in New York City provided a screening program, including chest x-rays, for construction workers who were exposed to asbestos. The x-rays were read by staff radiologists. In addition, separate arrangements were made for the x-rays to be read by specialists in occupational medicine. Table C8–2 presents a comparison of the x-ray interpretations by staff radiologists and specialist readers with respect to the presence of signs typical of asbestosis (Zoloth et al., 1986). The value of kappa is .27. What conclusions can you draw about validity? Can you measure sensitivity and specificity?

Table C8–2. Presence of Typical Signs of Asbestosis* in 775 X-rays, According to Staff Radiologists and Specialist Readers

	Staff Radiologists		
Expert Reader	Absent	Present	Total
Absent	660	39	699
Present	54	22	76
Total	714	61	775

*Small opacities (grade 1/0 or higher on the International Labor Organization scale) or comments indicative of interstitial marking.

Question C8–2

What is the prevalence rate, in these workers, of x-ray signs typical of asbestosis?

Question C8–3

There have been many studies of concordance with respect to the presence of various clinical signs and symptoms and electrocardiographic, radiographic, and other findings, based on comparison between examiners or between repeated examinations by the same observer. How high do you think kappa generally is in these studies?

Question C8–4

Suppose that a comparison of repeated examinations yielded a kappa of .95. What would you conclude about the validity of the measure?

Question C8–5

Suppose that replicate examinations are not feasible; and instead, interobserver variation is studied by comparing the findings of two physicians who examine different groups of patients. What condition or conditions must be met to make such a study of reliability satisfactory?

Question C8–6

The blood pressures of residents of nine homes for the elderly in Nottinghamshire, England, were examined, and people with diastolic pressures of ≥ 100 mm Hg were randomly divided into two groups, one of which received medication for hypertension, while the other did not. Six months later, the mean diastolic pressure in the control group had decreased by 6.5 mm Hg (Sprackling et al., 1981). How can this change in an untreated group be explained?

Notes

C8–1. The computation of kappa is explained by (*inter alia*) Altman (1991, pp. 404–408) and Fleiss (1981, chap. 2). Kappa can be used not only for dichotomies, but also for multiple categories (nominal or ordinal), and for multiple ratings. A word of warning: kappa may be misleading if the marginal totals in a table like Table C8–2 show a marked discrepancy between the numbers in the two categories, or if the marginal totals in the two sets of ratings are very different (Byrt et al., 1993). The value of kappa can be adjusted to counter these problems. (See Note A3–7.)

C8–2. According to Table C7–2, the expected number of chance agreements is (5/1,000) × 6 = 0.03 for advanced cataract and the number is (995/1,000) × 994 = 989.03 for the absence of advanced cataract. Total chance agreements = 0.03 + 989.03 = 989.06. Observed agreements = 5 (advanced cataract

present) plus 815 + 85 + 85 + 9 = 994 (advanced cataract absent); total, 999. Kappa for diagnosis of advanced cataract = (999 − 989.06)/(1,000 − 989.06) = 91%. Kappa for overall agreement is calculated after subtracting [(900/1,000) × 900 + (95/1,000) × 94 + 5/1,000 × 6] from both the numerator (815 + 9 + 5) and the denominator (1,000); its value is 5.6%.

■■■■■■■■■■ Unit C9

Appraisal of Reliability (Continued)

Validity cannot be high if reliability is low. The very low concordance between the two sets of x-ray interpretations (*Question C8–1*) points to the low validity of one or the other or both of the sets of readings. The specialists were more familiar with occupational diseases, and it is probably right to assume that their readings were more valid (face validity). If we take their results as a "gold standard," we can calculate the sensitivity and specificity of the staff radiologists' readings (sensitivity = 22/76 = 29%; specificity = 660/699 = 94%).

In the face of this low concordance, we cannot be sure of the prevalence rate of x-ray signs of asbestosis (*Question C8–2*). A tempting solution is to accept the specialist readers' interpretations—which is what Zoloth et al. (1986) did. The rate is then 76/775 = 9.8 per 100. But there are other possibilities: we can insist on a positive finding by both readers (in which case the rate is 22/775 = 2.8%), or we can be less strict and accept a positive finding by either reader (in which case the rate is 115/775 = 14.8%). If we wanted to compare the prevalence in this group with the rate in other workers, based on readings by other radiologists, we would have a problem.

In answer to *Question C8–3*, most comparisons of clinical examinations, as well as interpretations of x-rays, ECGs, and microscopic specimens yield kappa values in the 40–74% range ("fair to good" agreement).

A high kappa value (*Question C8–4*) means high reliability, but alone it tells us nothing about validity. The findings may be consistent without measuring what they purport to measure.

A reliability study based on a comparison of two physicians' findings in separate groups of patients (*Question C8–5*) can be satisfactory only if there is no selection bias: the two groups must be similar. The allocation of subjects should preferably be random, so that the only differences to be expected are those occurring by chance. If the purpose was to study interphysician reliability with respect to a specific examination procedure, it would be important to know whether they had agreed to use a standard procedure and had in fact adhered to it.

The above exercises have focused on the reliability of categorical measures

(e.g., "absent" or "present"). We will not deal with the reliability of metric measures (see Note C7); for example, blood pressure measurements. This requires use of a variety of statistical indices (Note C9), different ones being appropriate in different circumstances.

Regression Toward the Mean

Whenever there is a "random" element in measurements—whether this is because the characteristic is unstable or its measurement is unreliable—a repeated measurement in the same subject will tend to give a lower value if the initial value was high, and a higher value if the initial value was low. This is called "regression toward the mean." Whatever other suggestions you may have offered for the decrease in the mean blood pressure of untreated people with high blood pressures (*Question C8–6*), you should not have omitted this possible explanation.

This phenomenon may mimic the result of treatment and sometimes presents a problem when one is interpreting the results of trials of therapies and health programs. It may be countered by a comparison with the change seen in an appropriate control group (as in the study cited), or by statistical procedures that measure or compensate for regression to the mean. Sometimes one measurement is used to select the subjects for a trial or follow-up study, and a subsequent one is used as the baseline for measuring change.

Taking Account of Validity and Reliability

A short recap may be useful at this stage, putting what we have done into the framework of the basic procedure for appraising data (as outlined in Unit A16). When we want to interpret data, what do we do about validity and reliability?

First, we should always ensure that we know how the variables were measured. This is part of the process of "determining what the facts are"—the initial step in the basic procedure for appraising data. We can then appraise the face validity of the measures. Before or after inspecting the data, we should review any available evidence of criterion validity (sensitivity and specificity or, for metric-scale variables, correlation coefficients, mean discrepancies from criterion values, etc.). In studies where we are interested in associations, it is important to know whether validity is differential. If evidence of criterion validity is lacking, we should review evidence of predictive, construct, and content validity. Information about reliability and internal consistency–reliability may be important if clear evidence of validity is lacking, or for other reasons, as when regression toward the mean is suspected.

With this information, we can consider the role of validity and reliability when we seek explanations for the findings; specifically, we can give thought to the possibility that rates, means, or other summary statistics may be biased, or that the presence, absence, or strength of observed associations may be artifacts. Consideration of possible explanations may lead us to seek additional information about how the data were obtained and the accuracy of the methods.

We may be able to infer the direction and degree of bias in prevalence or incidence rates, mean values, or other summary measures. If we are interested in associations between variables, we can appraise the possibility that the association is spurious, or spuriously strong or weak; the effects of misclassification are most easily estimated if validity is nondifferential.

In some instances, it may be possible to compensate for the effects of low validity or reliability by appropriate statistical manipulations. In others, the best we can do is to allow for these effects when drawing conclusions from the findings, and to consider them when deciding whether, what, and how additional information should be collected.

Screening Tests

The purpose of a *screening test* is to identify individuals or groups who have a high probability of having a particular disease or other attribute.

Screening was defined in 1951 by the U.S. Commission on Chronic Illness as, "The presumptive identification of unrecognized disease or defect by the application of tests, examinations or other procedures which can be applied rapidly. Screening tests sort out apparently well persons who probably have a disease from those who probably do not. A screening test is not intended to be diagnostic." (Last, 2001)

The next two exercises deal with the validity of screening tests and the appraisal of their results.

Sensitivity and specificity are the main measures of the validity of a screening test.

Exercise C9

Question C9–1

You will remember that we have two tests for the detection of TV dementia—test A (sensitivity 80%, specificity 98%) and test B (sensitivity 99%, specificity 86%). Which would be a better screening test, and why?

Question C9–2

What other information (besides sensitivity and specificity) would be helpful in appraising the value of a screening test?

Note

C9. Indices of the *reliability of metric-scale measurements*, based on duplicate observations, include the intraclass correlation coefficient, the concordance correlation coefficient, 95% limits of agreement, the standard error of measurement, the components of variation according to one-way analysis of variance, regression coefficients, and the mean, frequency distribution, and quan-

tiles of discrepancies. See, for example, Bartko (1994), Lin (1989), and Shoukri (2000) and statistics textbooks—for example, Shoukri and Pause (1998, chap. 2). (See Note A3–7.)

■■■■■■■■■■Unit C10

Appraisal of a Screening Test

The aim of population screening is usually to detect as many cases as possible. Test B can be expected to identify 99% of cases, and test A only 80%. In answer to *Question C9–1,* Test B seems therefore to be a more useful screening test. But we cannot ignore its lower specificity. People with positive results will presumably be submitted to definitive diagnostic examinations, and if test B is used there will be a great deal of unnecessary expense, anxiety, and inconvenience. This may or may not be an important consideration. The cost of diagnostic tests and the availability of the personnel and other resources they require cannot be ignored.

If the purpose of screening is not to detect as many cases as possible, but merely to detect *some* cases—for example, to find subjects for a clinical trial to compare two treatments—test A may be an appropriate one.

A number of other measures may be helpful in appraising the value of a screening test (*Question C9–2*). The *predictive value of a positive result* is probably the most useful. This is the proportion with the disease (or other attribute) among people with a positive test result. It measures the probability that a person with a positive result has the disease, and gives an indication of what cost and effort the screening program will require. Other indices of this effort are the number of positive tests per case identified (which is also the number of definitive diagnostic examinations required per case identified), and the total number of screening tests per case identified. Multiplied by the average costs of the respective investigations, these figures provide an index of the average cost of finding a case. The *predictive value of a negative test,* which is the proportion free of the disease among people with a negative test result, is another measure of validity.

In your answer to *Question C9–2,* you may rightly have listed additional criteria of the value of a screening test. These include the extent to which there is a need for the test (taking account of the prevalence of undiagnosed cases, the impact of the condition, and the probability that detection will lead to effective action and a substantial impact on health), the side effects of the test (including anxiety caused by false positive results), practicability, acceptability, and the cost both of the test and of the more elaborate diagnostic examinations that are required if the result is positive.

Table C10–1. Results of Test A* in Relation to Presence of TV Dementia in Pepi (Prevalence, 21%)

	Disease		
Test Result	Absent	Present	Total
Positive	158	1,680	1,838
Negative	7,742	420	8,162
Total	7,900	2,100	10,000

*Sensitivity 80%, specificity 98%.

Exercise C10

Question C10–1

Table C10–1 (a copy of Table C3–1) shows the results of test A in Pepi. Use these data to calculate the predictive value of a positive test, the predictive value of a negative test, the number of positive tests per case identified, and the total number of tests per case identified.

Question C10–2

Now again calculate these indices for test A, this time using the results in Outer Shepi, where TV transmissions were only recently introduced, and the prevalence of TV dementia is only 1%, not 21% as in Pepi. To do this you may first need to construct a table like Table C10–1, based on your knowledge that the prevalence rate is 1%, the sensitivity is 80%, and the specificity is 98%. (If you have any difficulty, see note C10, which also provides formulae for the calculation of predictive values.) Compare the results and explain the findings.

Note

C10. Each 10,000 people in Outer Shepi include 100 (1%) with TV dementia. When test A is used, 80 (80%) of these have positive and 20 (20%) have negative results. There are 9,900 people without TV dementia, of whom 9.702 (98%) have negative and 198 have positive results. If you wish to use formulae, the predictive value of a positive test is $SeP/[SeP + (1 - Sp)(1 - P)]$ and the predictive value of a negative test is $Sp(1 - P)/[(1 - Se)P + Sp(1 - P)]$, where Se = sensitivity, Sp = specificity, and P prevalence (pretest probability) of the disease (all expressed as proportions). As will be seen in Unit C11, the predictive value of a positive test can also be calculated from the likelihood ratio.

■■■■■■■■■■Unit C11

Appraisal of a Screening Test (Continued)

In answer to *Question C10–1,* the predictive value of a positive test in Pepi was 1,680/1,838 or 91%. The predictive value of a negative test was 7,742/8,162, or 95%. The number of positive tests per case identified (which is the reciprocal of the predictive value of a positive test) was 1,838/1,680, or 1.1; and the total number of tests per case identified was 10,000/1,680, or 6.0.

The sensitivity and specificity of the test were the same in Outer Shepi (*Question C10–2*) as in Pepi. But the other indices differed, as shown by the figures in Table C11–1 (based on a prevalence rate of 1%). The predictive value of a positive test was only 80/278, or 29%. The predictive value of a negative test was 9,702/9,722, or 99.8%. The number of positive tests per case identified was 278/80, or 3.5, and the total number of tests per case identified was 10,000/80, or 125.

Clearly, these indices are determined not only by sensitivity and specificity, but also by the prevalence of the disease or attribute in the population in which the test is used: the lower the prevalence, the lower the predictive value of a positive test will be. To estimate these indices, we must know—or guess–the prevalence rate (see the formulae stated in Note C10).

The value of a screening test can be judged only by considering the results to be expected in the population in which it will be used.

Exercise C11

Question C11–1

For what purposes would a diagnostic test with a high sensitivity be useful, even if its specificity is low?

Question C11–2

For what purposes would a diagnostic test with a high specificity be useful, even if its sensitivity is low?

Table C11–1. Results of Test A* in Relation to Presence of TV Dementia in Outer Shepi (Prevalence, 1%)

	Disease		
Test Result	Absent	Present	Total
Positive	198	80	278
Negative	9,702	20	9,722
Total	9,900	100	10,000

*Sensitivity 80%, specificity 98%.

Table C11–2. Probability of Positive and Negative Results Among People With and Without TV Dementia, When Test A Is Used in Pepi

	Disease		
Result	Present	Absent	Likelihood Ratio*
Positive	0.80	0.02	40
Negative	0.20	0.98	0.204
Total	1.00	1.00	

*The ratio of the probability of the given result among people with the disease to the corresponding probability among people free of the disease.

Question C11–3

Go back to Table C10–1, which shows the results of test A in Pepi. On the basis of the prevalence rate, what is the probability that a member of this population (who has not yet been tested) has TV dementia? (This is called the *pretest probability.*) What are the odds in favor of the disease (the *pretest odds*)? If we now do test A and it turns out to be positive, what is the probability that the subject has the disease? If the test is negative, what is the probability that the disease is present? (These are the *posttest probabilities* of the disease.) What are the corresponding odds? (These are the *posttest odds.*)

How useful would test A be in clinical practice in Pepi?

Question C11–4

The facts about test A (sensitivity 80%, specificity 98%) are presented in another way in Table C11–2. Make sure you understand what the figures mean. Then multiply the pretest odds (0.266—is this the result you got in *Question C11–3?*) by each of the likelihood ratios in turn, and compare the answers with the posttest odds (which you also calculated in *Question C11–3*). What do you find?

Question C11–5

This and the following questions deal with a diagnostic test that yields a range of results. It is a supposititious test for TV dementia, acronymously named the BLIP test. The subject is shown a 1-hour video film titled "Bird Life in Patagonia," and the time that elapses before his or her eyes close in sleep is measured. The shorter this period of wakefulness (POW) is, the higher the probability of the disease. Table C11–3 is based on the results of a trial in two samples, one with and one without the disease. The results are shown as probabilities. The sensitivity and specificity of the BLIP test have been computed for each of the cutting-points shown in Table C11–3, and they are plotted against each other in

Table C11–3. Probability of Various Results of BLIP Test Among People With and Without TV Dementia

	Disease		
POW* (minutes)	Present	Absent	Likelihood Ratio†
Under 2	0.20	0.0025	80
2–4.9	0.30	0.005	60
5–9.9	0.20	0.01	20
10–14.9	0.15	0.025	6
15–19.9	0.10	0.1	1
20–29.9	0.02	0.2	0.1
30–44.9	0.02	0.35	0.06
45–59.9	0.01	0.3	0.03
60	0	0.0075	0
Total	1.0	1.0	

*POW = period of wakefulness.

†The ratio of the probability of the given result among people with the disease to the corresponding probability among people free of the disease.

Figure C11. This is called a *ROC curve.* How can the curve be used to tell whether the test is a good one (in terms of sensitivity and specificity)?

Question C11–6

If the BLIP test is to be used as a dichotomous (positive/negative) test, what point on the ROC curve represents the best cutting-point (i.e., the cutting-point that minimizes errors)?

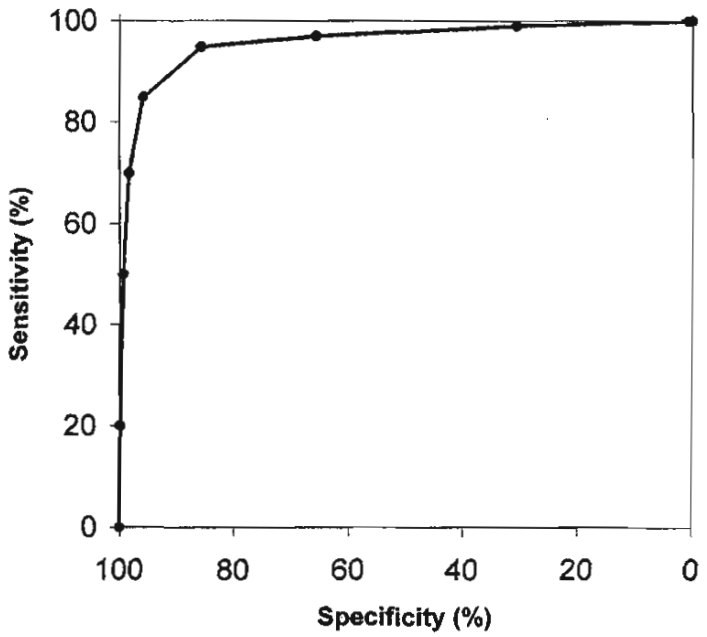

Figure C11. ROC curve (for data in Table C11–3).

Question C11–7

If false negative results are regarded as more important than false positives (because, for example, identified cases can be treated and cured) or if more weight is given to false positives (because, say, of the anxiety, expense, or inconvenience occasioned by a positive test), would this alter the optimal cutting-point?

Question C11–8

The previous two questions took no account of the prevalence of TV dementia in the group in which it is to be used (the pretest probability). Would you expect the BLIP test to have different optimal cutting-points in groups with different prevalences of TV dementia?

Question C11–9

Using the information about the BLIP test in Table C11–3, can you specify the "normal range" of results for this test? What does "normal" mean?

■■■■■■■■■■ Unit C12

Appraisal of Diagnostic Tests

Diagnostic tests are used for at least three purposes: to discover the presence of a disease, to confirm its suspected presence, and to exclude its presence.

A test with a high sensitivity (*Question C11–1*) may obviously be useful as a discovery test, as it will not miss many cases. If its specificity is low, there will be many false positives, but this will not matter much if the additional tests needed to make a firm diagnosis can easily be done. A test with a high sensitivity may also be useful as an exclusion test (however low its specificity): the higher the sensitivity, the more certainly a negative result means absence of the disease.

The higher the specificity of a test (*Question C11–2*), the more useful the test may be as a confirmation test: a specificity of 100% means that a positive result is pathognomonic of the disease. However, a negative result does not mean absence of the disease.

These rough-and-ready rules are not very useful in practice. It is more helpful to see how the test affects our assessment of the probability that the disease is present. This is what you did in *Question C11–3*. The probability of the disease before test A is done is 21% (because the prevalence rate is 21 per 100). The pretest probability may be based on a clinician's appraisal rather than on a known prevalence. The pretest odds are 2,100/7,900 = 0.266 to 1; odds can also be calculated from the probability P by the formula $P/(1 - P)$, as we saw in Unit B11; that is, $.21/(1 - .21) = .266$. If the test is positive, the posttest probability

becomes 1,680/1;838 = 91%, and the posttest odds are 10.6. If the test is negative, the posttest probability is 420/8,162 = 5.1%, and the odds are 0.05.

The results of the test have a big influence on our assessment of the likelihood that the disease is present. Test A would therefore be a useful diagnostic tool (it does not appear to be too inconvenient, expensive, or hazardous to use).

As you saw in *Question C11–4,* multiplying the pretest odds by the likelihood ratio provides the posttest odds. If we know the likelihood ratios for the results of a test, it is thus easy to calculate the posttest odds and probabilities; remember that probability = odds/(1 + odds).

To use this procedure for converting the result of the test into a meaningful statement about the certainty of a diagnosis, one requires (a) an estimate of the pretest probability, and (b) information about the likelihood ratios when the test is applied to patients similar to the patient under consideration. The procedure can be used both for tests that have dichotomous results (as was demonstrated in *Question C11–4*) and for tests that give a range of results. If the test is a dichotomous one, the likelihood ratio for a positive result is the sensitivity divided by the false positive rate.

The procedure can also be used before a test is done, to see how the result can affect the probability of the disease. This may help the clinician to decide whether the test is worth doing (Note C12–1).

As an exercise, suppose that a 55-year-old woman is given a BLIP test (Table C11–3), and that you know that the specific prevalence rate of TV dementia in women of her age is 20%. What is the posttest probability of the disease if she falls asleep in 1 minute? in 6 minutes? in 50 minutes? Is the test useful? (For answers, see Note C12–2.)

The appraisal of screening and diagnostic tests can be simplified by using nomograms or other aids (Sackett et al., 1985, 1997) or an appropriate computer program (see Note A3–7).

ROC Curves

The ROC (*receiver operator characteristics* or *relative operating characteristics*) curve displays the relationship between the sensitivity and specificity of a test. Sometimes the false positive rate is used instead of specificity, but this does not alter the appearance of the curve, for the scale is then reversed (0% to 100% instead of 100% to 0%). All the points for which data are shown in Table C11–3 are plotted in Figure C11.

In answer to *Question C11–5,* the higher the curve is (because of high sensitivity) and the farther it is to the left (because of high specificity), the better the test is. A test is therefore good if the curve comes close to the top-left corner, as it does in Figure C11. As a measure of this feature, the area under the ROC curve is often calculated, in terms of the percentage it occupies of the total area in the 0% to 100% rectangle. This percentage expresses the probability that the test will correctly rank a randomly chosen person with the disease (TV dementia) and a randomly chosen person without it. Its value is 50% if the test does not

discriminate. The area under the curve in Figure C11 is 95.8% (95% confidence interval, 95.6% to 96.1%).

The best cutting-point for the test, if it is to be used as a dichotomous (positive/negative) test (*Question C11–6*) is the point closest to the top-left corner (i.e., the point at which errors are minimal because both sensitivity and specificity are high). In Figure C11 this is the point representing a result of 15 minutes, where sensitivity is 85% and specificity is 96% (very closely followed by the point representing a result of 20 minutes).

The choice of an optimal cutting-point can, of course, be influenced by the relative importance attached to false positive and false negative results (*Question C11–7*). If twice as much weight is given to false negatives as to false positives, appropriate calculations indicate that in this instance the optimal cutting-point will be not 15 minutes, but 20; whereas if twice as much weight is given to false positives as to false negatives, the best cutting-point remains 15 minutes.

Because the numbers of false positives and false negatives are determined not only by sensitivity and specificity but also by the prevalence of the disease, the optimal cutting-point is obviously influenced by prevalence (*Question C11–8*). The choice of a cutting-point should be based not only on the sensitivity and specificity data shown in the ROC curve, but also on prevalence and the relative importance of false positive and false negative results; this usually requires a computer (see Note A3–7).

The Meaning of "Normal"

The "normal" range of response to the BLIP test (*Question C11–9* is not easy to define. "Normal" is used in at least three different ways:

- "What is usual." In this sense, a normal range can be defined in unequivocal terms—for example, "from two standard deviations below the mean to two standard deviations above the mean" or "between the 10th and 90th percentiles." But "abnormal" then only means "unusual."
- "What is desirable"—that is, a range of values that indicate or predict good health. But there may be no sharp dividing line between "healthy" and "unhealthy" findings. In the present instance (Table C11–3), the monotonically decreasing likelihood ratios show that there is a gradient of normality, not a dichotomy; no finding occurs only in disease-free people, and no finding occurs only in people with the disease. Any dividing line must be arbitrary. We can decide, for example, that any result with a likelihood ratio of 1 or less is "normal"; but this "normal" range will include some—and maybe many—people with the disease.
- "What requires no action"—that is, there is no need for further investigations, for surveillance, or for curative or preventive measures. This use of "normal" requires information not only about associations with health and disease, but also about the likely benefits of intervention.

Notes

C12–1. For detailed discussions of the selection and interpretation of diagnostic tests, see Sackett et al., 1985, 1997). Additional measures of the degree to which tests produce a gain in the certainty of the diagnosis are available (Connell and Koepsell, 1985).

C12–2. The pretest probability that a 55-year-old woman has TV dementia is .2. The pretest odds are $.2/(1 - .2) = 0.25$. If the subject falls asleep in 1 minute the likelihood ratio (Table C11–3) is 80. The posttest odds are therefore $0.25 \times 80 = 20$, and the posttest probability of the disease is $20/(1 + 20) = 95\%$. If the POW is 6 minutes, the posttest odds are $0.25 \times 20 = 5$, and the posttest probability is $5/6 = 83\%$. If the POW is 50 minutes, the posttest probability is 0.7%. The test is obviously a useful one.

■■■■■■■■■■Unit C13

Test Yourself (C)

To wrap up this section, see if you can do the following (Unit numbers in parentheses):

- List various ways of appraising the validity of a measure (C2).
- Calculate
  sensitivity and specificity of a measure (C2).
  false positive and negative rates (C2).
  predictive values of positive and negative results (C10).
  kappa (C8).
- Explain what is meant by
  criterion validity (C2).
  predictive validity (C2).
  construct validity (C2).
  content validity (C2).
  face validity (C2).
  consensual validity (C2).
  convergent validity (C2).
  responsiveness of a measure (C2).
  zero preference (C2).
  misclassification bias (C3).
  reliability (C7).
  a screening test (C10).
  a ROC curve (C12).

- Explain the difference between
  differential and nondifferential misclassification (C3).
  interobserver and intraobserver reliability (C7).
  percent agreement and kappa (C8).
- Explain
  how a low sensitivity will affect an estimate of prevalence (C3).
  how a low specificity will affect an estimate of prevalence (C3).
  how use of a measure of low validity affects the estimated prevalence of a rare disease (C3).
  why the predictive value of a positive test varies with the prevalence of the disease (C3).
- List
  ways of measuring the criterion validity of metric-scale measures (C7).
  different kinds of scale of measurement (C7).
- State how an association between two variables may be affected by
  nondifferential misclassification of one variable (C3).
  nondifferential misclassification of both variables (C6).
  differential misclassification of one variable (C6).
  differential misclassification of both variables (C6).
- Appraise
  a screening test (C10, C11)
  a ROC curve (C12)
- State what factors influence the predictive value of a positive screening test (C11).
- Interpret a kappa value (C8, C9).
- Explain what is meant by
  dichotomy (C7).
  nominal, ordinal, interval, and ratio scales (C7).
  metric scale (C7).
  continuous and discrete scales (C7).
- Explain (in general terms) what is meant by
  Berksonian bias (C5).
  internal consistency—reliability (C7).
  regression toward the mean (C9).

(The following items refer to diagnostic tests.)

- Compare the importance of sensitivity and specificity in determining the usefulness of a diagnostic test (C12).
- Explain what is meant by
  pretest probability and odds (C12).
  posttest probability and odds (C12).
  likelihood ratio (C12).
  a "normal" result (C12).
- Calculate the posttest probability from the pretest probability and a likelihood ratio (C12).

Section D

Making Sense of Associations

"I know what you're thinking about," said Tweedledum: "but it isn't so, nohow."

"Contrariwise," continued Tweedledee, "if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic."

(Carroll, 1872)

Unit D1

Introduction

Section D deals with the appraisal of associations between variables, using the approach described in Unit A16. By way of a reminder, here is a list of basic questions that may be asked about an association:

- Actual or artifactual? (selection bias? information bias?)
- Strength (rate ratio, odds ratio, rate difference, etc.) and other qualities (direction? monotonic? linear?)
- Nonfortuitous?
- Consistent? (influence of modifying factors?)
- Influence of confounding factors?
- Causal?

We have already done a number of exercises on the detection and examination of associations, the appraisal of selection and information bias, confounding and modifying effects, the use of stratification and standardization to control confounding effects, and other specific aspects.

Topics that will receive special attention in this section include statistical significance, methods of appraising the possibility and likely direction of confounding effects, measures of the strength of associations, synergism, the appraisal of associations in stratified data, and multivariate analysis. The appraisal of causation will be dealt with in more detail in Section E.

Table D1. Incidence Rate of Coronary Heart Disease* (CHD) per 1,000 Person-Years, by Presence of Varicose Veins at Entry into Study

Varicose Veins	No. of Men	Rate of CHD
None	5,477	2.9
Mild	1,217	4.4
Moderate	731	5.7
Total	7,425	3.4

*Myocardial infarction and deaths from CHD.

Exercise D1

Are people with varicose veins especially likely to develop coronary heart disease? This was one of the questions investigated in a prospective study of Paris policemen (Note D1). After an initial examination, 7,432 men (French-born, aged 42–53) with no evidence of coronary heart disease or certain other atherosclerotic diseases were followed up for an average of 6.6 years, to identify new cases and deaths of coronary heart disease. The results are shown in Table D1. The rates are person-time incidence rates.

Question D1–1

Summarize the facts about the association between varicose veins and coronary heart disease.

Question D1–2

What are the possible explanations for the association between varicose veins and coronary heart disease? (Ignore Occam's razor.)

Question D1–3

What additional information would you like? (Use Occam's razor.)

Note

D1. The study is by Ducimetière et al. (1981). The exercises use derived data, which may not completely conform with the actual findings.

■■■■■■■■■■ Unit D2

Explanations for an Association

In answer to *Question D1–1,* there is a positive association between the presence of varicose veins and the subsequent incidence of coronary heart disease (CHD). Men with mild varicose veins had a higher rate of CHD than men with no varicose veins, and men with moderate varicose veins had a still higher rate. One way of expressing the strength of this association is to calculate rate ratios, using one group (say, the men without varicose veins) as the *reference category.* The rate ratios are then 4.4/2.9 = 1.5 for mild varicose veins and 5.7/2.9 = 2.0 for moderate varicose veins. The rate ratio for the reference category (no varicose veins) is, of course, 1.0.

Note that some epidemiologists reserve the term "rate ratio" for ratios of incidence rates that are based on person-time denominators, as they are in this instance. They use the terms *risk ratio* or *relative risk* for the ratio of incidence rates based on number-of-individual denominators (see "Incidence rates" in Unit B5). For simplicity's sake we will not be strict about these terms in this book; we may even use "relative risk" for a ratio of incidence rates with person-time denominators. As previously stressed (Unit B5), it is important to know whether we are dealing with incidence rates based on number-of-individuals or person-time denominators; readers who consider it necessary, for this reason, to be strict in the use of the terms "rate ratio" and "risk ratio" can have fun changing our wording.

The possible explanations for the association (*Question D1–2*) are as follows:

1. The association may be an artifact resulting from selection bias, differential misclassification, or other shortcomings in the study methods.
2. The association may be a chance one.
3. The association may reflect the confounding effects of age, social class, fatness or other variables.
4. Varicose veins may be a cause of CHD (rather unlikely).

In seeking additional information (*Question D1–3*), it would be wise to start with information about the methods used in the study. This will give us a better understanding of what the numbers in the table represent, and enable us to appraise the likelihood of selection bias or information bias. We should ask such questions as: How was the study sample chosen? Were there many nonresponders or losses to follow-up? How were varicose veins and CHD measured? Is there information on validity or reliability?

The exercises that follow deal with possible information bias, statistical significance, confounding, and the uses of the study. We will assume that there is no reason to suspect selection bias.

Exercise D2

The report on the study states that

> during the examination, the clinician visually inspected and palpated the legs of each subject and noted any venous enlargement or tortuosity. The severity of the varicosities when present were coded as mild or moderate. . . . [There were] significant differences in the observations of individual clinicians. Among the 12 physicians who [each] examined at least 200 patients . . . the observed prevalence varied from 14% (of which 5% were moderate) to 40% (15% moderate). [The men] were followed up by annual examinations or in the case of retirement by mailed questionnaires, and new cases of atherosclerotic diseases and deaths were identified. . . . All events were confirmed by a medical committee from documents available . . . [indicating] appearance of new Q waves on the electrocardiogram . . . or clinical symptoms with electrical changes. Enzymatic data were evaluated when available.

Question D2–1

Can you reach a conclusion about the validity of the diagnoses of varicose veins and CHD?

Question D2–2

How may possible misclassification affect the association between varicose veins and CHD?

Question D2–3

How may possible misclassification of cases affect the association between CHD and other variables?

■■■■■■■■■■ Unit D3

Effects of Misclassification

In answer to *Question D2–1,* we cannot be certain that the differences in the findings of the 12 physicians occurred only because of interobserver variation in the diagnosis of varicose veins, as there may have been real differences in prevalence among the groups they examined. But it is probably correct to conclude that reliability was low, particularly in the absence of information about any efforts to standardize the examination methods or diagnostic criteria. The investigators themselves inferred that the diagnosis of varicose veins was "partially subjective" and "far from satisfactory."

If we conclude that reliability was not high, we must also conclude that valid-

ity was not high. The term used by the investigators was "uncertainty of diagnostic accuracy." As the presence of varicose veins was measured at the outset of the study, misclassification was probably nondifferential; that is, sensitivity and specificity were probably similar in men who subsequently developed CHD and men who did not. If this is so, the effect would be to reduce the strength of the association between varicose veins and CHD (Unit C3). We cannot, however, be absolutely sure that misclassification was nondifferential: possibly the diagnosis was less valid, for example, in fat subjects, who may also have been more likely to develop CHD.

The diagnoses of CHD cannot be completely valid; cases may well have been missed, especially among pensioners (who were not examined). There is no reason, however, to suspect that the validity of the diagnosis was related to the presence of varicose veins; information was obtained about all subjects annually, and the same methods and criteria were used for men with and without varicose veins. We may conclude that this misclassification, too, probably weakened the association between CHD and varicose veins. The true association is thus probably stronger than the observed one.

In answer to *Question D2–3,* the validity of the diagnoses of CHD probably differed in nonpensioners (who were examined) and pensioners (who were not), resulting in differential misclassification. This might strengthen, attenuate, or reverse the association between CHD and age or any other variable closely linked with retirement.

Statistical Significance

We test the statistical significance of an association to enable us to decide whether to regard the finding as nonfortuitous (that is, not occurring by chance). The test provides a *P* value, which tells us the probability that, if no association actually exists, chance processes alone would produce an association as strong as, or stronger than, the one actually observed (see note D3).

A critical value ("alpha") of 0.05 is often used for appraising significance. That is, a *P* value of under 1 in 20 is often regarded as justification for regarding an association as nonfortuitous. Lower critical values of *P*—for example, .01 or .001—may be used.

In the present example, the value of *P* was .0042; that is, the likelihood that chance processes alone would produce the observed association between varicose veins and CHD was 42 in 10,000 or 1 in 238. The association was highly significant.

Exercise D3

Question D3–1

Compare the make-believe data in Table D3–1 with the data in Table D1. In Table D3–1 the sample size is half that in Table D1, but the incidence rates are identical. Which table shows a stronger association? Which set of data will yield

Table D3–1. Incidence Rate of Coronary Heart Disease (CHD) per 1,000 Person-Years, by Presence of Varicose Veins at Entry Into Study: Imaginary Data

Varicose Veins	No. of Men	Rate of CHD
None	2,738	2.9
Mild	608	4.4
Moderate	365	5.7

a higher *P* value? Which set of data will yield more precise estimates of the rate ratios (i.e., narrower confidence intervals)?

Question D3–2

Are the following statements true or false?

1. When we detect an association that is of interest, we should always test its statistical significance.
2. A test of statistical significance will tell us whether an association is present.
3. A test of statistical significance will tell us whether an association is strong.
4. A test of statistical significance will tell us whether an association is causal.
5. If an association is statistically significant, it is not a chance association.
6. If an association is not statistically significant, it is a chance association.

Question D3–3

If you had to choose between a significance test and the confidence interval of a measure of association, which would you prefer?

Question D3–4

A well-designed trial in which a new treatment and a conventional treatment were compared in similar groups of patients shows that the new treatment is more effective. The *P* level is .045, according to a one-tailed significance test. Do you know what a one-tailed test is? What hypothesis was tested in this trial? How would you appraise the finding of the trial?

Question D3–5

Before returning to Paris, we take a brief look at a study in Cambridge, England, where Davies et al. (1986) compared the mothers of boys with undescended testes with the mothers of normal boys born on the same day in the same hospital, in order to test the hypothesis that undescended testis is caused by an excess of maternal estrogen in pregnancy. The specific hypothesis was that the

Table D3–2. Comparison of Pregnancies of Mothers Whose Boys Had Undescended Testes and Mothers of Normal Boys

Variable	Odds Ratio	*P*
Mean age at conception	—	NS*
Mean length of gestation	—	NS
Mean birth weight	—	NS
Birth weight <2,500 g	—	NS
Threatened abortion	4.9	.04
Breech presentation	0.5	NS
Nausea	1.3	NS
Consultation for nausea	1.1	NS
Antiemetics prescribed	1.4	NS
Vomiting	1.1	NS
Consultation for vomiting	1.1	NS
Hypertension	1.3	NS
Proteinuria	0.5	NS
Any of the above seven	1.1	NS
Any x-rays	0.8	NS
Any ultrasound	1.0	NS
Cigarette smoking (≥1/day)	1.4	NS
Alcohol (≥1 unit/day)	0.8	NS
Iron preparation taken	0.8	NS
Hypnotics	0.2	NS
Analgesics	1.8	NS

*NS: not significant ($P \geq .05$).

mothers of boys with undescended testes would have had a higher prevalence, during pregnancy, of nausea, vomiting, and hypertension (believed to be associated with a high estrogen level). The findings are shown in Table D3–2. Assume that these are the only results of the study. Would you regard the difference with respect to threatened abortion as a finding not attributable to chance?

Note

D3. *Significance tests* ("hypothesis tests") can be said to appraise the plausibility of the observed findings by calculating the probability (*P*) that these data would have occurred by chance if some "null hypothesis" (see Note A15–3; e.g., that no association is present) were true. *P* is then the probability of concluding that there is a real association when actually there is none. A low *P* throws doubt on the null hypothesis, whereas a high *P* means that the null hypothesis cannot be rejected. "Chance" usually means random sampling variation, but it may refer to random measurement error or some other unexplained variability.

■■■■■■■■■■ Unit D4

Statistical Significance (Continued)

In answer to *Question D3–1,* the incidence rates are the same in both tables. This means that the associations are equally strong. But the sample size is smaller in Table D3–1. Therefore the data in Table D3–1 will yield a higher *P* value: that is, there is a higher probability that chance processes alone would produce the association seen in this sample. The data in Table D1 will provide more precise estimates of the rate ratios.

All the statements in *Question D3–2* are false:

1. We may sometimes be interested in an association without caring whether it occurred by chance or not. If the immunization rate is lower in one neighborhood than in another, this may require special action, whatever the reason for the difference; statistical significance is irrelevant.
2. A significance test cannot tell us whether there is an association. What it does is to help us decide whether to regard an observed association as nonfortuitous.
3. One of the factors determining statistical significance is sample size. Even a trivial association may be statistically significant if the sample is large enough.
4. Statistical significance does not tell us whether an association is causal. A statistically significant association may be an artifact or a consequence of confounding.
5. A verdict of significance does not prove that the association is not a chance one; it tells us only that the association is unlikely to be due to "chance" processes alone (see note D3), so that we can have some degree of confidence in regarding it as nonfortuitous.
6. A "nonsignificant" result does not prove that the association is a chance one. It tells us only that "chance" processes might easily produce such an association. The verdict is "not proven." (But a "nonsignificant" result in a very large sample indicates that there is probably no *strong* nonfortuitous association.)

There is no simple correct answer to *Question D3–3.* Significance tests and confidence intervals carry overlapping messages; if a 95% confidence interval for a difference does not include zero, or if a 95% confidence interval for a ratio is wholly below 1 or above 1, significance at $P < .05$ can generally be inferred. But the confidence interval does not tell us the probability of a chance association—is it 1 in 20 or 1 in a million? On the other hand, a significance test gives no information on the precision of the findings—what range of values for the true effect is compatible with the observed findings? Confidence intervals, it has been said, are "almost always wider than one would wish" and thus "introduce an appropriate note of caution into the interpretation of 'clear' findings" (Walker, 1986). The advice given in a widely used set of guidelines for writers of biomedical articles (International Committee of Medical Journal Editors, 1997) is:

When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of *P* values, which fails to convey important quantitative information.

A *one-tailed (one-sided) significance test* tests for the presence of a difference in a specified direction, unlike the "ordinary" (two-tailed) test used in most epidemiological studies, which ignores the direction of the difference. The hypothesis tested in the trial described in *Question D3–4* was that the new treatment was better than the conventional one (the null hypothesis being that it was not better). A two-tailed test would have tested the hypothesis that the two treatments differed in their effectiveness (the null hypothesis being that there was no difference, in either direction).

One-tailed tests are quite valid, and their results can be taken at their face value, provided the test has not been misused. On this condition, we can compare the *P* value with whatever critical level (say, .05) we choose to use, and decide whether to regard the superiority of the new treatment as nonfortuitous.

There may be a temptation to use one-tailed tests inappropriately, because the one-tailed *P* value is generally half the two-tailed value: in this trial, the two-tailed *P* value would have been .09 ("not significant"). In the planning stage of a study, temptation may arise because one-tailed tests require smaller sample sizes. Statisticians agree that the decision to use a one-tailed test must be made before the data are examined (no data-snooping!). Such a test should obviously be used only if there is interest in a difference in a specific direction. An extreme, but "safe" (i.e., conservative) view is that "one should decide to use a one-sided test only if it is quite certain that departures in one particular direction will always be ascribed to chance, and therefore regarded as nonsignificant, however large they are. This situation rarely arises in practice" (Armitage and Berry, 1994). If the original intention was to use a one-tailed test but when the data became available a switch was made to a two-tailed test because of a surprising difference in the unexpected direction, Cochran (1983) suggests that the *P* value be multiplied by 1.5.

Significance tests have "built-in" errors. If a critical level of .05 is used, chance processes will produce a verdict of "statistically significant" in about five of every 100 tests performed, even if no real associations exist (Note D4). In *Question D3–5*, where 21 differences were tested and one of them was found to be (just) significant (in the absence of a prior hypothesis), it is difficult to be confident that this difference was not a "statistically significant" fluke.

On the other hand, most epidemiologists would agree that if the study had been undertaken in order to test the hypothesis of an association between threatened abortion and undescended testes, the significant result should not be ignored. In the present instance there was no such prior hypothesis.

This quandary in the interpretation of significance tests exists whenever many tests not based on prior hypotheses are done in a single study, or when the selection of associations for testing is based not on prior hypotheses but on eye-

catching differences discovered in the data. In such situations, we can play safe by lowering the critical level—for example, if 21 tests are done, by dividing .05 by 21, and demanding a P value of $< .0024$; alternatively (which comes to the same thing) we could multiply each P value by 21 before comparing it with our critical level of .05. Less stringent methods of adjusting P values are available, as are special tests for use in other circumstances in which multiple comparisons may lead to misleading results (e.g., when a number of samples are compared with one another, when a number of groups are compared with the same control group, or when the results of a trial are tested repeatedly as findings accumulate). (See Note A3–7.)

If no statistically significant difference is found between two samples, use is sometimes made of an *equivalence test,* which (unlike ordinary significance tests) tests the null hypothesis that there is more than a specified "negligible" difference (Armitage and Berry, 1994, pp. 195, 201–202; see Note A3–7). A significant result indicates equivalence (i.e., a negligible difference between the values that are compared). Put simply (if not quite accurately), the usual significance test tells us whether there is a difference, whereas an equivalence test tells us whether there is no difference. Equivalence tests may be used to compare the effects of different pharmaceutical preparations ("bioequivalence" tests), or (in a clinical trial) to determine whether a new treatment is at least as effective as the standard treatment. Equivalence tests require large samples; nonsignificant results may be attributable to small sample size.

Exercises D4

We have decided that the association between varicose veins and CHD is probably a real one (underestimated by our data), and can be regarded as nonfortuitous. We now consider possible confounding.

Table D4 shows the prevalence of varicose veins in police of different ranks.

Question D4–1

Summarize the facts shown in Table D4 concerning the difference between police of different ranks. Use ratios.

Question D4–2

May the association between varicose veins and CHD be confounded by rank?

Table D4. Prevalence (%) of Varicose Veins by Rank

Varicose Veins	Officers (N = 1,270)	Subofficers (N = 1,895)	Policemen (N = 4,260)
Mild	13.6	17.2	16.9
Moderate	7.8	9.7	10.5
Total	21.4	26.9	27.4

Question D4–3

The association between rank and varicose veins is highly significant: $P = .000013$. How does this finding affect the probability that rank may confound the association between varicose veins and CHD?

Question D4–4

If there were no association between rank and varicose veins, could rank confound the association between varicose veins and CHD?

Question D4–5

If rank is a confounder, in what direction will it bias the results?

Question D4–6

How can we determine whether rank is actually a confounder?

Question D4–7

Can you suggest other possible confounders of the CHD–varicose veins connection?

Note

D4. Spurious "statistically significant" results (indicating that there is a real association when actually there is none) are called "type I" errors. A type II error is the erroneous failure to find a true association. The *power* of a test is its capacity to avoid type II errors.

■■■■■■■■■■ Unit D5

Confounding Effects

In answer to *Question D4–1*, there is an inverse relationship between rank and varicose veins. The main difference is between officers and other ranks; both mild and moderate varicose veins are slightly less prevalent in officers than in other ranks. The differences between subofficers and policemen are small. Table D5–1 shows rate ratios. In a table of this sort, the reference category, with which the other groups are compared, has a rate ratio of 1.0.

The conditions necessary for confounding were considered in Units A10, A11, and A14: the association between an independent and dependent variable can be confounded by a third variable that influences the dependent variable and is

Table D5–1. Association Between Varicose Veins and Rank: Rate Ratios

Varicose Veins	Officers*	Subofficers	Policemen
Mild	1.0	1.3	1.2
Moderate	1.0	1.2	1.3
Total	1.0	1.3	1.3

*Reference category.

associated with the independent variable (without being an intermediate link in the chain of causation connecting the other two variables). In answer to *Question D4–2,* therefore, confounding by rank is a possibility; to meet the conditions completely, rank must also affect the incidence of CHD. However, a confounding effect of any importance is possible only if the associations between the confounder and the other variables are strong ones. As Table D5–1 shows, the association between rank and varicose veins is weak. Rank can have a substantial confounding effect only if the association between rank and CHD is very strong indeed.

The confounding effect is determined by the presence, direction and strength of the associations between the potential confounder and the other variables. The statistical significance of these associations (*Question D4–3*) is irrelevant. Weak associations—even if statistically highly significant—are unlikely to produce an important confounding effect, whereas strong associations that are not statistically significant (usually because the sample is small) may produce a substantial confounding effect. (Despite this, significance testing may have a role as a strategy for deciding which potential confounders to control; see Note D5.)

A variable can confound the association between two other variables only if it is associated with both of them. The simple answer to *Question D4–4,* then, is no: if rank is not associated with both varicose veins and CHD, it cannot confound the association between varicose veins and CHD.

This forms the basis for a strategy frequently used when considering possible confounders: we know the conditions that must be met if confounding is to occur, and can see whether they are met. If they are definitely not met, we can decide to disregard the possibility of confounding.

This *exclusion test* is useful, but unfortunately not foolproof. Confounding may occur even when the crude data do not demonstrate associations between the suspected confounder and the other variables, since conditional associations (see Unit A9) may be present; that is, an association with the dependent variable may exist when the independent variable is held constant in the analysis, or vice versa. An association between rank and CHD, for example, might exist in men without varicose veins, and this association might easily be missed if we looked only at the data as a whole, ignoring the presence of varicose veins. These con-

ditional associations may satisfy the requirements for confounding (see Note A10–4). What this means, in effect, is that an exclusion test based on the easily observed "crude" associations may be misleading; not only may the crude data fail to reveal an existing conditional association between the suspected confounder and the dependent variable (if the suspected confounder is also a modifier), it may also obscure a conditional association between the suspected confounder and the independent variable (for a fictional example, see Kahn and Sempos, 1989, p. 86). In these exercises, we will generally ignore this complication, remembering only that the exclusion test, as usually applied, is not foolproof. This is a calculated risk that many epidemiologists take in real life.

The direction of a confounding effect can be predicted by a simple and useful although not always reliable *Direction Rule*. If the associations of *C* (the confounder) with *A* and *B* are both in the same direction (i.e., if both are positive or both are inverse), confounding will tend to produce a positive association between *A* and *B*. Conversely, if the associations of *C* with *A* and *B* are in opposite directions (one positive and one inverse), confounding will tend to produce an inverse association between *A* and *B*. (This rule may be misleading if *C* is also a modifier, such that the direction of the association between *A* and *B* differs in the categories of *C*: the effect will depend on the relative size of these categories; paradoxical situations may occur.)

In this instance (*Question D4–5*), the direction of the possible confounding effect of rank cannot be predicted, as we have no information on the direction of the association between rank and CHD.

To determine whether rank is actually a confounder (*Question D4–6*), we can compare the crude rate ratios—that is, the ratios based on the crude rates (Table D5–1)—with the rate ratios seen when rank is controlled by stratification, standardization, or some other procedure. In the next exercise, we will see rates standardized for rank.

The candidates for inclusion in a list of possible confounders (*Question D4–7*) are variables that are known or suspected to be causally related to the dependent variable, and that may be associated with (but not affected by) the independent variable as well; consideration should always be given to the "universal variables" (see Unit A11). Your list probably includes age, smoking, blood pressure, obesity, diabetes, and other known risk factors for coronary heart disease.

Exercise D5

Question D5–1

The incidence rates of CHD were standardized for rank, using the indirect method. The rates in the total study sample were used as the standard. The results are shown in Table D5–2, together with the crude rates. According to these figures, was the association between varicose veins and CHD confounded by rank?

Table D5–2. Incidence of CHD by Presence of Varicose Veins

		Standardized for Rank	
Varicose Veins	Crude Rate*	SMR	Rate*
Absent	2.9	0.86	2.9
Present	4.9	1.37	4.7

*Mean annual rate per 1,000.

Question D5–2

Are the following statements true or false?

1. A variable can confound the association between two other variables only if it is associated with both of them.
2. Confounding often produces very strong associations.
3. If no association is detected between the variables that interest us, there is no point in considering possible confounding effects.
4. If the association between two variables becomes weaker or disappears when a third variable is controlled, this shows that the third variable is a confounder.
5. A confounding effect is always completely controlled by stratification.
6. A confounding effect is always completely controlled by standardization.

Question D5–3

You may remember that in a previous exercise (B12), we found that fractures of the femur were more common in Oxford than in Epiville, and considered the possibility that age might be a confounder. Older people had a higher incidence of fractures, and people in Oxford were older than in Epiville. Use the Direction Rule to predict how controlling for age will affect the association between fractures and place of residence.

Question D5–4

Is there any way of appraising the possible confounding effect of a variable that was not measured in the study under consideration?

Question D5–5

Table D5–3 shows an association between eating chocolate and acne in teenagers. (No need for alarm! The data are completely imaginary, and to the best of our knowledge chocolate has no specific real-life effect on acne.) According to the figures in the table, is the association seen in the total sample confounded by sex?

Table D5–3. Relationship of Eating Chocolate to Acne, by Sex (Far-Fetched Fictional Data)

	Chocolate		No Chocolate		
Sex	Acne	No Acne	Acne	No Acne	Odds Ratio
Both sexes	54	146	21	176	3.1
Females	50	50	20	80	4.0
Males	4	96	1	96	4.0

Note

D5. Experts disagree on the role of significance testing in the identification of possible confounders. Many view statistical significance as irrelevant. As pointed out by Fleiss (1986a, 1986b), however, significance testing provides explicit rules and hence a reproducible method for use in appraising the relative importance of potential confounders and deciding which to control. A suggested compromise is the use of a critical level of $P < .20$ (or higher) for the purpose of selecting possible confounders (Dales and Ury, 1978); computer simulations have provided justification for this approach (Rothman and Greenland, 1998, p. 257).

■■■■■■■■■■ Unit D6

Confounding Effects (Continued)

A change in the strength of an association when a suspected confounder is controlled is suggestive of confounding. To answer *Question D5–1,* we must know the strength of the association according to both the crude and standardized results. The crude rate ratio was 4.9/2.9, that is, 1.7, and the standardized rate ratio was 1.37/0.86 or 4.7/2.9, that is, 1.6. There was thus a very slight—and hence unimportant—confounding effect.

The answers to the "true–false" questions (*D5–2*) are:

1. True. However, the associations with the other variables may not be obvious; they may be conditional ones.
2. False. Even if the confounder is strongly associated with the other variables, "the spurious effect is only a relatively weak echo" (Note D6).
3. False. The apparent absence of an association may be due to confounding.
4. False. The third variable may be a confounder, but it may also be an intervening cause that mediates the causal relationship between the two variables.

5. False. Stratification controls the confounding effect completely only if the categories are homogeneous. If we were controlling for systolic blood pressure, and used broad categories such as "< 140," "140–159," and ≥ 160 mm Hg, there would still be much variation *within* the strata: blood pressure would not be altogether "held constant," and some of its confounding effect might remain.
6. False. In the same way, the use of broad categories may also impair the value of standardization.

To use the Direction Rule (*Question D5–3*), we must be able to designate associations as positive or negative. This may require the choice of reference categories (the choice is arbitrary, and does not affect the conclusions). In this instance, let us choose "Epiville" as the reference category for place of residence. The facts, then, are that age is negatively associated with the independent variable (residence in Epiville) and positively associated with the dependent variable (incidence of fractures). As these associations are in opposite directions, we can predict that if age is a confounder it will probably tend to produce a negative association between residence in Epiville and fractures of the femur. If the confounding effect is controlled, the association will therefore become "more positive." Because the crude incidence rates showed a negative association between residence in Epiville and fractures, we can expect that if age is controlled the negative association will become weaker or disappear, or even change to a positive one—as it actually did when we controlled for age by stratification (Table B14–1) or standardization (Table B14–2).

In answer to *Question D5–4*, it is sometimes possible to make inferences about a confounding effect even if the suspected confounder was not measured. This requires knowledge (from other studies) of the strength and direction of the suspected confounder's associations with other variables. It is then possible to apply the "exclusion test" and the Direction Rule, and even to estimate the magnitude of the possible confounding effect (Note D6).

In *Question D5–5* the crude odds ratio expressing the association between chocolate and acne, taking no account of sex, is 3.1. This is lower than its value, 4.0, in each separate sex. A difference between what we see in crude data and what we see when we neutralize or eliminate the effect of a suspected confounder is indicative of confounding (Unit A11). If the overall odds ratio were standardized by sex, the adjusted value would also obviously be 4.0. The figures thus suggest confounding by sex. Experts may say that this is an instance not of true confounding but of the "noncollapsibility" of odds ratios, which "is usually confused with confounding, although it has nothing to do with the latter phenomenon" (Rothman and Greenland, 1998, pp. 52–53, 60). Noncollapsibility means that the odds ratio in a total group may fall outside the range of the values in separate strata, because (unlike other common measures of association) it is not a weighted average of the values in separate strata. In practice, there is no harm in calling this phenomenon "confounding"; since whatever we call it,

the practical implication is that we can reach useful conclusions only if we control the factor (in this instance, sex) by stratification, standardization, or some other method.

Exercise D6

In this exercise we glance at multivariate analysis. (We will return to this topic later.)

A multivariate analysis was used in the study of Paris police, to control simultaneously for the possible confounding effects of six variables known or suspected to be associated with CHD. These were age, number of cigarettes smoked per day, systolic blood pressure, serum cholesterol, the presence of diabetes, and Quetelet's body mass index. The adjusted relative risks (rate ratios) of CHD when these variables were controlled (i.e., held constant) are shown in Table D6, together with the relative risks based on the crude data. The association between varicose veins and CHD remained statistically significant ($P = .0053$) when these six variables were controlled.

Question D6–1

According to Table D6, can the association between varicose veins and CHD be attributed to the confounding effects of the six variables controlled in this analysis?

Question D6–2

The following explanation was provided for the method of multivariate analysis used in this study. (Don't worry if you don't understand it.)

Multivariate analysis of the relationship between annual incidence rates and different variables was performed by an exponential model with covariates which allowed for unequal follow-up durations (Lellouch, J. and Rokotovao, R., 1976). During follow up, the hazard rate for illness is assumed to be constant (r) for each subject. This assumption is equivalent to stating that the probability that the subject will get the illness before the in-

Table D6. Relative Risk of CHD by Presence of Varicose Veins

Varicose Veins	Crude*	Adjusted†
None	1.00	1.00
Mild	1.52	1.34
Moderate	1.97	1.78

*Based on rates in Table D1.

†Controlling for six variables (see text).

stant t is $1 - \exp(-rt)$, the classical exponential survival model. The individual hazard rate, r, is chosen as an exponential function of the covariates $x_i \ldots x_k$:

$$r = r_0 \exp(b_i x_i \ldots + b_k x_k)$$

Writing the likelihood of observations for cases and noncases and maximising this quantity by an iterative technique gives an estimate of r_0 and the b_j's as well as their asymptotic standard error, allowing a test of the significance of the b_j's by a t test.

Just for argument's sake, pretend you don't understand this explanation. Do you feel that, despite this, you can safely use the results?

Note

D6. See Bross (1966, 1967), who explains how to find whether a possible confounder's association with two other variables are strong enough to account for the observed association between these other variables.

■■■■■■■■■■ Unit D7

Multivariate Analysis

The use of multivariate analysis to control six possible confounders (Table D6) reduces the strength of the association between varicose veins and CHD, but the association remains apparent. The answer to *Question D6–1*, therefore, is that the association can be only partly explained by the confounding effects of these factors.

Question D6–2 poses a real dilemma. We have seen how even a simple statistical manipulation like standardization may, under some circumstances, yield misleading results (Units B14 and B15). How much more likely is it that a complicated procedure—especially one that we do not understand—may mislead us.

We cannot avoid this dilemma. Multivariate analysis provides a short-cut way of handling the effects of a number of variables at the same time, and of looking at complicated interrelationships. With ready access to computers and ready-made computer programs, such analyses are easy to do and increasingly popular. But this does not make their results easier to appraise. Must we just take them on trust?

Ideally, we should understand the procedures well enough to know when they are appropriate, and how to relate to the findings. But what if we don't, and can't find a friendly statistician to ask? There are many forms of multivariate analysis: multiple linear regression, analysis of variance and covariance, discriminant analysis, log-linear analysis, logit analysis, multiple logistic regression, Poisson

regression, proportional hazards regression, and others. Each uses its own mathematical model (Note D7–1) and is based on its own set of assumptions, which are not always clearly spelled out, and may or may not be justified.

A basic general understanding of the main multivariate methods is not difficult to acquire (see note D7–2). But if we lack this and cannot obtain help, we should not ourselves use a multivariate procedure; and if we come across one in a published paper, we should see whether the investigators present a plausible case for the validity of the method: are the assumptions explained and justified, and has the model as a whole been tested to see how well it fits the observed facts? Failing this, the best we can do may be to consider the qualifications and stature of the investigators and the reputation of the journal, and decide whether these inspire us with confidence. (Maybe this is a cop-out, but there may be no alternative.)

In any case, it is prudent to regard the results of any multivariate analysis as providing only an *approximate* picture of the truth. A mathematical model rarely fits the facts perfectly. It is probably wise not to take the findings too literally; associations may be somewhat weaker or stronger than they appear, adjustment for confounding effects may be incomplete, and levels of statistical significance may be misleading. Clear-cut findings are probably correct, but borderline ones—associations that are weak or of marginal statistical significance—should be taken with a pinch of salt.

Exercise D7

In this exercise, we review possible explanations for the association between varicose veins and CHD, and consider the possible uses of the findings.

Question D7–1

This study has shown an association between varicose veins and CHD which (because of misclassification) is probably stronger than it appears.

1. In the light of what you now know, is it possible that the association is a chance finding?
2. Is it possible that the association is a consequence of confounding?
3. May the association be explained by an effect of CHD on the occurrence of varicose veins?
4. May the association be explained by an effect of varicose veins on the occurrence of CHD?
5. Is it possible that varicose veins and CHD are associated because they share a common cause or causes?

Question D7–2

Summarize the additional information that Table D7 provides about the varicose veins–CHD association. Can you suggest an explanation for the new findings?

Table D7. Occurrence of CHD by Presence of Varicose Veins and Rank (Numbers of Cases and Mean Annual Rates per 1,000)

	Rank					
	Officers		Subofficers		Policemen	
Varicose Veins	Cases	Rate	Cases	Rate	Cases	Rate
Absent	21	3.3	28	3.1	54	2.9
Present	5	3.1	11	3.4	44	5.9
P		NS*		NS*		.0005

*NS = not significant ($P \geq .05$).

Question D7–3

The title of the paper on which these exercises were based asks "Varicose veins: a risk factor for atherosclerotic disease?" What is your answer to this question?

Question D7–4

Brandishing the results shown in Table D7, the health officer of the Paris police force excitedly announces that he intends to institute a program using varicose veins as a risk marker. In order to reduce the incidence of CHD, all rank-and-file policemen with varicose veins will be identified and subjected to intensive health surveillance and risk factor intervention, including advice on diet and smoking, and treatment of blood pressure where necessary. Do you have any reservations about his decision? What criteria would you use for appraising the value of a risk marker (i.e., an indicator of increased risk)?

Question D7–5

What are the possible other uses of what we have learned about the association between varicose veins and CHD in Paris policemen?

Notes

D7–1. "*Mathematical model.* A representation of a system, process or relationship in mathematical form in which equations are used to simulate the behaviour of the system or process under study"—*A Dictionary of Epidemiology* (Last, 2001).

D7–2. Multiple linear regression and multiple logistic regression are explained in most statistics textbooks; see, for example, Daniel (1995, chaps. 10 and 11). For a 32-page "brief introduction" to proportional hazards regression analysis, see Selvin (1996, chap. 12); a shorter explanation is offered by Altman (1991, pp. 387–393); this procedure is often called "Cox regression," although

the proportional hazards model is only one of the models described by Cox for use in survival analysis (Cox and Oakes, 1984).

■■■■■■■■■■ Unit D8

Explanations for the Findings

In answer to *Question D7–1:*

1. Yes, the association may be a chance finding. The probability that it is due to chance is .0053 (according to the multivariate analysis), or 1 in 189.
2. Yes, the association may be a consequence of confounding by factors that we have not yet examined, or maybe thought of.
3. No, the association cannot be due to an effect of CHD on the risk of incurring varicose veins—an impossibility if we accept the investigator's assurance that the men were free of CHD at the outset of the study. An effect cannot precede its cause.
4. Yes, the association may be explained by an effect of varicose veins on the occurrence of CHD. The "dose–response" relationship shown in Table D1—that is, the monotonic increase in CHD incidence when men with no varicose veins, mild varicose veins, and moderate varicose veins were compared—is consistent with a causal explanation. The only argument against this explanation is that it is difficult to suggest a plausible etiological mechanism. This low biological plausibility may lead us to regard a causal explanation as improbable, but we may be wrong: maybe the explanation is correct, and current biological knowledge is defective.
5. Yes, it is possible that varicose veins and CHD have a common cause (or causes), even if we cannot identify it. A common cause may have a confounding effect (Fig. A14–2). Finding a variable that confounds the association between varicose veins and CHD because of its effect on both these disorders would add to our understanding of etiology; a confounder is not always just a "nuisance variable."

In answer to *Question D7–2,* stratification of the data (Table D7) shows that the association between varicose veins and CHD is modified by rank. There is no noteworthy association in officers (relative risk = 3.1/3.3 = 0.9) or suboffi-cers (relative risk = 1.1); however, in rank-and-file policemen the relative risk is 2.0, and this is statistically highly significant. In other words, the presence of varicose veins is a risk marker for CHD, but only in rank-and-file policemen.

To explain why the association between varicose veins and CHD is restricted to rank-and-file policemen, we must consider how these men differ from police

of higher ranks—in the nature of their work, the conditions they are exposed to, their lifestyle, or the characteristics or experiences that led to their being rank-and-file policemen and not officers or subofficers. We need to identify some factor whose presence is a condition for the processes (which we do not yet understand) that link varicose veins and CHD. The factor we are seeking must, of course, be one that is associated with the incidence of CHD (see Unit A13). It need not, of course, be associated with the independent variable (varicose veins); this is a requirement for a confounding effect, but not for a modifying effect.

No explanation for the effect modification was suggested by the investigators. You may have been more successful. If so, check that the factor you have named meets the above condition. Your suspected factor may, for example, be excessive standing. It is not enough to know that (as the investigators tell us) the average Paris policeman spends a large amount of time standing relatively motionless; we must also know, or at least believe it plausible, that prolonged standing is associated with CHD. If these conditions are met, we can proceed to seek facts that will test the hypothesis that excessive standing accounts for the findings seen in Table D7. (To do this, we will need data on the amount of standing.) Note that the possible association between excessive standing and varicose veins (found in other studies) is not relevant to the hypothesis that excessive standing modifies the association between varicose veins and CHD.

Risk Factors and Risk Markers

"Yes," "no," and "don't know" are all acceptable answers to *Question D7–3*, depending mainly on how "risk factor" is defined. There is unfortunately no agreed definition. To cite the *Dictionary of Epidemiology* (Last, 2001):
The term *risk factor* is rather loosely used, with any of the following meanings:

1. An attribute or exposure that is associated with an increased probability of a specified outcome, such as the occurrence of a disease. Not necessarily a causal factor. A risk marker.
2. An attribute or exposure that increases the probability of occurrence of disease or other specified outcome. A determinant.
3. A determinant that can be modified by intervention, thereby reducing the probability of occurrence of disease or other specified outcomes. To avoid confusion, may be referred to as a "modifiable risk factor."

If we use definition 1, the answer to the question is "yes." If we use one of the other definitions, our answer may be "no" (not proved by the study) or "don't know" (not disproved).

In the interests of clarity, it is probably best to use the term "risk factor" only if we know that the factor is causal—that is, that it increases the risk (definition 2) and does not merely *point to* an increased risk (definition 1). Men with low semen quality may be more likely to develop cancer of the testis in later years

(Jacobsen et al.; 2000); but their increased risk is obviously not caused by their low semen quality. If a factor points to—but does not necessarily bring about—an increased risk, it is advisable to call it a risk marker. These are the terms we will use in these exercises. If we thought that varicose veins were a cause of CHD and that treating them would reduce the incidence of CHD, we could use the term "modifiable risk factor" (definition 3).

Appraising a Risk Marker

A risk marker should be appraised in the same way as a screening test (Units C10 and C11). The only difference between them is that screening tests identify people with a high probability of *having* a disease, whereas risk markers identify people with a high probability of *developing* the disease. Before deciding to use varicose veins as a risk marker in his program (*Question D7–4*), the police health officer should review statistical indices such as sensitivity and predictive value, and compare them with the corresponding indices for alternative risk markers—as well, of course, as having satisfactory evidence for the effectiveness of preventive intervention.

The sensitivity of varicose veins as a predictor of CHD in rank-and-file policemen was 45%. (Do you know where this figure comes from? If not, see Note D8–1). The risk marker would have identified under half of those who incurred CHD by the end of the study. If cases in all ranks are taken into account, we see from Table D7 that only 60/163, or 37%, of cases would have been identified in the program. The health officer should certainly take these facts into consideration. Even if the proposed intervention can completely prevent CHD (which is unlikely), the program will prevent only part of the cases. Maybe the health officer should consider the provision of preventive care to the whole police force (irrespective of individual risk), or seek a more sensitive risk marker.

The predictive value of a risk marker (equivalent to that of a screening test) is the risk associated with the marker. The health officer knows that in rank-and-file policemen this risk is 5.9 per 1,000 per year (Table D7), or about 3.5% in 6 years, and has presumably decided that this provides sufficient justification for his program.

Additional factors to be taken into account in appraising the value of a risk marker in a program of this sort include the risk marker's prevalence. If this is very high, so that the high-risk group requiring special attention is very large, it may be more effective and efficient to give extra care to the total population. (Do you know the difference between *effectiveness* and *efficiency?* If not, see Note D8–2.) In this instance, the prevalence rate of varicose veins in rank-and-file policemen is 27% (Table D4). Also, the use of the risk marker must be practicable in terms of cost, resources, acceptability, and convenience. Obviously, there must also be good reason to believe that the detection of vulnerability will lead to an appreciable reduction of risk, and the expected benefit must outweigh any harm that may be done by labeling apparently healthy people as being "at risk" and involving them in surveillance and preventive activities.

Uses of the Findings

In considering the possible uses of knowledge about the association of varicose veins with CHD in Paris police (*Question D7–5*), we should take account of the various categories of users (Unit A17).

First, for users whose chief interest is in the health care of Paris police, the results point to a way of identifying men with an especially high risk of CHD, who may merit special surveillance and preventive care. This may be applied not only in a special program, but in the clinical care of individual policemen. Second, the results may possibly serve the same purpose for those who want to identify high-risk individuals or groups in other populations. And third, for users whose basic interest is in "research," the association may provide clues that will lead, in the long run, to a better understanding of etiological processes and methods of prevention. This is probably the most important potential contribution of the study. Why does the association exist? Do varicose veins and coronary heart disease have common etiological factors, such as dietary factors or decreased blood fibrinolytic activity (Ducimetière et al., 1981) or hitherto unsuspected causes? In particular, why is the association strongest in rank-and-file policemen? What clues to etiology does this provide? *Unexplained effect modification—like any other unexplained or unexpected finding—should always be regarded as a possible clue to etiology.*

We now bid adieu to the Paris gendarmèrie.

Exercise D8

Question D8–1

Using the terms "risk factor" and *risk marker" in the way recommended above, are the following statements true or false?*

1. Every risk marker is a risk factor.
2. A factor cannot be both a risk marker and a risk factor.
3. Every risk factor is useful as a risk marker.
4. Every factor that brings about a change in the probability of a disease is a risk factor.
5. Removing a risk factor does not necessarily remove the risk attributable to the factor.

Question D8–2

A large-scale follow-up study of army veterans, initiated in the United States in 1954, demonstrated strong relationships between smoking and mortality (Kahn, 1966). The findings in Table D8 show that in the veterans aged 65-74 (as in other age groups) cigarette smoking was an indicator of an increased risk of dying.

According to these data, what is the approximate risk of dying within the next 5 years, for a 68-year-old man in each of the three smoking categories?

Table D8. Annual Probability of Death* for Veterans Aged 65–74 Years by Smoking Category

Smoking Category	Annual Probability of Death (%)	Relative Risk
Never smoked (or occasional only)	2.4	1.0
Ex-cigarette smokers (who stopped for reason other than "doctor's orders")	3.1	1.3
Current cigarette smokers	4.0	1.7

*Equivalent to the annual cumulative mortality rate.

Question D8–3

For geniuses only. A study of a large sample of 7-year-old boys showed that 4.77% had been diagnosed as having inguinal hernia, and 8.1% of the boys with such diagnoses had low birth weights (< 5 lb). A representative sample of 7-year-old boys without hernias was investigated, and in this control group the proportion with low birth weights was 2.1%. Can you estimate the risk of having an inguinal hernia diagnosed by the age of 7, for a live-born boy who weighs < 5 lb at birth and survives to the age of 7 years? (See Note D8–3.)

Notes

D8–1. The *sensitivity of a risk marker* is the proportion of incident cases in whom the risk marker was previously present. Table D7 tells us that 98 cases of CHD occurred in rank-and-file policemen during the period of the study. Of these, 44 had varicose veins at the outset. In these circumstances, sensitivity was thus 44/98 = 45%.

D8–2. *Effectiveness* refers to the extent to which desirable effects are achieved. *Efficiency* refers to the balance between these effects and the expenditure (in time, effort, money, and other resources) required to achieve them.

D8–3. Data from Depue (1984); modified slightly.

■■■■■■■■■■ Unit D9

Risk Factors and Risk Markers (Continued)

The following are the answers to the "true–false" questions in *Question D8-1.*

1. False. Varicose veins may point to an increased risk of CHD, without being responsible for the increased risk.

2. False. Hypertension, for example, points to an increased risk of CHD, and is also a reason for the increased risk.
3. False. Considerations such as low sensitivity, low predictive value, and the cost or inconvenience of examinations to determine the presence of a given risk factor may render it of little practical value as a marker.
4. False. A factor that affects the probability of occurrence of a disease is, of course, a risk factor only if it *increases* the probability of the disease: "risk" is generally used to refer to the probability of an unfavorable outcome. If the factor *reduces* the probability of the disease, it is a *protective* or *preventive* factor.
5. True. Hypertension, for example, is unquestionably a risk factor for myocardial infarction, but treating it does not reduce the incidence of myocardial infarction to the level found in nonhypertensives (Poulter and Sever, 1992). A risk factor can have irreversible effects.

In answer to *Question D8–2,* we can make a rough estimate of the risk of dying within 5 years by multiplying the annual probability of death by five. This gives a risk of 12% for the "never smoked" group, 15.5% for ex-smokers, and 20% for cigarette smokers (see Note D9–1).

Question D8–3 (skip this paragraph if you didn't try the question) is difficult; you probably were not able to do it if you skipped the exercise on diagnostic tests (C11). The risk that is required is the "exposure-specific" risk, for individuals exposed to a specific factor (a low birth weight). This is analogous to the predictive value of a positive test—that is, the disease probability associated with a positive test result (a low birth weight), or the posttest probability (see Unit C12)—and it can be computed in the same way. Calculate the likelihood ratio (8.1/2.1 = 3.86), and then multiply the pretest odds in favor of a hernia diagnosis—that is, 0.0477/(1 − 0.0477) = 0.050—by the likelihood ratio (3.86), to obtain the posttest odds of 0.193. The posttest probability—which is what we require—is 0.193/(1 + 0.193), or 16.2%. You may have reached this answer in a different way (Note D9–2).

Measures of the Strength of an Association

A wide variety of indices may be used to measure the strength of associations between variables. They include absolute differences (e.g., between rates, proportions, or means), ratios (e.g., risk ratios and other rate ratios, the odds ratio, and other measures of relative differences), and other statistical indices (e.g., correlation and regression coefficients). (See Note D9–3.)

The choice of a measure of strength depends, *inter alia,* on the scales of measurement of the variables (Note C7), the purpose of the study (are we more interested in absolute or relative differences?—see Unit A3), and the kind of study.

The next two exercises test your ability to interpret and use some of these measures.

The *relative risk* or *risk ratio* is the ratio of two incidence rates (or, if the terms are defined strictly, of two incidence rates based on number-of-persons denominators). The ratio of two incidence rates based on person-time denominators may be called the *incidence density ratio* or the *incidence rate ratio.* An odds ratio is sometimes referred to as the *estimated relative risk,* since if the risk is low the odds ratio and risk ratio are very close to each other (Note B11–1).

Exercise D9

Question D9–1

The incidence rate of disease *A* is twice as high in vegetarians as in nonvegetarians. The incidence rate of disease *B* is 0.2 times as high in vegetarians as in nonvegetarians. Which disease is more strongly associated with eating habits?

Question D9–2

A large follow-up survey showed that the mortality rate from cancer of the lips, tongue, and mouth was 4.1 times as high in cigar smokers as in people who had never, or only occasionally, smoked (Kahn, 1966). Does this show that cigar smoking is a modifying factor?

Question D9–3

Is this association (relative risk = 4.1) likely to be due solely to confounding?

Question D9–4

Assuming you had no other information, could you conclude from this association that preventive activities with respect to these cancers should center on efforts to reduce the smoking of cigars?

Question D9–5

What does a relative risk of 1 mean?

Question D9–6

If we conduct a follow-up study and obtain a relative risk by comparing the incidence of a disease in a cohort (group) of smokers and a cohort of nonsmokers, will this tell us what the relative risk is in the total population?

Question D9–7

If we compare the previous smoking habits of people who have a certain disease (cases) and people who do not (controls), will the results tell us (a) the relative risk (i.e., the ratio of incidence rates based on number-of-persons denominators); (b) the incidence density ratio (i.e., the ratio of incidence rates based on person-time denominators)? Can the results of such a study be generalized to the population as a whole?

Question D9–8

One of the findings of a 19-year follow-up study of 5,135 male Japanese physicians (Kono et al., 1986), in which the relationship between drinking habits and mortality was investigated, was that the age-adjusted death rate from coronary heart disease per 10,000 person-years was 26.3 in nondrinkers and 16.2 in occasional (less than daily) drinkers. The difference between the rates was 10.1 deaths per 10,000 person-years, and the ratio of the rates was 1.6 (or 0.6). Is the difference or the ratio a better measure of the strength of the association?

Question D9–9

More findings from the study of Japanese physicians are shown in Table D9. Are any of the associations shown in the table statistically significant? What do you think may explain the finding in ex-drinkers?

Question D9–10

The response rate in the above study was low. Only 51% of the physicians in the region participated. The investigators discuss the possibility that this may have biased the associations between drinking and mortality. What kind of bias are they referring to?

Question D9–11

If a risk ratio is statistically significant, does this mean it is significantly different from 0, from 1, or from some other value? If a rate difference is statistically significant, does this mean it is significantly different from 0, from 1, or from some other value? If an odds ratio is statistically significant, does this mean it is significantly different from 0, from 1, or from some other value?

Table D9. Association Between Occasional Drinking and Mortality from Coronary Heart Disease: Relative Risks Adjusted for Age and Smoking Habits

Drinking Habits	Relative Risk (With 95% Confidence Interval)
Nondrinker	1.0
Occasional drinker	0.6 (0.4–0.9)
Daily drinker	
<2 go* of *sake*	0.7 (0.5–1.1)
≥2 go* of *sake*	0.7 (0.4–1.1)
Ex-drinker	1.5 (1.0–2.4)

*One go of *sake* contains about 27 ml of alcohol.

Notes

D9–1. Better estimates of the 5-year risk, using the formulae in Note B5–4, are 11.5% (never smoked), 14.6% (ex-smokers) and 18.5% (current smokers). For the "never smoked" group, for example, the person-time rate is 0.024/[1 − (0.024/2)] = 0.0243, and the 5-year cumulative rate is (0.0243 × 5)/[(0.0243 × 5/2) + 1] = 11.45%. Alternatively, we could use the method described in Unit B8: multiply together the survival rates in each period, and subtract the answer from 100%. For the "never-smoked" group, the survival rate in each year is 1 − .024 = 0.976. To obtain the 5-year survival rate we then calculate 0.976 × 0.976 × 0.976 × 0.976 × 0.976 (i.e., 0.976 to the power of 5) = 0.8856, and obtain a 5-year risk of 1 − 0.8856 = 0.1144 = 11.44%.

D9–2. Another method of calculation is to divide the prevalence of a low birth-weight history plus hernia in 7-year-olds (8.1% × 4.77%, or 0.386%) by the total prevalence of a low birth-weight history in 7-year-olds, which is 0.386% plus the prevalence of a low birth-weight history without hernia (2.1% × [100 − 4.77]%, or 2.000%). In other words, 0.386%/2.386%, which is 16.2%.

D9–3. The concept that differences as well as ratios and other indices can serve as measures of the strength of an association is a useful one, although not consistent with a narrow statistical definition of "strength," which requires "free" (nondimensional) measures.

■■■■■■■■■■ Unit D10

Measures of Strength

In *Question D9–1,* disease *B* exhibits a stronger association with eating habits than disease *A*. The risk of disease *A* is only twice as high in one group as in the other, whereas the risk of diseases *B* is five times as high in one group as in the other. Whether the ratio of two rates is 0.2 or 5 depends only on which rate we decide to divide by which; this decision does not affect the strength of the association.

A relative risk of 4.1 (*Question D9–2*) tells us that cigar smoking is strongly associated with the disease, but a single relative risk can tell us nothing about effect modification. Effect modification is detected by comparing the associations found in different groups or different circumstances. If we found that the relative risk was 5 in older men and 2 in younger men (and if this difference was statistically significant, not an artifact, and not caused by confounding), we would conclude that age modified the association between cigar smoking and the disease—or, as a corollary, that cigar smoking modified the association between age and the disease (Unit A13).

A relative risk as high as 4.1 (*Question D9–3*) is unlikely to be due solely to

confounding, except in unusual circumstances. The stronger an association is, the more likely it is (if not an artifact) to be a causal one.

Decisions about the institution of preventive activities (*Question D9–4*) do not depend solely on the strength of an association. Other considerations would come into play even if cigar smoking was to be used only as a risk marker, as we saw when we considered a proposed preventive program based on the presence of varicose veins (Unit D8). In this instance, we are considering preventive activities that center on the reduction of cigar smoking. Such activities presuppose that cigar smoking is causal and that its reduction will have an important impact on the incidence of mouth cancers in the population. More evidence is required.

A relative risk of 1 (*Question D9–5*) means that there is no association: the rates under comparison are identical.

A comparative study of groups of smokers and nonsmokers (*Question D9–6*) will tell us the relative risk in the total population only if the groups are representative samples of all smokers and nonsmokers, respectively, in the population.

A case-control study can provide an odds ratio and a rate ratio—in this instance (*Question D9–7*) the ratio of smoking rates—that can serve as measures of the association. But the study does not tell us the incidence rates in smokers and nonsmokers. A case-control study therefore does not permit direct calculation of the ratio of incidence rates, unless ancillary information is available, such as the incidence of the disease in the total population, which permits the computation of incidence rates, and hence of ratios of incidence rates (we had an example in *Question D8–3*). But in most case-control studies the odds ratio can be used as an estimator of the ratio of incidence rates using person-time denominators (the incidence density ratio) [Note D10–1], and if the frequency of the disease is low the odds ratio is also a good estimator of the ratio of incidence rates using number-of-persons denominators (the relative risk) (Note D10–2).

Application of the findings of a case-control study to a total population is, of course, warranted only if the samples are drawn from this population and are representative.

The choice of an absolute or relative difference as a measure of association (*Question D9–8*) depends on the use we want to make of the finding. If we wish to study processes of causation, the rate ratio will serve our purpose well. If we believe that occasional drinking saves lives, and want to know how many lives it saves, we should use the absolute difference.

In answer to *Question D9–9*, if the 95% confidence interval of a rate ratio lies wholly above 1 or wholly below 1, it is generally safe to conclude that P is under .05. The association with occasional drinking is thus statistically significant, and the association with being an ex-drinker *may* be statistically significant: the unrounded value of the lower confidence limit may be below 1 (e.g., 0.951) or above 1 (e.g., 1.049). The investigators' comment on the high CHD rate in ex-drinkers is: "It is possible that ex-drinkers may have drunk heavily before they abstained, but it seems more likely that ex-drinkers stopped drinking because of their illnesses" (Kono et al., 1986).

The possibility of biased associations (*Questions D9–10*) in this study does not arise from the low participation rate itself, but from the possibility that participation rates may differ in people with different drinking habits and also in people with different probabilities of dying, and that the interplay of these selection factors may produce associations in the sample that differ from those outside the sample and in the population as a whole (see Berksonian bias, Unit C5).

In answer to *Question D9–11,* statistical significance means a significant difference from 1 in the case of risk and odds ratios, and a significant difference from zero in the case of a rate difference.

Exercise D10

In this exercise we look at some other measures of the strength of an association.

Question D10–1

Table D10–1 shows the correlation of diastolic blood pressure with age and weight in a random population sample in the West Indies (Khaw and Rose, 1982).

Are the correlations strong? What does the value "0.00" mean?

Question D10–2

What modifying effects are shown in Table D10–1?

Question D10–3

Can you tell whether the association between diastolic pressure and weight in the older age group is confounded by age?

Question D10–4

Do you know a simple way to see whether the association with weight in the younger age group is confounded by age?

Table D10–1. Association of Diastolic Pressure With Age and Weight in Two Age Groups: Correlation Coefficients

Age Group (yr)	Correlation With Age	Correlation With Weight
30–44	0.24*	0.36*
≥45	0.00	0.24*

*$P < .01$.

Table D10–2. Relationship of Melanoma Mortality to Latitude

	Correlation Coefficient	Regression Coefficient of Mortality on Latitude (Deaths per Million)*
Male	−0.79	−0.056 (0.044–0.068)
Female	−0.72	−0.034 (0.026–0.042)

*95% confidence intervals shown in parentheses.

Question D10–5

The association between malignant melanoma and geographical latitude was examined, using the age-standardized mortality rates from melanoma in 1950–1967 in the states of the United States and the provinces of Canada, and the latitude of the largest city in each state or province (Elwood et al., 1974). Are the results (Table D10–2) consistent with the hypothesis that exposure to sunlight plays a part in the etiology of malignant melanoma (as it does in other skin cancers)? Do you know how to calculate what proportion of the variation in melanoma mortality can be explained by the association with latitude?

Question D10–6

What do the regression coefficients in Table D10–2 tell us? Does sex have a statistically significant modifying effect?

Question D10–7

In a follow-up study of a population sample in Wales, it was found that between 1957 and 1966 the mean height of a sample of men aged 25–34 (in 1957) declined by 2.24 cm, whereas the mean height of men aged 55–64 declined by 3.13 cm (Cole, 1974). The difference between these differences (0.89 cm) was highly significant ($P < .001$). What association is measured by the difference between the differences?

Question D10–8

In this Welsh study, there was apparently an error in the measurement of height in 1966, when the measuring pole was fitted to the wall in the wrong place—about 2.5 cm too high—so that the measured heights were lower than the true values. How would this error affect the difference between the differences in the two age groups?

Table D10–3. Purchase of Raw Milk by Cases and Matched Controls

	Purchased		Not Purchased		Total	
	No.	%	No.	%	No.	%
Cases	51	67	25	33	76	100
Controls	29	38	47	62	76	100

Question D10–9

During an investigation of an outbreak of gastroenteritis in a rural community, 76 patients and 76 controls (individually matched for age, sex, and street) were questioned about their food purchases and consumption (Tillett, 1986). Data on the purchase of raw (unpasteurized) milk are shown in two different ways in Tables D10–3 and D10–4. Make sure you understand the tables.

What was the reason for using matching? Which table makes fuller use of the information? Do you know how to calculate an odds ratio from these data? Do you know what significance tests you could use?

Notes

D10–1. The odds ratio can be used as an estimator of the incidence density ratio (the ratio of incidence rates using person-time denominators) in case-control studies in which new (incident) cases are compared with controls who at the time they are studied can be regarded as possible future cases, and in case-control studies based on existing (prevalent) cases (provided that the disease is not lethal and its duration is not affected by exposure). This assumes that the controls are drawn from the same source as the cases, and that they were selected independently of exposure status; the disease need not be rare. For an algebraic explanation, see Rothman and Greenland (1998, pp. 95–96).

D10–2. The odds ratio can be used as an estimator of the risk ratio (the ra-

Table D10–4. Purchase of Raw Milk by Cases and Matched Controls

	Cases		
Controls	Purchased	Not Purchased	Total
Purchased	19	10	29
Not purchased	32	15	47
Total	51	25	76

tio of cumulative incidence rates—i.e., incidence rates using number-of-persons denominators) if the disease has a low frequency. Selvin (1996, p. 205) suggests that "low" here means a rate of under 10% in each of the groups that are compared.

■■■■■■■■■■ Unit D11

Measures of Strength (Continued)

A *correlation coefficient* (r) measures the linear relationship between two variables. A coefficient of 1 means that a higher value of one variable is always associated with a higher value of the other, and a coefficient of -1 means that a higher value of one is always associated with a lower value of the other. A zero coefficient means there is no association between the variables (*Question D10–1*). The correlation coefficient does not indicate *how much* each variable changes when the other changes; this is what a regression coefficient tells.

The best way of assessing the strength of a correlation is to calculate r^2, which expresses the proportion of the variance of each variable that is "explained" by its linear relationship with the other. The values of r^2, based on the values of r in Table D10–1, are 0.057, 0.130, 0, and 0.057, or (expressed as percentages), 5.7%, 13%, 0%, and 5.7%. The correlations are not strong.

In answer to *Question D10–2*, the correlations of blood pressure with both age and weight appear to be modified by age, since the coefficients differ in the two age groups. The correlations with age are significantly different from each other, but we do not know whether the differences between the correlations with weight are larger than might easily occur by chance: the P values refer to differences from zero, not to the differences between the coefficients.

The exclusion test for possible confounding (Unit D5) indicates that the correlation between blood pressure and weight in the older age group (*Question D10–3*) is not confounded by age (because age is not correlated with blood pressure in this group).

A simple way to see whether the association with weight in the younger age group is confounded by age (*Question D10–4*) is to compute a *partial correlation coefficient*, which expresses the linear association between two variables (blood pressure and weight) when a third variable (age) is held constant. Its calculation is based on the correlation coefficients among the three variables. In this instance, we do not know the correlation between age and weight.

In *Question D10–5* the correlations between melanoma mortality and latitude are fairly strong, and are negative. The higher the latitude (i.e., the farther from the equator and the less the exposure to sunlight), the lower the mortality. The findings are thus consistent with the hypothesis that sunlight is a cause of

this disease. The square of the correlation coefficient tells us what proportion of the variation (variance) of one variable can be explained by the linear correlation with the other; for males this is $(-.79)^2$, or 62%; for females it is 52%.

A *regression coefficient* tells us the mean change in one variable when there is a change of one unit in the other. The answer to *Question D10–6* is that an increase of one degree in latitude is associated, on average, with a decrease in melanoma mortality of .056 per million (in males) and .034 per million (in females). The statistical model is the linear regression equation $y = a + bx$, in which y is the melanoma mortality rate, x is the latitude, a (the *intercept*) is the value of y when x is zero, and b is the regression coefficient of the mortality rate on latitude. If melanoma mortality rates are plotted against latitude on graph, the correlation coefficient measures how close the points are to a straight line, and the regression coefficient b measures the slope of this line.

The regression coefficients are different in the two sexes (Table D10–2), and there is no overlap between their confidence intervals, clearly showing that sex has a statistically significant modifying effect on the regression coefficients. (Could there be a statistically significant modifying effect if there was an overlap between the confidence intervals? See Note D11.)

In answer to *Question D10–7,* the difference between the differences observed between 1957 and 1966 in the two age groups is an index of the association between age and the change in height. The systematic error in measurement (*Question D10–8*) does not bias this association. The error can be corrected by adding 2.50 cm to all 1966 heights; the mean changes are then + 0.26 cm (25–34 years) and − 0.63 cm (55–64 years), and the difference between the differences is still 0.89 cm.

Matched Samples

When a matching procedure is used in the selection of samples that are to be compared, the purpose is to prevent confounding. If these samples (cases and controls, in *Question D10–9*) are similar with respect to certain variables, these variables cannot have a confounding effect.

The samples may be selected by choosing individuals who are similar in defined respects (individual matching), or just by ensuring that the groups as a whole are similar in certain respects (group matching). When individual matching is used, the findings are best tabulated as in Table D10–4, where each entry represents a *pair* of observations: it indicates the findings for each member of the pair (both members consumed raw milk, neither did, etc.). This table makes fuller use of the information than does a table like Table D10–3, which shows the data as if the two samples were unrelated. The observations in a table like Table D10–4 need not relate to cases and controls. They may, for example, relate to matched pairs whose one member was exposed and the other not exposed to a suspected risk factor, or to paired observations (e.g., before and after treatment) in the same subjects. We used this sort of table when we compared the diagnoses of two ophthalmologists who examined the same eyes (Table C7–1).

Table D11–1. Death Rates From Suicide, United States 1996–98 (Age-Standardized Rates per 100,000), With Rate Differences and Rate Ratios

	Rate		Difference	Ratio
	Black	White	(Black—White)	(Black : White)
Male	11.2	18.6	−7.4	0.60
Female	1.9	4.4	−2.5	0.43
Difference (male − female)	+9.3	+14.2		
Ratio (male : female)	5.9	4.2		

In such studies the odds ratio is the ratio of the two numbers of pairs with discrepant findings (Rothman and Greenland, 1998, p. 286). In Table D10–4, the discrepant pairs are those whose one member purchased raw milk whereas the other did not. There were 32 such pairs in which it was the case who bought raw milk, and 10 in which it was the control. The odds ratio is 32/10, that is, 3.2, or 10/32, that is, 0.31. The appropriate significance test, which uses the same two numbers, is a McNemar test or an exact binomial probability test.

Exercise D11

This exercise deals with synergism.

Table D11–1 shows death rates from suicide in the United States in 1996–1998 (National Center For Health Statistics, 2000), by race and sex. It also shows rate differences and rate ratios, as two measures of the strength of the associations with race and sex.

Question D11–1

Does Table D11–1 show effect modification?

Table D11–2. Effects of Race and Sex on Death Rate From Suicide: Rate Differences

	Black	White
Male	+9.3	+16.7
Female	0*	+2.5

*Reference category.

Table D11–3. Effects of Race and Sex on Death Rate From Suicide: Rate Ratios

	Black	White
Male	5.9	9.8
Female	1.0*	2.3

*Reference category.

Question D11–2

Table D11–2 shows the strength of the same associations by comparing each mortality rate with the rate in Black females (the group with the lowest rate). It shows the rate differences. Is there evidence of a synergistic effect on the death rate from suicide? That is, is the effect of being both male and White greater than the combined separate effects of being male and being White?

Question D11–3

Table D11–3 again shows the strength of the associations, this time in terms of rate ratios. Is there evidence of a synergistic effect in this table?

Question D11–4

Table D11–4 shows lung cancer death rates by smoking habits and occupational exposure to asbestos. It is based on a large study in the United States (Hammond et al. 1979). Do smoking and exposure to asbestos have a synergistic effect on the risk of the disease? (You will find it helpful if you first construct tables like Tables D11–2 and D11–3, showing the strength of the associations with the incidence rate.)

Table D11–4. Age-Standardized Death Rates (per 100,000 Man-Years) From Lung Cancer, by History of Cigarette Smoking and Occupational Exposure to Asbestos Dust

	Exposure to Asbestos	
Cigarette Smoking	No	Yes
No	11.3	58.4
Yes	122.6	601.6

Question D11–5

Why is synergism based on rate ratios worth detecting?

Question D11–6

Why is synergism based on rate differences worth detecting?

Note

D11. The difference between two values can be statistically significant even if there is some overlap between their separate confidence intervals. When two values are compared, more informative confidence intervals are those of their difference and their ratio.

■■■■■■■■■■ Unit D12

Synergism

Table D11–1 shows that the strength of the association between the death rate from suicide and race differs in men and women (whether use is made of rate differences or rate ratios), and the strength of the association between the death rate and sex differs in Blacks and Whites. Thus in answer to *Question D11–1*, there is clear evidence of effect modification: there is interaction between race and sex in their effects on the death rate from suicide.

Synergism refers to *positive interaction*—a situation where the joint effect of two or more factors is greater than their combined separate effects. (Sometimes the use of the term is confined to situations where the factors act together in a biological or mechanistic sense.) *Question D11–2* refers to the absolute differences connected with race and sex. The separate effect of being male is to increase the death rate (in comparison with the rate of Black females) by 9.3 per 100,000 (Table D11–2). The separate effect of being White is to increase the rate (again in comparison with the rate of Black females) by 2.5 per 100,000. A combination of these factors may therefore be expected to raise the rate to a value that is higher than the rate in Black females by (9.3 + 2.5), or 11.8 per 100,000. In fact, the rate was higher by 16.7 per 100,000. The findings therefore indicate a synergistic effect.

This conclusion is based on an *additive* model, wherein effects are measured as rate differences, and combined by adding them to one another.

In *Question D11–3* we use a *multiplicative* model: effects are measured as ratios, and must be combined by multiplying them by one another. Table D11–3 shows that being male multiplies the rate (of Black females) by 5.9 and being

Table D12–1. Effects of Smoking and Exposure to Asbestos on Lung Cancer Deaths: Rate Difference

	Exposure to Asbestos	
Cigarette Smoking	No	Yes
No	0.0*	+47.1
Yes	+111.3	+590.3

*Reference category.

white multiplies the rate by 2.3. The predicted combined effect is to multiply the rate by (5.9 × 2.3), or 13.6. In fact, the rate in White males was only 9.8 times the rate in Black females. Using this model, there is no synergism.

The data on smoking and asbestos (*Question D11–4*) yield a similar conclusion. When rate differences are examined (Table D12–1), the joint effect of these factors on lung cancer deaths is an increase of 590.3 per 100,000 person-years, which is greater than the combined separate effects (47.1 + 111.3 = 158.4). But when rate ratios are examined (Table D12–2), the joint effect is a 53.2-fold increase, which is less than the combined separate effects (5.2 × 10.8 = 56.2). There is synergism only if an additive model is used.

The occurrence of multiplicative synergism (*Question D11–5*) has etiological implications, and may provide useful clues to causal processes. Additive synergism (*Question D11–6*) is meaningful if we are interested in the absolute magnitude of a public health problem or an individual's risk. In the instance of asbestos and smoking, the findings provide no clue to etiological processes, but the fact that asbestos workers who smoke have especially high lung cancer death rates has obvious practical implications.

Table D12–2. Effects of Smoking and Exposure to Asbestos on Lung Cancer Deaths: Rate Ratio

	Exposure to Asbestos	
Cigarette Smoking	No	Yes
No	1.0*	5.2
Yes	10.8	53.2

*Reference category.

Table D12–3. Use of Oral Contraceptive ("Pill") by Women With Myocardial Infarction (MI) and Controls (Ctl)

Pill	MI	Ctl
Yes	29	135
No	205	1,607

Odds ratio: 1.7 (95% confidence interval, 1.1–2.8). *P* (by chi-square test) = .011.

The fact that we found effect modification with respect to one measure of an association (the rate difference) but not with respect to another (the rate ratio) should not surprise us. *Whenever we examine modifying effects—or, for that matter, confounding effects—our findings relate to a specific measure of association that we have chosen as a suitable one for our purposes.* If we use a different measure, we may come to different conclusions.

Exercise D12

This exercise introduces a procedure commonly used in the appraisal of associations when stratified data are available.

The association of oral contraceptives with myocardial infarction was investigated in a case-control study in 155 hospitals in the United States (Note D12). The cases were women admitted to a coronary-care unit for a first episode of definite myocardial infarction (MI) and the controls were women who had never had a myocardial infarction. The women, who were aged 25–49 and premenopausal, were asked whether they had used oral contraceptives in the previous month. The crude findings are shown in Table D12–3, and the findings stratified by age appear in Table D12–4.

Question D12–1

Is the association between oral-contraceptive use and MI confounded by age?

Table D12–4. Use of Oral Contraceptive ("Pill") by Women with Myocardial Infarction (MI) and Controls (Ctl), by Age

	25–29 yr		30–34 yr		35–39 yr		40–44 yr		45–49 yr	
Pill	MI	Ctl	MI	Ctl	MI	Ctl	MI	Ctl	MI	Ctl
Yes	4	62	9	33	4	26	6	9	6	5
No	2	224	12	390	33	330	65	362	93	301
Odds ratio	7.2		8.9		1.5		3.7		3.9	

Question D12–2.

Is the association between oral-contraceptive use and MI modified by age?

Question D12–3

Can you suggest a simple way of using the data in Table D12–4 to obtain a single odds ratio that circumvents possible confounding by age?

Note

D12. This exercise is based on data from Shapiro et al. (1979), using the Cornfield-Gart procedure (Fleiss, 1981) for confidence intervals and heterogeneity tests. The same example is treated in more detail by Schlesselman (1982).

■■■■■■■■■■ Unit D13

Appraising Stratified Data

The discrepancy between the findings based on crude and age-stratified data is clear evidence of confounding by age (*Question D12–1*). The odds ratio that expresses the strength of the association between "the Pill" and myocardial infarction is 1.7 in the sample as a whole, but much higher than this in all but one of the age strata.

There is also evidence that the association is modified by age (*Question D12–2*), because the odds ratios in the various age strata differ. The differences may, however, be due to sampling variation (Note B3–2). We can, if we wish, do a significance test to determine the probability that heterogeneity of this degree might occur by chance (see Note D13–1). If we do so, we will find that $P = .17$; that is, there is no significant heterogeneity.

The odds ratios in the separate age strata are not confounded by age, as the strata have such narrow age spans (5 years) that there cannot be much age variation within them. Therefore, if (in answer to *Question D12–3*), we can combine the stratum-specific odds ratios to obtain some sort of average, this too will be an odds ratio that is not affected by age confounding. The method most often used for this purpose is the Mantel-Haenszel procedure (Note D13–1), which in this instance provides a value of 4.0—much higher than the crude odds ratio of 1.68. This value, 4.0, is a *point estimate* of the common odds ratio; the confidence interval is 2.4–6.7. Unlike standardization, this and similar procedures do not require the use of a standard reference population. The Mantel-Haenszel chi-square test, which is often used to test the significance of an asso-

Table D13–1. Common Odds Ratio (Controlling for Age) Based on Data in Table D12–4

Estimator	Odds Ratio	95% Confidence Interval
Mantel-Haenszel	3.97	2.43 to 6.49
Maximum-likelihood		
Conditional	3.98	2.34 to 6.65 (Fisher's)
		2.41 to 6.48 (mid-*P*)
Unconditional	4.00	2.37 to 6.71 (Cornfield-Gart)

$P < .00001$ by Mantel-Haenszel test, Fisher's exact test, and exact mid-*P* test.

Heterogeneity test: $P = .172$.

ciation when effects connected with suspected confounders are controlled, yielded a *P* of less than one in a million.

A procedure that pools the stratum-specific findings in this way provides an odds ratio that controls for possible confounding by the stratifying variable. This may be regarded as the common "underlying" odds ratio, in instances where the absence of significant variation between the findings in the various strata makes this an acceptable concept. Stratification is widely used to control for confounding effects on other measures, as well as the odds ratio. For example, extensions of the Mantel-Haenszel method and similar procedures can compute and test estimates of rate and risk ratios, differences between rates, kappa, and hazard ratios (based on Kaplan-Meier life table analysis).

When the findings are clear-cut, different statistical procedures for the analysis of stratified categorical data (Note D13–1) generally yield similar results (Kahn and Sempos, 1989, chap. 9), as is illustrated in Table D13–1. (Do you know what *Fisher's* and *mid-P* are? See Note D13–2.)

The data may be stratified by two or more variables. Each of the five age strata in Table D12–4, for example, may be divided into three cigarette-smoking categories, producing 15 two-by-two tables, and the Mantel-Haenszel procedure can be applied to these. When this is done, the common odds ratio is 3.3. (What does this tell us? For answer, see Note D13–3.)

The data can also be rearranged so as to study a different independent variable. For example, we could stratify the same data by age and the use of oral contraceptives, and then use the Mantel-Haenszel procedure to examine the association between smoking and myocardial infarction (controlling for the other variables).

Making Sense of a Multivariate Analysis

The last three exercises in Section D are devoted to multivariate analysis. Multiple linear regression analysis, multiple logistic regression analysis, and proportional hazards regression analysis will be used as illustrations.

As was stressed in Unit D7, a basic general understanding of multivariate pro-

cedures (see note D7–2) is an essential condition for their intelligent use. The following brief notes are no substitute for this, but serve only as reminders of some salient features. If at present you know nothing at all about these procedures, you should probably leave these exercises until you do (go to Unit D18).

Multivariate analysis looks at a number of variables at the same time (generally in relation to a single dependent variable), using a mathematical model to represent the processes being studied. The model may be additive or multiplicative (using these terms as they were defined in Unit D12).

Multivariate analysis has two main purposes in epidemiology:

- For appraising the strength and significance of the relationships of a number of variables (separately or jointly) with the dependent variable, paying attention both to the variables' "main effects" and to their interactions (modifying effects). The relationship of each independent variable with the dependent variable can be examined while controlling for effects connected with other variables, by holding the other variables constant in the analysis. Multivariate analysis is a way of controlling for confounding.
- For making predictions of risk, based on the effects of multiple factors.

Multiple linear regression analysis, which generally has a metric-scale dependent variable, is based on an additive model:

$$y = a + b_1x_1 + \ldots + b_kx_k$$

where y is the predicted value of the dependent variable. In this and the subsequent formulae, the independent (predictor) variables are numbered from 1 to k, k being the number of independent variables; each b is the coefficient (estimated from the data) by which the value x of the corresponding variable is multiplied; a (the intercept) is a constant estimated from the data.

Multiple logistic regression uses a model that is multiplicative with respect to odds (it is additive with respect to log odds; adding the logarithms of numbers is the same as multiplying the numbers). The variable of interest is generally a disease or other "yes–no" characteristic. The model is expressed in terms of the log odds of the disease (i.e., the natural logarithm of the predicted odds in favor of the disease):

$$\text{log odds of disease} = a + b_1x_1 + \ldots + b_kx_k$$

The variables whose values are represented by x may be categorical or metric-scale. If the variable is dichotomous, the values 0 for "absent" and 1 for "present" are commonly used. If the variable has more than two categories, one is generally designated as a reference category, and the others become "dummy variables." For example, if there are three categories of cigarette-smoking ("none," "moderate," and "heavy"), two of these would appear as variables in the model—probably "moderate" and "heavy"—each of them being scored (say) 0

for "not in this category" or 1 for "in this category." The probability of the disease can be estimated by the formula

$$\text{Probability of disease} = 1/\{1 + \exp[-(\text{Log odds of disease})]\}$$

Proportional hazards regression analysis (regression using the Cox proportional hazards model), which appraises relationships with survival, is used for time-to-event data (see Unit B9); it can cope with censored data. The procedure may be univariate, appraising the relationship of one variable with survival, or multivariate, appraising several variables. An important assumption is that the relationships with survival do not vary with time; that is, if at one time smoking doubles the risk of occurrence of the event, it should do so at other times also. The model is expressed in terms of the *hazard function,* which can be interpreted as the risk of the event at any given time:

$$\text{Log of hazard} = \log(a) + b_1x_1 + \ldots + b_kx_k$$

The model is additive with respect to the log of the hazard, and multiplicative with respect to the hazard. The probability of survival (i.e., nonoccurrence of the event) up to any specified time t can be estimated by the formula

$$\text{Probability of survival} = \exp[-\exp(\log(H_t) + b_1x_1 + \ldots + b_kx_k)]$$

where H_t is the cumulative underlying hazard function at time t, estimated from the data.

In the above formulae, the regression coefficient b expresses the strength of the association of a specific variable x with the dependent variable when the other variables in the model (often called *covariates*) are held constant. In a multiple linear regression analysis, it is similar to the simple regression coefficient we encountered in Unit D12. It "indicates an average change in y for a unit change in x_i after their linear association with all other x variables has been removed from both y and x" (Kahn and Sempos, 1989). In a multiple logistic analysis, the coefficient b is the natural log of the odds ratio; the exponential ("antilog") of b is the odds ratio for the variable's association with the disease, adjusted for effects connected with other variables; this odds ratio indicates the change in the disease odds when there is a change of one unit (e.g., from 0 to 1) in the independent variable. In a proportional hazards regression analysis, the coefficient b is the natural log of the hazard ratio; its exponential (the "antilog" of b) is the hazard ratio expressing the effect of the variable, adjusted for the effects of other variables. This hazard ratio, or "relative risk," indicates the change in the risk of the event when there is a change of one unit (e.g., from 0 to 1) in the independent variable. For a dichotomous variable (given values of 0 and 1) it is analogous to the hazard ratio provided by the Kaplan-Meier life table method (Note B9–4), except that it is adjusted for the effects of the other variables in the model.

Once the coefficients are available, the effects of a specific constellation of factors can be estimated by inserting the appropriate values of each x in the formula and calculating y (for linear regression), the log of the odds or the probability of the disease (for logistic regression) or the log of the hazard or the probability of survival (for proportional hazards regression). The analysis generally provides P values and standard errors or confidence intervals for the b coefficients. The P values indicate whether the coefficients are significantly different from zero—that is, whether the relevant association with the dependent variable (controlling for effects connected with other variables) is statistically significant.

A multivariate analysis may include additional terms that express interactions of specified variables.

There may be information on the validity of the multivariate model; without this, use of the findings may be subject to reservations. Methods of appraising validity are available, but they often receive no mention in the reports of studies that use multivariate analysis. The validity of an equation for estimating y or the probability of a disease or of survival is most convincing if the model is developed and tested in one sample (or part of a sample) and then retested in another.

In multiple linear regression, a crude indication of the validity of the model is provided by R^2 (the square of the multiple correlation coefficient R), which is the proportion of the variation (variance) of the dependent variable that is explained by the total set of independent variables. For a fuller appraisal, the observed values of the dependent variable can be compared with those predicted by the regression equation (see Note D13–4).

In multiple logistic regression, simple comparison or a goodness-of-fit test can be used to see how well the values predicted by the regression equation conform with observed data (Kahn and Sempos, 1989, pp. 151–153), as we will see in the coming exercises. Also, the analysis generally provides a likelihood-ratio chi-square statistic that can indicate the suitability of the model (Note D13–5). Other indices are also available (Note D13–6). By trying models that include smaller or larger sets of variables and interactions, and comparing the above indices, it is possible to see whether specific variables or interactions contribute appreciably to the validity of the model.

Appraisal of the appropriateness of a proportional hazards model is not easy (see Note D13–7).

Exercise D13

Table D13–2 shows some results of a multiple logistic regression analysis of the same study of oral contraceptives and myocardial infarction (MI) that we looked at in the last exercise.

Question D13–1

Explain in words the meaning of the figure 8.47 in Table D13–2; do you know how this figure was obtained?

Table D13–2. Associations With Myocardial Infarction: Multiple Logistic Regression Analysis*

Variable	Coefficient	S.E.†	*P*	Odds Ratio (With 95% Confidence Interval
Oral contraceptive (0 = no, 1 = yes)	1.188	0.206	.032	3.28 (1.97–5.47)
Age (years)	0.152	0.014	.0010	1.16 (1.13–1.20)
1–24 cigarettes/day (0 = no, 1 = yes)	1.125	0.209	.020	3.08 (2.04–4.64)
≥25 cigarettes/day (0 = no, 1 = yes)	2.137	0.208	.0013	8.47 (5.64–12.74)
Constant	−9.283	0.629		

*Likelihod ratio statistic (4 degrees of freedom): 272.8.

†Standard error of coefficient.

Question D13–2

Which is more strongly associated with MI: age or taking oral contraceptives?

Question D13–3

Do the results in Table D13–2 tell us whether the association between the Pill and MI is confounded by smoking? If not, what extra information do you need? (Can you guess what the likelihood ratio statistic tells us? See Note D13–8.)

Question D13–4

Do the results in Table D13–2 tell us whether the association between the Pill and MI is modified by smoking—that is, whether this association is the same in nonsmokers and women who smoke various numbers of cigarettes per day? If not, what extra information do you need?

Question D13–5

According to the results in Table D13–2, what (controlling for effects connected with age) is the ratio of the odds in favor of MI among women who use oral contraceptives and smoke ≥25 cigarettes a day, to the corresponding odds among women who do neither?

Notes

D13–1. Methods for testing significance and estimating a common odds ratio, rate ratio, or rate difference from stratified data include the Mantel-Haenszel, precision-based, and maximum-likelihood procedures. The measure com-

puted by these methods is variously called the common, underlying, overall, summary, pooled, or uniform measure. In this book the term "Mantel-Haenszel" refers not only to the original Mantel-Haenszel procedure for odds ratios, but to similar procedures now used for rate and risk ratios and differences and other extensions of the Mantel-Haenszel procedure (Landis et al., 2000). All these methods for estimating a common measure, as well as methods for testing the heterogeneity of the findings in the various strata, are described by Rothman (1986, chap. 12) and Rothman and Greenland (1998); methods using the odds ratio are explained by Fleiss (1981, chap. 10); formulae are summarized by Kleinbaum et al. (1982, pp. 359–361). See note A3–7.

D13–2. "Exact tests" are defined by the *Dictionary of Epidemiology* (Last, 2001) as tests "based on the actual null probability distribution of the study data, rather than, say, normal approximation." These tests and their corresponding confidence intervals may be especially appropriate if data are sparse. The usual procedure (*Fisher's*) produces conservative results, and many experts prefer the *mid-P* procedure, which yields lower *P* values and narrower confidence intervals (Berry and Armitage, 1995).

D13–3. A Mantel-Haenszel odds ratio of 3.3 when the data are stratified by age and smoking habits tells us how strong the Pill–MI association is when age and smoking are controlled; it also tells us that this association was to some extent confounded by smoking, since the value is now lower than it was when only age was controlled (4.0).

D13–4. For methods of examining discrepancies between observed values and those predicted by a multiple regression equation, see (e.g.) Kahn and Sempos (1989, pp. 140–143) or Altman (1991, pp. 346–347).

D13–5. The chi-square statistic for a multiple logistic regression analysis may test how well predictions based on the model fit the observed data. The goodness-of-fit chi-square test provided by the SPSS logistic regression program is an example. A low *P* value (say $< .05$) indicates a poor fit; the higher the *P* value is, the more confidence we can have in the model's validity. The same interpretation is given to a chi-square statistic that appraises how well the findings estimated from the logistic coefficients fit with the observed findings. The "− 2 log likelihood" chi-square provided by the SPSS program is an example. On the other hand, a chi-square statistic may be used that tests whether the independent variables, considered jointly, are associated with the dependent variable; in this instance, a *low P* value points to the model's validity. The "model chi-square" provided by the SPSS program is an example. The contribution of specific variables or interactions to the model's validity can be appraised by doing the analysis with and without them, and comparing the chi-square values. The difference between the chi-square values—sometimes called the "partial chi-square"—tests the significance of the effect of these added variables or interactions (using the difference between the degrees of freedom in the two analyses).

D13–6. In multiple logistic regression analysis, the square of the correlation coefficient between the observed values of the dependent variable (0 or 1 = "no" or "yes") and the probability (of "yes") predicted by the logistic equation is an

estimate of the proportion of the variation of the dependent variable that is explained by the independent variables (Mittlboeck and Schemper, 1996). The "pseudo-R^2" value that is often provided by logistic regression programs may be helpful, although it is not really a measure of goodness-of-fit (Selvin, 1996, p. 266).

D13–7. To appraise the suitability of a proportional hazards model, a suggested first step is to plot and compare "log-minus-log" curves for different subgroups of subjects (e.g., cases and controls, or cases and controls with high and low blood pressures). The values to be plotted against time are transformations of the survival probabilities predicted by the model; for each survival probability S, the transformed value to be plotted is $\log[-\log(S)]$. The suitability of the model may be questioned if the curves are not more or less parallel (Selvin, 1996, pp,. 388–400; McNeil, 1996, pp. 213–216). Some computer programs offer log-minus-log plots as an option.

D13–8. A likelihood-ratio statistic is a kind of chi-square statistic. As explained in Note D13–5, different chi-squares are used for testing multiple logistic models. In this instance, chi-square = 272.9 with 4 degrees of freedom, so that $P < .000001$. If this were a goodness-of-fit test, it would indicate a very poor fit. It is actually a test of the association between the Pill, age and smoking (considered together), and myocardial infarction, and the null hypothesis (of no association) can be rejected.

■■■■■■■■■■ Unit D14

Multiple Logistic Regression

In answer to *Question D13–1,* the odds ratio of 8.47 is the odds ratio when women who smoke ≥25 cigarettes a day are compared with women who smoke none (i.e., the ratio of the odds in favor of MI among women who smoke ≥25 cigarettes day to the odds in favor of MI among women who smoke none), with the other variables (age and oral contraceptives) held constant. Alternatively, it is the ratio of the odds in favor of smoking ≥25 cigarettes (rather than none) among women with MI, to the odds in favor of smoking ≥25 cigarettes among women without MI (you will remember from Unit B11 that the disease odds ratio and exposure odds ratio are identical). The figure was obtained by taking the exponential (antilog) of the coefficient 2.137; $e^{2.137}$ is 8.47.

The coefficient and odds ratio for age express the effect of a 1-year difference in age when the other variables in the analysis remain unchanged. A comparison of these values with those for oral contraceptives, as requested in *Question D13–2,* is meaningful only if a specific age difference is stated. For a 20-year dif-

ference, for example, the coefficient 0.152 may be multiplied by 20 to obtain 3.04. This is the natural log of 20.9, so the appropriate odds ratio for comparison with that for oral contraceptives (3.28) would be 20.9. The *P* values can, of course, not be used to measure the strength of the associations.

The odds ratios in the table are adjusted for effects connected with smoking. The only way to tell whether the association between the Pill and MI is confounded by smoking (*Question D13–3*) is to compare the findings with those when smoking is *not* controlled in the analysis. We could do another analysis, excluding smoking from the list of variables. This hardly seems worth doing, as we have already controlled for possible confounding.

The table tells us nothing about modifying effects (*Question D13–4*). We can examine the modifying effect of smoking on the association between the Pill and MI by repeating the analysis after adding a term or terms that express the interaction of smoking and the Pill. We can then see how this changes the findings (we will do this in the next exercise), and can appraise the strength and significance of the interaction effect. Alternatively, we could conduct separate analyses in nonsmokers and moderate and heavy smokers, using only oral contraceptives and age as independent variables, and compare the strength of the associations shown in the three analyses.

The multiple logistic model is a multiplicative one, in the sense that we obtain the odds ratio for a combination of two factors (*Question D13–5*) by multiplying their separate odds ratios. The odds ratio for use of the Pill is 3.28, and the odds ratio for smoking ≥25 cigarettes a day is 8.47. The odds ratio for both factors together is therefore 3.28 × 8.47, or 27.8.

Exercise D14

Question D14–1

Logistic regression models that included different sets of variables yielded different odds ratios for the association between oral contraceptives and MI, as shown in Table D14–1. How do you account for this? Compare the figures in the table with the corresponding Mantel-Haenszel odds ratios (Unit D13).

Table D14–1. Odds Ratios Expressing Association Between Oral-Contraceptive Use and Myocardial Infarction in Three Logistic Regression Models

Variables Included in Model	Odds Ratio
Oral contraceptive	1.68
Oral contraceptive, age	3.81
Oral contraceptive, age, cigarettes	3.28

Table D14–2. Age-Adjusted Odds Ratios Expressing Association Between Use of Oral Contraceptives and MI, by Contraceptive Use and Smoking Habits: No-Interaction Model

	Oral Contraceptives	
Cigarettes/Day	No	Yes
None	1.0	3.6
1–24	3.3	10.1
≥25	8.5	27.8

Question D14–2

When contraceptive–cigarette interaction is included in the logistic model used in Table D13–2 (i.e., in addition to contraceptives, age, and cigarettes), the overall validity of the model (as appraised by likelihood-ratio chi-square statistics) does not change significantly, and the coefficients for the interaction terms are not statistically significant. The odds ratios for the Pill–MI association are different, however, from those based on the no-interaction ("main effect") model. Odds ratios based on the two models are shown in Tables D14–2 and D14–3. In their summary of their results, the investigators say that the combined effect of oral contraceptives and smoking

> was appreciably larger than could be accounted for by the separate effects of cigarettes and oral contraceptives, and this suggests a considerable accentuation by cigarette smoking of the effect of oral contraceptive use on myocardial infarction. (Shapiro et al., 1979)

Do the results of the multiple logistic analyses support this conclusion?

Table D14–3. Age-Adjusted Odds Ratios Expressing Association Between Use of Oral Contraceptives and MI, by Contraceptive Use and Smoking Habits: Interaction Model

	Oral Contraceptives	
Cigarettes/Day	No	Yes
None	1.0	3.6
1–24	3.1	3.7
≥25	8.0	40.3*

*Calculated by multiplying the odds ratios for contraceptives (3.6), ≥25 cigarettes (8.0), and their interaction (1.4).

Table D14–4: Odds Ratios for Relationships of Low Social Class and Low Educational Level With Obesity in Four Logistic Regression Models: Imaginary Data

Variables Included in Model	Odds Ratio	
	Social Class	Education
Social class	0.30	—
Education	—	0.30
Social class, education	0.50	0.50
Social class, education, social class–education interaction	0.50	0.50

Question D14–3

The relationships of social class and educational level with obesity were examined in an imaginary population. Social class and education, which were treated as dichotomies ("low" and "high"), were strongly correlated; 90% of the people in the "low" category of one were also in the "low" category of the other, and 90% of those in the "high" category of one were also in the "high" category of the other. The results of logistic regression analyses are shown in Table D14–4. How can the differences be explained?

Question D14–4

For the purpose of this question, assume that Table D13–2 was based on a 10-year follow-up study of the incidence of MI in a representative population sample, so that it can be used as a basis for predictions of incidence (it cannot actually be so used). Do you know how to calculate the risk of having an infarction in the next 10 years, for a 30-year-old woman who uses oral contraceptives and smokes 30 cigarettes a day? How could we appraise the validity of the model as a predictor of risk?

■■■■■■■■■■ Unit D15

Multiple Logistic Regression (Continued)

Different logistic models may provide different odds ratios for the same association (*Question D14–1*) because the odds ratios express the strength of the association after controlling for effects connected with other variables in the model. The results thus vary, depending on what other variables are included. The

odds ratios in Table D14–1 are very close to the Mantel-Haenszel odds ratios, which were 4.0 (controlling for age only) and 3.3 (controlling for age and smoking).

Similarly, the addition of interaction terms may appreciably change the results, as Tables D14–2 and D14–3 show. It is probably wise to treat the results of any multiple logistic analysis with reserve if the possible importance of interactions (effect modification) has not been investigated. If interaction is unimportant, the results of a main-effect analysis will fit the data accurately and the meaning of the odds ratios will be straightforward. However, if there is important interaction and it is ignored, the results may be misleading (Note D15–1).

Question D14–2 is not easy to answer. The fuller model, including interactions, shows a definite synergistic effect. However, the interaction term was not statistically significant, so that we cannot be confident that this is not a chance finding. In a detailed discussion of this study, Schlesselman (1982) suggests that the interpretation based on the no-interaction model is preferable, because the analysis using the interaction model (Table D14–3) indicates that oral contraceptives increase the risk of MI markedly in nonsmokers and heavy smokers but not in moderate smokers, which is "biologically implausible"; there may be uncontrolled confounding factors.

In Table D14–4, we again see that the strength of an association in a logistic regression model may change when the model is changed. The specific answer to *Question D14–3* is that the inclusion of highly correlated independent variables in a single model may have a marked effect on the findings (this is referred to as *multicollinearity*). The associations with both social class and education became weaker (odds ratios closer to 1) when the other variable was included.

To use multiple logistic regression for predicting the probability of a disease, we must substitute the appropriate values in the equation. In this instance (*Question 14–4*) the log odds (the natural logarithm of the odds) in favor of myocardial infarction is

$$-9.283 + (1.188 \times 1) + (0.152 \times 30) + (1.125 \times 0) + (2.137 \times 1)$$

or -1.398. The risk of the disease is $1/[1 + \exp(1.398)]$, or $1/(1 + 4.047)$—that is, 0.198 or 19.8%.

The model's validity as a predictor of risk—that is, the degree to which the model conforms with observed facts—can be tested in the sample from which the coefficients were derived or (more convincingly) in other samples. One method is illustrated in Table D15–1, (from Kahn and Sempos, 1989). Each individual's probability of developing the disease was calculated from the model, the individuals were divided into quartiles according to their level of risk, and the predicted number of cases in each group was calculated (by adding together the probabilities of the members of the group) and then compared with the observed number. Does Table D15–1 show a good fit with the data? (For answer, see Note D15–2.) If we have them, we can, of course, also use the chi-square results described in Note D13–5.

Table D15–1. Fit of Multiple Logistic Risk Function to Data: Comparison of Predicted and Observed Incidence of Diabetes

	Cases of Diabetes	
Risk (Quartile)	Number Expected	Number Observed
1	72.1	70
2	31.3	28
3	19.5	23
4	10.5	10

Source: Data from Kahn et al. (1971).

Exercise D15

This exercise deals with *proportional hazards regression analysis* (Cox regression), which is used for time-to-event data.

Question D15–1

In a study of publication bias, clinical research projects submitted to a hospital ethics committee for approval between 1979 and 1988 were followed. "Significant" studies (those that subsequently yielded statistically significant results at $P < .05$) were compared with "nonsignificant" ones $P = .1$ or more). The proportions of these studies that were published by 1992 were 68% and 44%, respectively. The time lapse between committee approval and publication was analyzed by proportional hazards regression analysis (Table D15–2). The year of study approval, performance of the study as a degree requirement, and other variables that were not significantly associated with the hazard ratio (according to univariate analyses) were excluded from the multivariate analysis (Stern and Simes, 1997). Is proportional hazards regression analysis appropriate for this study?

Question D15–2

What happened to the unpublished studies in this analysis?

Table D15–2. Hazard Ratios for Publication, 146 Significant Studies Compared With 53 Nonsignificant Studies; Cox Regression

	Hazard Ratio for Publication
Univariate analysis	2.32 (95% C.I., 1.47 to 3.66)
Multivariate analysis	2.34* (95% C.I., 1.47 to 3.43)

*Adjusted for research design (observational study, clinical trial, other experiment) and funding by an external source.

Question D15–3

What does a hazard ratio of 2.32 mean?

Question D15–4

Were research design and external funding confounding factors?

Question D15–5

Could Kaplan-Meier life table analysis have been used instead of multivariate proportional hazards regression?

Question D15–6

A follow-up study of over 40,000 babies, a sample of those born in England and Wales between 1976 and 1997, showed that the lower the birth weight, the higher was the mother's risk of subsequently dying of cardiovascular disease. According to proportional hazards regression analysis, the hazard ratio was 2.26 (95% confidence interval [C.I.], 1.48 to 3.41) for a birth-weight difference of 1 kilogram (kg). Adjusting for socioeconomic status and marital status at birth (by including these variables in a multivariate analysis), the hazard ratio was 2.22 (95% C.I., 1.46 to 3.38) for a birth-weight difference of 1 kg (Smith *et al.* 2000). What information was required for the purpose of this analysis?

Question D15–7

In a comparison of smaller babies with those weighing 2 kg more at birth, how much higher was their mothers' risk of cardiovascular mortality (controlling for socioeconomic status and marital status)?

Question D15–8

What explanations can you suggest for the findings, which confirmed those of an earlier small-scale study? The hazard ratios for other important causes of death were considerably lower: 1.33 for deaths from cancer and 1.06 for accidental and violent deaths.

Question D15–9

About 4,000 children aged 16 or under whose dietary intake was studied in a survey in England and Scotland between 1957 and 1969 were followed until mid-1996 to identify deaths and causes of death (Frankel et al., 1998). Proportional hazards regression analysis showed positive associations between childhood energy intake and the risk of cancer. Which of the hazard ratios shown in Table D15–3 are statistically significant? Approximately what percentage increase in the risk of dying of cancers unrelated to smoking is associated with an increased energy intake of 1,000 kcal per day?

Table D15–3. Associations Between Childhood Energy Intake and Adult Mortality; Hazard Ratios* per 1 mJ/day (239 kcal/day)

Cause of Death	Hazard Ratio	95% C.I.
All causes	1.04	0.99 to 1.09
All cancers	1.15	1.06 to 1.24
Cancers not related to smoking	1.20	1.07 to 1.34
Cancers related to smoking **	1.09	0.86 to 1.23
Causes other than cancer	0.99	0.93 to 1.05

*Adjusted for age at time of survey, household food expenditure, social class, number of siblings, time since survey, and neighborhood deprivation score.

**Cancers of the lip, tongue, mouth, pharynx, esophagus, pancreas, and respiratory and urinary tracts.

Question D15–10

In a study in Finland, possible risk factors for myocardial infarction (MI) were studied in a population sample of about 3,000 middle-aged men free of coronary disease who were examined between 1983 and 1989 and followed until December 1992 to determine whether acute MI occurred (Nyyssonen et al., 1997). The hazard ratios that were statistically significant (by proportional hazards regression analysis) are shown in Table D15–4. Men with vitamin C deficiency, for example, had an increase risk of MI. According to these hazard ratios, which risk factor was most strongly associated with the risk of MI?

Notes

D15–1. For a detailed discussion of the impact of effect modification on the results of multiple logistic regression analysis, with examples, see Lee (1986).

Table D15–4. Associations Between Selected Risk Factors and Risk of Myocardial Infarction: Hazard Ratios

Risk Factor	Hazard Ratio*	95% C.I.	*P*
Pack-years of smoking**	1.40	1.15 to 1.70	.0008
Plasma vitamin C (<2 mg/L vs. >2 mg/L)	2.55	1.26 to 5.17	.0095
Maximal oxygen uptake (ml/min×kg)	0.65	0.47 to 0.92	.0137
Family history of coronary heart disease (yes vs. no)	1.86	1.14 to 3.02	.0129
Hair mercury (>2.0 μg/g vs. <2 μg/g)	1.68	1.01 to 2.81	.0448
Serum apolipoprotein B (g/L)	1.29	1.01 to 1.66	.0454

*Controlling for the other independent variables included in the analysis, namely the other risk factors listed in the table, 12 other possible risk factors, age, season, year of examination, and intakes of tea, fiber, and saturated fats.

**A measure of lifetime exposure to smoking.

D15–2. Yes (by visual inspection). This may be confirmed by a goodness-of-fit test (see Note F2–1). An appropriate test (Lemeshow and Hosmer, 1982; described simply by Schlesselman, 1982, p. 264) yields a high *P* value (.58), indicating that there is no significant difference between the observed and predicted distributions.

■■■■■■■■■■ Unit D16

Proportional Hazards Regression

Proportional hazards regression would seem to be well-suited to the analysis of the study described in *Question D15–1*. The follow-up periods for different research studies were started in different years and varied in length, so that it was preferable to compare the intervals until the studies were published, rather than only whether publication occurred. However, as pointed out in Unit D13, this procedure assumes that the hazard ratio remains constant at different times after the start of follow-up; but we have no evidence for this; the suitability of the model was apparently not appraised. We are told that the hazard ratio was not affected by the year in which follow-up started; but this is not quite the same thing. The results should therefore be treated with some reserve. Regrettably, this reservation also applies to the use of proportional hazards regression in the other studies cited in Exercise D15.

Proportional hazards regression analysis can handle censored data (*Question D15–2*). Data for all the studies, including the unpublished ones, were entered and taken into account in the analysis. For published studies, the time from approval until publication was entered; for unpublished studies, the time until the end of the follow-up period was entered.

A hazard ratio or "relative risk" of 2.32 (*Question D15–3*) means that the "risk" of publication was 2.32 times higher for significant studies than for nonsignificant ones, irrespective of the time lapse since approval of the study. This value was virtually unchanged (2.34) when research design and external funding were controlled in the analysis, so that (in answer to *Question D15–4*) these variables cannot be regarded as confounders.

Question D15–4 is not as simple as it seems. The investigators decided that the year of study approval was not a potential confounder, and they therefore excluded it from the multivariate analysis. But they apparently based this decision (as do many researchers) on the absence of a statistically significant association (between year of approval and the hazard ratio). But this can be misleading, because even large effects may be nonsignificant if sample size is small; it is generally preferable to base decisions about potential confounders on the strength of associations, rather than on their significance.

The Kaplan-Meier life table method is usually used for analyzing the survival of a single group. But if stratified data are entered, the procedure can combine the results to produce an overall result that controls for possible confounding by the stratifying variable or variables. The Kaplan-Meier procedure could therefore have been used in this study instead of multivariate proportional hazards regression (*Question D15–5*), by first stratifying the data by research design and funding. The log-rank test for the difference between survival curves can also be applied to stratified data.

Proportional hazards regression analysis requires a survival time (uncensored or censored) for each subject and also information on the independent variable or variables. In the study of babies' birth weights and their mothers' mortality (*Question D15–6*), what was required or each baby–mother pair was (a) the survival time (from the date of birth until the mother's death or, if she remained alive, until the end of the study—December 1997); (b) whether death from cardiovascular disease occurred (uncensored) or not (censored); and (c) information on birth weight, socioeconomic status, and marital status.

The adjusted hazard ratio was 2.22 for a birth-weight difference of 1 kg. Because the model is multiplicative, the hazard ratio for a birth-weight difference of 2 kg (*Question D15–7*) is 2.22 times 2.22, or 4.93.

The investigators suggested three possible explanations for the strong inverse relationship between birth weights and mothers' mortality from cardiovascular disease (*Question D15–8*): "First, poor social circumstances could lead to both lower birth weight and higher mortality risk. Secondly, maternal health, nutritional, and behavioural profiles could influence both birth weight and cardiovascular mortality. Thirdly, intergenerational factors—such as genomic and epigenetic processes that lead to a positive correlation between the birth weights of mothers and their offspring—could influence cardiovascular risk" (Smith *et al.*, 2000).

In the study of energy intake and cancer mortality (*Question D15–9*), the associations with all cancers and those not related to smoking were statistically significant; these were the only hazard ratios whose confidence intervals did not straddle 1. The association with all cancers was mainly attributable to the association with cancers not related to smoking, as associations with other cancers were weak. For an increased daily energy intake of 239 kcal, the hazard ratio for cancers unrelated to smoking was 1.20. For an increased daily energy intake of 1,000 kcal, which is about four times 239, the hazard ratio would be 1.2 times 1.2 times 1.2 times 1.2, which is about 2.07, or an increase of about 107%.

In answer to *Question D15–10*, the results shown in Table D15–4 do not permit a decision as to which one of the six risk factors has the strongest association with myocardial infarction. Obviously, the *P* values do not measure the strength of the associations. If the hazard ratios are compared (the ratio 0.65 would obviously have to be converted to its reciprocal, i.e., 1/0.65, or 1.54) the difficulty is that they are based on contrasts of different kinds—between categories, or between measurements with differences of 1 unit, using various scales of measurement (pack-years, ml/min×kg, or g/L). If the hazard ratio for smoking were

expressed per 3 pack-years, it would be 1.4 times 1.4 times 1.4, or 2.7; and if the hazard ratio for plasma vitamin C were expressed per mg/L, it would obviously be much less than 2.55.

Exercise D16

Multiple linear regression, with its simple additive model, is easier to use and understand than multiple logistic regression. We will take a single example. The indices used in this example are the regression coefficient *b* (see formula in Unit D13, p. 209) and the proportion of total variation (variance) explained by a variable or set of variables.

Data from the National Study of Health and Growth in England and Scotland were analyzed to appraise the relationship between parents' smoking and children's growth. Children, aged 5–11 years, in a stratified random sample were examined, and their parents were asked to fill in self-administered questionnaires. Information was available for 5,903 children out of 8,120 (Rona et al., 1985).

Question D16–1

The dependent variable in the analysis was the difference between the child's height and the mean height of children of the same age, sex, and country (England or Scotland), divided by the standard deviation for that group. It was denoted the *standard deviation score.* Why was this score, rather than the height itself, used as the dependent variable?

Question D16–2

The following independent variables were initially included in the multiple linear regression model. Why were variables *c* to *i* included?

a. Smoking at home: the sum of cigarettes currently smoked at home in a day, by the father and the mother; this was used as a measure of passive smoking by the child.
b. Smoking in pregnancy: the number of cigarettes smoked a day during the pregnancy with the given child.
c. Birth weight.
d. Father's height.
e. Mother's height.
f. Number of older siblings.
g. Social class (based on father's occupation).
h. Duration of pregnancy.
i. Household crowding index (number of persons per room).

Question D16–3

A multiple regression analysis that included a similar set of factors yielded a multiple correlation coefficient (R) of .56 (Rona et al., 1978). What does this tell us about the validity of the model?

Question D16–4

The proportion of the total variation in the child's height that was explained by parents smoking, according to two different regression models, is shown in Table D16–1. What does the discrepancy between the figures in the first two columns (totalled) and the third column tell us?

Question D16–5

What does the discrepancy between the figures in the two rows of Table D16–1 tell us? Can we always conclude that such discrepancies are due to confounding effects?

Question D16–6

Social class and duration of pregnancy were omitted from the analyses summarized in Table D16–1, on the grounds that "they did not explain a significant amount of variation in height." "Significant" may refer either to statistical significance, or to a "meaningful," "substantial," or "appreciable" effect. Which would be a more valid reason for omitting these variables?

Question D16–7

Regression coefficients expressing the relationship of parents' smoking to their children's height, based on four different linear regression models, are shown in Table D16–2. Explain what the coefficients tell us. ("What are the facts?")

Question D16–8

Can we conclude that smoking in pregnancy does not affect the child's height?

Table D16–1. Proportion of Variation in Height Explained by Parents' Current Smoking at Home, Mother's Smoking in Pregnancy, and Both These Factors Combined; Multiple Linear Regression

	Proportion Explained by:		
Variables Included in Model	Smoking at Home	Smoking in Pregnancy	Smoking at Home and in Pregnancy
Smoking at home, smoking in pregnancy	1.34%	0.67%	1.41%
Smoking at home, smoking in pregnancy, birth weight, father's and mother's height, number of older siblings, crowding index	0.23%	0.14%	0.26%

Table D16–2. Relationship of Parents' Smoking (Number of Cigarettes per Day) to Child's Height (Standard Deviation Score): Linear Regression Coefficients

	Smoking at Home		Smoking in Pregnancy	
Variables Included in Model	Coeffic.	*P*	Coeffic.	*P*
Smoking at home	−0.0099	<.001		
Smoking in pregnancy			−0.0122	<.001
Smoking at home, smoking in pregnacy	−0.0086	<.001	−0.0045	NS
Smoking at home, smoking in pregnancy, birth weight, father's and mother's height, number of older siblings, crowding index	−0.0034	<.01	−0.0028	NS

Question D16–9

What explanations can you suggest for the association between passive smoking and child's height?

Question D16–10

What use or uses does this study serve?

■■■■■■■■■■ Unit D17

Multiple Linear Regression

In Unit A15, we discussed the control of confounding by use of a dependent variable that incorporates, and neutralizes the effect of, the confounder. The illustrations included the use of the IQ, which is a test score expressed as a percentage of the average score of children of the same age in order to neutralize the effect of age. In *Question D16–1,* the replacement of height by its discrepancy from the mean height of children of the same age, sex, and country similarly obviates possible confounding by age, sex, and country. Dividing this discrepancy by the standard deviation to obtain a standard deviation score (often called a *z score*) takes this a step further, by controlling for the spread as well as the cen-

tral tendency of the distribution: the same discrepancy may have different meanings in narrow and wide distributions. (This method also has other statistical advantages.)

Regression analysis is sometimes used as a way of "purging" unwanted influences from a variable for this purpose. If we have a valid regression model for predicting blood pressure from age, sex, and other biological attributes, for example, we can calculate each subject's expected blood pressure and determine the discrepancy between the actual and predicted values. This discrepancy (the "*residual,*" or "what is left after the model is fit") is a measure that is not influenced by these biological attributes; using it as a dependent variable in other analyses will therefore control for confounding by these attributes.

Residuals may also be used to see how well a multiple regression model fits the observed facts. For example, Table D17 (from Kahn and Sempos, 1989) presents a simple test of a model that used age and weight to predict systolic blood pressure. (Would you conclude that the fit was good? See Note D17.)

The independent variables in the model used for parents' smoking and children's height (*Question D16–2*) were included because it was thought they might have a confounding effect on the association between smoking and height. In each instance there was reason to believe there might be a relationship with smoking, height, or both.

The square of the multiple correlation coefficient is the proportion of the variation of the dependent variable that is "explained" by the total set of independent variables. In *Question D16–3,* the square of .56 is .31, or 31%. This is higher than the explained proportion in most epidemiological studies.

The discrepancy between the proportions of variation explained by the smoking factors when considered separately and together (*Question 16–4*) obviously points to an overlap between their effects. We can compute from the figures in the top line that when nonsmoking variables are not taken into consideration (1.41 − 0.67)%, or 0.74%, of the variation is attributable only to smoking at home and (1.41 − 1.34)%, or 0.07%, only to smoking in pregnancy; the remaining (1.41 − 0.74 − 0.07)%, or 0.60%, is a shared effect. When other variables are included, the proportions are 0.12% (smoking at home), 0.03% (smoking in pregnancy), and 0.11% (shared). This overlap means that the number of ciga-

Table D17. Agreement Between Observed and Predicted Blood Pressure (mm Hg)

Age (yr)	Weight (lb)	Mean Residual (Observed BP Minus Predicted* BP)
<53	<172	−0.3
<53	≥173	−4.6
≥53	<172	−4.0
≥53	≥173	+3.8

*Predicted from age and weight.

rettes currently smoked at home and the number smoked during pregnancy are correlated; the correlation coefficient (for smoking by mothers) was in fact .64. We cannot determine which part of the overlap is attributable to current smoking, and which to smoking during pregnancy. This is another example of multicollinearity (Unit D15).

The reduction in the proportion of variation explained by an independent variable when other factors are included in a model (*Question 16–5*) may mean that the other factors (or some of them) are confounders, or it may mean that the other factors (or some of them) are intermediate causes. The statistical constellations in the two instances are the same (Unit A14). In this analysis there is one factor that may be an intermediate cause. This is birth weight: smoking in pregnancy is known to reduce the mean birth weight, and small size at birth may be one of the factors leading to low stature.

Absence of a statistically significant association (*Question D16–6*) does not prevent a variable from being a confounder. Strong associations can produce important confounding effects whatever their statistical significance. However, because no explicit criteria exist for deciding whether an association is sufficiently strong to produce confounding, opinions are divided about the use of significance tests for the purpose of deciding which potential confounders to control (Note D5).

A multiple linear regression coefficient indicates the average change in the dependent variable when there is a change of one unit in the relevant independent variable, with no change in the other variables in the model. The figure −0.0099 (*Question D16–7*) means that every additional cigarette currently smoked in the home, by mother of father, is associated with an average decrease in height of 0.0099 standard deviations. This is true if other variables are held constant. When smoking in pregnancy is added to the model, the specific ("unique") effect connected with smoking in the home (i.e., excluding the area of overlap) becomes slightly smaller, and it becomes still smaller (height decreases by only 0.0034 standard deviations for every cigarette) when other variables are added to the model and adjustment is made for their effects. But the association with smoking in the home remains statistically significant. Smoking a cigarette in pregnancy has a stronger effect than currently smoking one in the home, when other factors are held constant. But when the latter are taken into account, the effect is smaller and not statistically significant.

We cannot, however, conclude that smoking in pregnancy does not affect the child's height (*Question D16–8*). First, absence of statistical significance does not mean that an association is necessarily a chance finding. Second, one of the variables whose control weakened the association was birth weight, and (as pointed out above) small size at birth may be a link in a causal chain connecting smoking in pregnancy with low stature in childhood. Holding an intermediate cause constant weakens the statistical association between cause and effect. Such a finding supports a causal explanation; but we do not have data to enable us to separate the effects of controlling for birth weight and for other (confounding) variables. Third, as we have seen, there is a correlation between their

effects. The coefficients for smoking in pregnancy when current smoking is controlled express the effect that is "unique" to smoking in pregnancy, and may underestimate the true total effect of smoking in pregnancy. Our conclusion must be that the results do not tell us whether smoking in pregnancy affects height in childhood.

The association between passive smoking and the child's height (*Question D16–9*) is statistically significant, and remains apparent when variables expressive of genetic and other biologic attributes and social circumstances are held constant in the analysis. But adjustment for these factors may be incomplete: controlling for social class, number of older siblings, and household crowding may not hold socioeconomic factors completely constant. This is the first of the competing explanations considered by the investigators. Second, there may be an indirect causal association, mediated by other changes attributable to smoking, such as changes in family food consumption resulting from the effects of smoking on appetite or the family budget, or an increase of respiratory diseases in children exposed to the smoke. And third, tobacco smoke may have components that have a more direct effect on growth. You may have thought of other explanations—for example, the possibility of Berksonian bias, particularly because information was available for only 5,903/8,120, or 73%, of the study sample.

In answer to *Question D16–10*, this study may serve at least two purposes. First, an endeavor to identify the associated or intermediate reasons for the association may lead to new insights into factors affecting growth. Second, the results may serve pragmatic purposes. The effect of smoking on the child's height may or may not be thought important: assuming that the association is causal, parents who between them daily smoke 50 cigarettes in the home reduce their children's height by an average of (50×0.0034), or 0.17 standard deviations, which is approximately a centimeter. But even if this specific effect is regarded as unimportant, the study's additional evidence of the hazards of passive smoking may, if properly used, help to reduce the prevalence of smoking.

Note

D17. Table D17 shows that the mean residuals differ in different subgroups of the study sample. This would not happen if the model had a perfect fit with the observed facts. But we might well decide that the mean discrepancies are so small that they do not matter.

■■■■■■■■■■ Unit D18

Test Yourself (D)

Check that you can do the following:

- Judge whether the possibility of confounding can be excluded (D4).
- Predict the probable direction of a confounding effect (D4).
- Detect synergism (D12).
- Calculate
  the sensitivity of a risk marker (Note D8–1).
  the predictive value of a risk marker (D8).
  an odds ratio from paired data (D11).
  an odds ratio from a logistic regression coefficient (D14).
  risk from multiple logistic regression coefficients (D15).
- Explain
  when statistical significance should be tested (D4).
  the various meanings of "risk factor" (D8).
  when to use a rate difference and when to use a rate ratio (D10).
  the difference between additive and multiplicative models (D12).
- Explain what is meant by
  an equivalence test (D4).
  a reference category (D5, D6).
  noncollapsibility (D6).
  a risk ratio (D2, D9).
  a relative risk (D2, D9).
  a risk marker (D8).
  a statistically significant risk ratio or odds ratio (D10).
  a statistically significant rate difference (D10).
  exact tests (Note D13–2).
  a z score (D17).
  an intercept (D11).
  a statistically significant correlation coefficient (D11).
- Infer statistical significance from a confidence interval (D4, D10).
- Infer relative risk from an odds ratio (Notes D10–1, D10–2).
- Appraise a risk marker (D8).
- Appraise the validity of a multivariate model (D13).
- Make sense of
  a P value (Note D3, D4).
  a correlation coefficient (D11).
  a partial correlation coefficient (D11).
  a simple regression coefficient (D11).
  a multiple regression coefficient (D13, D17).
  a logistic regression coefficient (D13, D14).

- Explain (in general terms) what is meant by
  a conditional association (D4).
  a dose–response relationship (D8).
  synergism (D12).
  the Mantel-Haenszel procedure (D13).
  multiple logistic regression (D13).
  Cox proportional hazards regression analysis (D13).
  multiple linear regression (D13).
  residuals (D17).

Section E

Causes and Effects

"Don't let us quarrel," the White Queen said in an anxious tone. "What is the cause of lightning?"

"The cause of lightning," Alice said very decidedly, for she felt quite certain about this, "is the thunder—oh, no!" she hastily corrected herself. "I meant the other way."

"It's too late to correct it," said the Red Queen: "when you've once said a thing, that fixes it, and you must take the consequences."

(Carroll, 1865)

Unit E1

Introduction

This set of exercises deals with three main topics—the kinds of epidemiological study used to investigate causal processes, criteria for the appraisal of causal associations, and ways of measuring the impact of causal factors.

Kinds of Study

Epidemiological studies of causal processes can be broadly divided into *experiments* (in which the researcher decides which subjects or groups will be exposed to, or deprived of, the factor whose effect is under study) and *analytical surveys* (where surveys are defined as nonexperimental or "observational" studies). There is also a gray zone of *quasi-experiments,* which do not meet all the requirements of a well-designed experiment. We need not here concern ourselves with *descriptive surveys,* which aim to describe a situation rather than explain it; we have had examples in previous exercises, such as the studies of fractures of the femur in Oxford (Exercise B8) and suicide death rates in the United States (Exercise D11).

Analytic surveys can be classified in different ways (Note E1). The main types are:

- *Cross-sectional studies* (sometimes called *"prevalence studies"*). These are studies of total populations or population groups (or representative samples of them), in which information is collected about the present and (sometimes) the past characteristics, behavior, or experience of individuals. Examples in

previous exercises are the studies of correlations with blood pressure in a population sample in the West Indies (Exercise D10) and of children's height and parents' smoking in England and Scotland (Exercise D16).

- *Case-control studies,* which compare cases and controls with respect to their present or past characteristics, behavior, or experience. Examples are the studies of cancer of the lip and previous herpes (Exercise C5), gastroenteritis and food consumption (Exercise D10), and myocardial infarction and the use of oral contraceptives (Exercise D12).
- *Cohort studies,* in which a total population group, a sample, or samples of people with known differences in their exposure to a supposed causal factor are followed up to determine the subsequent development of a disease or other outcome ("follow-up" or "prospective" studies). Examples are the studies of electrocardiographic abnormalities (Exercise C5), varicose veins (Exercise D1), and drinking (Exercise D9) in relation to subsequent coronary heart disease, and the study of smoking and mortality (Exercise D8).
- *Group-based studies,* which compare groups (e.g., countries) and not individuals; these are sometimes called "ecologic" studies, or "studies of groups of groups" (Friedman, 1980). The study of the relationship between melanoma mortality rates and latitude (Exercise D10) is an example.

Each kind of study has its own special features, which affect the use of its results. These relate especially to the use of measures of association, sources of bias (artifactual findings), confounding, and the study's external validity.

We start with a cross-sectional study.

Exercise E1

The association between caffeine consumption and indigestion, palpitation, and other symptoms was investigated in a cross-sectional survey of 4,558 Australians (Shirlow and Mathers, 1985). The subjects were volunteers aged 20–70 years collected "off the street" by a voluntary screening clinic, and by a mobile unit that visited places of employment. Questions were asked about the usual intake of coffee, tea, cola drinks, chocolate, and medications, the kind of coffee drunk, and the strength of the tea or coffee. Caffeine consumption was calculated, using standard figures for the caffeine content in different sources. The frequen-

Table E1–1. Mean Daily Caffeine Intake (mg) by Frequency of Indigestion (Men)

Indigestion	No.	Caffeine Intake
Never/rare	1,370	233
Sometimes/frequent	754	251
P		<.001

Table E1–2. Prevalence Rates of Indigestion in Low, Medium, and High Caffeine Consumption Groups, With Rate Ratios (Men)

Caffeine Consumption	Rate %	Rate Ratio
Low (0–150 mg/day)*	33.2	1.0
Medium (151–250 mg/day)	33.0	0.99
High (>250 mg/day)	39.3	1.18

*Reference group.

cy of symptoms was reported as "never or rarely," "sometimes" (1–3 times a month), or "frequently" (once a week or more). Selected findings in men are shown in Tables E1–1 and E1–2 (the findings in women were similar).

Question E1–1

Two different approaches to the examination of associations are used in Tables E1–1 and E1–2. Do you know what these approaches can be called? Summarize the facts shown in the tables. Are the rate ratios in Table E1–2 risk ratios?

Question E1–2

Table E1–2 shows rate ratios, and Table E1–3 shows odds ratios calculated from the same raw data. Which are preferable?

Question E1–3

May the respondents' or interviewers' awareness that symptoms were present have influenced the association with caffeine consumption?

Question E1–4

The data were submitted to multiple logistic regression analyses in which indigestion and other symptoms were dependent variables. The independent vari-

Table E1–3. Association Between Indigestion and Caffeine Consumption: Odds Ratio

Caffeine Consumption	Odds Ratio
Low (0–150 mg/day)*	1.0
Medium (151–250 mg/day)	0.99
High (>250 mg/day)	1.30

*Reference group.

Table E1–4. Relationships of Caffeine Consumption to Prevalence of Symptoms (Men): Odds Ratios Based on Multiple Logistic Regression

Symptom	Odds Ratio*	*P*
Indigestion	1.1	NS
Palpitation	1.3	<.01
Headache	1.4	<.0001
Tremor	1.2	<.05
Insomnia	1.3	<.0001

*The odds ratios indicate the change in the odds when daily caffeine consumption increases by 200 mg.

ables were caffeine consumption, age, Quetelet's body mass index, smoking, and alcohol consumption. Odds ratios derived from the results are shown in Table E1–4. Summarize the findings. Can you conclude that caffeine consumption produced these symptoms?

Question E1–5

Would you have any hesitation in applying the results of this study to Australian adults in general?

Question E1–6

Suppose that the association between caffeine consumption and congestive heart failure had also been investigated in this cross-sectional survey. What other bias or biases would you suspect?

Note

E1. The different kinds of study and their pros and cons are explained in all epidemiological texts. See, for example, Abramson and Abramson (1999), and Rothman and Greenland (1998, chap. 5). There are many hybrid designs.

■■■■■■■■■■ Unit E2

Appraising the Results of a Cross-Sectional Study

The two approaches used in Tables E1–1 and E1–2 (*Question E1–1*) may conveniently be termed *retrospective* and *prospective*, despite a confusing lack of

consensus about the use of these terms (see Note E2). Table E1–1 uses what can be called a *retrospective approach,* in that the subjects are classified according to the supposed outcome (indigestion), and we see whether the groups differ in their exposure to the supposed cause (caffeine); we move from the postulated outcome to the postulated cause. In Table E1–2, we start at the other end: the subjects are classified according to their exposure, and we see whether they differ in the frequency of the outcome. This can be called a prospective approach. Both approaches are feasible in a cross-sectional study, in contrast to a case-control study (which is characterized by a retrospective approach) or a cohort study or experiment (where the approach is prospective).

Both tables show positive associations between caffeine intake and indigestion; the association is statistically significant. The prevalence rate of indigestion is similar in the low and medium caffeine consumption groups, and higher in the high-caffeine group.

If we use the usual definition of risk (Note A6), the rate ratios in Table E1–2 are not risk ratios; they are not ratios of incidence rates. A cross-sectional study cannot provide a direct measure of risk.

There is no compelling reason to prefer either rate ratios or odds ratios (*Question E1–2*). Both are good measures of the strength of an association.

In answer to *Question E1–3,* respondents who thought their symptoms were caused by coffee drinking might tend to report a higher consumption; and interviewers inclined to this view might try harder to get full information about caffeine consumption. The investigators say, however, that "the questionnaire . . . did not indicate to the subject that an association was expected. . . . The possibility of such a bias was lessened by the questions forming part of a general health screening examination aimed primarily towards the identification of cardiovascular risk factors" (Shirlow and Mathers, 1985).

The subjects' awareness of their symptoms may have influenced the association in another way: it may have led them to drink less coffee. (But the investigators report that only 2.6% of the subjects said they avoided coffee because of palpitation.) Or coffee might have been used to alleviate headaches. A frequent problem in cross-sectional studies is that it may be difficult to know which came first, the supposed cause or the supposed outcome.

These possibilities of effects arising from the fact that the postulated outcome occurs before the study is done also apply to case-control studies.

In answer to *Question E1–4,* all the symptoms except indigestion showed statistically significant, if weak, associations with caffeine consumption when possible confounders were controlled. The investigators report that the association with indigestion was accounted for by strong correlations with adiposity (Quetelet body mass index), and disappeared when adiposity was controlled in the analysis. We have no grounds for concluding that caffeine consumption produces indigestion. The findings are consistent with the hypothesis that it produces the other symptoms; but there may be unidentified confounders. The investigators concluded "that this study presents *suggestive* evidence that habitual caffeine consumption causes palpitations, tremor, headaches, and sleep disturbances" (Shirlow and Mathers, 1985).

The main reservation about the external validity of the study (*Question E1–5*) is the possibility of Berksonian bias. The associations observed in this volunteer sample may be different from those in the community at large. This might happen, for example, if people who drank a lot of coffee, and those with symptoms, were especially prone to volunteer.

If congestive heart failure had been included in this study (*Question E1–6*) the findings would have been relevant to mild cases only. People with severe disease would either have died before the study, or (if alive) would be in places other than the streets and workplaces in which the sample was collected. Also, extremely mild cases might have tended to be excluded, either because of the absence of clear signs and symptoms or because mild cases tend to have remissions.

Exercise E2

Exercises D12 to D14 were based on a case-control study that showed a strong association between the use of oral contraceptives and myocardial infarction. The study was done in 155 hospitals in a region of the United States. The cases consisted of all premenopausal women aged 25–49 who were admitted to a coronary-care unit during a 2-year period for a first episode of definite myocardial infarction (by WHO diagnostic criteria). Five potential controls were interviewed for each case of definite or possible myocardial infarction admitted. The controls were premenopausal women who had never had a myocardial infarction and were admitted to the surgical, orthopedic, or medical service of the same or a nearby hospital for conditions judged to be unrelated to oral-contraceptive use or cigarette smoking; controls who turned out not to meet these criteria were disqualified. The women in both groups were asked (*inter alia*) whether they had used oral contraceptives in the last month. Permission for the interview was refused by the patient or physician in 6% of cases and 6% of controls (Shapiro et al., 1979).

Question E2–1

Can a case-control study (like this one) measure

- risk?
- relative risk?
- a risk difference?
- a rate ratio?

(Remember that we are not strict in our use of the term "rate" and may apply it to proportions.)

Question E2–2

An obvious problem with case-control studies is that the samples of cases and controls may not be closely comparable, and the differences between them may

confound the associations with the disease. What, therefore, do you think should be one of the first steps in the analysis?

Question E2–3

In a case-control study, the occurrence of the postulated effect (in this instance, myocardial infarction) precedes the collection of information about the postulated cause (oral contraceptives). How might this produce bias?

Question E2–4

As we have seen, the results of this study are consistent with the hypothesis that oral contraceptives increase the risk of myocardial infarction. Are they consistent with a completely different hypothesis—that oral contraceptives protect the lives of women who have an infarction?

Note

E2. The terms "retrospective" and "prospective" are often used to indicate whether a study is based on already available data. Rothman and Greenland (1998, pp. 74–75) advocate the use of these terms to indicate whether information about the putative cause was obtained after or before the occurrence of the outcome (cases of the disease). To avoid confusion, Feinstein (1977) has suggested use of the term *"retrolective"* for a study based on previously recorded data, and *"prolective"* for one in which data collection is planned in advance. These terms use the Latin root of the word "collect."

■■■■■■■■■■ Unit E3

Appraising the Results of a Case-Control Study

A case-control study cannot generally provide a direct measure of risk (*Question E2–1*): the number of cases in the study is determined by the investigator, not by the incidence of the disease. Thus, the study cannot yield a direct measure of relative risk, or a risk difference. A case-control study can, of course, provide a rate ratio that is not a risk ratio—in this instance, the ratio of the rate of contraceptive use in cases to that in controls (the "exposure rate ratio"). The odds ratio, however, may under certain conditions be an estimator of the incidence rate ratio based either on number-of-persons denominators or on person-time denominators. (Can you remember these conditions? They were listed in Notes D10–1 and D10–2.)

In certain circumstances, and if ancillary information is available, the risk associated with a specific factor *can* be estimated; we had an example in Question

D8–3. Risk can be estimated in a *nested case-control study,* in which new cases of a disease are identified during a follow-up study of a cohort and are then compared with controls drawn from the same cohort.

An obvious early step in the analysis of data from a case-control study (*Question E2–2*) must be a comparison of the characteristics of the two samples. The controls in a case-control study should be representative of the population "base" from which the cases were drawn. In this study, the cases and controls were found to be similar in ethnic group, religion, marital status, parity, and education, but they differed in geographical area (Boston, New York, or Philadelphia), cigarette smoking, obesity, and a history of diabetes, hypertension, lipid abnormality, angina pectoris, and preeclamptic toxemia. The latter variables were controlled by including them in a multiple logistic regression model; the adjusted odds ratio for the association between oral contraceptives and myocardial infarction was then 4.1.

In answer to *Question E2–3* in this case-control study (as in cross-sectional studies), obtaining information about the "cause" only after the "effect" has occurred may produce bias in various ways. Those listed by Sackett (1979) in a catalogue of biases are "rumination bias" (cases may ruminate about causes for their illnesses and thus report different prior exposures than controls), "obseqiousness bias" (subjects may alter their responses to fit what they believe the investigator wants), and "exposure suspicion bias" (a knowledge of the subject's disease status may influence the intensity and outcome of a search for exposure to the putative cause). In this study, the investigators could not rule out the possibility of information bias, since the nurses who did the interviewing and many of the patients were aware of the hypothesis.

If a postulated causal factor (in this instance, oral-contraceptive use) affects the chance of inclusion as a case or control in the study, this will produce selection bias. This study did not include women who died immediately after having an infarction, before admission to hospital. If oral contraceptives protect infarction patients from death (*Question E2–4*), the lucky women who stayed alive and entered the study would include a high proportion of users of the Pill—producing the observed association. The results are therefore consistent with the hypothesis that oral contraceptives keep infarction patients alive. (The investigators refute this interpretation by citing studies of patients with fatal infarction.)

Exercise E3

Our example of a cohort study is a follow-up study conducted in a rural district of southern India, in which the association between tobacco-chewing and mortality was investigated (Gupta et al., 1984). In that part of the world, tobacco is chewed in the form of "pan"—that is, with betel leaf, areca nut, and slaked lime. A random sample of villagers aged 15 years and over—about 5,000 males and 5,000 females—were questioned about their tobacco habits. Deaths were ascertained through follow-up household interviews conducted 3 years later, and then annually until 10 years had passed.

Table E3. Mortality Rates per 1,000 Person-Years and Relative Risks, by Tobacco-Chewing Habit (Females)

	Crude		Age-Standardized	
	Rate	Relative Risk	Rate	Relative Risk
Tobacco chewers	12.8	3.4	8.3	1.3*
Nonchewers	3.8	1.0	6.2	1.0

*$P < .05$.

Table E3 shows the results in females, 41% of whom chewed tobacco; tobacco smokers (1%) are excluded. Rates were age-standardized by the direct method, using specific rates for 10-year age intervals and the world standard population (Note B14–3).

Question E3–1

Person-time mortality rates were calculated, not cumulative mortality rates. Can you guess why? Does this study provide measures of risk?

Question E3–2

What is the explanation for the difference between the crude and age-standardized relative risks?

Question E3–3

May the statistically significant association shown by the age-standardized data be a spurious one caused by confounding?

Question E3–4

Do you want to know anything about losses to follow-up? If so, why and what?

Question E3–5

Can you conclude that tobacco-chewing increased the risk of dying?

Question E3–6

In men, the age-standardized relative risk of mortality in tobacco chewers was 1.2. Does this alter your reply to *Question E3–5*?

Question E3–7

Should the validity of this study be questioned because of a possibility of diagnostic suspicion bias (bias caused by knowledge of the subjects' prior exposure to a putative cause)?

Question E3–8

If the confidence intervals of the age-standardized relative risks were computed, would these express the association between tobacco-chewing and mortality in the population from which the sample was drawn?

■■■■■■■■■■ Unit E4

Appraising the Results of a Cohort Study

In answer to *Question E3–1,* some people were lost to follow-up before the end of the 10 years of the study, so that direct measurement of cumulative mortality rates was not possible. By using person-time denominators it was possible to utilize all the available information about each subject, until loss of contact. Cumulative mortality rates can, of course, be estimated from the person-time rates (Note B5–4). With rates as low as those reported, the person-time and cumulative mortality rates would be almost identical. Both can be used as measures of risk. One of the advantages of a cohort study, with its prospective approach, is that it provides measures of risk.

The difference between the crude and age-standardized relative risks (*Question E3–2*) shows that there is confounding by age. (Can you say whether tobacco chewers were older or younger than nonchewers? See Note E4 for answer). If the chewers and nonchewers were very different in age, some degree of confounding may remain even after age standardization (*Question E3–3*), because there may be substantial age differences between chewers and nonchewers *within* the broad (10 year) age groups used for standardization. There may also be other confounders. The only other variable mentioned by the investigators is socioeconomic status, which was not measured because of practical difficulties and because it was estimated that 90–95% of the population were in the lower socioeconomic strata. (But if the other 5–10% did not chew tobacco and had a low death rate, this could account for part of the association seen in Table E3.) We thus cannot exclude the possibility that the association is, at least in part, spurious.

Information about losses to follow-up (*Question E3–4*) is important in any cohort study. If people whose traces are lost have a different risk from those whose fate is known, the observed risk will be biased; and if this bias is different in the groups under comparison, the relative risk will be biased. We should therefore seek information about losses to follow-up and their reasons. The report tells us that most losses were due to leaving the district, probably because of marriage. Since nubile women tend to be healthy, these losses probably produced an upward bias of the death rate. Losses were more frequent in nonchewers, whose mean follow-up period was shorter (7.7 years) than that of chewers (8.8 years).

This suggests that bias due to losses would tend to reduce rather than produce a difference in mortality.

It is difficult to be confident that tobacco-chewing increased the risk of dying (*Question E3–5*), as confounding can easily produce a weak association such as the one seen in this study, and it is not certain that age and other possible confounders were adequately controlled. If similar results were obtained in another sample or study, this would support the inference that the association was causal, and not due to chance, bias, or confounding. But the similar relative risk found in men in this study (*Question E3–6*) may mean only that the same confounding factors were operative in both sexes.

Diagnostic suspicion bias, which is one of the hazards besetting cohort studies, seems unlikely here (*Question E3–7*). This bias may be suspected if the people responsible for measuring the outcome (the putative effect) know what hypothesis is being tested and which subjects were exposed to the putative cause, and if this knowledge can influence the methods used to determine the outcome. In this study, deaths were ascertained during household interviews. We do not know whether the interviewers were "blinded" to prior tobacco-chewing habits, or whether they knew what hypothesis was being tested. But it seems unlikely that their knowledge could affect the responses to a simple question about the survival of household members. The results of any cohort study can be applied to a target population if the exposed and unexposed individuals in the sample (i.e., those exposed or not exposed to the suspected cause) are representative of the exposed and unexposed, respectively, in the population. This study was based on a random sample, and the relative risks should therefore be applicable to the population; their confidence intervals would express the uncertainty attributable to random sampling variation. But the answer to *Question E3–8* is not that simple, as confidence intervals do not express the possible uncertainty attributable to confounding or losses to follow-up.

Exercise E4

As an example of a group-based ("ecologic") survey, we will use a study of correlations between the infant mortality rate and other national statistics in 18 developed countries—the United States, Canada, Australia, New Zealand, and 14 European countries—in 1970 (Cochrane et al., 1978). These countries were chosen because they met criteria based on population size, the gross national product (GNP) per caput, and the availability of data.

Multiple regression analysis showed that 97% of the variation (variance) of infant mortality was explained by seven variables: the GNP per head, population density, the percentage of health expenditure covered by public expenditure, the number of doctors per 10,000 population, the annual cigarette consumption per head, the annual alcohol consumption per head, and the annual consumption of sugar per head. Other variables—the number of nurses, pediatricians, midwives, hospital beds, protein and fat consumption, and education—made little additional contribution.

Question E4–1

There was a negative correlation ($r = -.46$) between infant mortality and the GNP per head; that is, richer countries had lower infant mortality rates. This correlation was statistically significant. The GNP per head alone explained 21% of the variation of infant mortality. According to the multiple regression analysis, the infant mortality rate decreased by 16%, on average, for a rise of one standard deviation in the GNP per caput, when the other six factors in the analysis were held constant. How would you explain the association between infant mortality and the GNP per head?

Question E4–2

There was a positive association ($r = .67$) between infant mortality and the number of doctors per 10,000 population; that is, countries with a higher prevalence of doctors had higher infant mortality rates. This correlation was statistically significant. The number of doctors alone explained 45% of the variation of infant mortality. According to the multiple regression analysis, the infant mortality rate increased by 17%, on average, for a rise of one standard deviation in the number of doctors per 10,000 population when the other factors in the analysis were held constant. An analysis of data for 1960 revealed similar results, suggesting that the findings "cannot too easily be dismissed as a chance curiosity." How would you explain the association between infant mortality and the number of doctors per 10,000 population?

Note

E4. Confounding by age produced spurious strengthening of the association, and mortality obviously has a positive association with age. According to the Direction Rule (Unit D5), therefore, tobacco-chewing, too, was probably positively associated with age.

■■■■■■■■■■ Unit E5

Appraising the Results of a Group-Based Study

There are two kinds of explanation for the negative correlation between infant mortality and the GNP per head (*Question E4–1*). Richer countries may have lower rates because they are richer (better hospital facilities, better food, better sanitation, etc.), or the correlation may be due to confounding factors that are correlates but not necessarily consequences of wealth, such as differences in knowledge, attitudes, and practices with respect to infant care.

Similarly, the positive correlation with the prevalence of doctors (*Question E4–2*) may be causal or due to confounding. As an iatrogenic explanation is implausible, confounding seems likely. But by what? The investigators were not able to find an explanation: "We must admit defeat and leave it to others to extricate doctors from their unhappy position" (Cochrane et al., 1978).

Two main problems beset the appraisal of group-based studies. The first is the influence of confounding factors that, especially in studies based only on official statistics, may be difficult to investigate. The second is the "ecological fallacy" of concluding that an association found at a group basis also exists at the individual level. (There is more malaria in poor countries than in rich ones; but this does not necessarily mean that poor people are at higher risk than rich people in the same country.)

Exercise E5

This exercise deals with three studies of the effects of health care procedures.

Question E5–1

The first is a clinical trial of the effect of acupuncture (Godfrey and Morgan, 1978). The subjects were patients with chronic, dull, moderate pain at any site, attending outpatient clinics in a Toronto hospital; 57% volunteered for the study in response to a public announcement, and 43% were referred by physicians. The most frequent diagnoses were osteoarthritis (24%), degenerative disk disease (20%), and lumbosacral strain (8%). Patients found to have inflammatory conditions were excluded. The subjects were randomly allocated to two groups: one whose members received acupuncture (i.e., needling at the sites where, according to acupuncture theory, this was most likely to relieve their pain) and a control group who received sham acupuncture (needling at the sites least likely to reduce their pain). The study was double-blind: the acupuncturist (a Chinese expert) did not know whether he was administering true or sham acupuncture, nor did the subject. The patients used a 6-point scale to measure the level of pain. Table E5–1 shows the results after five treatments.

Do the results prove that acupuncture does not work—that is, that "appro-

Table E5–1. Reduction of Pain After Five Treatments: Double-Blind Randomized Trial of Acupuncture

	Acupuncture	Controls
Number of subjects	84	84
Number with reduction of pain	53	45
Sucess rate (%)	63	54
$P = .21$		

priate" acupuncture does not relieve pain better than sham acupuncture? If not, why not? What extra information would you like?

Question E5–2

If there had been 8,400 subjects in each group and the *P* value was the same (.21), would this affect your appraisal of the results?

Question E5–3

If the trial had showed a clearly beneficial effect, could the results be applied to everyone with pain?

Question E5–4

What kinds of bias are reduced by "blinding" experimental subjects or observers?

Question E5–5

The effect of breast cancer screening on mortality from breast cancer was examined in a randomized trial (Shapiro et al., 1982). Women aged 40–64 who were members of the Health Insurance Plan of New York were randomly allocated to two groups: a "study group," whose members were offered four annual screening examinations (clinical examination and mammography); and a control group, who continued to receive their usual medical care. About 31,000 women were in each group. The groups were very similar with respect to a wide range of demographic and other characteristics.

Mortality rates from causes other than breast cancer are shown in Table E5–2. How can the findings be explained?

Question E5–6

Table E5–3 shows the numbers of breast cancer deaths in the 9 years following entry to the study (Shapiro, 1977). (Because the denominators in the two groups are almost identical, we can use numbers instead of rates.) What would you conclude from these results? You may assume that the differences are not fortuitous.

Table E5–2. Deaths From All Causes Other Than Breast Cancer: Ten-Year Follow-up After Entry to Study

	Death Rate*
Members of study group who were screened	54.9
Control group	64.8

*Deaths per 10,000 person-years.

Table E5–3. Deaths From Breast Cancer: Nine-Year Follow-up After Entry to Study

Age (yr) at Diagnosis	No. of Deaths		Ratio
	Study Group	Control Group	
40–49	30	27	1.1
50–59	42	67	0.6
≥60	19	34	0.6
Total	91	128	0.7

Question E5–7

A multicenter randomized trial was conducted to determine the value of treatment for mild hypertension in the elderly (Amery et al., 1985). The trial was double-blind, the subject's allocation to the treatment or control (placebo) group remaining undisclosed until the end, unless an event occurred—such as a severe increase in blood pressure—that necessitated "breaking the code." The mortality rates in the treatment and placebo groups are shown in Table E5–4, using two different methods of analysis. The "intention to treat" analysis is based on deaths during the entire follow-up period, in the subjects originally allocated to each group—whether they persisted with their allotted treatment or not. The "on randomized treatment" analysis is confined to the findings while the subjects were in the double-blind part of the study, on their allocated treatment. Which form of analysis is better?

Question E5–8

A randomized controlled trial of low-dose aspirin for primary prevention, conducted in 108 group practices in the United Kingdom, among men aged 45–69 who were at increased risk of coronary heart disease, demonstrated a beneficial effect in men with lower systolic blood pressures (Table E5–5). Using the data in Table E5–5, can you compute how many men with blood pressures (a) below 130 and (b) of 130–145 mm Hg must be treated for 1 year in order to prevent one major cardiovascular event?

Table E5–4. Mortality From Cardiovascular Diease in Treated and Control Groups: Rates per 1,000 Person-Years

Type of Analysis	Group		Ratio
	Treatment	Placebo	
"Intention to treat"	34	47	0.72
"On randomized treatment"	30	48	0.63

Table E5–5. Aspirin Trial: Incidence of Major Cardiovascular Events (Coronary Heart Disease and Stroke) in Treated and Control Groups, by Systolic Blood Pressure

Systolic Blood Pressure (mm Hg)	Rate per 1,000 Person-Years		Rate Ratio*
	Aspirin	No Aspirin	
<130	7.7	12.2	0.59
130–145	9.0	14.0	0.66
>145	20.5	17.9	1.08

*Adjusted for age and seven cardiovascular risk factors.

Source: Meade and Brennan (2000).

■■■■■■■■■■ Unit E6

Appraising the Results of an Experiment

The trial of acupuncture did not demonstrate a statistically significant effect. The slight benefit observed could easily be a chance finding. The absence of statistical significance does not, however, mean that the benefit *was* a chance finding; we have no way of telling. The study does not prove that acupuncture works, but (in answer to *Question E5–1*) neither does it prove that it does not.

Randomization (random allocation into treatment groups, based on tossing a coin, using random numbers, etc.) minimizes the likelihood of confounding, but it cannot completely prevent it. Substantial differences may occur between the groups, just by chance, and these may exaggerate or weaken the apparent effects of treatment. Information on the characteristics of the groups (age distribution, diagnoses, sites of pain, etc.) would satisfy us that confounding was unlikely. We should also have information on withdrawals from the study, for the same reasons as in a nonexperimental cohort study.

In answer to *Question E5–2*, a statistically nonsignificant result based on large numbers—that is, where the power of the test (Note D4) is high—may be taken as evidence that no real effect of any importance exists.

Clinical trials are never conducted on random samples; the requirement that subjects must give their informed consent is enough to ensure this, let alone the trial's specific eligibility criteria. The result can be generalized only to a reference population that the subjects are believed to represent. In this instance (*Question E5–3*), the subjects were certainly not representative of all people with pain. We do not know just what the selective factors were. At best, we might decide that the results can be applied to the sort of hospital outpatient with chronic pain who is likely to request acupuncture or be referred by a physician for acupuncture.

The use of blind methods (*Question E5–4*) reduces the chance that the sub-

jects' reactions or reports, or their readiness to remain in the study, will be influenced by their knowing what treatment they are having. Keeping clinicians and other observers in the dark prevents them from communicating this knowledge to the subjects and from handling the experimental groups differently, and it keeps their own findings unbiased.

Randomization ensures that the subjects in a trial are divided into groups that have only chance differences. But if after randomization we remove people who refuse to participate or are withdrawn from the study (because the treatment is inappropriate, etc.), the groups may no longer be comparable. This is illustrated in *Question E5–5,* where the reason for the difference in mortality is that members of the study group who refused the offer of screening were omitted. The fuller facts (Table E6–1) show that the study and control groups did not differ in their mortality from causes other than breast cancer.

In answer to *Question E5–6,* Table E5–3 shows fewer breast cancer deaths among women allocated to the study group. As this difference cannot easily be attributed to bias or confounding, the results, indicate that screening decreases mortality from breast cancer. This benefit is not apparent below the age of 50.

The stratification by age in Table E5–3 represents one of the procedures commonly used in the analysis of trials (Note E6–1). Prognostic factors that are associated with the outcome are identified. It is then possible, by appropriate analyses, to examine their modifying and possible confounding effects. The term *post-stratification* may be used to distinguish this method from stratified allocation to treatment and control groups (i.e., stratification of the potential subjects according to supposed prognostic factors, followed by random allocation of the members of each stratum, so as to obtain matched treatment and control groups).

Excluding randomized subjects of a therapeutic trial from the analysis may lead to bias, and the correct answer to *Question E5–7* is that an *"intention to treat" analysis,* comparing the outcomes in all the subjects originally allocated to each group (including those who did not have or who stopped having the specified treatment) is preferable. This stringent approach may sometimes, however, underestimate the efficacy of the treatment. This probably happened in this

Table E6–1. Deaths From All Causes Other Than Breast Cancer: Five-Year Follow-up After Entry to Study

	Death Rate*
Study group	
Screened	42
Refused	86
Total	57
Control group	58

*Deaths per 10,000 person-years.

study, where a proportion of the subjects in the treatment group stopped treatment, and a proportion of those in the placebo group received antihypertensive treatment: 15% of the subjects in the placebo group (but only 1% of those in the treatment group) were removed from the double-blind part of the study because of a severe increase in blood pressure.

Question E5–8 is surprisingly easy. The "number needed to treat" is 1 divided by the difference between the rates. The rates for men with systolic pressures under 130 mm Hg are 7.7 and 12.2 per 1,000 person-years—that is, 0.0077 and 0.0122 per person-year; the difference is thus 0.0045, and the number needed to treat is 1/0.0045, or 222. That is, 222 men must receive aspirin for 1 year in order to prevent one case. For men with pressures of 130–145 mm Hg, the number is 1/0.005, or 200. The rationale is as follows. In the <130 mm Hg group, where the numbers of cases in 1,000 person-years are 12.2 in the untreated sample and 7.7 in the treated sample, it can be inferred that 1,000 person-years of treatment reduce the number of cases from 12.2 to 7.7 (i.e., by 4.5). By simple proportion, the number of person-years of treatment required to prevent a single case (i.e., 4.5/4.5) is therefore 1,000/4.5, which is the same as 1/0.0045.

A confidence interval for the number needed to treat (see Note E6–2) can be calculated in the same way, by using the reciprocals of the upper and lower confidence limits of the difference between the rates, instead of the reciprocal of the difference itself.

Exercise E6

This exercise deals with another two studies of health care.

Question E6–1

An "early stimulation" program for promoting children's development (by encouraging mothers to speak and play with their infants) was instituted and tested at two maternal and child health (MCH) clinics operated by a university department in two neighborhoods of Jerusalem. It was decided not to allocate mothers to the program randomly, partly for practical and ethical reasons, and partly because dissemination to other mothers living in the same neighborhoods and using the same clinics (i.e., "contamination" of controls) was inevitable.

It was therefore proposed to appraise effectiveness by comparing the development of infants served by these clinics with that of infants served by two clinics in neighborhoods where there was no such program. This plan was abandoned when it was found that, mainly because of poor attendance, it would not be possible to measure the status of the control children. Instead, a "before–after" design was chosen, comparing the development of two birth cohorts of infants served by the intervention clinics—those born after implementation of the program and those born before. At 2 years of age, the mean developmental quotient (DQ) turned out to be higher in children born after implementation of the program (Palti, 1983).

If the first study plan had been practicable, would this have been a good ex-

periment? (And if not, why not?) After the change of design, was this a good experiment? If the two designs had been combined (so as to compare the "before–after" differences in the intervention and control communities), would this have made a good experiment?

Question E6–2

What precaution would be needed when appraising the results?

Question E6–3

Some years later, another evaluative study was done, by comparing the IQs of two groups of 5-year-olds who were attending nursery schools in the neighborhoods in which the experimental clinics were situated: children who as infants had received care in these clinics, and control children who had received care at other MCH clinics (Palti et al., 1986). The controls were individually matched by ethnic group, mother's education, and birth rank. The groups were found to be similar with respect to mother's age, mother's work outside the home, father's education, social class, number of years in nursery school, number of languages spoken in the home, and other variables. Would you call this an experiment?

Question E6–4

Selected results are shown in Table E6–2. Summarize the findings. What would you conclude?

Question E6–5

The effect of obstetric care on the outcome of pregnancy was appraised in a hospital in Oxford by comparing fetal deaths ascribed to asphyxia or trauma with randomly selected live-born control infants (Niswander et al., 1984). By use of the clinical records, "blind" assessments were made of the quality of care in pregnancy, and of the complexity of the pregnancy and labor (poor obstetric history, intrauterine growth retardation, abnormalities of fetal heart rate, preterm delivery, etc.). Selected results are shown in Table E6–3.

Table E6–2. Mean IQ of Five-Year-Olds Exposed to Early Stimulation Program and Matched Controls, by Mother's Education

	Mean IQ			
Mother's Education	Exposed	Control	Difference	*P*
5–8 years	106.3	92.0	14.3	.021
9–12 years	111.7	104.6	7.1	.012
>12 years	121.9	121.6	0.3	NS
Total	114.4	108.6	5.8	.003

Table E6–3. Relationship of Fetal Deaths Ascribed to Asphyxia or Trauma to Quality of Care in Pregnancy

Quality of Care	Fetal Deaths	Controls	Odds Ratio (With 95% Confidence Interval)	*P*	Adjusted Odds Ratio*
Suboptimal	8	17	3.7 (1.6–8.6)	<.01	3.4
Satisfactory	45	355			

*Controlling for complexity of the pregnancy and labor (by use of the Mantel-Haenszel procedure).

What conclusion can you reach about the effect of the quality of antenatal care on the outcome of pregnancy?

Question E6–6

The above study was obviously not an experiment. An experiment to study the effect of suboptimal antenatal care would have serious ethical objections. Was it a quasi-experiment or a survey? If a survey, what kind? A cross-sectional, case-control, or cohort study?

Notes

E6–1. The design, conduct, and analysis of trials are explained in many textbooks. For a simple but thorough exposition, see Peto et al. (1976, 1977). Design and analysis are dealt with in detail by Fleiss (1986c).

E6–2. If the confidence interval of a difference between rates is (say) from 2 to 4 per 1,000, the confidence interval of the number needed to treat is from 1/0.004 to 1/0.002 (i.e., from 250 to 500). A difficulty arises if the difference is not significant (i.e., if the lower confidence limit of the difference is negative). If this confidence interval is from −2 to 4 per 1,000, the confidence interval of the number needed to treat is from 250 to −500. The latter figure means that, at the upper extreme of the confidence interval, 500 person-years of treatment will produce (not prevent) one case; this has been termed the *"number needed to treat for harm"* (Altman, 2000). One way of thinking about this is that the confidence interval for the number needed to prevent one case extends from 250 to infinity in the treatment group, and then up to 500 in the untreated group.

■■■■■■■■■■ Unit E7

Appraising the Results of a Quasi-Experiment

Quasi-experiments, which do not fully satisfy the criteria of a sound experiment, are usually performed because a better design is not feasible (Note E7–1).

All three of the studies described in *Question E6–1* are quasi-experiments. In the first—the comparison of children served by intervention and control clinics—there was no randomization of clinics (because the investigators were able to implement the program only in their own clinics). Also, the design took no account of the possibility that children living in the different neighborhoods might have differed in their development before the program was started: there were "after" measurements but no "before" measurements. In addition, it can be claimed that there were too few sampling units. In effect, two clusters of children (in different neighborhoods) were compared with two others If children in different neighborhoods differ much in their development, a good experiment would require a fair number of clusters—certainly more than two—in each group.

The second design—a "before–after" comparison based on the findings in different birth cohorts in the neighborhoods where the program was implemented—makes no allowance for the possibility that a change might have occurred even without the program. Observations in control neighborhoods over the same period might have demonstrated a similar change. To mitigate the problem of a possible "secular trend" (a change with time), the investigators actually used a *time series* instead of a simple "before–after" comparison. They included two birth cohorts born before the program was started, and showed that there was no evidence of a change before the program was instituted (Palti, 1983).

A combination of these two designs—that is, a comparison of "before–after" changes in intervention and control communities—would remedy some of these drawbacks. But here, too, there is no randomization.

The main precaution to be taken when appraising the results of a quasi-experiment (*Question E6–2*) is that the same careful attention to the possibility of confounding is needed as in an analytic survey.

The design described in *Question E6–3* is also quasi-experimental. It is again a comparison of children served by different clinics, this time using matching to control selected confounders, but still with no randomization or "before" measurements.

The main findings (*Question E6–4*) were that children in the exposed group had a significantly higher mean IQ, that this difference was apparent only in the children of mothers with 12 or fewer years of education, and that there was a positive association between maternal education and the child's IQ (in both the exposed and control groups).

Since some possible confounders were controlled by matching, and others could be disregarded because of the results of the "exclusion test," the findings

suggest that the program was probably effective. This interpretation is supported by the interaction with maternal education, since if early stimulation works it might be expected to work best with the disadvantaged children of less educated mothers. The findings conform with this expectation. The program appears to reduce the gap in development between the children of less educated and better educated mothers.

The results of the study in Oxford (*Question E6–5*) suggest that suboptimal antenatal care is a cause of fetal death. The association was strong and statistically significant, it was based on appraisals that were apparently unbiased (because they were "blind"), and it was only slightly attributable to the confounding effect of the complexity of the pregnancy or labor. There is, however, a reservation: the control of confounding may not be as good as it appears. The appraisals of complexity may not have provided sufficient control of prognostic factors. The investigators admit that "failure to achieve adequate control of confounding factors . . . may have led us to overestimate some of the risks associated with suboptimal care. In future studies we shall try to match cases and controls more closely by the clinical problems for which the quality of care is to be assessed" (Niswander et al., 1984).

In answer to *Question E6–6,* this is, of course, a case-control study. Case-control studies in which the case is a person with a condition that may be due to poor care may be used to evaluate health care procedures and programs.

Exercise E7

In this exercise we appraise causal associations in three studies.

Question E7–1

A study of all infants born in Michigan from 1950 to 1964 showed a strong positive association between birth rank and the rate of Down's syndrome (Note E7–2). There was a threefold variation in rates. Do the findings in Table E7–1 indicate that birth rank influences the risk of the disease?

Table E7–1. Downs Syndrome in Michigan by Birth Rank: Rates, Relative Risks, and Standardized Morbidity Ratios (SMR)

Birth Rank	Rate per 100,000 Live Births	Relative Risk	SMR*
1	56.3	1.0	1.0
2	67.6	1.2	1.0
3	83.3	1.5	1.1
4	115.5	2.1	1.0
≥5	167.1	3.0	1.1

*Maternal age controlled by indirect standardization, using the "birth rank 1" group as the standard.

Question E7–2

An English study of over 2,500 patients who were treated for hypertension showed that 6% died during 4 years of follow-up (Bulpitt et al., 1979). Patients were entered into the study at presentation to a hospital hypertension clinic (86%) or when seen in general practice with hypertension (14%). The cumulative mortality rate after 4 years was 12% for smokers and 5% for nonsmokers. This difference was statistically significant ($P < .001$).

Can you think of any reason to suggest that the difference may be an artifact?

Question E7–3

The investigators compared the characteristics of the hypertensive patients who subsequently died and those who stayed alive. Weight, serum cholesterol, pulse rate, and a history of angina pectoris were not associated with mortality, and could be exonerated from suspicion as confounders. Characteristics that were related to mortality were included, together with smoking, in multiple regression and multiple logistic regression models. The multivariate analyses (in which mortality was the dependent variable) showed significant associations with smoking, age, systolic blood pressure level, and plasma urea; doubtfully significant associations with retinal hemorrhages, proteinuria, and a history of myocardial infarction; and no significant associations with diastolic blood pressure before treatment, serum uric acid, and other variables.

The multiple logistic analysis showed that, controlling for other variables, the odds ratio for the association between smoking and death was 3.6 ($P < .001$). Can we conclude that smoking increased the risk of dying in this group of treated hypertensive patients?

Question E7–4

If we conclude that the patients who smoked had a higher risk of dying because of their smoking, can we infer that their risk would have been reduced if they had stopped smoking?

Question E7–5

The next two questions are based on a study of the association between the use of artificial sweeteners and weight change, in which women who said they added sweeteners (mainly saccharin) to beverages or food were compared with women who said they did not (Stellman and Garfinkel, 1986). The dependent variable was weight change during a 1-year period.

The information was obtained from a single questionnaire, which included questions about the use of sweeteners, current weight, and weight 1 year previously. The difference between these two weights was the dependent variable. The questionnaire was administered during the baseline investigation of subjects enrolled in a prospective mortality study in the United States, in which over a million people were enlisted.

"Rather than attempt to adjust for a multitude of factors," this analysis was confined to 78,694 white women aged 50–69 with at least a high school education, with no history of diabetes, heart disease, or cancer, who said there had been no major change in their diet in the past 10 years and that they had not changed their smoking status for at least 2 years. To simplify the analysis, only two groups were compared: women who said they had used sweeteners for 10 or more years, and women who said they had never used them.

How would you classify this study? Cross-sectional? Case-control? Cohort?

Question E7–6

There were no differences between users and nonusers of sweeteners with respect to the mean number of times per week they reported eating beef, pork, liver, ham, smoked meats, franks or sausages, carrots, squash, citrus fruits or juices, cereal or oatmeal, ice cream, and chocolates. Users ate green leafy vegetables, tomatoes, cabbage, chicken, and fish more frequently than did nonusers; and ate butter, white bread, and potatoes less frequently. Information on quantities was not available.

The percentages who reported losing and gaining weight during the previous year are shown in Tables E7–2 and E7–3. The results are stratified by relative weight at the start of the year. The percentages are age-standardized by the direct method, using 5-year age intervals.

Do the findings show that artificial sweeteners cause a gain in weight? What other explanations may there be?

Question E7–7

A study of dog bites showed that dogs kept chained were much likelier than unchained dogs to bite nonhousehold members. This suggested that "owners may be able to . . . modify risk by . . . not keeping them chained," according to an abstract printed in the proceedings of a scientific meeting (Gershman, 1992). Do you agree with this inference?

Table E7–2. Percentage of Women Who Lost Weight During a One-Year Period, by Use of Sweeteners and Relative Weight* at Start

Relative Weight	Percentage Who Lost Weight: Sweeteners Used for ≥10 Years	Percentage Who Lost Weight: Sweeteners Never Used	Ratio	*P*
Very Low	11.9	12.0	0.99	NS
Low	14.9	16.0	0.93	NS
Average	18.5	19.2	0.96	NS
High	22.2	23.8	0.93	NS
Very High	28.2	25.6	1.10	NS

*Quetelet's body mass index (quintiles).

Table E7–3. Percentage of Women Who Gained Weight During a One-Year Period, by Use of Sweeteners and Relative Weight* at Start

Relative Weight	Percentage Who Gained Weight: Sweeteners Used for ≥10 Years	Percentage Who Gained Weight: Sweeteners Never Used	Ratio	*P*
Very Low	32.3	29.6	1.09	<.001
Low	39.0	33.5	1.16	<.001
Average	41.5	35.0	1.19	<.001
High	41.5	32.4	1.28	<.001
Very High	31.9	26.3	1.21	<.001

*Quetelet's body mass index (quintiles).

Notes

E7–1. Quasi-experimental designs and their strengths and weaknesses are described by Campbell and Stanley (1966), Campbell (1969), and Cook and Campbell (1979).

E7–2. Stark and Mantel (1966). For a detailed explanation of standardization, using this example, see Fleiss (1981, chap. 14).

■■■■■■■■■■ Unit E8

Artifact, Confounding or Cause?

When an association is found, a causal explanation can be seriously considered only if the association cannot readily be explained as an artifact or a consequence of confounding.

In answer to *Question E7–1,* the association between birth rank and Down's syndrome virtually disappears when maternal age is controlled by indirect standardization. The findings thus provide no support for the hypothesis that birth rank influences the risk of the disease. The strong association shown by the crude data can be attributed to the confounding effect of maternal age. Confounding does not usually produce strong associations. But it can, as these findings show.

The cohort study of hypertensive patients (*Question E7–2*) showed a higher 4-year mortality for smokers than for nonsmokers. The difference may, however, be due to lead time bias (Unit B10), since the starting-point for follow-up was the beginning of treatment—in most cases, the first attendance at a hypertension clinic. It is possible that the smokers were people who tended to take less care of themselves, and began to get treatment for their hypertension at a later

stage in the natural history of the disease than did nonsmokers. Their mortality may have been higher because their disease was more advanced.

The results of the subsequent analysis (*Question E7–3*) suggest that the association was not an artifact caused by lead-time bias, since the variables controlled in the multivariate analyses include some that are indicative of the stage of the disease at the outset (the initial blood pressure level and the presence of cardiac, renal, and eye complications of hypertension at entry into the study). The results also show that the association was not caused by the confounding effects of the other variables examined. It is probably safe to infer that smoking increased the risk of dying.

It does not follow, however, that giving up smoking would necessarily have lessened the risk of dying (*Question E7–4*), because some etiological factors have irreversible effects that remain after the factor is removed. We would require other evidence, based on observational or experimental comparisons of the mortality of hypertensives who cease and continue to smoke.

The study of artificial sweeteners (*Question E7–5*) is best classified as a cross-sectional study in which information was obtained about past as well as present characteristics. A prospective approach was used in the analysis. It is not a typical cohort study—although a cohort study can be based on historical data (a *historical prospective study*)—because the information about the use of sweeteners was not collected before the occurrence of the outcome. The study has the potential biases of a cross-sectional study.

A causal relationship between sweeteners and weight gain (*Question E7–6*) is not inconceivable. The mechanism might be pharmacological or psychological—for example, a tendency to regard the addition of sweeteners as a substitute for caloric restriction. However, we should consider other explanations. First, the data concerning weight change (calculated from reported weights) may be biased. It can be claimed that "since changes in weight between two points in time are used . . . any bias due to systematic under-estimation by individuals will tend to be minimized" (Stellman and Garfinkel, 1986). But the validity of the information may be different in users and nonusers. Women who are "weight-conscious"—and therefore take sweeteners and avoid butter, white bread, and potatoes—may, because of this awareness, be especially likely to report that they are gaining weight. Second, there may be confounding by some factor not controlled by the procedures used (these were: limiting the study to a homogeneous group of subjects, stratifying for relative weight, and standardizing for age). One possible confounder is weight change prior to the year under consideration. Women who had previously been gaining weight (and were therefore using sweeteners) may have tended to continue their weight gain during the year of the study, producing the association that was found. Weight gain may have *preceded* the use of sweeteners.

You may have thought of other explanations.

In answer to the question about dog bites (*Question E7–7*), the association with being kept chained may be due to confounding. As subsequently stated in the full report of the study (Gershman et al., 1994), "a dog may be chained as

the result of having exhibited aggressive behavior, which itself may be a risk factor for biting, rather than chaining somehow causing a dog to bite."

Exercise E8

Question E8–1

An association cannot be regarded as causal if it can be completely explained by confounding—that is, if it disappears when other variables (that cannot be regarded as intermediate causes) are held constant. We have encountered many ways of dealing with confounding in these exercises. How many can you list?

Question E8–2

It often happens that a study has more potential confounders than can be handled simultaneously in a multivariate analysis. You may come across studies using the following ways of deciding which variables to control when analyzing the association between a risk factor and a disease. What do you think of them?

1. Select variables whose confounding effects have been shown to be important in other studies of the topic.
2. Select variables that are significantly related to both the risk factor and the disease.
3. Select variables that are strongly associated with the risk factor and the disease (using odds ratios or other measures of strength).
4. See how the strength of the association between the risk factor and the disease (measured by, say, the odds ratio) is affected when each variable in turn is controlled, and select the variables that make the most difference.
5. Do a multivariate analysis, starting with a simple set of potential confounders (e.g., age and sex); then, by trial and error, find the variable whose addition has the biggest effect on the strength of the association, and add it; repeat this until the change becomes negligible.

■■■■■■■■■■ Unit E9

Coping with Confounding

In answer to *Question E8–1,* confounding may be handled in various ways. The following methods have been mentioned or used in previous pages.

1. Confounding may be reduced or prevented by the manner of selecting the study sample or samples:
   - individual and group matching (Unit D11).

- restriction of the study to a homogeneous group (Question E7–5).
- random allocation to experimental groups (Unit E6).
- stratified allocation to experimental groups (Unit E6).

2. In the analysis, confounders may be held constant by stratifying the data and then using the stratum-specific findings (Unit A11). Post-stratification may be used when analyzing the results of a trial (Unit E6).
3. Other methods that may be used in the analysis include
   - direct standardization (Unit B14).
   - indirect standardization (Unit B13).
   - Mantel-Haenszel and similar procedures based on stratified data (Unit D13).
   - multivariate analysis (Units D7, D13)—for example, multiple linear regression (Unit D17), multiple logistic regression (Units D14, D15) and proportional hazard regression (Unit D16).
   - current life table analysis (Note B9–3).
   - partial correlation coefficients (Unit D11).
4. Use is sometimes made of dependent variables that incorporate, and thus neutralize the effect of, the confounder(s) (Unit A15)—for example, use of the intelligence quotient (IQ) as a way of controlling for the effect of age on test achievement. These include "residuals" based on regression analysis (Unit D16).
5. Confounding is sometimes handled by reasoning, based on the (non-foolproof) logic of the exclusion test (Unit D5), the Direction Rule (Unit D5), and estimates of the magnitude of the possible confounding effect (Note D6).

In answer to *Question E8–2,* all these methods of selecting potential confounders to be controlled have their advocates. It is common practice to start with variables that have been shown to be important in other studies of the disease—for example, age and sex and (say) smoking (option 1). If this is not done, readers may mistrust the study. Other variables are then selected by appraising the findings and either selecting the potential confounders that are most likely to be actual confounders (options 2 and 3) or selecting those that have most effect on the association between the risk factor and the disease (options 4 and 5).

Option 3 is preferable to option 2, since it is based on the strength of associations rather than on statistical significance. An important confounding effect is likely only if the associations with the risk factor and disease are strong. Large effects may be nonsignificant if sample size is small, and unimportant effects may be significant if sample size is large. If significance tests are used, it has been suggested that variables should be rejected only if $P > .20$.

Option 4 may be used as a preliminary to option 5 so as to exclude nonconfounders and weak confounders before seeing whether the confounding effects persist in a multivariate setting. Option 5 is a "forward selection" strategy, and the number of variables may become too large for the analysis to handle. A symptom of this is the appearance of a very high or very low measure of association (e.g., an odds ratio of over 10 or under 0.1), and this should excite suspicion. The

counterpart of option 5 is a "backward selection" strategy: the analysis starts with as many variables as possible, and these are then pruned by repeatedly removing the variable with the smallest effect on the measure of association, until the measure becomes appreciably different from what it was at the start. See Note E9–1.

Delving into Causes

We cannot "prove" a causal relationship. The best we can hope for is that new facts will consistently conform with what we would expect to find if the association were causal. The key to the study of causation is the development of hypotheses that can be subjected to empirical testing (Units A6, A15, A16). Clues, ideas, and new specific hypotheses often arise during the analysis, in the form of inferences that emerge when associations are elaborated and variables are refined. Hypotheses may be tested in the framework of a single study, by subjecting the available data to additional analyses, or may need new data.

In the long run, judgments about causal relationships are based on evidence that comes from many studies, including nonepidemiological ones (Note E9–2). Studies may be reviewed and appraised in an informal way, or their results may be subjected to an integrated statistical analysis (*meta-analysis;* see Section F).

A great deal has been written about methods and criteria for the appraisal of causality (Note E9–3).

Exercise E9

What would persuade you that one variable is causally related to another? List as many criteria as you can.

Notes

E9–1. The selection of confounders for controlling and the biases that may arise are discussed by Rothman and Greenland (1998, pp. 256–259).

E9–2. For examples of the way that etiological knowledge has evolved from the complementary contributions of population studies, clinical observations, and laboratory experiments, see Morris (1975, pp. 250–261).

E9–3. Methods of deciding whether an association is causal are discussed in all epidemiology textbooks. For fuller discussions, see Susser (1973, pp. 140–162), Susser (1986), and Rothman and Greenland (1998, pp. 24–28). See Note A6–1.

Evidence for a Causal Relationship

A well-designed experiment can provide better evidence for a causal relationship than a survey can, and the evidence is strongest if the findings are replicated in other experiments.

Whatever kind of study the evidence comes from, there are four basic conditions that must be met before a causal relationship between two variables can be seriously contemplated. These prerequisites are that

- The variables are associated with one another.
- The association cannot readily be explained as an artifact.
- The association cannot readily be explained as an effect of confounding.
- The "cause" precedes the"effect," or (at a minimum) there is no evidence that the "effect" precedes the "cause."

A number of additional criteria that, taken together and not individually, may strengthen or weaken the case for a causal association, although they cannot provide absolute proof that the causal hypothesis is true or false. The following list (based in part on Susser, 1986) states what evidence may be regarded as supporting or weakening the case for a causal association. "Indeterminate" findings that neither strengthen nor weaken the case—such as the *absence* of a dose–response relationship—are not specified.

- *Probability.* Statistical significance supports the case for a causal association. Absence of statistical significance or a significant equivalence test (see Unit D4) weakens it, but only if the test is powerful (large numbers).
- *Strength of the association.* A strong association (e.g., a high or low risk ratio) supports the case. The stronger the association is, the more likely that it is causal, and not produced by bias or confounding; but a weak association may also be (weakly) causal.
- *Dose–response relationship (biological gradient).* If there is a monotonic association between the amount, intensity, or duration of exposure to the "cause" and the quantity or severity of the "effect," this supports the case. There may also be an *all-or-none* response that appears only when the causative factor reaches a threshold level, or a relationship that is U- or J-shaped (or inverted U- or J-shaped) other than linear, suggesting a more complex causal relationship.
- *Time–response relationship (temporality).* If the incidence of the "effect" rises to a peak some time after a brief exposure to the "cause" and then decreases, this supports the case.
- *Predictive performance.* If information about the "cause" is predictive of the occurrence of the "effect," this supports the case (but it may be a risk marker and not a cause); if it is not, it weakens it. The case for a new *a priori* causal

hypothesis can be supported or weakened by the results of an experiment or survey that tests predictions based on the hypothesis.

- *Specificity.* The finding that the "effect" is related to only one of a set of alternative "causes" (e.g., exposure to different microorganisms), or that the "cause" is related to only one "effect," may be regarded as supporting the case. But a lack of specificity in no way negates a causal relationship.
- *Consistency on replication* (in different populations, circumstances, and studies). If the same association is found repeatedly, this strongly supports the case. If results are inconsistent, and the variation cannot be attributed to modifying factors or differences in study methods, this weakens the case.
- *Coherence* with current theory and knowledge (plausibility) supports the case. Incompatibility with known facts weakens it.

Exercise E10

Question E10–1

Table E10–1 shows the association between beer drinking and rectal cancer in men, according to a case-control study in the United States (Kabat et al., 1986). The odds ratios are based on a multiple logistic regression analysis in which suspected confounders were controlled. Are the results consistent with a causal explanation?

Question E10–2

The authors of the paper on beer and rectal cancer provided the review of epidemiological studies shown in Table E10–2. On the basis of this evidence, does beer drinking (in your judgment) increase the risk of rectal cancer?

Question E10–3

A cohort study of 361,662 men aged 35–57 years revealed an association between smoking (the number of cigarettes smoked per day, at the outset) and suicide during a 12-year follow-up period, as shown in Table E10–3 (Smith et al., 1992). The relative rates are adjusted by proportional hazards regression analysis, to control for possible confounding by age, race, socioeconomic status (as measured by the median family income, and the postal Zip code for area of res-

Table E10–1. Association Between Beer Drinking and Rectal Cancer

Beer Consumption	Odds Ratio	95% Confidence Interval
Never	1.0	—
Occasional	1.4	0.8–2.6
1–7 oz/day	1.4	0.7–2.6
8–31 oz/day	1.6	0.8–3.1
≥32 oz/day	2.7	1.3–5.7

Table E10–2. Evaluation of Studies of Beer Drinking and Rectal Cancer Risk

Criteria	Fit*	Comments
Strength	+	The relative risks, where elevated, are small or borderline.
Specificity	+	Two correlation [group-based] studies have found significant positive correlations between beer and a number of cancers other than the rectum and colon.
Consistency	+	Five of ten case-control or prospective studies showed no association. Several correlation [group-based] studies showed . . . an association, but one did not.
Dose–response	+	None of the published studies, except the present one, provides evidence of a dose–responsive relationship.
Temporal sequence	++	Three published prospective studies showed a positive association . . . ; one found no association . . .
Biological rationale	+	Ethanol by itself has not been shown to be a carcinogen. Furthermore, no epidemiological studies have reported an association of wine or whiskey with rectal cancer . . .

*Fit is defined as how well the existing evidence fulfills each of the criteria. +++ = good, ++ = fair, + = poor.

Source: Kabat et al. (1986) (table abbreviated).

Table E10–3. Association Between Smoking and Suicide

Cigarettes per Day	Suicide Rate per 10,000 Person-Years	Adjusted Relative Rate (With 95% Confidence Interval)
0	1.09	1.00
1–19	1.47	1.36 (1.00–1.84)
20–39	2.00	1.86 (1.54–2.26)
40–59	2.46	2.27 (1.76–2.92)
60+	3.78	3.33 (2.01–5.52)

Chi-square test for trend: $P < .0001$.

Table E10–4. Association Between Smoking and Being Murdered

Cigarettes per Day	Adjusted Relative Murder Rate (With 95% Confidence Interval)
0	1.00
1–39	1.71 (1.29–2.28)
40+	2.04 (1.32–3.15)

idence), previous myocardial infarction, and diabetes (taking of medication). The investigators cite two previous studies that yielded a similar result. This study also showed an association between smoking and being murdered (Table E10–4); the adjusted relative rates control for possible confounding by age and socioeconomic status. On the basis of this evidence, does smoking (in your judgment) increase the risk of suicide and murder?

■■■■■■■■■■ Unit E11

Evidence for a Causal Relationship (Continued)

In answer to *Question E10–1,* the results shown in Table E10–1 are consistent with a causal relationship between beer drinking and rectal cancer. There is evidence of a dose–response relationship: the association is strongest in men who drink most beer. As the confidence intervals show, only in this group is the association statistically significant.

Their review of the available epidemiological evidence on beer and rectal cancer (*Question E10–2*) led Kabat et al. (1986) to the conclusion that

> it is clear that the existing studies, at best, provide weak support for a causal association, . . . Two explanations can be proposed to explain the conflicting results. . . . The first is that some component of beer itself is a weak initiator or promoter of rectal cancer. The alternative explanation is that the association . . . is indirect [i.e., due to confounding] and that beer consumption is associated with an as yet unknown factor, possibly dietary in nature, that is itself a rectal carcinogen . . . ; we are inclined to favour the second explanation.

You may or may not agree with this appraisal. The interpretation of the criteria of causality is a matter of judgment, and judges may disagree.

In answer to *Question E10–3,* the results presented are consistent with causal relationships between smoking and suicide and murder. The temporal sequence is correct; the associations are strong and statistically significant; there are dose–response relationships; and possible confounders have been controlled. Other studies have shown similar results.

After considering this evidence, you may have decided that it is not plausible that smoking is a causal factor, and the associations are probably explained by inadequate control of the confounders or by the study's failure to take additional confounders into account. In other words, your judgment may be that smoking is probably correlated with other factors that lead to an increased risk of suicide or being murdered. You may even have considered the possibility that (since $P < .001$) the findings represent a 1 in over-1,000 long-shot chance occurrence.

Conversely, if you were able to think of mechanisms whereby smoking might

lead to an increased risk of suicide and murder, you may have opted for a causal relationship.

This is a strange quandary—does one's acceptance or rejection of epidemiological evidence for causality depend on plausibility? After all, plausibility—the availability of a possible explanation that is coherent with current theory and knowledge—may depend solely on one's inventiveness. As the authors of the cited study point out, investigators who have found opposite associations (e.g., between oral contraceptive use and a low risk of HIV infection, or between oral contraceptive use and a high risk of HIV infection) have had no difficulty in suggesting plausible mechanisms. The ability to think of a plausible mechanism may lead to a decision that a noncausal association, actually attributable to defective study methods or confounding, is causal.

To add to the dilemma, there have been numerous examples of causal relationships—subsequently confirmed by experiments or intervention studies or programs—that were brought to light by epidemiological studies at a time when their biological mechanisms were unknown. Examples are the relationships of smoking to lung cancer and other diseases, and of putting babies to sleep on their abdomens to the sudden infant death syndrome (SIDS).

A plausible biological mechanism is not a condition for the acceptance of a causal relationship demonstrated in an experiment, although it may explain it. The postulated mechanism is not necessarily correct, and in this sense an experiment is not a foolproof test of a causal hypothesis. As an example (from Rothman and Greenland, 1998, p. 27), an observed drop in the incidence of malaria after the draining of swamps, in an experiment conducted to test the hypothesis that the disease is caused by swamp gas (methane), may be incorrectly interpreted as support for this hypothesis.

The Impact of a Causal Factor

We now leave causes and pass on to consider their *effects*. Our last topic is the measurement of *impact* on morbidity. Once we have decided that a factor is causal, there are several ways of expressing the magnitude of its influence on the occurrence of a disease in a given population or population group.

For example, we can say how much disease a given factor causes, expressed as a number of cases (the *attributable number*) or as an incidence or prevalence rate; if an incidence rate is used, this is the *attributable risk* or *excess risk*. Alternatively, we can say what *proportion* of the total incidence or prevalence can be attributed to this cause. This is the *attributable* or *etiologic fraction;* it may refer to the impact on the total population (the *population attributable fraction*) or only to the impact on people exposed to the causal factor—that is, the *attributable fraction (exposed).*

If the factor is a protective one (not a risk factor), we can speak of the amount of potential disease it *prevents*—that is, the *prevented fraction* in a total population or in people exposed to the factor.

We can also speak of the *preventable fraction*—the proportion of the ob-

Table E11–1. Prevalence of Varicose Veins in Male Workers Aged 20–64 in Jerusalem, by Work Posture

Work Posture	Prevalence Rate %
Standing*	12.3
Other	7.7
Total	8.3

*For at least half the working time.

served incidence that *could* be prevented by removal of a given risk factor or exposure to a given protective factor.

The exercises will use simple calculations only. Depending on what data are available, the calculation of measures of impact—and especially of their confidence intervals—may be more complicated (Note E11).

Exercise E11

Watch out for at least one "trick question" in this exercise.

Question E11–1

There is much evidence that prolonged standing is a cause of varicose veins. An association between standing and varicose veins is shown in Table E11–1, which is based on a population study (Abramson et al., 1981).

Using these data, what proportion of the varicose veins in men who work standing can you attribute to their standing? This is the attributable fraction (exposed). To calculate it, assume that if these men had not worked standing, their prevalence of varicose veins would have been 7.7% instead of 12.3%.

Question E11–2

What proportion of the varicose veins in this total male working population can be attributed to standing? This is the population attributable fraction. (Assume

Table E11–2. Prevalence of Varicose Veins in Male Workers Aged 20–64 in Epiville, by Work Posture

Work posture	Prevalence Rate %
Standing	12.3
Other	7.7
Total	9.7

that if men had not worked standing, the rate would have been 7.7% instead of 8.3%.)

Question E11–3

Table E11–2 presents fictional data from a similar study in Epiville. (This is Epiville's swan song; farewell, Epiville.) Note that the exposure-specific rates of varicose veins are identical to those in Jerusalem.

Using the data in this table, calculate the attributable fraction (exposed) and the population-attributable fraction. Compare your answers with the figures for Jerusalem. How is the difference explained?

Question E11–4

In Table D7, we saw that the annual incidence of CHD was 5.9 per 1,000 in Paris policemen with varicose veins, and 2.9 per 1,000 in those without varicose veins. What proportion of the incidence of CHD in policemen with varicose veins can be attributed to their varicose veins?

Question E11–5

In Table D8, we saw that the annual mortality rate was 4.0% in cigarette-smoking men aged 65–74, and 2.4% in men who had never (or only occasionally) smoked. What proportion of the mortality in the smokers can be attributed to their smoking? (This is the attributable fraction in the exposed.) Do you have any reservations about your answer?

Question E11–6

Suppose that in Question E11–5 you were not told the rates, but only the relative risk in cigarette smokers, which was 1.67. Could you have calculated the attributable fraction in the exposed?

Question E11–7

For what purposes may attributable fractions be used?

Note

E11. Basic measures of impact are explained in all epidemiology textbooks. For statistical procedures (see Note A3–7), see Kahn and Sempos (1989, chap. 4), Kleinbaum et al. (1982, chap. 9), or Rothman and Greenland (1998, pp. 53–58, 295–297). There is considerable confusion about nomenclature, and you may encounter the same terms used differently.

■■■■■■■■■■Unit E12

The Attributable Fraction

A cause–effect relationship has been established between standing and varicose veins. The difference between the rates of varicose veins in men who stand when at work and those who do not can therefore be used as a measure of the impact of standing. We answer *Question E11–1* by assuming that the men who stood would have had a prevalence rate of 7.7% if they had not stood, instead of 12.3%. The difference, 4.6%, can be attributed to their standing. (If this were a difference between incidence rates, it could be called the "attributable risk.") Expressed as a proportion of the total prevalence in men who stand at work, this is 4.6/12.3, or 37%. In other words, 37% of the prevalence of varicose veins among workers who stand can be attributed to their standing. This is the attributable or etiological fraction (exposed).

Similarly, the prevalence of varicose veins in the men as a whole would have been 7.7% if no one had stood when at work, instead of 8.3%. The population attributable fraction (*Question 11–2*) is therefore (8.3 − 7.7)/8.3, or 7%.

In Epiville (*Question 11–3*) the attributable fraction (exposed) is again 37%, but the population attributable fraction is now (9.7 − 7.7)/9.7, or 21%, which is considerably higher than in Jerusalem, despite the identical exposure-specific rates. The reason, of course, is that in Epiville more men worked standing. Clearly, a population attributable fraction depends not only on the exposure-specific rates, but on the prevalence of the causal factor in the population. It cannot be applied to populations other than the one in which it was calculated.

The attributable fraction is meaningful only if the factor is a causal one or can be regarded as a proxy for a closely correlated, causal, factor. *Question 11–4* therefore cannot be answered. (This is the trick question.)

In *Question 11–5* the proportion of the smokers' mortality attributable to their smoking is (4.0 − 2.4)/4.0, or 40%. The main reservation (and this applies to standing and varicose veins also) is that the difference may be partly attributable to confounding factors. *This possibility should be kept in mind whenever attributable fractions are used* (although somehow it often remains unvoiced when they are used to convince decision-makers of the urgency of a problem).

The attributable fraction (exposed) can easily be calculated from the relative risk (RR). It is (RR − 1)/RR. In *Question 11–6*, it is 0.67/1.67, or 40%.

The population attributable fraction can be calculated from the relative risk, provided that the relative risk was derived from a study of representative samples, and additional information is available (Note E12). The odds ratio can often replace the relative risk in these calculations (see Notes D10–1 and D10–2).

In answer to *Question 11–7*, attributable fractions are of use mainly to those concerned with practical aspects of health care. The attributable fraction is based on the absolute difference between rates, and it measures the magnitude of the problem produced by a specific risk factor. The attributable fractions in the population and the exposed are easily understood measures, useful as a ba-

sis for determining priorities and for communicating epidemiological findings to nonepidemiologists.

Exercise E12

This exercise deals with *prevented* and *preventable* fractions.

Question E12–1

A follow-up study in a community in Jerusalem showed that the mortality attributable to hypertension was 23%. This was the population attributable fraction, based on a comparison of 10-year mortality in adults who had raised and normal blood pressures at the outset of the study (Goldbourt and Kark, 1982). Can we infer that this is also the preventable fraction in the population—that is, the proportion of deaths that would be prevented by appropriate intervention with respect to hypertension?

Question E12–2

So far we have considered *risk* factors. This question and the following ones deal with the impact of *protective* factors. Table E12 presents the results of a trial of a whooping cough vaccine performed in England in the 1940s, when this vaccine was still new. Children were randomly allocated to the "vaccinated" and "unvaccinated" groups, and they were followed up for 2 to 3 years (Hill, 1962).

What proportion of the incidence was prevented, in children who were vaccinated? This is the *prevented fraction in the exposed* (i.e., in those exposed to this protective factor).

Question E12–3

Fictional data: In England as a whole, the incidence of whooping cough at that time was 6 per 100 child-years. The use of the vaccine throughout the country was patchy, and the number of children who were vaccinated was unknown. Assume that the data in Table E12 refer to representative samples of the vaccinated and unvaccinated children in England. Using these figures, what was the impact of vaccination on incidence in the total child population? That is, what

Table E12. Incidence of Whooping Cough per 100 Child-Years

Group	Incidence Rate
Vaccinated	1.74
Unvaccinated	8.07

proportion of the potential incidence of whooping cough was prevented by vaccination? (This is the prevented fraction in the population.)

Question E12–4

Using the same figures, what proportion of the actual incidence of whooping cough in the child population would be prevented if all children were vaccinated? (This is the *preventable fraction in the population.*)

Question E12–5

What was the preventable fraction in unvaccinated children?

Question E12–6

As we have previously seen (Table E5–4), a randomized trial of treatment for mild hypertension in the elderly showed that the mortality rate per 1,000 person-years was 34 in the treated group and 47 in the control (placebo) group. On the basis of these figures, how efficacious was the treatment in preventing cardiovascular deaths? The *P* value was .037. Do you think your measure of efficacy has a wide or narrow 95% confidence interval?

Question E12–7

For what purposes may prevented fractions be used?

Question E12–8

For what purposes may preventable fractions be used?

Note

E12. The population attributable fraction can be estimated from the relative risk (RR) if we know the proportion (F) of the population exposed to the risk factor. The formula is $F(RR - 1)/[F(RR - 1) + 1]$. An alternative formula is $F'(RR - 1)/RR$, where F' is the proportion of *cases* who were exposed to the factor. If the risk is low, the odds ratio (OR) can replace RR in these formulae.

■■■■■■■■■■ Unit E13

Prevented and Preventable Fractions

The attributable fraction is a ceiling estimate of the preventable fraction. To predict what fraction of the mortality can be prevented by controlling hypertension

(*Question E12–1*), we also need to know how effectively hypertension can be controlled, and the influence of blood pressure reduction on mortality. We should also consider possible confounding effects: associated risk factors may partly account for the magnitude of the attributable fraction. We might conclude that the preventable fraction is appreciably less than the attributable fraction.

To estimate the prevented fraction in children exposed to vaccination (*Question E12–2*), we can assume that their incidence rate would have been 8.07% instead of 1.74%, if they had not been vaccinated. The difference (6.33%) can be attributed to the preventive effect of vaccination. The prevented fraction is therefore 6.33/8.07—that is, 78% of what the incidence would have been, had they not been vaccinated. This may be termed the *efficacy* of the vaccine, or the "percentage reduction." (Does it matter whether person-time or cumulative incidence rates are used in studies of vaccine efficacy? See Note E13–1.)

The incidence rate in the total child population (*Question E12–3*) would (hypothetically) have been 8.07 per 100 child-years (Table E12), had no children been vaccinated. The actual incidence was 6%. The difference (2.07%) can be attributed to the preventive effect of vaccination. The prevented fraction in this population is therefore 2.07/8.07 = 26%.

If all children were vaccinated (*Question E12–4*), the expected incidence would be 1.74% (Table E12). In fact, it was 6%. The difference (4.26%) tells us what part of the actual incidence would have been prevented. Expressed as a proportion, the preventable fraction in the population is 4.26/6, or 71%.

The preventable fraction in unvaccinated children (*Question E12–5*) is 6.33/8.07, or 78%. This is, of course, the same as the prevented fraction in vaccinated children (*Question E12–2*).

In answer to *Question E12–6,* the prevented fraction in the exposed (treated) sample is a measure of the efficacy of treatment. It is (47 − 34)/47, or 28%. This can also be derived from the relative risk (RR): it is (1 − RR). The relative risk is 34/47 = 0.72, and 1 − 0.72 = 0.28. The "high" *P* value of .037 suggests a wide 95% confidence interval, because the lower confidence limit cannot be far from zero; the 95% interval of the prevented fraction was in fact 1–46%.

In answer to *Question E12–7,* the prevented fraction in people exposed to a preventive procedure is, as we have seen, a measure of efficacy. It is an index commonly used when procedures are tested and compared, both for primary preventive procedures like vaccination and for therapeutic procedures that aim to prevent complications. The prevented fraction in the population measures the effectiveness of a preventive program. (What is the difference between "efficacy" and "effectiveness"? See Note E13–2.)

Preventable fractions (*Question E12–8*) provide both a guide and a stimulus to action. The preventable fraction in people exposed to a risk factor can be applied to individuals as well as to groups, to dramatize the likely effect of change or intervention: "If you stop smoking you will reduce your risk of so-and-so by such-and-such per cent." The preventable fraction in the population is of value to decision-makers who are planning health services, as it provides an estimate of the impact that intervention is likely to have on the public's health.

Notes

E13–1. Vaccines are commonly used for diseases with a high incidence. Person-time and cumulative incidence rates may therefore be dissimilar, and give different estimates of vaccine efficacy. Cumulative incidence rates are more appropriate if vaccination is believed to render a proportion of people completely immune (Smith et al., 1984).

E13–2. *"Efficacy"* and *"effectiveness"* are often used synonymously, but are sometimes distinguished from each other. "Efficacy" often refers to the benefits when a procedure is applied as it "should" be, with full compliance by all concerned (as in a clinical trial subjected to "on randomized treatment" analysis); and "effectiveness," to the benefits at the population level, or among people to whom the procedure or service is offered. According to this usage, a program for the control of hypertension in a community would use drugs known to be efficacious; the program might or might not be effective.

■■■■■■■■■■ Unit E14

Test Yourself (E)

- Explain the difference between
  experiments and surveys (E1).
  descriptive and analytic surveys (E1).
  cross-sectional, case-control, and cohort studies (E1).
  a retrospective and a prospective approach (E2, Note E2).
  retrolective and prolective studies (Note E2).
  an attributable risk and an attributable fraction (E11).
  population attributable risk and attributable risk (exposed) (E11).
  efficacy and effectiveness (Note E13–2).
- Say whether a direct measure of risk can be provided by
  a cross-sectional study (E2).
  a case-control study (E3).
- State some of the possible biases of
  a cross-sectional study (E2).
  a case-control study (E3).
  a cohort study (E4).
- Explain what is meant by
  a group-based study (E1).
  a quasi-experiment (E1, E7).
  a nested case-control study (E3).
  diagnostic suspicion bias (E4).
  randomization (E6).

post-stratification (E6).
a time series (E7).
a historical prospective study (E8).
dose–response relationship (E10).
time–response relationship (E10).

- Calculate
  the number needed to treat (E6).
  an attributable fraction (exposed) from rates and from a relative risk (E12).
  a population attributable fraction (E12).
  a prevented fraction (exposed) from rates and from a relative risk (E13).
- State the main drawbacks of group-based studies (E5).
- Say to whom the following can be applied:
  the results of a clinical trial (E6).
  a population attributable fraction (E12).
- Explain the advantages of
  "blind" studies (E6).
  "intention-to-treat" analysis (E6).
- Explain how to use a case-control study to evaluate care (E7).
- Provide a list of
  ways of selecting the potential confounders to be controlled (E9).
  ways of handling confounding (E9).
  criteria for the appraisal of causality (E10).
- State the uses of
  attributable fractions (E12).
  the prevented fraction (exposed) (E13).
  the preventable fraction (exposed) (E13).
  the preventable fraction (population) (E13).
  the prevented fraction (population) (E13).
- State the conditions for using the following to estimate the relative risk in a target population:
  an odds ratio from a case-control study (E3).
  a relative risk from a cohort study (E4).

Section F

Meta-Analysis: Putting It All Together

"That's the most important piece of evidence we've heard yet," said the King, rubbing his hands; "so now let the jury—"

"If any of them can explain it," said Alice, "I'll give him sixpence. I don't believe there's an atom of meaning in it."

"If there's no meaning in it," said the King, "that saves a world of trouble, you know, as we needn't try to find any."

(Carroll, 1865)

Unit F1

Introduction

Meta-analysis refers to the critical review and integration of the findings of separate studies (Note F1). Its specific features are a systematic approach to avoid bias, and (where possible) the use of quantitative methods rather than reliance on judgment alone. These features distinguish it from most traditional literature reviews. The rapidly increasing volume of research, often with discrepant findings, has led to an increase in the need for and performance of meta-analyses.

This section has two main aims: to help you to adopt reasonable precautions when reviewing the results of a set of studies, by applying the basic principles that underlie good meta-analytic studies, and to help you to appraise published reports of meta-analyses and decide whether to use their results.

Some researchers prefer to speak of "systematic reviews" or "overviews" of research rather than "meta-analysis."

Exercise F1

Question F1–1

Meta-analysis is commonly used to integrate the results of different trials of a specific treatment or other intervention (curative or preventive). Can you think of other kinds of study that might be subjected to meta-analysis?

Question F1–2

Table F1 displays the results of 23 randomized controlled trials of long-term treatment with beta-blockers after myocardial infarction (Yusuf et al., 1985). The

Table F1. Results of 23 Randomized Controlled Trials of the Long-Term Use of Beta-Blockers After Myocardial Infarction; Comparison of Deaths in Subjects Allocated to Treatment and Control Groups

	Treatment Group		Control Group		Comparison of Fatality Rates (%)		
Trial	No.	Deaths	No.	Deaths	Ratio*	Difference**	*P*
1	11	1	11	1	1.00	0.0	NS†
2	38	3	39	3	1.03	0.2	NS
3	59	4	52	6	0.59	−4.8	NS
4	69	5	93	11	0.61	−4.6	NS
5	114	7	116	14	0.51	−5.9	NS
6	154	25	147	31	0.77	−4.9	NS
7	151	8	154	6	1.36	1.4	NS
8	174	6	134	3	1.54	1.2	NS
9	251	28	122	12	1.13	1.3	NS
10	207	33	213	38	0.89	−1.9	NS
11	209	32	218	40	0.83	−3.0	NS
12	263	45	266	47	0.97	−0.6	NS
13	278	25	282	37	0.68	−4.1	NS
14	291	9	293	16	0.57	−2.4	NS
15	355	28	365	27	1.07	0.5	NS
16	391	27	364	43	0.58	−4.9	.02
17	632	60	471	48	0.93	−7.0	NS
18	680	22	674	39	0.56	−2.6	.02
19	873	64	583	52	0.82	−1.6	NS
20	858	57	883	45	1.30	1.5	NS
21	945	98	939	152	0.64	−5.8	.0002
22	1,533	102	1,520	127	0.80	−1.7	NS
23	1,916	138	1,921	188	0.77	−2.6	.004
Total	10,452	827	9,860	986	0.79	−2.1	.0000002

*Ratio of rate in treatment group to rate in control group.

**Rate in treatment group minus rate in control group.

†NS = not significant ($P \geq .05$).

rate ratios comparing the occurrence of deaths in the treatment and control groups show considerable variation, ranging from 0.56 (i.e., a death rate lower by 44% in the treatment group) to 1.54 (a death rate higher by 54% in the treatment group). What reasons can you suggest for this variation?

Question F1–3

When reviewing the results of a set of studies of the same topic, would you expect to find more differences between the results of randomized controlled trials, case-control studies, or cohort studies?

Question F1–4

What advantages might there be in drawing conclusions from a series of studies, rather than a single study?

Note

F1. Quantitative methods of combining the results of studies were first described in the early 1930s. Interest grew in the 1970s, stimulated by the work of Glass (who coined the term "meta-analysis") and his colleagues (Glass et al., 1981). Meta-analyses in the health field began to appear in the 1970s, and started to flourish in the mid-1980s, largely because of the enthusiasm of Peto and his colleagues at Oxford. Methods are described by Chalmers and Altman (1995) and, in more detail, by Petitti (1994); statistical methods are described by Hedges and Olkin (1985) and Greenland (1998b); see Note A3–7. For a compendious review of principles and methods, see Yusuf et al. (1987). Problems are discussed by Abramson (1990/91) ["pros and cons"], Boden (1992) ["has a tool become a weapon?"], Chalmers (1991) ["problems induced by meta-analysis"], Eysenck (1995) ["problems with meta-analysis"], Felson (1992) ["bias in meta-analytic reearch"], Goodman (1991) ["have you ever meta-analysis you didn't like?"], Jenicek (1989) ["where we are and where we want to go"], Naylor (1988) ["two cheers for meta-analysis"], Spitzer (1991) ["unanswered questions about aggregating data"], and Thompson and Pocock (1991) ["can meta-analyses be trusted?"].

■■■■■■■■■■ Unit F2

The Scope of Meta-Analysis

In answer to *Question F1–1,* meta-analysis may in principle be applied to quantitative studies of any sort, including clinical trials and other experiments, quasi-experiments, and observational studies (e.g., cohort and case-control studies). Most meta-analyses deal with studies that aim to clarify causal associations; but meta-analysis may also be applied to studies of associations that are not necessarily causal (e.g., studies of risk markers) and to descriptive surveys (e.g., of the magnitude of a health problem). Meta-analysis may be applied to studies of screening or other diagnostic methods for use in individual or community health care (e.g., of their validity and reliability) or to studies of feasibility and cost, factors affecting the feasibility of interventions, and other topics.

Differences between the findings of separate studies (*Question F1–2*) may be due to chance, to differences in the design, execution or circumstances of the studies, or to differences between the people studied. This applies to both ex-

perimental and nonexperimental studies. Possible differences between controlled clinical trials include:

1. Differences in the criteria for inclusion in or exclusion from the trial, including differences in diagnostic criteria.
2. Differences in the baseline status of subjects, even if selection criteria are identical.
3. Differences in the manner of allocation to treatment and control groups (randomization or other methods).
4. Differences in the treatment under test, including differences in dosage and timing.
5. Differences in the management of controls (no treatment? placebos? other treatments?).
6. Differences in general management, including the diagnosis and treatment of other conditions, supportive care, responses to complications, etc.
7. Differences in outcome measures (e.g., due to differences in criteria).
8. Differences in follow-up times.
9. Variations in analysis—for example, the use of "intention to treat" or "on randomized treatment" analysis (see *Question E5–7*).
10. Differences in the quality of the study's design or execution—for example, in the precautions taken to avoid bias (e.g., the use of blind methods), in the criteria for withdrawing a subject from the allocated group or from the trial, in the efforts made to trace lost subjects, and in attention to the accuracy of measurements.

The trials listed in Table F1 used different drugs, doses, and exclusion criteria, and they differed in the time at which treatment was started and the duration of treatment and follow-up, which ranged from 6 weeks to 4 years.

In answer to *Question F1–3,* randomized control trials and cohort studies are less likely to yield divergent results than case-control studies. The use of randomized controls minimizes the likelihood of confounding, since the only differences in the initial status of the groups compared are those that occur by chance. Divergent results are more likely in nonexperimental studies (or nonrandomized trials or quasi-experiments), where it may be difficult to prevent or adequately control for differences between the groups compared. The possible biases in case-control studies—especially those resulting from an inappropriate selection of controls and from recall bias, exposure suspicion bias, and other forms of information bias (Unit E3)—are in general more difficult to avoid or control than the biases in cohort studies, especially if the cohort studies compare subgroups of the same population.

Possible advantages in drawing conclusions from a series of studies rather than a single study (*Question F1–4*) include the following:

1. If the studies have similar results, this consistency will reinforce the validity of whatever inferences are drawn (unless, of course, all the studies have the same bias).

2. Individual studies may be too small to yield statistical significance, especially if the effect is a weak one, but this may be overcome if the results of several studies are combined. As an example, increased infection rates (generally sepsis or pneumonia) were reported in seven randomized control trials of total parenteral (i.e., intravenous) nutrition in cancer patients undergoing chemotherapy, but in no trial was the effect statistically significant; but the combined results showed that the hazard was appreciable: there was a highly significant ($P < .0001$) fourfold increase in the odds in favor of developing infection (Klein et al., 1986).
3. If the findings are similar, combining them will provide better estimates of the effect of an intervention or the strength of other associations studied. Larger numbers will result in narrower confidence intervals.
4. Consideration of a series of studies may reveal that a result observed in an isolated study is an artifact or chance finding.
5. If there are differences between the findings, inquiry into the reasons for these differences may lead to new knowledge or the formulation of new hypotheses.
6. It may be possible to compare the effects of various interventions (applied in different studies).
7. It may be possible to compare various effects (examined in different studies) of an intervention.

Exercise F2

Question F2–1

This question deals with techniques of combining the results of separate studies. Suppose that we wish to use the findings shown in Table F1 as a basis for an overall conclusion about the value of the treatment tested in these trials, and that this is a reasonable thing to do. (In a later unit we consider the precautions that should be taken before results are combined.) What do you think of the following summary statements? What are the main advantages or disadvantages of the techniques used?

1. Altogether there were 827 deaths among the 10,452 subjects in the treatment groups (fatality rate, 7.9%) and 986 among the 9,860 controls (fatality rate, 10.0%). The pooled data thus show a rate ratio of 0.79 and a rate difference of −2.1%. A simple chi-square test shows that the difference between the pooled treatment and control groups is highly significant ($P = .0000002$). Before saying what you think of this method of analysis, look at the fictional data in Table F2.
2. Of the 23 trials, 16 showed a favorable effect (rate ratio less than 1 and rate difference less than zero) and 7 did not. This difference points to the value of the treatment.
3. Of the 23 trials, 16 showed a favorable effect and 7 did not. A chi-square goodness-of-fit test shows no significant difference ($P = .06$) between this

distribution (16 and 7) and the 50:50 distribution that might be expected by chance. (If you do not know what a goodness-of-fit test is, see Note F2–1.) The effect of the treatment is thus not statistically significant. Can we conclude that the treatment does not reduce the risk of dying?

4. Significance was tested by computing an overall *P* value from the 23 separate *P* values (for simplicity, most of these are not specified in the table). Several methods are available for this purpose (see Note F2–2). The overall *P* value was .0000005, showing that the difference in fatality between the treatment and control groups is highly significant.
5. Significance was tested by the Mantel-Haenszel chi-square test, which controls for effects connected with a stratifying variable (see Unit D13). The 23 trials were treated as separate strata; the data used were the numbers who died and survived in the treatment and control groups in each trial. The *P* value was .0000002.
6. The mean value of the rate ratio, computed by summing the 23 rate ratios and dividing by 23, was 0.87. This suggests that treatment prevented 13% of deaths.
7. The mean difference between rates, computed by summing the 23 differences and dividing by 23, was −2.3 per 100. On average the fatality rate was thus lower in the treatment group.
8. Using the Mantel-Haenszel procedure for stratified data (see Unit D13) the point estimate of the common rate ratio was 0.79, with a 99% confidence interval of 0.70 to 0.89. This procedure treats each trial as a separate stratum, as in the Mantel-Haenszel chi-square test, and the findings in the strata are combined, giving an appropriate weight to each stratum (greater weight is given to findings that have narrower confidence intervals). These results indicate that in general, allocation to a treatment group reduces the chance of dying by about 21%, and this reduction can be estimated with 99% confidence to lie between 11% and 30%. (Is 21% the prevented or preventable fraction? See Note F2–3.)
9. By the Mantel-Haenszel procedure, the point estimate of the common difference between the fatality rates in treatment and control groups was −2.1 per 100, with a 99% confidence interval of −1.1 to −3.1 per 100. When

Table F2. Results of Two Randomized Controlled Trials of the Effectiveness of Fresh Water in the Prevention of Deaths Among Shipwreck Victims: Fictional Data

	Treatment Group			Control Group			
Trial	No.	Deaths	Rate	No.	Deaths	Rate	Rate Ratio
A	50	10	20%	80	32	40%	0.5
B	450	45	10%	80	16	20%	0.5
Total	500	55	11%	160	48	30%	0.37

might the rate difference be preferable to the rate ratio as a measure of the effect of treatment?

10. The Mantel-Haenszel odds ratio expressing the difference in fatality between the treatment and control groups was 0.77, with a 99% confidence interval of 0.68 to 0.88. Which is preferable—the rate ratio or the odds ratio? In a similar meta-analysis based on studies of mixed types (trials, cohort studies, and case-control studies), which would be a preferable measure—the rate ratio or the odds ratio?

Question F2–2

What importance may effect modification have in this meta-analysis? May confounding have any relevance?

Question F2–3

Data on nonfatal recurrences of myocardial infarction were available for 19 of the 23 trials listed in Table F1. The Mantel-Haenszel rate ratio, based on these 19 trials, was 0.75 (99% confidence interval, 0.65 to 0.87). The corresponding rate ratio for fatality was 0.79 (99% confidence interval, 0.70 to 0.89). Can it be concluded that treatment prevented nonfatal recurrences about as well as it prevented deaths? Or do you want other information before deciding?

Notes

F2–1. Goodness-of-fit tests assess the agreement between an observed distribution and a specified expected distribution. A significant result means that the null hypothesis (of a good fit) can be rejected. The closer the agreement with the expected distribution (i.e., the better the fit), the higher the *P* value.

F2–2. Various methods of combining *P* values from independent tests of essentially the same hypothesis are described by DeMets (1987), Hedges and Olkin (1985, chap. 3), and Wolf (1986). Some use the *P* values, others use the corresponding normal deviates (*Z* values). In the present instance, the *Z* values were used (Stouffer et al., 1949), after weighting them by the square root of the sample size (the total number of subjects), a method that may give results close to those of the Mantel-Haenszel and similar tests (Canner, 1987); but there is no agreement on whether weighting should be used or what weights are best. These methods are not always valid; the main condition is that the *P* values must be one-tailed (Unit D4), and must test the same direction of effect (two-tailed *P* values should first be halved; if the observed effect in a specific test is opposite to that of the study hypothesis, the halved value should be subtracted from 1); in the present instance the combined one-tailed *P* was computed and then doubled to yield a two-tailed *P*.

F2–3. Both! On the assumption that the difference in fatality is attributable to treatment, 21% is the prevented fraction among those exposed to treatment, whose fatality rate is 79% of what it would have been had they not been exposed

to treatment. It is also the preventable fraction among those not exposed to treatment, whose deaths would be reduced by 21% if they were exposed to treatment. (See Unit E13.)

■■■■■■■■■■ Unit F3

Measures Used in Meta-Analysis

In *Question F2–1* the aim is to bring together the results of 23 trials, on the assumption that it is legitimate to do this. Statement (1) is based on simple pooling of the 23 sets of basic data; the numbers are lumped together, as if only one large study had been done, and differences in study design and execution are ignored. This is inadvisable. Not only may the results of large studies overwhelm the results of small ones, but the overall results may be distorted, as clearly shown in Table F2, where the two trials differ in the relative sizes of their treatment and control groups (the pooled results yield a lower rate ratio than was seen in either trial). It is preferable to use techniques that treat each trial as a separate stratum, by comparing each treatment group with its own control group, and then bringing together the stratified findings. This is essentially what is done in statements (2) to (10).

Statements (2) and (3) are based on what has been called "vote counting" (how many in favor? how many against?). Its main drawback is that equal weight is given to each study, however small, and however weak or strong the association, so that the conclusions may be misleading. The significance test used in statement (3) has an extremely low power; it is based on a sample of only 23. A significant result might be meaningful, but is hard to achieve (the test has a low power); in this instance it would be attained ($P < .05$) only if at least 17 of the 23 trials showed favorable effects. "Not significant" means only that chance processes might easily account for the observed results, and not that they do; the verdict is "not proven," not "disproved."

Combining the *P* values, as in statement (4), is an appropriate method, although not often used. Its advantages are that it can be applied to *P* values based on different kinds of significance tests, and that it is feasible even if only the *P* values are available, without the basic data on which they were based. It is used much less often than the Mantel-Haenszel test utilized in statement (5) or similar tests for stratified data. Both methods are appropriate, and the difference between the *P* values—.0000005 versus .0000002—is, of course, negligible. But significance tests alone are of limited value, for they tell us nothing about the strength of the association or its confidence interval.

Calculating an average rate ratio in the way described in statement (6) is not permissible. Imagine two trials: Trial *A* has fatality rates of 4% and 16% in its

treatment and control groups (ratio = 0.25), and trial *B* has the reverse findings—fatality rates of 16% and 4%, respectively (ratio = 4). The mean treated control group rate ratio is (0.25 + 4)/2, or 2.125; that is, on average the fatality rate is over twice as high in the treatment group. (Verdict: Shun the treatment like the plague.) Now leave the findings unchanged, but instead of the treated/control group rate ratio, use the controls/treated rate ratio; in trial *A* this is 4, and in trial *B* it is 0.25. The mean rate ratio is again 2.125, but this time the fatality rate is over twice as high in the control group. (Verdict: Welcome the treatment with open arms.) The method is obviously faulty; rate ratios (like percentages) cannot be averaged unless they are based on the same denominators.

On the other hand, it is permissible to take an average of rate differences, as in statement (7). But simple averaging gives every trial the same weight, so that small studies have an unduly large effect on the average.

The Mantel-Haenszel procedure, used in statements (8), (9), and (10), brings together the results of the various studies so as to estimate a common rate ratio, rate difference, or odds ratio. This is one of several techniques available for this purpose (see Note F3–1). Each study is treated as a separate stratum (which, in a meta-analysis of trials, means that the treatment group in each trial is compared only with the control group in the same trial), and the findings in the strata are combined, giving an appropriate weight to each stratum. The assumption is that there is in fact a uniform effect, each study providing a different estimate of this effect (this is called the *fixed-effect model*); the results are valid if this is a reasonable assumption. Alternative methods are available for use in meta-analyses where this assumption is questionable; they will be dealt with in Unit F8.

The rate difference (statement 9) might be used in preference to (or as well as) the rate ratio if we wanted to estimate the absolute number of deaths that the treatment might prevent (see Unit A3).

Both the rate ratio and the odds ratio (statement 10) are satisfactory measures (see Unit B11), but the rate ratio is easier to understand and explain. Case-control studies do not provide direct measures of relative risk, and a meta-analysis including case-control studies would necessarily use the odds ratio.

Effect modification may, of course, be important in any meta-analysis of studies of an association (*Question F2–2*). In a meta-analysis of clinical trials, the distinctive features of each trial may modify the association between treatment and outcome observed in the trial, resulting in differences between the results of the trials. The uniform measure estimated by the Mantel-Haenszel or similar procedures may not be very meaningful if these modifying effects are marked. In such instances, the factors that affect the association may be of more interest than estimation of an imaginary uniform measure.

Confounding may be of relevance in two contexts. First, the results of the individual studies may be distorted by confounding. This is relatively unlikely in a meta-analysis of randomized controlled trials; but even in such trials, differences between cases and controls (e.g., in the severity of the disease or in other prognostic factors), possibly caused by randomization errors, may distort the results.

Second, there may be distortion (as shown in Table F2) when the results are combined, as a consequence of imbalances between the sizes of the treatment and control groups. The Mantel-Haenszel and similar procedures guard against this latter kind of confounding.

Because of the differences between trials, the results of a meta-analysis obviously depend on which trials it covers. The exclusion of four trials from the meta-analysis of nonfatal recurrences (*Question F2–3*) may influence the findings, and this possibility should be explored. Do the missing trials differ in any obvious way from the others? If so, a comparison of the rate ratio for nonfatal recurrences, based on 19 trials, with the corresponding rate ratio for fatality, based on all 23 trials, might be misleading. The simplest approach is to perform a meta-analysis of fatality in the 19 trials covered by the meta-analysis of recurrences, and then compare the results. The rate ratio for fatality in these 19 trials is, in fact, 0.79 (95% confidence interval, 0.70 to 0.80), confirming that treatment prevented nonfatal recurrences about as well as it prevented deaths.

The measures of association used in Unit F2 are obviously not the only ones available, and they are not always appropriate. Use is often made of what is called the *effect size*. This is generally defined as the difference between the mean values in the two groups compared, divided by the standard deviation in the control group; a result of 2 means that the magnitude of the difference is 2 standard deviations. The effect sizes in the various trials may then be averaged, for use as an overall measure. The mean effect size can be made more meaningful by looking it up in a table of the normal distribution and translating it to a statement that the average member of one group has a higher value (or a lower value, depending on how the difference was calculated) than a specific proportion of the members of the other group (see Note F3–2). The mean effect size is used in the next three questions. Assume that the necessary precautions were taken before the results were combined.

Question F3–1

Eleven controlled trials of the psychological treatment of asthma all showed a favorable effect. The trials used various outcome measures; these included lung function (the peak expiratory flow), the number of asthma attacks, the amount of medication required, the number of emergency room visits, and so on. The average effect size was 0.86 (Glass et al., 1981); according to a table of the normal distribution, this result indicates that the average patient in a treatment group had a better outcome than did 81% of controls. The mean effect size was significantly greater than zero. What advantages of using the effect size does this example illustrate? Can you think of any disadvantages?

Question F3–2

The results of a meta-analysis of controlled trials of patient education for people with chronic medical problems are summarized in Table F3–1. The out-

Table F3–1. Meta-Analysis of 27 Controlled Trials of Patient Education

Outcome	No. of Studies	Mean Effect Size
Compliance	18	0.67**
Therapeutic progress	13	0.13**
Long-term outcome	5	0.06*

*Significantly greater than zero ($P < .05$).

**Significantly greater than zero ($P < .01$).

Source: Mazzucca (1983).

comes that were measured were compliance with medical advice, physiological progress toward therapeutic goals, and long-term health outcomes. The trials did not use the same outcome measures. Therapeutic progress, for example, was appraised in only 13 trials, and it was measured by changes in blood pressure, body weight, or other characteristics; the long-term outcome was measured in terms of return to work, hospitalization, and so on. The mean effect sizes pointed to a much greater effect on compliance than on therapeutic progress, and a relatively small long-term effect. What advantage of using mean effect sizes is illustrated by this example? What obvious possible source of bias do you see in this study, and how could it be explored?

Question F3–3

In both the above meta-analyses, there were often two or more outcome measures in the same trial, and these were included when the effect sizes were averaged. In the meta-analysis of patient education, for example, the result in the five trials with data on the long-term outcome is the average of 11 estimates of effect size (1–4 per trial). How may the inclusion of a variable number of outcome measures per trial affect the findings?

Table F3–2. Prevalence Rates (%) of Four Symptoms in Women in Two Surveys in California

Symptom	Survey 1 (n = 234)	Survey 2 (n =170)	Pooled Data (n =404)
Symptoms	15.0	11.2	13.4
Eye irritation	30.0	25.3	28.0
Sleep disturbance	15.8	17.2	16.5
Fatigue	15.9	18.9	17.4

Source: Lipscomb et al. (1992); figures modified slightly.

Question F3–4

This question deals with a meta-analysis of descriptive surveys. Information on various symptoms was collected in two surveys in California. The samples comprised all adults residing in two small neighborhoods, which had no nearby hazardous waste disposal sites; the populations had very different demographic characteristics. The questions were almost identical, and were administered by interviewers. Table F3–2 shows the prevalence of selected symptoms in women in each survey and in the pooled data. The authors of the meta-analysis suggest that the pooled results might be used as reference (control) rates in studies of communities exposed to suspected environmental hazards. What do you think of this? What precautions would you suggest?

Question F3–5

In the same meta-analysis, associations between the symptoms and other variables were examined by pooling the data of the two surveys and then performing multiple logistic regression analyses. The variables in the logistic regression model included age, sex, race, education, smoking status, and "study" (i.e., whether the person was included in survey 1 or survey 2). As an example of the findings, Caucasians reported more fatigue than did Asians (odds ratio, 2.7), Hispanics (odds ratio, 1.5), or other race groups (odds ratio, 2.3); the difference from Asians was statistically significant. Would it have been better to examine the associations between symptoms and other variables by using the Mantel-Haenszel procedure, treating each survey as a separate stratum? Or are there advantages to the use of multiple logistic regression?

Question F3–6

Enough of statistics for now; quantitative methods are a key feature of meta-analysis, but are not its main problem. This question serves as an introduction to basic principles and predicaments.

The following meta-analysis, dealing with the effect of treating hypertension on coronary mortality, was skimpily described in three sentences in the middle of a narrative review in 1976:

> Most trials have shown little or no effect on the incidence of coronary complications. The combined results of a number of studies [9 references] indicate that the risk of coronary mortality among treated hypertensives is about 0.7 times that among the untreated cases; this is a weighted average [using the Mantel-Haenszel procedure] of the relative risks in these trials. Since [this does] not differ significantly from unity ($P = .18$) an effect . . . must be regarded as "not proven," although it cannot be ruled out. (Abramson and Hopp, 1976)

What additional information would you like in order to be convinced that the result is not an artifact attributable to flawed methods? Do not go into detail, but try to list the most important questions.

Notes

F3–1. See Note D13–1 for references to the Mantel-Haenszel, precision-based, and maximum-likelihood methods. In these and other commonly used procedures for estimating a common measure of association, larger numbers increase the weight assigned to a stratum. Unless numbers are small, the procedures generally provide fairly similar results; for numerical examples, see Kahn and Sempos (1989, chap. 9). One technique frequently used in meta-analyses, the "O minus E" method (Peto, 1987b) is particularly simple; but it may give misleading results if the association is strong and the groups compared are very different in size (Greenland and Salvan, 1990).

F3–2. Since the effect size is expressed in terms of standard deviations (this is what is sometimes called a "Z score"), it can be made more meaningful by referring to a table showing the area in the tail of the normal distribution. Suppose the mean effect size in a meta-analysis of trials is 0.86. The value shown for $Z = 0.86$ in a table of the normal distribution (such as Table A1 of Armitage and Berry, 1994) is 0.1949, indicating that the average patient in a treatment group has a better result than 80.51% of controls. If no table is handy, this percentage can be fairly accurately calculated by the formula $49.32 + 45.23es - 10.56es^2$, where *es* is an effect size between 0.1 and 2; for an effect size of 0.86, the calculated result is 80.41. This way of expressing the results is, of course, valid only if the variable has a normal or near-normal distribution.

■■■■■■■■■■ Unit F4

Measures Used in Meta-Analysis (Continued)

Each of the effect sizes used in a meta-analysis is based on a comparison of groups (e.g., treatment and control groups) in the same study. The effect sizes can be weighted before averaging, giving more weight to larger studies (Hedges and Olkin, 1985), although in this instance (*Question F3–1*) the same weight was given to every study.

A special advantage illustrated in *Question F3–1* arises from the fact that the effect size is "unitless"—that is, it is expressed in terms of standard deviations (of whatever variable is measured) rather than in terms of number of attacks, number of visits, and so on. This permits the calculation of average effect sizes based on various dependent variables, in meta-analyses where the latter can be regarded as indicators of a general effect.

This may be misleading, however, if the effect sizes are different for different dependent variables. Moreover, a measure based on standard deviations has little meaning in terms of health implications, and a change of 0.86 standard deviations in one variable may not have the same importance as a change of 0.86

standard deviations in another. This problem remains if the effect measures are interpreted in terms of the normal distribution (as explained in Note F3–2): The statement that the average member of a treatment group has a better result than 84% of controls may have a different health relevance for one variable than for another. Unless logic or necessity dictates otherwise, it may be preferable to conduct a separate meta-analysis for each dependent variable.

Even if effect sizes based on a single outcome variable (such as peak expiratory flow) are used, their magnitude depends on the standard deviations in the studies, which may vary because of population differences or for other reasons. Effect sizes should therefore be used with circumspection; some experts deprecate their use (Greenland, 1998b).

As illustrated in *Question F3–2,* the unitlessness of effect sizes also permits comparisons of different kinds of outcome. Here too there may be a problem of interpretation, as there is no simple way of comparing the importance (in terms of health relevance) of one standard deviation of different outcome measures.

Also, a comparison of mean effect sizes based on different sets of trials may be misleading. The differences in mean effect sizes shown in Table F3–1 may be partly or wholly due to differences (e.g., in the mode of education or in the nature of the medical problem) between the trials included in the three sets. This problem may be approached by comparing the descriptions of the trials in the tree sets, and/or by comparing mean effect sizes (for each pair of outcomes) based on the same trials, if there is enough overlap to permit this.

If different trials contribute different numbers of outcome measures to the calculation of a mean effect size (*Question F3–3*), trials with more outcome measures may have an undue influence on the mean. If they differ from other trials, this may produce a bias. In the study of patient education, trials of behavioral (rather than didactic) educational methods tended to have more "compliance" than "therapeutic progress" or "long-term outcome" measures per study. As a result, over two-thirds of the effect sizes in the "compliance" set were based on behavioral methods, as compared with half of the effect sizes in the other two sets—a difference that may have contributed to the contrast seen in Table F3–1.

When prevalence findings are combined by simple pooling (*Question F3–4*), the weight given to each study is determined by the size of the sample studied; the "pooled" prevalence rate is therefore to some extent arbitrary, since it reflects the relative sizes of the study samples. A more important consideration—and this applies to any meta-analysis of descriptive studies of characteristics whose frequency varies—is that generalizations to other populations may be of uncertain validity, whatever technique of combination is used. Unless the studies were performed in samples that represent a total population, generalizations to a total population may be debatable. The authors of this meta-analysis advise caution in applying the pooled results to populations that differ demographically from the populations studied. They also warn that the results should be compared only if the questions are identical and are administered by interviewers; in a third survey, which used self-administered questionnaires asking the same questions, symptom rates were two- to fivefold higher.

In answer to *Question F3–5*, there is no compelling reason to prefer the Mantel-Haenszel procedure to multiple logistic regression in this analysis. Both methods can control for confounding, and the inclusion of "study" as an independent variable in the logistic regression model serves the same purpose as handling the studies as separate strata in the Mantel-Haenszel procedure. The two methods generally yield very similar odds ratios; if these differ, the Mantel-Haenszel value is probably preferable because it does not depend on the validity of the logistic model (Kahn and Sempos, 1989, p. 156). But the Mantel-Haenszel procedure may be awkward in studies where there are many uncontrolled potential confounders, so that elaborate substratification (e.g., by study, age, sex, race, and education, etc.) is required. It is also awkward if the independent variable has more than two categories, since a separate analysis is required for each comparison (Caucasians vs. Asians, Caucasians vs. Hispanics, etc.). Multiple logistic regression has the advantages that it permits the simultaneous study of several independent variables and the exploration of interactions (effect modification), and provides a risk-predicting equation (see Unit D13).

Linear regression methods may also be used in meta-analyses (Greenland, 1998b).

Basic Information

In reply to *Question F3–6*, every meta-analysis should include the answers to at least the following basic questions:

1. How were the studies found? There is a possibility of bias if the meta-analysis does not include all relevant studies.
2. How were studies selected for inclusion? (What were the inclusion and/or exclusion criteria?)
3. What are the distinctive features of the studies, with respect to their design, execution, study populations, and other characteristics, and are these features sufficiently similar to justify combining the studies' results?
4. How well were the studies designed and executed?
5. What are the results of the studies, and are the results consistent enough to justify combining them?

Exercise F4

How to find studies, and how to select studies for the meta-analysis—these are the topics of this exercise and the next.

Question F4–1

A meta-analysis obviously requires a systematic search of the literature, using (for example) *Index Medicus*, *Current Contents*, or a computerized database (MEDLINE, MEDLARS, etc.). Can you guess what proportion of published randomized controlled trials related to vision research were detected by a MED-

LINE search? The "gold standard" included studies detected by hand searches of journals or reported at a meeting of investigators, as well as those located by MEDLINE (Dickersin et al., 1995).

Can you suggest any other ways of finding published studies? What method would you recommend?

Question F4–2

Do you think that the omission of unpublished studies might bias the results of a meta-analysis? If so, what would you expect the direction of the bias to be?

Question F4–3

Should unpublished studies be sought and included? If not, why not?

Question F4–4

How can the results of unpublished studies be sought?

Question F4–5

Can you suggest a way of assessing how important the omission of unidentified unpublished studies may be, with respect to a specific meta-analysis? This is a difficult question. Clue: See statement (4) in *Question F2–1.*

■■■■■■■■■■ Unit F5

Finding the Studies

In the test described in *Question F4–1,* a MEDLINE search found 48% of the published trials. A second much more elaborate MEDLINE search (using 34 search terms) revealed 82% of the studies; the price for this high sensitivity was a "false positive" rate of 87%. According to a meta-analysis of 15 studies in various fields of health and health care, MEDLINE's sensitivity in detecting randomized clinical trials was 51% on average, with a range of 17–82% (Dickersin et al., 1995). Clearly, reliance should not be placed on any single method of searching the literature, and use of a combination of methods is recommended.

In the meta-analysis of trials of beta-blockers (Table F1), a systematic literature search for published studies (including those listed in conference abstracts) was supplemented by an informal search for studies known to the investigators and their colleagues, and perusal of reference lists in the reports found. The meta-analysis of trials of patient education (Table F3–1) used a MEDLARS search and two annotated bibliographies on the subject.

There is much evidence for publication bias in the health field. In general

(*Question F4–2*) it is the negative or inconclusive studies that are rejected or remain "tucked away in file drawers." In Oxford, for example, a survey of 487 clinical research projects approved in 1984–87 showed that the odds in favor of publication by 1990 were over twice as high if the results were statistically significant (Easterbrook et al., 1991). Omission of unpublished studies from a meta-analysis may thus be expected to bias the results by making overall effects stronger and exaggerating their statistical significance. Bias in the opposite direction has also been occasionally reported, with smaller effects found in published studies; studies that contradict conventional wisdom may be less likely to appear in print, even if they show strong effects, unless they are especially newsworthy. In principle, unpublished studies should be included if possible, to avoid bias in any direction (*Question F4–3*), although some investigators oppose this on the grounds that unreported studies are likely to be of poor quality; surprisingly, however, follow-up studies of medical researches have shown no independent relationship between the quality of research design and the likelihood of publication (Chalmers et al., 1990, Easterbrook et al., 1991).

Unpublished studies may be sought in several ways (*Question F4–4*). In the beta-blocker meta-analysis, the investigators interrogated colleagues. Other methods include the scanning of conference proceedings and lists of dissertations, and contacts with funding organizations. Emphasis has recently been placed on registers of clinical trials; in one instance, where there was a relatively complete register for comparison, a MEDLINE search revealed only 28 of 96 known trials (Dickersin et al., 1985). If unpublished studies are identified, it becomes necessary to ask the investigators for information on their methods and results.

An easy way of handling possible bias due to the omission of unidentified studies (*Question F4–5*) is to calculate the number of studies showing no effect (the "fail-safe *N*") that would be needed to change the observed overall *P* value to a nonsignificant level or reduce the observed overall effect to a trivial value (Note F5). In the beta-blocker meta-analysis (Table F1), the number of null studies required to push the observed overall *P* value (.0000005) up to .05 turns out to be 108. Because is it is very unlikely that there are 108 unreported null randomized controlled trials of beta-blockers, the possibility that the finding is attributable to this source of bias can be disregarded. By contrast, the fail-safe *N* was only 2 in a meta-analysis of trials of total parenteral nutrition in cancer patients undergoing surgery, which showed a significant reduction in operative mortality (Klein et al., 1986).

Exercise F5

Question F5–1

A MEDLARS search, together with screening of *Current Contents*, review articles and reference lists, revealed 12 controlled trials showing the effect of vitamin A supplements on child mortality. Four of the trials were conducted in hospitals and dealt with children with measles. Eight were community-based trials in which children living in different villages, districts, or households were as-

signed to treatment or control groups (Fawzi et al., 1993). Can all 12 trials be included in a meta-analysis?

Question F5–2

A meta-analysis of studies of age trends in the prevalence of senile dementia was limited to studies conducted since 1980 (Ritchie et al., 1992). It was also restricted to studies of moderate and severe (not mild) dementia. Can you suggest reasons for these limitations?

Question F5–3

Can you suggest why old studies might be excluded from a meta-analysis of clinical trials?

Question F5–4

The selection of studies for a meta-analysis obviously influences the results; a biased selection may lead to biased results. Assuming that an appropriate search has been conducted, can you suggest what precautions should then be taken to make the selection of studies as objective as possible?

Question F5–5

It is plainly advisable that findings that are to be combined in a meta-analysis should be independent of each other. If two papers report the same study (one of them perhaps including additional cases and controls), it is obviously wrong to include both in the meta-analysis. But what should be done if one paper reports the results of a short-term follow-up and a later paper describes a long-term follow-up of the same subjects; should both papers be included? What should be done if the short-term and long-term results are reported in the same publication; should both sets of findings be included?

Question F5–6

Controlled trials of work-site smoking cessation programs were identified by searches of MEDLINE and 11 other literature databases, an index of theses and dissertations, and reports of meetings of two associations, and by contacts with other investigators (Fisher et al., 1990). Twenty trials were found; because some programs were conducted in 2–4 different treatment groups (e.g., in different companies), 34 experimental–control comparisons were available. The outcome variable was the long-term quit rate—that is, the proportion (of smokers who were exposed to the program) who quit smoking, as measured 12 months later. The 34 effect sizes were calculated and averaged, after weighting them by a method that gives more weight to larger samples. A weak mean effect size of 0.21 (95% confidence interval, 0.16 to 0.26) was found, indicating that (by the method described in Note F3–2) the average smoker who was exposed to a program had a better result (i.e., was more likely to quit) than about 56–60% of

smokers who were not ($P < .01$). How might the inclusion of all 34 comparisons affect the mean effect size? What solutions can you suggest?

Question F5–7

A MEDLINE search was performed for controlled clinical studies of acupuncture for chronic pain, supplemented by screening of *Excerpta Medica,* the *Journal of Traditional Chinese Medicine,* and bulletins from a documentation service for alternative medicine, as well as correspondence with and visits to colleagues. The search revealed 71 reports that met the following criteria: (1) Needles were used; studies in which only surface electrodes or laser acupuncture were used were excluded; (2) the word "chronic" was mentioned in the title or abstract, or the duration of pain was stated to be at least 6 months; (3) a reference (control) group was used, which was exposed to another treatment or sham treatment (placebo). Some reports were excluded because they turned out not to deal with chronic pain or because they replicated descriptions of the same studies or patients, and one because it was totally uninterpretable. This left 51 studies for analysis. Can you suggest why the above three criteria were used?

Question F5–8

The 51 trials of acupuncture were of uneven quality, as shown in Table F5; only six were randomized and double-blind. Should some of the studies be excluded

Table F5. Methods Used in 51 Controlled Trials of Acupuncture

	Blind?		
Randomized?	Patients	Evaluator	Number of Trials
Yes	Yes	Yes	6
Yes	?*	Yes	3
Yes	No	Yes	7
Yes	Yes	No	1
Yes	?	?	3
Yes	?	No	1
Yes	No	?	2
Yes	No	No	11
?	?	?	1
?	No	Yes	1
?	No	No	4
No	Yes	Yes	1
No	?	?	1
No	No	Yes	2
No	No	No	7

*?=maybe; the study report is unclear.

Source: Ter Riet et al. (1990).

from the meta-analysis? What arguments might be offered in favor of including studies that are not of the best quality in a meta-analysis?

Note

F5. Formulae for the fail-safe *N* (if the meta-analysis yields a significant effect) are provided by Rosenthal (1979), Orwin (1983), Klein et al. (1986), and Wolf (1986). More elaborate statistical approaches are considered by Iyengar and Greenhouse (1988) and discussants of their paper.

■■■■■■■■■■ Unit F6

Selecting Studies

The inclusion or exclusion of a study should be determined, in the first instance, by the objective of the meta-analysis. If the question asked is a general one, broad selection criteria may be used; if it is a more specific one—for example, what is the effect of a particular drug on a particular outcome with respect to a particular disease in a particular kind of patient?—stricter criteria must be used.

The answer to *Question F5–1* thus depends on what we want to learn from the meta-analysis. If the question of interest is the value of vitamin A in the treatment of children with measles, the first four trials should be selected. If interest lies in the prophylactic administration of vitamin A supplements to children living in the community, the eight community-based trials should be selected. If both these questions are of interest, both sets of trials can be included in the meta-analysis, but they should be analyzed separately; this might also permit comparison of its value in the two situations. And if the question asked is a general one—can vitamin A supplements reduce child mortality? (without reference to any specific situation)—all 12 trials can legitimately be included in a single analysis. But if the effects of vitamin A in the two situations are very different, the overall effect in the 12 trials will of course depend on the relative numbers of trials in the two sets (4 and 8), and may not be a very meaningful quantitative measure.

Appropriate rules may be applied for the inclusion or exclusion of studies so as to reduce excessive differences that may make it difficult to integrate findings in a meaningful way. If it is known or suspected that there have been changes over time, for example, a time limitation may be built in. Plausible reasons for the exclusion of pre-1980 studies from a meta-analysis of senile dementia (*Question F5–2*) might include known changes in the epidemiology of this condition or in diagnostic methods. The actual reason was the development of the *DSM-III* and other standardized sets of diagnostic criteria in and after 1980. Widely varying definitions were used in earlier years.

Studies of mild senile dementia were excluded from the meta-analysis because of the questionable reliability of diagnoses; reported rates range from 2.6% to 52.7%.

The exclusion of old studies might be advisable in a meta-analysis of clinical trials (*Question F5–3*) if there have been changes in medical or nursing care that may influence the outcome, changes in diagnostic methods, or changes in the natural history of the condition under study.

In answer to *Question F5–4,* precautions for ensuring an objective selection of studies center around the formulation and application of inclusion and/or exclusion criteria. These should be explicitly stated, should be as clear and specific as possible, and (if feasible) should be applied "blind": Decisions concerning the inclusion of specific studies should not be influenced by knowledge of their results.

Short-term and long-term outcomes (or any different outcomes) in the same subjects (*Question F5–5*) can be included in a systematic review—whether they appear in the same or different reports—provided that each outcome is analyzed separately. If some studies or individuals contribute more outcome measures than others to the estimation of an overall effect, they may have an undue influence on the overall effect (as illustrated in *Question F3–3*).

Question F5–6 presents another example of overrepresentation of some trials in a meta-analysis. Because the trials differed in educational methods and other respects, the mean effect size may well be biased. Also, it may have a spurious degree of precision (i.e., an unduly narrow confidence interval) because it includes clusters of results that are similar because they come from the same trial. These problems could be avoided by using only one effect size for each trial. This might be the average effect size for the experimental–control comparisons in the trial, or a single one of the effect sizes in the trial, randomly or systematically selected. These methods obviously entail the loss of specific information. The investigators found that the mean effect size was 0.27 (95% confidence interval, 0.22 to 0.33) if it was based on the average results in the 20 trials, and it was 0.26 (95% confidence interval, 0.20 to 0.32) if it was based on a single result per trial (the result that the author of the study regarded as the strongest). They decided, as a calculated risk, to use all 34 measures in subsequent analyses aimed at investigating modifying factors affecting the success of the programs, so as not to waste data.

Criteria for the inclusion of studies should reflect the objectives of the meta-analysis. The reason for the first two criteria used in the overview of acupuncture studies (*Question F5–7*) is obviously the investigators' interest in the effectiveness of needle acupuncture (and not laser acupuncture) in chronic pain (and not in other conditions). The third criterion (the use of a control group) relates to study quality and represents an effort to restrict the meta-analysis to studies with a potential for giving an adequate answer to the research question.

Opinions differ as to whether studies of a poor quality should be excluded from a meta-analysis (*Question F5–8*). Some experts suggest that acceptable standards should always be set in advance, in the form of criteria for inclusion,

and studies that do not meet them should not be accepted. The extreme view is that in a meta-analysis of clinical trials "it is important to restrict inclusion to randomized trials, ideally with intention-to-treat analysis, complete follow-up information, and objective or blinded outcome assessment" (Thompson and Pocock, 1991). Some suggest that the best of the available studies should be used (Slavin, 1986, 1987). Others advise the inclusion of all studies except those that are really bad: "[I]f it is clear that a certain study is fundamentally flawed, say with obvious numerical errors, I find it hard to argue for its inclusion. I do not believe that wrong information is better than no information" (Light, 1987).

The main arguments for including studies that are not of the best quality are that increasing the number of studies permits the examination of the topic in more circumstances, and that it boosts numbers. If an effect truly exists in all circumstances, this consistency may be demonstrated more convincingly if more studies are included. On the other hand, if the effect is not consistent, the inclusion of more studies may make it easier to detect this inconsistency and explore its sources. Other things being equal, larger numbers will increase the power of statistical tests and make confidence intervals narrower.

In some instances the appraisal of study quality is the main purpose or main contribution of the meta-analysis, and all studies must, of course, then be included. An overview of papers on the effectiveness of health education programs in developing countries, for example, revealed that only 3 of 67 studies met four simple criteria, and it led to specific recommendations for improvements in health education research (Loevinsohn, 1990). A meta-analysis of studies of lumbar spine fusion revealed widely variable results and numerous flaws in study design, leading to the conclusion that the indications for this surgical procedure were not scientifically established, and that randomized controlled trials were required (Turner et al., 1993).

The advantages of including more studies in an analysis must, of course, be balanced against the obvious disadvantages of including questionable results. If poor studies are included, the differences in quality should be taken into account; it may be possible to control bias in the analysis, or make allowances for it.

There is thus no "correct" answer to the question: Should studies that are not of the best quality be included in a meta-analysis? The best answer is probably a qualified yes: They should be included, but only on condition that due regard is paid to the possible problems.

One way of taking account of differences in study quality is to pay separate attention to studies that are of higher and lower quality. This is what was done by the authors of the acupuncture meta-analysis, who scored the trials by giving points for randomization, blinding, and other features, and found that even the better studies (which were mediocre) gave contradictory results. They concluded that "the efficacy of acupuncture in the treatment of chronic pain remains doubtful," and called for research of a higher quality. Their summary table shows that only 2 of the 17 trials that used randomization and blind methods gave "positive" results (i.e., better results for acupuncture, according to the

investigators' own statements), as compared with 22 of the 34 other trials ($P < .0001$).

Exercise F6

Question F6–1

The importance of appraising the scientific quality of the individual studies is clear. This may be done as part of the selection procedure or after the studies have been selected for inclusion, or even after the analysis. Can you suggest what precautions might be taken to make the appraisal of quality as objective as possible?

Question F6–2

A meta-analysis of 375 controlled evaluations of psychotherapy, using various outcome measures, revealed a mean effect size of 0.68, indicating that the average patient receiving treatment had a better outcome than 75% of controls (Smith and Glass, 1977). A critic called this study an "exercise in mega-silliness," and fulminated against "the abandonment of critical judgments of any kind. A mass of reports—good, bad, and indifferent—are fed into the computer in the hope that people will cease caring about the quality of the material on which the conclusions are based. . . . The notion that one can distill scientific knowledge from a compilation of studies mostly of poor design, relying on subjective, unvalidated, and certainly unreliable clinical judgments, and dissimilar with respect to nearly all the vital parameters, dies hard. . . . 'Garbage in—garbage out' is a well-known axiom of computer specialists; it applies here with equal force" (Eysenck, 1978). Assuming that the studies were uneven in their quality, can you suggest any additional analyses that might counter this criticism?

Question F6–3

Guess whether the following statements are true or false:

1. In the meta-analysis of trials of work-site smoking cessation programs (*Question F5–6*) the effect was largest in trials in which reported smoking habits were not verified by biochemical tests.
2. In an overview of studies of anticoagulants for acute myocardial infarction (Gifford and Feinstein, 1969) the benefit of anticoagulant therapy (in comparison with no therapy) was more often observed in studies that failed to meet defined quality standards.
3. In a meta-analysis comparing coronary artery bypass surgery with nonsurgical intervention (Wortman and Yeaton, 1983), the results were better in nonrandomized than in randomized trials.
4. A meta-analysis of the effect of physical activity in preventing coronary heart disease (Berlin and Colditz, 1990) showed a larger preventive effect in methodologically stronger studies than in less well designed ones.

5. In a meta-analysis of trials of the treatment of mild hypertension, in which trials using different treatments were compared (Andrews et al., 1982), the effect was largest in the trials of higher quality.
6. In a meta-analysis of cohort studies of the relationship between mammary dysplasia (dense areas seen in a mammogram) and subsequent breast cancer (Goodwin and Boyd, 1988), the association was strongest in the studies of higher quality.
7. In a meta-analysis of randomized controlled trials of antibiotic prophylaxis in biliary tract surgery (Meijer et al., 1990), the effect was unrelated to study quality.

Question F6–4

In which (if any) of the following situations might it be worthwhile to contact the investigators and ask for further information:

1. The study report does not describe the methods clearly (as in the acupuncture meta-analysis, Table F5).
2. The study is reported in an abstract, with incomplete information on methods and results.
3. In a meta-analysis of trials, some of the reports use "while-on-randomized-treatment" analysis only, and provide no information on what happened to subjects after they abandoned the allocated regimen.
4. In a meta-analysis of trials of intravenous beta-blockers in acute myocardial infarction, it is found that three studies report significant reductions in cardiac arrest, with prevented fractions of 61–79%, but many other trials provide no information on this outcome.

Question F6–5

Suppose that a meta-analysis of epidemiological studies of the association between calcium intake and fractures includes a case-control study (Kreiger et al., 1992) in which postmenopausal women with hip and wrist fractures were compared with a group of controls. The report of this study provides separate findings for hip and wrist fractures, and for each kind of fracture there are five measures of the association with calcium intake: (1) the crude difference in mean daily calcium intake; (2) an odds ratio (adjusted by multiple logistic regression for age, obesity, and other variables) comparing women with a low and moderate dietary calcium intake; (3) a similar odds ratio comparing women with a low and high dietary calcium intake; (4) an odds ratio (adjusted for the above factors and dietary calcium intake) expressing the association with long-term calcium supplements; and (5) a similar odds ratio expressing the association with recent calcium supplements. When an overall measure is calculated in the meta-analysis, would the inclusion of more than one of these ten measures carry any disadvantages? Would restriction of the meta-analysis to only one of the measures carry any disadvantages?

Question F6–6

What solutions can you suggest to the problem of multiple measures, illustrated in the previous question?

■■■■■■■■■■ Unit F7

The Quality of the Studies

Appraising the scientific quality of a study is not easy. Judgments of the same study may differ, and there is no "gold standard" for comparison. The best approach is to ask questions about the presence of a number of features generally regarded as important determinants of the internal or external validity of studies. The studies can then be classified and ranked according to their quality (e.g., randomized controlled trials with blinding, randomized controlled trials without blinding; nonrandomized controlled trials; uncontrolled trials). If points are allotted to the various features, a quality score can be assigned to each study (see Note F7–1).

To make the appraisal as objective as possible (*Question F6–1*), the questions should be explicitly stated, and they should be phrased as clearly and specifically as possible. If a quality score is used, the method of scaling should be explicitly specified. Since expertise both in research methods and in the study topic may be needed to answer some of the questions (e.g., is the statistical analysis appropriate?), it may be advisable to have each study appraised by two reviewers, who can then compare their verdicts and reach a consensus. One recommendation is that material that might bias the reviewers (e.g., the names and affiliations of the researchers) should first be blotted out, and the methods appraised without knowledge of he study results; the appraisal can then be completed by looking at the results section (Chalmers, 1991); these precautions are often not feasible.

When there are differences in the quality of the studies, as in *Question F6–2*, the possible approaches are:

- To discard studies of poor quality before combining the results.
- To compare the combined results in all studies with the results when poor-quality studies are excluded; this is a form of "sensitivity analysis" (a term used for examinations of the extent to which the results of an analysis are affected by changes in methods or assumptions).
- To compare the results of studies that differ in quality; the relationship between quality and the result can be shown graphically, especially if quality scores are used.
- To give each study a weight determined by its quality, before combining results, so that better studies have a larger impact on the overall result.

- If regression analysis is employed, to use a measure of study quality as an independent variable in the model, and to control statistically for the effect of differences in quality.
- If few or no studies reach an acceptable standard, to abandon the meta-analysis and issue a call for better research.

In their reply to the criticism of their meta-analysis of controlled trials of psychotherapy, the authors said that in their analysis, "good, bad, and indifferent" studies showed almost the same results; "such features as use of randomization versus matching and double versus single versus no-blinding had virtually no correlation with study findings." Also, their findings were confirmed by multiple regression analyses using a model that included a score measuring the subjectivity of each outcome measure (Glass and Smith, 1978). A careful reanalysis, restricted to trials that used randomization and whose control groups received placebo or no treatment, yielded almost the same findings as the overall analysis: The mean effect size was 0.78 which, using a table of the normal distribution, suggests that the average patient receiving treatment had a better outcome than 78.2% of untreated controls (Landman and Dawes, 1982).

All the statements in *Question F6–3* are true. The influence of methodological shortcomings may not be easy to guess. In comparison with better studies, studies of a poorer quality can show an enhanced, reduced, or similar effect.

Often, of course, studies with methodological weaknesses (and there are those who put all nonexperimental studies in this category) are the only ones available. For example, a meta-analysis of studies of bone marrow transplantation in acute nonlymphocytic leukemia was based only on nonrandomized controlled trials and uncontrolled follow-up studies, because no randomized controlled trials had yet been done; after adjustment for a number of biases, the analysis showed a consistent advantage over chemotherapy in long-term disease-free survival (Begg et al., 1989).

Requests to investigators for further information about their methods or results may often add to the value of a meta-analysis, and might be worthwhile in all the situations listed in *Question F6–4*. They are least likely to be successful if they demand further action by the investigators, as they probably do in situation 3. In a meta-analysis in which unreported means and correlation coefficients were required, "letters were sent to 10 authors, but the necessary information was provided by 1 of those 10" (Gray et al., 1991). Collecting complete data on all subjects is "often . . . the most difficult and time-consuming aspect of doing overviews and can take 3 to 4 years!" (Yusuf, 1987a). Additional data on unreported outcomes, obtained by correspondence, reduced the summary prevented fraction for cardiac arrest in situation 4 from over 60% to 15% (Yusuf, 1987a). (Can you suggest why the unpublished data changed the overall result? See Note F7–2.) In a meta-analysis of randomized controlled trials comparing two forms of chemotherapy for advanced cancer of the ovary, it was found that the difference was larger (and significant) when the analysis was based on published reports than when it used full data on all randomized patients, including those ex-

cluded from published analyses and those studied in unpublished trials; the full data showed a nonsignificant difference (Stewart and Parmar, 1993).

Extracting the Findings

Question F6–5 exemplifies a common situation: A single study offers more than one measure of effect, and a decision must be made as to what should be used when an overall measure is calculated in the meta-analysis. This is a difficult problem. If more than one measure is included, extra weight is given to that study. Moreover, the different measures in a single study cannot be completely independent of each other; this dependence is particularly obvious in the study under consideration, where both fracture groups are compared with the same set of controls, whose calcium intake will influence each one of the measures. If only one measure is included, the specific choice may affect the overall result; in this study, fractures of the wrist were associated with a low intake of calcium, but fractures of the femur were not.

This dilemma occurs most often in observational studies. Not only may there be alternative dependent and independent variables and a choice between crude and adjusted measures, as in the present instance, but there may also be a choice between different adjusted or specific measures (controlling for different sets of potential confounders, or using different methods of adjustment), and there may be comparisons with different control groups. A choice may also be required in trials, as the same trial may provide two or more outcome measures; we saw this in the meta-analyses of asthma treatment and patient education (*Question F3–3*), smoking cessation programs (*Question F5–6*), and beta-blockers (*Question F2–3*).

There is no simple universal way of handling the problem of multiple measures (*Question F6–6*), and it must be considered anew in each meta-analysis. To reduce possible bias, explicit rules or guidelines should be formulated (preferably in advance), and (unless these are very simple) two or more reviewers should independently extract the findings and later resolve their disagreements by discussion. Preference should, of course, be given to measures that control for confounding. Sometimes an average of the measures is used, or a randomly or systematically selected one—for example, "the largest estimate, the smallest estimate, or, in order to be fair, the estimate closest to the average of the individual ones" (Fleiss and Gross, 1991). Analyses may be conducted using different choices, and the results compared. It is sometimes decided to take a calculated risk and use more than one measure from the same study; this was done in the study of work-site smoking cessation programs (*Question F5–6*) in order to have more data for the examination of modifying factors.

Apples and Oranges

Most writings on meta-analysis are packed with fruity metaphors (Note F7–3). A recurrent theme is the disadvantage of "adding apples and oranges"—that is, combining the results of studies that are so different that "it may be uncertain

to which fruit or to what specific combination of fruits, the results apply." "By mixing apples and oranges and an occasional lemon, one may end up with an artificial product." "Interpreting a weighted average of different odds ratios can be like describing an 'average fruit'." If different dependent variables are used, "a good meta-analysis capitalizes on this by coding apples as apples and oranges as oranges." If important differences between studies are overlooked, "the problem of 'apples' and 'oranges' . . . may render the entire exercise—dare one say it?—fruitless."

A meta-analysis can, of course, include both apples and oranges if what is wanted is general information about fruit. But even then, "the trials used . . . are not likely to be a random sample of all the trials that might conceivably have been done to provide information about 'fruit'; that is, the proportion of 'apples' and 'oranges' may be wrong." There will accordingly remain some uncertainty about the accuracy of the picture provided by "the 'fruit salad' created in our meta-analysis."

Also, "a miscellany of fruit is not necessarily a disadvantage. . . . Comparisons of apples and oranges may . . . provide useful additional information."

With respect to study quality: "Rotten apples in the basket may invalidate the results. . . . Some apples are healthy, some are slightly spoiled, some have a moderate degree of decay, and some are really rotten. Due account should be taken of the quality of whatever studies are included in the analysis."

And finally, publication bias: "Only the big apples may get to the market."

Studies are usually selected in the expectation that their results can be combined, and narrowly defined inclusion criteria may be used for this purpose. But the results may still turn out to be so dissimilar that it does not make sense to combine them as if they were all expressions of a single overall result. It is therefore essential to appraise the consistency of the results. A systematic review of 125 meta-analyses of clinical trials revealed heterogeneous odds ratios in 33% of the meta-analyses, and heterogeneous rate differences in 45% (using a statistical criterion, $P < .01$); in only 50% of the meta-analyses were both the odds ratios and the rate differences "homogeneous" (Engels et al., 2000). Heterogeneous findings should be combined only with reservations, or by using statistical methods that are specifically designed for this situation.

The next exercise deals with the appraisal of combinability. Both a statistical test for heterogeneity (Note F7–4) and visual inspection of the results (preferably after plotting them graphically) should be performed before combined results are used.

Exercise F7

Question F7–1

A meta-analysis was performed of epidemiological studies of the association between passive smoking in the home and lung cancer (in nonsmokers) in the United States (Fleiss and Gross, 1991). Different methods were employed in the nine studies that were identified. For example, one study was a cohort study and the

others were case-control studies. Some of the case-control studies used hospital patients as controls; others used healthy people living at home. In only two studies were the interviewers explicitly "blinded" as to whether the subject was a case or control. The proportion of proxy informants about smoking habits ranged from zero (in a study confined to living cases) to nearly 70%. Definitions of "nonsmoker" and exposure to smoking varied: In one study the subjects who reported occasional smoking were classified as nonsmokers, and in another the comparison was not between subjects exposed and not exposed to smoking, but between those exposed for more or less than 4 hours a day. Can the odds ratios provided by these nine studies be combined?

Question F7–2

A recent meta-analysis of epidemiological studies showed a weak but statistically significant association between cigarette smoking and leukemia (Brownson et al., 1993). In cohort studies the summary risk ratio was 1.3 (95% confidence interval, 1.3 to 1.4). In case-control studies the summary odds ratio was 1.1 (95% confidence interval, 1.0 to 1.2). Can you suggest why cohort studies provided a higher estimate of the increased risk associated with smoking? Can the cohort and case-control studies be combined in a single analysis?

Question F7–3

Statistical tests for the heterogeneity of findings have a low power; that is, they may fail to show heterogeneity when actually it exists, unless the number of studies is very large. How, therefore, would you interpret a *P* value of .001? .04? .09? .15? .8?

Question F7–4

In the meta-analysis of 23 trials of beta-blockers (Table F1), the Mantel-Haenszel summary rate ratio was 0.79, indicating that "on average" (controlling for the differences between the trials) the beta-blockers prevented 21% of deaths. A heterogeneity test yielded a *P* value of .38, indicating that the differences between the trial results could easily be accounted for by chance variation. Now suppose that the heterogeneity test had yielded a *P* value of .001 instead of .38, so that the differences could not be regarded as fortuitous. How would this modify the interpretation of the summary rate ratio?

Question F7–5

In the same meta-analysis the overall *P* value, based on the results of one-tailed significance tests in each study (see Note F2–2), was .0000005, indicating that the effect on fatality could safely be regarded as nonfortuitous. The Mantel-Haenszel test gave a similar *P* value, .0000002. Would the interpretation of these results be different if the heterogeneity test had yielded a *P* value of .001 instead of .38?

Question F7–6

The results of 14 randomized controlled trials of hypertension treatment, showing the effect on the occurrence of strokes, are summarized in Table F7. The reduction in strokes was highly significant; by the Mantel-Haenszel test, $P = 2.82E-13$ (what does this number mean? See Note F7–5). The rate ratios are shown graphically in Figure F7(A), together with their 95% confidence intervals. They are plotted on a logarithmic scale (Why? And what scale would you use for plotting rate differences? odds ratios? effect sizes? See Note F7–6). For convenience, the results are arranged in decreasing sequence. Can you tell which of the 14 values were statistically significant (i.e., significantly different from 1)? Do you think the results are acceptably consistent? Would you expect a test to show significant heterogeneity?

Question F7–7

Figure F7(B) shows the same results, but here the sequence is determined by the size of the study sample (treatment and control groups combined). The smallest trial (on the left) had 87 subjects, and the largest (on the right) had 17,354 subjects. What do you observe, with respect to (1) the point estimates of

Table F7. Results of 14 Randomized Controlled Trials of Antihypertensive Drugs; Comparison of Stroke Rates in Subjects Allocated to Treatment and Control Groups

	Strokes Per 100			
Trial	Treatment Group (a)	Control Group (b)	Ratio (a/b)	Difference (b − a)
1	1.51	2.24	0.67	0.73
2	0	1.32	0	1.32
3	0.76	1.29	0.59	0.53
4	0.69	1.26	0.55	0.57
5	2.69	10.31	0.26	7.62
6	0.52	3.06	0.17	2.54
7	2.39	3.59	0.67	1.20
8	18.46	23.74	0.78	5.29
9	1.47	4.76	0.31	3.29
10	4.44	2.38	1.87	−2.06
11	20.41	43.75	0.47	23.34
12	3.37	6.43	0.52	3.06
13	7.69	11.32	0.68	3.63
14	4.77	8.39	0.57	3.61
Pooled*			0.61	0.72

*Using precision weighting (i.e., each value was weighted by the reciprocal of its variance). The smaller the variance, the greater the weight.

Source: Based on Collins et al. (1990).

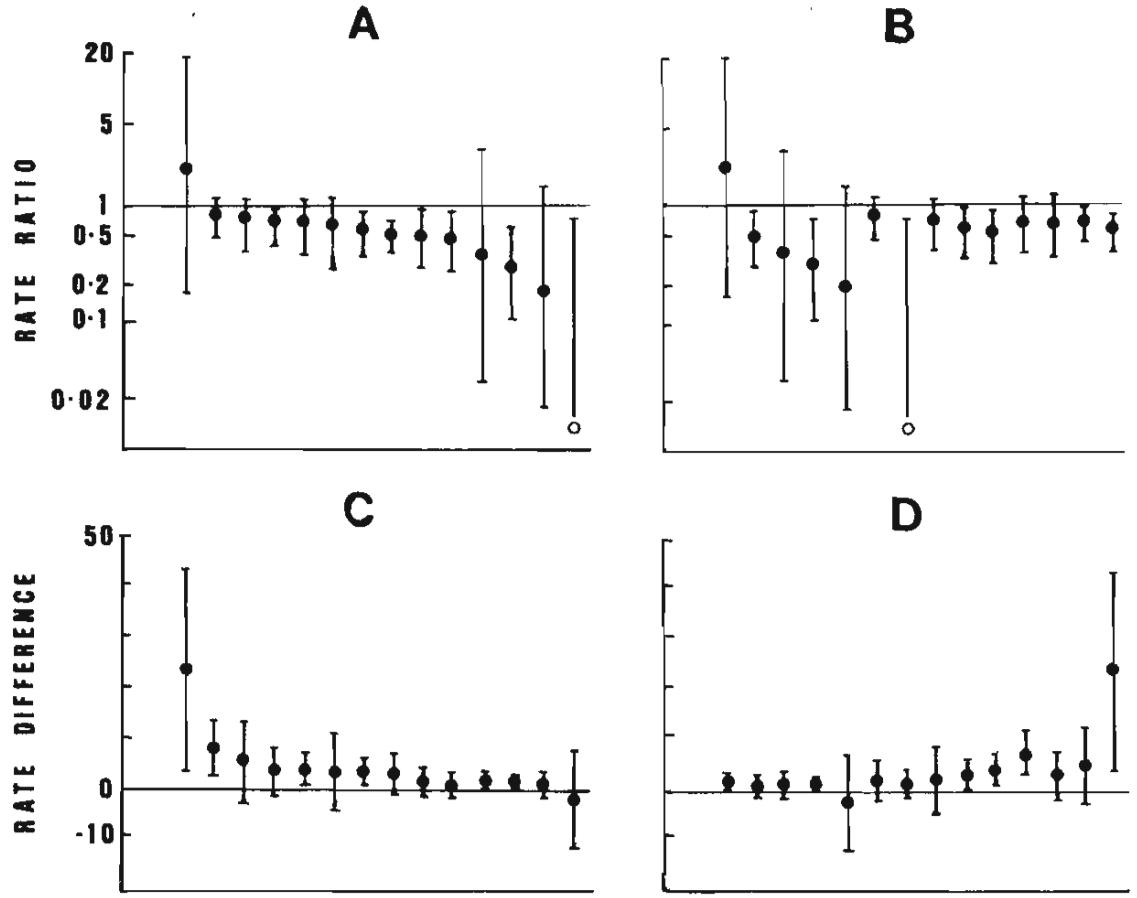

Figure F7. Comparisons of stroke rates in 14 trials of hypertension treatment: rate ratios (ratio of rate in treatment group to rate in control group) and rate differences (rate in control group minus rate in treatment group), with 95% confidence intervals. The rate ratio marked "0" is zero; only its upper confidence limit is shown. (A) Rate ratios. (B) Rate ratios, arranged in sequence of sample sizes (smallest on left). (C) Rate differences. (D) Rate differences, arranged in sequence of stroke rates in control group (lowest on left).

the rate ratios, and (2) the confidence intervals? Can you suggest an explanation?

Question F7–8

The rate differences listed in Table F7 are charted in Figure F7(C) [not in the same sequence as in Figure F7(A)]. What is your impression of the consistency of the values? Would you expect a test to show significant heterogeneity? When two different measures of an effect are used, is it possible for one to exhibit heterogeneity and the other not?

Question F7–9

In the meta-analysis of controlled trials it is generally advisable to see whether the outcomes in the control groups are similar. Why? Table F7 shows marked heterogeneity of the stroke rates in the control groups. What are the most likely explanations for this variation?

Question F7–10

In Figure F7(D) the trials are arranged in the sequence of the stroke rates in their control groups (with the lowest rate on the left), and rate differences are plotted. The graph shows a clear association (correlation coefficient = .91). Can this explain the heterogeneity of rate differences? There was no such association between the rate ratios and the stroke rates in the control groups (correlation coefficient = .08).

Notes

F7–1. *Quality scores* are arbitrary, for they depend on the items included and the weight given to each; but different systems tend to rank studies in the same order (Detsky et al., 1992). Greenland (1998b) believes that quality scores are too arbitrary and should be avoided; instead, he recommends the separate inclusion of each quality-relevant feature in a regression analysis. Methods of assigning quality scores to clinical trials are described by (*inter alia*) Chalmers et al. (1981), Klein et al. (1986), Liberati et al. (1986), Zelen (1983), and Detsky et al. (1992). The simplest method (Chalmers et al., 1991) examines three aspects: method of treatment (full marks for true randomization); control of selection bias after treatment assignment (full marks if both intention-to-treat and on-randomized-treatment analyses were done); and blinding of participants and investigators (full marks if subjects, caregivers and outcome assessors were all kept unaware of the treatment assignment). Criteria for case-control studies are listed by Feinstein (1985) and Lichtenstein et al. (1987). A formula for weighting by quality when results are combined is provided by Fleiss and Gross (1991). The use of quality scores as weights in multiple logistic regression analysis is described by Detsky et al. (1992).

F7–2. Although different outcomes may be measured in a trial, investigators may tend to report only those that show statistically significant findings; other results may be merely reported as "not significant," or omitted. One way of handling this "reporting bias" is to assume that unreported results were statistically null and give them effect sizes of 0 (Felson, 1992).

F7–3. The quotations come from Abramson (1990/91), Furberg and Morgan (1987), Goodman (1991), Naylor (1988), and Wolf (1986).

F7–4. For references describing the heterogeneity tests usually used, see Note D13–1. Heterogeneity tests for effect sizes are described by Hedges and Olkin (1985) and Wolf (1986), and testing of heterogeneity in a multiple logistic regression analysis by Detsky et al. (1992).

F7–5. 2.82E−13 is expressed in scientific notation. It means 2.82×10^{-13}; to convert the number to ordinary (fixed point) notation, multiply 2.82 by 10^{-13} (−13 is the exponent of 10). In other words, move the decimal point 13 spaces to the left, which gives 0.000000000000282. If there is no minus, move the decimal point to the right; 2.82E4 is 28200.

F7–6. Rate ratios and odds ratios are measures of relative difference, and a

logarithmic scale is therefore appropriate (see Unit A4). Absolute differences, such as rate differences and effect sizes, should be plotted on an ordinary scale.

■■■■■■■■■■ Unit F8

Appraising Combinability

Opinions differ as to whether nonexperimental studies in which nonrandomly selected groups are compared should be combined (*Question F7–1*). At one extreme is the view that these studies have so many possible biases that they should never be submitted to a meta-analysis; at the other is a permissive approach that accepts all studies that conform to the objectives and inclusion criteria of the meta-analysis (i.e., those that are relevant and have sufficiently similar features, according to the rules laid down for the meta-analysis), provided that their results are reasonably consistent. The results of the studies of passive smoking were fairly similar, and the summary odds ratio was 1.12 (95% confidence interval, 0.95 to 1.30)—a finding that did not convincingly support a relationship to lung cancer in the United States. The odds ratio of 1.17 for the cohort study was similar to the summary odds ratio of 1.07 for the case-control studies.

Question F7–1 serves to highlight the importance of knowing (and, in a published meta-analysis, of reporting) the methods and possible biases of the studies included in a meta-analysis. This knowledge may help to explain any heterogeneity of the results, and may influence the inferences from the findings of the meta-analysis. Sometimes, possible biases can be reduced before the findings are combined. In a trial, this generally involves collecting additional information—for example, about the outcomes in subjects who did not continue the regimen to which they had been randomized. In an observational study, it may involve statistical procedures to control for differences in sociodemographic or other characteristics between the groups compared or to compensate for misclassification of subjects with regard to the independent or dependent variable (see Note F8–1).

The authors of the meta-analysis of smoking and leukemia (*Question F7–2*) could not explain the difference between the results of the cohort and case-control studies. They suggested that it might be due to the biases commonly found in case-control studies, such as those arising from the manner of selection of controls. The use of different measures—the risk ratio for the cohort studies and the odds ratio for the case-control studies—is a red herring. It cannot explain the difference, since simple algebra (see Note F8–2) shows that an odds ratio would, in this instance, exceed a risk ratio based on the same data. If a summary odds ratio had been estimated from the cohort studies, it would have been

more than 1.3. Before combining the cohort and case-control studies in a single analysis (using odds ratios throughout), the heterogeneity of the odds ratios in the combined studies should be assessed. It is probably advisable not to combine the results of the cohort and case-control studies; if this is done, it should be done with reservations, because the overall result will depend on how many studies there are of each type.

The results of different studies are never identical. The issue is not whether differences exist, but whether they can reasonably be ignored. If a statistical test for heterogeneity yields a low *P* value (*Question F7–3*), the differences between study results should not be ignored. But these tests have a low power, and there is no clearly defined critical level; .05 may be regarded as too low a critical level. As an arbitrary rule of thumb, *P* values below .1 can usually be taken to mean that the differences should not be ignored; and if the number of studies is small, even higher *P* values than this should not be regarded as safe. However few the studies may be, a value exceeding .5 can usually be taken as convincing support for homogeneity; in the meta-analysis of passive smoking studies (*Question F7–1*) the *P* value was .71. Unless the *P* value is very high, possible heterogeneity should also be assessed by visual inspection.

In answer to *Question F7–4,* a heterogeneity test result of $P = .001$ would mean that the study results should be regarded as heterogeneous. The assumption that there is a single underlying true effect—a "fixed effect"—that can be estimated from the results in the separate studies, then becomes untenable. We have an "apples and oranges" situation, with different true effects, and the Mantel-Haenszel summary measure can be regarded only as a convenient weighted average of the measures, but not as an estimate of a single fixed effect. An average of discrepant results may be misleading (remember the statistician with his head in a freezer and his feet in an oven, who felt comfortable on the average).

Even if the results were heterogeneous (*Question F7–5*), overall tests showing a significant effect on fatality would mean that we can confidently reject the null hypothesis that there is no effect in any trial. In the presence of heterogeneity it would be wrong, however, to use the Mantel-Haenszel test result as an indication of the significance of a common measure of association.

Charting the results, as in Figure F7(A), often makes the appraisal of homogeneity or heterogeneity relatively easy. The values need not be arranged in sequence, as they are here. The confidence intervals show how precise each estimate is; but they may be confusing, as the studies with the largest and therefore most eye-catching confidence intervals are the most imprecise ones. The confidence intervals also show which results are statistically significant: If 1 is not included in the 95% confidence interval, *P* can be taken to be under .05. (In answer to *Question F7–6,* the 4th, 7th, 8th, 9th, 10th, 12th, and 14th values are significant). The impression provided by a diagram is, of course, subjective, and judgments may differ. But it is clear that most of the values are similar, with only one divergent result on the left and a few on the right. Also, the aberrant results have especially wide confidence intervals, and all the confidence intervals overlap. It is probably safe to decide that the results are consistent enough to war-

rant use of the overall measure. This combinability was confirmed by the test for the heterogeneity of the rate ratios, which yielded a *P* value of .73.

Figure F7(B) is an example of a "funnel display," in which results are plotted against a measure of precision (this is generally sample size or the reciprocal of the variance). The idea is that if all studies are in fact estimating a similar value, the spread of results should become narrow as precision increases, forming a funnel-like shape. This (in answer to *Question F7–7*) is what is seen here. The point estimates in the left half of the diagram vary, whereas those on the right almost form a straight line. The confidence intervals based on small samples are broad, whereas those associated with larger samples are narrow. With a little imagination the picture resembles a funnel, and this suggests that random variation is the main explanation for the inconsistencies among the values.

Close inspection of Figure F7(C), which displays the rate differences in the same 14 studies (*Question F7–8*), suggests more heterogeneity than Figure 7(1), with special reference to the two left-hand values, which are "outliers." The first confidence interval has no overlap with the 9th, 10th, 11th, 12th, or 13th confidence interval, and the second confidence interval has no overlap with the 10th, 11th, 12th, or 13th confidence interval. It would not be surprising if a test showed significant heterogeneity. The test result was, in fact, $P = .006$ indicating that the pooled value should not be used as an overall measure of effect.

As this example shows, one measure of an effect may manifest heterogeneity whereas another does not. The odds ratios in this meta-analysis, like the rate ratios, were not significantly heterogeneous ($P = .50$). We can therefore express the effect of antihypertensive treatment in terms of a single rate ratio or odds ratio, but not in terms of the absolute reduction in the rate of stroke occurrence. (Does this matter? See Note F8–3). An extreme discrepancy of this kind is unusual. For the meta-analysis of 23 trials of beta-blockers shown in Table F1, for example, the heterogeneity test results were $P = .40$, .38, and .14 for rate ratios, odds ratios, and rate differences, respectively.

Explaining Heterogeneity

Once it has been decided that there is more heterogeneity than can easily be attributed to chance, that it is too marked to be ignored, and that it cannot be avoided by switching to a different measure of effect, the next step is to consider and examine possible explanations. Differences should be brought to the surface and examined, rather than drowning them in a statistical pool.

Obvious possible explanations are that the methods or circumstances of the studies were very different, or that the people studied were very different. One simple way of exploring this possibility, in studies in which treatment groups, cases, or groups exposed to a supposed risk or protective factor are compared with control groups, is a comparison of the findings in the various control groups. The most likely explanation for the wide variation in stroke rates in control groups seen in Table F7 (*Question F7–9*) is that the follow-up periods differed; there may also have been differences (especially in age) between the samples

studied, or differences in definitions or in methods of ascertainment. Careful reading of the study reports would probably clarify the main reasons.

The heterogeneity of stroke rates in the control groups, together with the association between these rates and the rate differences, can easily account for the observed heterogeneity of rate differences (*Question F7–10*). The two "outliers" with respect to rate differences (trials 5 and 11 in Table F7) were among those with high stroke rates in their control groups. If antihypertensive treatment reduces the stroke rate to about 0.61 of the rate in the control group (as the pooled value in Table F7 suggests), the absolute rate differences can be similar only if the rates in the control groups are similar.

If the heterogeneity in this instance is accounted for by differences in the follow-up period, it may be regarded as an artifact, as also in meta-analyses where it can be attributed to methodological flaws (such as insufficiently objective methods of measurement) in some or all studies. If heterogeneity of this sort cannot be sidestepped (e.g., as in this study, by the use of odds or rate ratios rather than rate differences) or somehow controlled in the analysis, it may preclude any conclusions, or lead to qualified conclusions.

On the other hand, heterogeneity may be an expression of interesting effect modification. Comparisons of the results of different studies can be used for testing or developing hypotheses concerning factors that affect the association under study, so that heterogeneity becomes an asset rather than a liability. The strategy of meta-analysis is to "combine results if you can, compare them if you can't." Instead of the fixed-effect model (which assumes that the association under investigation is equally strong in every study, apart from random variation), use may then be made of a *fixed-effects* [plural] model, which assumes that there are different fixed effects in different sets of studies, or a regression model, which assumes that the effect is altered by a specific amount by each variable included in the model. (What is the assumption if the dependent variable in the regression model is the logarithm of the rate ratio or odds ratio? See Note F8–4.)

If there is unexplained heterogeneity, it is difficult to draw useful conclusions, since there are unknown biases or unknown modifying factors. In such circumstances a *random-effects model* is sometimes used to summarize the findings. This model is based on the assumption that the true effects in the different studies differ, and are randomly positioned about some central value. Allowance is made for the variation between studies as well as within studies. Some experts query the usefulness of the random-effects approach (see Note F8–5) on the grounds that its assumptions are difficult to justify. The random-effects model is sometimes used even when there is little heterogeneity; its results are then very similar to those provided by the Mantel-Haenszel procedure and other methods that use the fixed-effect model.

Exercise F8

Question F8–1

A meta-analysis of eight community-based controlled trials of vitamin A supplementation (previously referred to in *Question F5–1*) showed a beneficial effect

Table F8–1. Results of Eight Controlled Trials of Vitamin A Supplementation; Odds Ratios Showing Effect on Mortality in Children Aged 6 to 72 Months

Trial	Location	Odds Ratio	
		Point Estimate	95% Confidence Interval
1	Sarlahi, Nepal	0.70	0.57–0.87
2	Northern Sudan	1.04	0.81–1.34
3	Tamil Nadu, India	0.45	0.31–0.67
4	Aceh, Indonesia	0.73	0.56–0.95
5	Hyderabad, India	1.00	0.64–1.55
6	Jumla, Nepal	0.73	0.58–0.93
7	Java, Indonesia	0.69	0.57–0.84
8	Bombay, India	0.20	0.09–0.45

on mortality in children aged 6 to 72 months. The Mantel-Haenszel summary odds ratio was 0.72 (95% confidence interval, 0.66 to 0.79). However, the results, which are listed in Table F8–1, showed significant heterogeneity ($P = .0004$). [Do the results look heterogeneous? This may be easier to answer if you chart them, using the format of Figure F7(A); use a logarithmic scale, or plot the logarithms of the odds ratios and their confidence limits on an ordinary scale. When you have answered, see Note F8–6.] Suggest at least three possible reasons for the heterogeneity.

Question F8–2

The results of a sensitivity analysis, exploring the possibility that the heterogeneity was related to the quality of the studies, are listed in Table F8–2; combined results were recomputed after omitting first the worst study, then the two of lowest quality, then the three worst, etc. The criteria of study quality suggested

Table F8–2. Combined Results of Eight Controlled Trials of Vitamin A Supplementation; Summary Odds Ratios for Mortality in Children Aged 6 to 72 Months: Sensitivity Analysis

Studies Pooled	Heterogeneity Test (P)	Summary Odds Ratio (With 95% C.I.)
All eight	.0004	0.72 (0.66–0.79)
All but the poorest study	.01	0.74 (0.67–0.81)
All but the two poorest studies	.006	0.76 (0.68–0.84)
All but the three poorest studies	.005	0.76 (0.67–0.86)
All but the four poorest studies	.004	0.75 (0.66–0.85)
All but the five poorest studies	.001	0.75 (0.65–0.87)
All but the six poorest studies	.020	0.82 (0.70–0.97)

by Chalmers et al. (1981) were used. The numbers used in Table F8–1 indicate the ranking of the trials according to their quality; trial 1 was the best and trial 8 the worst. What light do the new results throw on the reasons for heterogeneity?

Question F8–3

Vitamin A was administered in small frequent doses in trials 3 and 7, and in large doses once or every 4–6 months in other trials. The summary odds ratio was lower (i.e., the apparent protective effect was larger) in the former two trials (odds ratio, 0.58; 95% confidence interval, 0.37 to 0.92) than in the latter trials (odds ratio, 0.81; 95% confidence interval, 0.68 to 0.97). What, in general terms, are the possible explanations for this difference? Would it be helpful if we knew that the difference was statistically significant? Why is it worth knowing that the two sets of trials yielded different summary odds ratios?

Question F8–4

If we wish to make this meta-analysis the basis for a policy decision concerning the use of vitamin A supplementation to reduce child mortality in developing countries, should we use the Mantel-Haenszel odds ratio, which is based on the fixed-effect model, or an odds ratio (the DerSimonian-Laird odds ratio) that is based on the random-effects model)? The respective 95% confidence intervals are 0.66 to 0.79 (Mantel-Haenszel) and 0.58 to 0.85 (DerSimonian-Laird). Or should we use neither? Can a certainty of 95% be ascribed to a 95% confidence interval?

Question F8–5

In another kind of sensitivity analysis, which is recommended if there are few studies, the impact of each study is examined by seeing how its removal influences the overall findings. As an example, a meta-analysis of six randomized controlled trials of the effectiveness of aspirin in preventing death after a myocardial infarction provided a Mantel-Haenszel odds ratio of 0.90 (95% confidence interval, 0.80 to 1.02), with a heterogeneity P of .076. The DerSimonian-Laird odds ratio (using the random-effects model) was 0.84 (95% confidence interval, 0.70 to 1.02). Table F8–3 shows the findings of each trial and the summary results after exclusion of each trial in turn. What conclusion can be reached about the value of aspirin in reducing the risk of death after myocardial infarction? There was no obvious reason for the discrepant result of trial F; should this trial be excluded?

Question F8–6

A subsequent meta-analysis (Fleiss and Gross, 1991) was able to include a large new randomized controlled trial of the effect of aspirin, in which the odds ratio was 0.89. The Mantel-Haenszel odds ratio for the seven trials was 0.90 (95% con-

Table F8–3. Results of Six Randomized Controlled Trials of the Effect of Aspirin in Preventing Deaths (From Any Cause) Within Two Years After Myocardial Infarction*: Sensitivity Analysis

Study	Aspirin: Deaths /Total	Placebo: Deaths /Total	Odds Ratio**	Combined Results, Excluding Specified Study	
				Heterogeneity Test (*P*)	Odds Ratio** (Mantel-Haenzel)
A	49/615	67/624	0.72 (0.48–1.08)	.075	0.93 (0.81–1.06)
B	44/758	64/771	0.68 (0.45–1.03)	.099	0.93 (0.82–1.06)
C	102/832	126/850	0.80 (0.60–1.07)	.058	0.93 (0.81–1.07)
D	32/317	38/309	0.80 (0.57–1.36)	.045	0.91 (0.80–1.03)
E	85/810	52/406	0.80 (0.54–1.17)	.050	0.92 (0.80–1.05)
F	246/2267	219/2257	1.13 (0.96–1.34)	.960	0.76 (0.65–0.90)

*A meta-analysis cited by Bailey (1987) and Fleiss and Gross (1991).

**95% confidence intervals shown in parentheses.

fidence interval, 0.84 to 0.96), and the DerSimonian-Laird odds ratio was 0.88 (95% confidence interval, 0.77 to 0.99). The heterogeneity *P* was .126. Do you want to change your conclusion about the value of aspirin in reducing the risk of death after myocardial infarction?

Notes

F8–1. Greenland (1998b) discusses methods of handling confounding, selection bias, and misclassification in a meta-analysis. Spitzer (1991) provides a list of "unanswered questions" about the combinability of nonexperimental studies, including those raised in *Question F7–1* and others (e.g., "Should control groups assembled by matching be combined with independent samples of . . . populations?"), and expresses the view that until they are answered the widespread application of meta-analysis to these studies is not warranted, except as methodological research.

F8–2. Call the risk in smokers $P1$ and the risk in nonsmokers $P2$ (both $P1$ and $P2$ are between zero and 1). The relative risk is $P1/P2$. As we saw in Unit B11, odds $= P/(1 - P)$. Therefore, the odds ratio is $P1/(1 - P1)$ divided by $P2/(1 - P2)$. This is the same as $P1/P2$ (the relative risk) multiplied by $(1 - P2)/(1 \times P1)$. Because in this instance $P2$ is less than $P1$, $(1 - P2)$ must exceed $(1 - P1)$, and $(1 - P2)/(1 - P1)$ must thus be more than 1. Hence, the odds ratio must be larger than the relative risk. In the beta-blocker example (statements 8 and 10 in *Question F2–1*) the odds ratio was lower than the risk ratio, because $P1$ was less than $P2$.

F8–3. A general rule of thumb for preferring the absolute difference or a relative difference (like a rate ratio) was suggested in Unit A3.

F8–4. In a regression equation with the format

$$\text{log rate ratio} = a + b_1x_1 + \ldots + b_ix_i,$$

each regression coefficient b_i tells us the mean increase in the log rate ratio associated with a rise of one unit in the value of independent variable x_i (see Units D11 and D13). Increasing the log rate ratio by b_i is equivalent to multiplying the rate ratio by b_i. The assumption is that each independent variable included in the model has a specific multiplicative effect on the rate ratio. This principle is illustrated in a meta-analysis of studies of senile dementia (Ritchie et al., 1992), in which simple regression analysis, using the log of the prevalence as the dependent variable, showed that prevalence increases exponentially with age, with a doubling of the rate for each 6-year increase in age.

F8–5. The *random-effects* model assumes that the studies are representative of a hypothetical universe of studies with a specific statistical distribution of effects, and it estimates the findings in this hypothetical universe. Allowance is made for the variation between as well as within studies, so that the summary measure has a wider confidence interval than that provided by the fixed-effects model, and its statistical significance is lower (Berlin et al., 1989); the results are similar, however, if heterogeneity is slight. The random-effects model gives more weight to small studies than the fixed-effects model. Methods are decribed (*inter alia*) by DerSimonian and Laird (1986), Petitti (1994), Whitehead and Whitehead (1991), and Fleiss and Gross (1991). Proponents of the random-effects model suggest that it is more appropriate than the fixed-effect model if the intention is (as usually it is) to make generalizations that go beyond the studies actually included (Berlin et al., 1989, Fleiss and Gross, 1991): "The fixed effects model leads to valid inferences about the particular studies that have been assembled. The random-effects model leads to inferences about all studies in the hypothetic *population* of studies" (Berlin et al., 1989). Other experts query the usefulness of the random-effects approach, on the grounds that it is based on assumptions that are difficult to justify (Hedges, 1987; Thompson and Pocock, 1991; Jones, 1992). Pocock and Hughes (1990) conclude that "neither the fixed effect nor the random effects model can be trusted to give a wholly informative summary of the data when heterogeneity is present.

F8–6. The point estimate of the odds ratio for study 8 is much lower than the other point estimates, and the 95% confidence interval for this study shows no overlap with any other confidence interval, except that for study 3. Also, there is no overlap between the confidence intervals for study 2 (which has the highest point estimate) and study 3.

■■■■■■■■■■ Unit F9

Explaining Heterogeneity (Continued)

Some of the most obvious possible reasons for the heterogeneity of the effects of vitamin A in preventing child mortality (*Question F8–1*) are differences between the populations studied (especially with respect to their nutritional and morbidity status), the use of different dosages, different durations of intervention, and differences in the quality of the trials.

When studies of poor quality were omitted from the analysis, the heterogeneity remained statistically significant. This indicates that the observed heterogeneity was not attributable, or at least not attributable solely, to the poor quality of some studies (*Question F8–2*).

The difference observed between the findings of studies using different dosage schedules (*Question F8–3*) may reflect the modifying effect of the dosage schedule (which may have various mechanisms), but it may also be due to an artifact, a chance occurrence, or confounding. We have previously (in Unit F3) considered two possible manifestations of confounding in a meta-analysis: distortion of the results of individual studies and (as a consequence of imbalances between the sizes of the groups compared) distortion of the combined results. Here we are concerned with a third possibility: that studies with different dosage schedules may differ in other respects too—such as the nutritional status of children in the communities studied—and the different results may be due to these other differences rather than, or in addition to, the differences in dosage.

The difference between the findings in the two groups of studies obviously explains part of the overall inconsistency of findings, whatever the reason for the difference—even if it is a chance occurrence.

The meaning of a significance test depends on whether the association tested (in this instance, the association with dosage schedules) was postulated as a study hypothesis before inspection of the findings (a priori). If so, the result can be taken at its face value. But if the test is done only because a difference was noticed when the findings were examined (a posteriori), a significant result may be misleading. Chance differences occur in every set of data, and any difference observed when the data are searched for interesting findings (a process of "data dredging" or "panning for gold") may be a chance one. The same data cannot then be validly used for a conventional test of significance, and hence "we are unable to separate the phantom effects from real ones" (Furberg and Morgan, 1987). As an illustration of the "statistically significant" chance associations that may be brought to light by data dredging, in a randomized controlled trial of the treatment of acute myocardial infarction with intravenous atenolol, with over 16,000 subjects, the percentage reduction in the odds of death was 48% among subjects born under Scorpio ($P < .04$), and only 12% (not significant) among subjects born under other astrological birth signs (Collins et al., 1987). The prob-

lem of spurious significance did not arise in the instance described in *Question F8–3*, as the significance test yielded a *P* value of .21.

The difference between the findings in the two groups of studies obviously explains part of the overall inconsistency of findings, whatever the reason for the difference—even if it is a chance occurrence. The value of knowing that the difference exists depends on whether the hypothesis was derived from the data. If the hypothesis was formulated in advance (and if we decide that the difference is probably not due to chance, flawed methods, or confounding), the finding has obvious practical implications. If not, detection of the difference—even if the role of chance or other factors cannot be excluded—is still of value because it permits the generation of a hypothesis for subsequent testing. This is an example of how a comparison of apples and oranges can raise new questions—which may be one of the main fruits of a meta-analysis.

In an "apples and oranges" situation, the Mantel-Haenszel odds ratio or any other summary measure based on the fixed-effect model can be used only as a weighted average of the results of the studies included in the meta-analysis, but not as an estimate of the effect that might be expected elsewhere (*Question F8–4*). A summary measure based on the random-effects model (see Note F8–3) is often regarded as a more appropriate basis for generalizations that go beyond the studies included, and for policy decisions. If nothing else, it provides a wider confidence interval, which seems to express the variety of the findings better. Neither measure is ideal; use of the random-effects model "exchanges a doubtful homogeneity assumption for a fictitious random distribution of effects" (Greenland, 1998b, p. 668). However, the attributable or preventable fractions computed from either summary measure and its confidence intervals can generally be used as guides for a policy decision.

Because the studies included in a meta-analysis do not constitute a random sample of all the situations in which the findings of the meta-analysis might be applied (although the random-effects model makes this assumption), the confidence intervals of summary measures cannot be taken too literally. It is prudent to regard them as underestimates of the range of variation in the real world, and to attach less than 95% certainty to a 95% confidence interval (Fleiss and Gross, 1991). One recommendation is that 99% confidence intervals should be used in meta-analyses (Peto, 1987b). In the vitamin A meta-analysis the 99% confidence intervals for the odds ratio were .64 to .82 (Mantel-Haenszel) and .54 to .90 (random-effects model). The latter figures may be translated into a preventable fraction of 10–46%.

In the meta-analysis of trials of aspirin after myocardial infarction (*Question F8–5*), the sensitivity analysis shows that the results were heterogeneous only if trial F was included. This was the only trial that did not show a reduction in the risk of death, and it was so large that it had a very appreciable impact on the summary odds ratio. If trial F is included, the effect on fatality is not statistically significant; whereas if it is excluded the preventive effect is stronger (odds ratio = 0.76) and significant. Trial F was actually reported some years after the other five; previously, the summary odds ratio had been significant; with the addition

of trial F to the meta-analysis, the effect became nonsignificant. If the discrepant result is obviously due to flawed methods, trial F can legitimately be excluded; but we have no evidence that this is so. The findings are therefore inconclusive. Maybe aspirin is helpful; maybe it is not.

The addition of a new trial (*Question F8–6*) changes the picture. The combined results now indicate that aspirin has a modest but statistically significant effect. But the basic heterogeneity remains; the heterogeneity *P* is now .126, a value that, in a comparison of only seven results, is not high enough to provide assurance that the differences can be ignored. The overall result may still be so fragile that another new trial might alter it. The authors' conclusion was expressed very cautiously:

> Aspirin seems to be a modestly effective agent . . . with a percentage reduction in the odds for dying relative to placebo equal to approximately 10%. The limits of uncertainty about this value are unsure, with the conservative random effects approach yielding a much wider confidence interval than the anticonservative fixed effects approach. . . . It would be prudent always to attach greater uncertainty than provided by traditional confidence intervals to the results of a meta-analysis of studies conducted to date. (Fleiss and Gross, 1991)

The kind of sensitivity analysis shown in Table F8–3 may be advisable in all meta-analyses of small numbers of studies, to appraise the influence of each study. At the very least, the analysis should be repeated with the largest study excluded to assess the influence of that study (Andersen and Harrington, 1992). It may be unwise to draw a conclusion that hinges on a single study.

Exercise F9

Question F9–1

In the meta-analysis of controlled trials of work-site smoking cessation programs referred to in *Question F5–5*, the mean effect size was significantly higher in the six comparisons conducted at work sites with under 250 employees than in the 28 conducted at larger work sites. What is the value (if any) of this finding? (Ignore the possible bias caused by the overrepresentation of trials that embraced more than one comparison, as explained in *Question F5–5*.)

Question F9–2

A meta-analysis of eight randomized controlled trials of intravenous streptokinase for acute myocardial infarction showed a summary risk ratio for fatality of 0.80, with significant heterogeneity. It was afterwards shown that the risk ratio was different in trials with different eligibility rules (Zelen, 1983). In the two trials that excluded patients whose symptoms had lasted for more than 72 hours, the risk ratio was 1.29. In the three trials where the maximum duration of symptoms was 24 hours, the risk ratio was 0.80. And in the three trials where the max-

Table F9–1. Results of 23 Randomized Controlled Trials of the Effect on Fatality of Long-Term Use of Beta-Blockers After Myocardial Infarction; Comparison of Trials of Beta-Blockers With and Without Intrinsic Sympathomimetic Activity (ISA)

Type of Beta-Blocker	No. of Trials	Heterogeneity Test (P)	Summary Rate Ratio* (With 95% C.I.)
Without ISA	12**	.70	0.72 (0.64–0.81)
With ISA	11†	.60	0.91 (0.81–1.02)
Total	23	.38	0.79 (0.73–0.87)

*The summary rate ratios are precision-based.

**Trials 1, 3, 6, 9, 13–16, 18, 19, 21, and 23 in Table F1.

† Trials 2, 4, 5, 7, 8, 10, 11, 12, 17, 20, and 22 in Table F1.

imum duration of symptoms was 12 hours, the risk ratio was 0.69. These differences were significant ($P = .01$). What is the value (if any) of these findings?

Question F9–3

Do older men have less testosterone (male sex hormone) in their blood? A meta-analysis of 88 studies (Gray et al., 1991) displayed heterogeneous findings, with age–testosterone correlations ranging from −.68 (a moderate decrease with age) to +.68 (a moderate increase). Weighted regression analyses showed that the direction and degree of change with age varied significantly with the health status of the subjects and the time of day at which blood was taken. For example, in studies that included ill men there was no decline with age, whereas in studies of healthy subjects there was a decline with age. How can the significance of modifying effects be tested in a regression analysis? How can possible confounders be controlled in a regression analysis? What is the value (if any) of the findings of this meta-analysis? Do not try to explain the findings.

Question F9–4

Let us return to the meta-analysis of 23 beta-blocker trials, with which we started (Table F1). There seemed to be no reason not to combine the results; the heterogeneity P was .38. However, the results of trials using different kinds of beta-blocker were compared to determine whether different beta-blockers differed in their preventive value. This revealed no differences related to the drug's cardioselectivity or its membrane-stabilizing difference, but there was a difference related to intrinsic sympathomimetic activity (ISA). As shown in Table F9–1, the effect on mortality was larger in the 12 trials that used beta-blockers without ISA than in the other 11 trials, where the effect was weak and not statistically significant. The two summary odds ratios were significantly different from each other ($P < .01$). The authors conclude that "it appears that [beta-blockers with appreciable ISA] may confer less benefit," but say this "remains uncertain, for the

distinction between these two categories of agent was a data-derived hypothesis" (Yusuf et al., 1985). Were they right to draw attention to the difference? Was it right to soft-pedal it?"

Question F9–5

A meta-analysis of 19 randomized controlled trials of measures to reduce plasma cholesterol (Holme, 1993) showed a significant reduction of coronary heart disease incidence, with a summary odds ratio of 0.91 (95% confidence interval, 0.87 to 0.96). The results were very heterogeneous ($P = .027$). So were the trials: Some used drugs, some used dietary or other measures; several trials were multifactorial ones that tried to control other risk factors also; some were concerned with preventing first episodes (primary prevention), others with recurrences (secondary prevention). A weighted regression analysis was performed, with the log odds ratio for coronary heart disease incidence as the dependent variable (see Note F8–4) and the mean percentage decrease in cholesterol in the study as the independent variable. This led to the conclusion that on average the incidence rate decreased by 2.5% (95% confidence interval, 2.0% to 3.0%) for each 1% reduction in cholesterol. The heterogeneity P rose to .14 when differences in cholesterol response were controlled in the analysis. This study suggests that the inconsistent effects on incidence are largely explained by differences in plasma cholesterol reduction. In the context of this meta-analysis, is plasma cholesterol reduction a modifying factor, a confounding factor, or what?

Question F9–6

In the meta-analysis of eight community-based trials of the value of vitamin A supplements in preventing child mortality, the modifying effect of age was examined not by comparing different trials or subsets of trials, but by comparing different subsets of individuals. The results are summarized in Table F9–2. Why

Table F9–2. Effects of Vitamin A Supplementation on Child Mortality in Controlled Community-Based Trials, by Age

Age (Months)	Studies Pooled*	Summary Odds Ratio** (with 95% C.I.)
0–11	1,2,3,4,6,7	0.76 (0.84–0.91)
12–23	1,2,4,6	0.90 (0.70–1.15)
24–35	1,2,4,6	0.89 (0.57–1.39)
36–47	1,2,4,6	0.94 (0.49–1.79)
48–59	1,2,4,6	0.80 (0.38–1.70)
≥60	2,4	0.55 (0.11–2.77)

*Numbered as in Table F8–1.

**DerSimonian-Laird method (random effects model).

do you think that different sets of studies are used for different age groups? Does this matter? Why do you think study 8 does not appear at all?

Question F9–7

Suppose that we wish to see whether smoking habits modify the effect of the long-term use of beta-blockers after myocardial infarction, by comparing the results in individuals with different smoking habits who were included in the trials listed in Table F1. What difficulties might we encounter?

■■■■■■■■■■ Unit F10

Effect Modification

The investigation of effect modifiers—factors that influence the outcomes of trials or the associations examined in nonexperimental studies—can be an important feature and sometimes the main contribution of a meta-analysis. It is usually done by comparing the results of different studies, either to test previously formulated hypotheses or to explain inconsistent findings. What is learned may have important theoretical and practical implications.

In the meta-analysis of smoking cessation programs (*Question F9–1*), the significantly greater success observed at small work sites is not necessarily attributable to the size of the workforce or factors related to the size of the workforce, such as (maybe) the degree of social interaction, integration, or support. It might be due to other (confounding) influences. But if confounders unrelated to the operation of the program (e.g., age and sex differences) can be excluded, careful examinations of how the programs at small and large work sites operated may point to ways of enhancing effectiveness.

An obvious possible explanation for the findings described in *Question F9–2* is that streptokinase is more effective if it is given early, and might even be harmful if given very late. The association may be attributable to other differences between the groups of trials, and the significance test may be misleading because the hypothesis was not formulated in advance. But the ostensible explanation may have practical importance, and can be tested in subsequent trials. In this meta-analysis it was not possible to compare the results in individuals with different duration of symptoms, because these data were not available (Stampfer et al., 1982).

Modifying effects can be studied not only by comparing mean effect sizes (as in *Question F9–1*), summary odds ratios (as in *Question F8–3*) or risk ratios (as in *Question F9–2*), or other measures of effect, but also by using regression analysis. In such an analysis the significance of a modifying effect can be tested in two ways (*Question F9–3*). First, if regression coefficients are calculated sep-

arately for different samples (e.g., for studies of healthy men and those including ill men), using the simple regression equation $y = a + bx$ (see Unit D11), where (in this instance) y = testosterone level, x = age, and b is the slope of the regression line, the difference between the b coefficients is an expression of a modifying effect (in this instance, of health status on the testosterone–age association), and its significance can be tested. Second, if multiple linear regression is used (Unit D17), the suspected modifier can be included in the model, together with a term that expresses its interaction with age. The significance of the latter term is the significance of the modifying effect. In this meta-analysis, a multiple regression analysis showed that the interactions between age and health status and between age and time of day were both significant ($P = .02$ and .01, respectively); that is, both health status and time of day modified the age–testosterone relationship. Suspected confounders can be controlled by adding them to the regression model.

The demonstration of factors that modify the age trend of testosterone in the blood may be a stimulus to research to explain these influences. It also has practical implications with respect to the performance of testosterone assays, the appraisal of their findings, and the way in which the results of future studies should be reported.

The comparison of two subgroups of beta-blocker trials—those using beta-blockers with and without ISA (*Question F9–4*)—raises the same problems as those considered in the comparison of vitamin A trials with different dosage schedules (in *Question F8–3*). There may be confounding because of other differences between the studies, and statistical testing may be misleading because the hypothesis was derived from the data and not formulated in advance.

One of the authors of the beta-blocker meta-analysis gave the following answer to *Question F9–4* some years after the report was published:

> This difference was conventionally significant at the 0.01 level. At the time this seemed rather impressive—and it did not take long to think up a biological "explanation" for it—but it is interesting that all the data that has turned up since has tended to contradict this finding. . . . More recently we have seen results of two more trials. . . . These two extra trials demolish that statistically significant interaction. In retrospect, we were wrong to give it as much credence as we did. It was right to observe it and report it; but it was wrong to believe it. (Peto, 1987a)

"You should look for subgroups, you should report what you find, and half the time you shouldn't believe it" (Peto, 1987b).

In the meta-analysis of trials with cholesterol-lowering measures (*Question F9–5*), heterogeneity with respect to one outcome (reduction of coronary heart disease incidence) is explained, or partly explained, by heterogeneity with respect to another outcome (plasma cholesterol reduction). Because the reduction in plasma cholesterol is presumably a link in the chain of causation between the cholesterol-lowering measures and the reduced incidence, it is an intermediate cause (see Unit A14) rather than a modifier or confounder.

When modifying effects are investigated by comparing the findings in various subsets of individuals, as in *Question F9–6,* it frequently happens that findings for specific subsets are not available from all studies. In this meta-analysis, studies 3 and 7 were apparently confined to children less than a year old, and studies 1 and 6 were apparently restricted to children less than 5 years old. Age comparisons based on the data shown in Table F9–2 might therefore be misleading. The published table does not include the ≥60-month age group. Study 8 apparently provided no age-specific information.

Comparisons of subsets of individuals—for example, those with different smoking habits (*Question F9–7*)—present numerous difficulties. As in the previous example, some studies may not provide the required information at all, and others (e.g., a study restricted to nonsmokers) may not supply it for all categories. If information is available, categories and definitions may differ in different studies. The information on different sets of individuals may thus be based on different sets of studies, raising possibilities of bias; exploration of this bias might require additional information at an individual level, which might not be available. Moreover, the shrinkage of sample sizes owing to the fact that separate categories of individuals are analyzed, and their further shrinkage due to incomplete information, may produce summary measures with very wide confidence intervals.

One suggestion made to overcome some of these difficulties is that, where possible, meta-analysis should be based not on study reports but on the collection and analysis of full data on all the individuals studied (Note F10); this is seldom possible.

Exercise F10

Question F10–1

A meta-analysis (cited in *Question F7–2*) showed a statistically significant association between cigarette smoking and leukemia; the summary risk ratio based on seven cohort studies was 1.3. Can it be concluded that smoking is a cause of leukemia; and if not, why not? What additional information from the meta-analysis might strengthen the case?

Question F10–2

Assuming that smoking is a cause of leukemia, what extra information is required to estimate how many cases of leukemia are caused by smoking each year in a given population?

Question F10–3

A clinician finds an up-to-date meta-analysis of randomized controlled trials showing that a particular treatment is effective in the treatment of a particular disease. The reported effect is statistically significant, and it is stronger, to a clinically meaningful degree, than the effect of the usual current treatment. Sup-

pose that there is no difference from the usual treatment in safety, side effects, cost, convenience of use, or acceptability to patients. What should the clinician look for in the report of the meta-analysis before deciding to apply the treatment to his or her own patients?

Question F10–4

In what ways can a meta-analysis help future research? If a meta-analysis of clinical trials clearly shows that a treatment is effective, does this mean that additional trials of the treatment are superfluous?

Question F10–5

Before the results of a meta-analysis are used, its quality should always be appraised. The author's eminence is no guarantee of validity. In fact, one study of review articles found that the greater the author's expertise in the content, the poorer the quality of the review (Oxman and Guyatt, 1993). Can you suggest a set of questions that might be asked about a meta-analysis in order to appraise its quality? List as many questions as you can.

Note

F10. Stewart and Parmar (1993) compare what they call MAP (meta-analysis of individual patient data) with MAL (meta-analysis of the literature). Using information collected by a group of investigators conducing cancer trials, they show that these two methods can provide different estimates of the effectiveness of a treatment, and they point out that MAP provides a less biased means of comparing results in different groups of patients.

■■■■■■■■■■ Unit F11

Using the Results

Although the meta-analysis of cohort studies (*Question F10–1*) showed an association between smoking and leukemia, this alone is not convincing evidence for a cause–effect relationship. The overall association is statistically significant, and the "cause" apparently precedes the"effect," but we do not know whether other criteria for the appraisal of causality (see Unit E10) are met. The observed association is not strong, and a weak association, especially one found in nonexperimental studies, can easily be caused by flawed methods or confounding.

The following additional information might be helpful: (1) How were the studies done? Can the association be readily explained as an artifact caused by flawed methods? (2) Were the smokers and nonsmokers in each study similar in

age, social class, ethnic or racial group, alcohol consumption, occupation, and other characteristics? And if not, were adequate measures taken to control for possible confounding? (3) Were the results of the studies consistent? What was the finding of each study, and were the findings tested for heterogeneity? Evidence of consistency and (if this is lacking) evidence of modifying effects are probably the main potential contributions of meta-analyses to etiological research. (4) Was there a dose–response relationship?

This meta-analysis did not include a systematic assessment of study quality. All the studies matched or controlled for at least age and sex, but no information is provided on the comparability of smokers and nonsmokers with respect to ethnic group or other characteristics; the authors say that "since the causes of leukemia are largely unknown . . . analyses cannot completely control for potential confounding." The findings of the studies were not very similar; two studies had risk ratios below 1 and confidence intervals that did not overlap with the confidence intervals of most other studies; a heterogeneity test was apparently not performed. An association with the number of cigarettes smoked per day was found in most but not all of the studies. The summary rate ratio was 1.4 (95% confidence interval, 1.3 to 1.6) for smokers of 1 to 19 cigarettes a day and 1.6 (95% confidence interval, 1.5 to 1.8) for smokers of ≥20 cigarettes a day; the report does not say how many studies supplied the data required for these comparisons of subsets of individuals. In the light of the evidence you now have, would you conclude that smoking is a cause of leukemia? (See Note F11.)

If it is assumed that smoking is a cause of leukemia, the population attributable fraction (*Question F10–2*) can be estimated from the risk ratio and the rate of smoking in the population (see Note E12 for the formula). If the number of new cases of leukemia per year in the population is also known, this fraction can be translated into an absolute number. The finding of this meta-analysis led to an estimate that about one in seven adult leukemia cases in the United States, or about 3,600 new cases per year, may be caused by smoking.

For a clinician who is deciding whether to adopt the findings of a meta-analysis of clinical trials (*Question F10–3*), the first prerequisite is confidence in the scientific quality of the meta-analysis and of the trials on which it is based. This requires careful reading of the report to see whether anything about the way the trials were found, selected, or analyzed gives rise to doubts about the validity of the results, and to see whether the quality of the trials was appraised and found satisfactory.

Next, the clinician should see whether the results of the trials were consistent: What were the findings, and were they tested for heterogeneity? If they were consistent, and the trials encompassed a varied collection of patients, the treatment can probably be considered for any patient. However, if the meta-analysis includes a summary measure of effect based on a subset of trials or individuals (e.g., trials conducted in a specific age group, or patients of a specific age group) that seems particularly relevant to a specific patient, the clinician may prefer to use this. (We will return to this issue later, in *Question G3–4.*)

If there is appreciable heterogeneity of the findings in the various trials, the

overall finding may have little relevance to a particular patient even if the random-effects model is used. The physician should then comb the report of the meta-analysis for descriptions of the kinds of patient included in each trial and the conditions in which each trial was administered, to see whether any trials are particularly relevant to his or her specific patient, and should rather use the results of these trials. If the meta-analysis provides a summary measure of effect based on a particularly appropriate subset of trials or individuals, this too may be preferred to the overall measure.

Whichever summary measure is used, its confidence interval should be sought and used as a guide to the decision on use of the treatment and to the prognosis if the treatment is used. There is, however, "invariably a leap of faith between formal statistical inference . . . and extrapolation to the true population of future patients." This uncertainty can be recognized by using broad confidence intervals if these are available: 99% rather than 95% intervals, and/or intervals based on the random-effects model rather than the fixed-effect model.

The bottom line in the answer to this question, as in the answer to the question (*F10–1*) on the meta-analysis of nonexperimental studies, is that the results of meta-analyses should be regarded at least as critically and applied with at least as much caution, as would the results of any individual study.

A meta-analysis can help future research (*Question F10–4*) in at least three ways. First, however inconclusive its results may be, it may draw attention to flaws in the design, conduct, or reporting of previous studies and thereby stimulate improved methods and reporting in future studies; "in order to do meta-analyses with a high degree of certainty tomorrow, one must do meta-analyses with a certain degree of uncertainty today!" (O'Rourke and Detsky, 1989). Second, it may resolve uncertainties and consolidate present knowledge, thus providing a firm basis for new research. And third, it can identify unexplained inconsistencies and unanswered questions, leading to the formulation of hypotheses for subsequent testing.

It is tempting to say that if a meta-analysis clearly shows that a treatment is effective, further trials of the treatment are unnecessary. Repeated (cumulative) meta-analyses of trials of many treatments for myocardial infarction have in fact shown that once a significant effect has been detected, the main consequence of adding new trials is narrowing of the confidence interval. For example, the summary odds ratio expressing the effect of intravenous streptokinase on mortality in myocardial infarction, based on eight trials involving 2,432 patients between 1959 and 1972, was 0.74 (95% confidence interval, 0.59 to 0.92). In 1988, 25 trials and 34,542 experimental patients later, the summary odds ratio was almost the same, but the confidence interval was much narrower (Lau et al., 1992).

But there may be surprises. We had an example in *Questions F8–5 and F8–6:* The addition of a sixth trial made a significant and consistent effect nonsignificant, and the addition of a seventh trial made it significant again; similarly, a meta-analysis of small trials of phenobarbital for intracranial hemorrhage in premature infants showed a positive effect, but a larger subsequent trial showed a detrimental effect (T. C. Chalmers, 1991). Also, new trials may permit a bet-

ter look at the effects of specific modes of treatment in specific kinds of patient. For example, a meta-analysis of trials of calcium antagonists in myocardial infarction "did not indicate any overall beneficial effect" (Held et al., 1989), but later meta-analyses showed a significant and marked benefit for patients with infarction of the non-Q-wave type who received calcium antagonists that reduce the heart rate (Yusuf et al., 1991; Boden, 1992). It is probably safe to conclude that if a meta-analysis based on many trials and several thousand subjects clearly shows a consistent and statistically significant effect, further trials are generally needed only if there is interest in questions (e.g., about modifying factors) that have not been adequately answered, or if there is reason to suspect that the effect may be modified by time-related or other differences.

Evaluating a Meta-Analysis

The questions that might be asked about a meta-analysis in order to appraise its quality (*Question F10–5*) include the following 30; see how many of them you mentioned. You may, of course, have thought of others.

Objective

Does the meta-analysis have a clearly defined objective?

Identification of Studies

Was the search for relevant published studies thorough?
Was a search made for unpublished studies?
Was the search unbiased?
Is the fail-safe *N* large?

Selection of Studies

Were explicit inclusion and exclusion criteria used?
If so, are they concordant with the objective of the meta-analysis?
Were precautions taken to avoid bias when selecting studies?

Quality of the Studies

Was the quality of the studies appraised?
Were explicit criteria used when appraising study quality?
Were precautions taken to avoid bias when appraising study quality?
Was appropriate attention given to study quality in the analysis?

Extraction of Results

Were precautions taken to avoid bias in the extraction of results?
Was missing information (if any) sought from the investigators?

Combining of Results

Are the studies similar enough (e.g., in design, study samples, definitions of variables, methods of data collection and analysis, and outcome criteria) to justify combining of their results?
Was the heterogeneity of the study results appraised?
Are the results similar enough to justify the combining of results?
Were appropriate statistical methods used to combine the findings?
Are confidence intervals presented?
Is the measure of effect concordant with the study objective?
Was sensitivity analysis used to appraise the effect of specific studies on the combined result?
Was sensitivity analysis used to appraise the effect of decisions about study eligibility or procedures used in the meta-analysis?

Comparison of Results

If the studies are dissimilar, were their results compared?
Were the study results compared, graphically or in other ways?
If the findings were heterogeneous, were the reasons explored?
If subgroups were compared, was possible confounding taken into account?
If subgroups were compared, were the pitfalls of tests of data-derived hypotheses taken into account?

Interpretation of Findings

Were possible biases in individual studies considered?
Were the results of the meta-analysis interpreted correctly?
Are the practical implications presented correctly, and with appropriate reservations?

As an example of the application of questions of this sort, the findings of a meta-meta-analysis of meta-analyses of published randomized controlled trials were summarized as follows: "We found indications of a written protocol in very few. Attempts to include all relevant trials seemed optimal in a minority and in none was the determination of suitability for inclusion made in a blinded manner (i.e., without knowledge of source or results of the trial). Interrater disagreement rates in the selection of papers and in the extraction of data were almost never reported. Statistical methods of combining the data were considered adequate in most, but only a few carried out sensitivity analyses by employing more than one method, or considered the problem of heterogeneity of results. . . . Publication bias was rarely considered. . . . Quality of the original trials was considered in few of the meta-analyses" (T. C. Chalmers et al., 1987; Sacks et al., 1987).

Over recent years, considerable efforts have been made to improve the qual-

ity of meta-analyses, particularly by the Cochrane Collaboration. This international network of interested individuals and institutions has set explicit standards for systematic reviews and provided a framework for the preparation and dissemination of reviews that meet these standards; it looks forward to an expansion of its efforts: "Relatively few health problems have been covered by systematic reviews so far. . . . It will take a concerted effort over many years to reach the point at which existing evidence about the effects of health care has been organized systematically and made available to the variety of people who need this information to help them make better decisions in health care and research" (Chalmers and Haynes, 1995).

Note

F11. In an editorial commenting on this meta-analysis, Severson and Linet (1993) say, "On balance, the evidence suggests a causal relationship between cigarette smoking and leukemia, but many questions remain." This is a matter of judgment, and you are entitled to disagree with this verdict.

■■■■■■■■■■ Unit F12

Test Yourself (F)

- Explain what is meant by
  meta-analysis (F1).
  a goodness-of-fit test (Note F2–1).
  a heterogeneity test (F8).
  a mean effect size (F3).
  unitlessness (F4).
  the fail-safe *N* (F5).
  a quality score (Note F7–1).
  sensitivity analysis (F7).
  a funnel display (F8).
  data dredging (F9).
  an a priori hypothesis (F9).
  an a posteriori hypothesis (F9).
- State arguments for and against the inclusion of unpublished studies in a meta-analysis (F5).
- Provide a list of
  possible explanations for differences between the results of clinical trials of the same topic (F2).
  possible benefits of drawing conclusions from a series of studies (F2).
  possible reasons for excluding old studies from a meta-analysis (F6).

possible reasons for including poor studies in a meta-analysis (F6).
possible procedures if the studies vary in quality (F7).
possible procedures if a study to be included in a meta-analysis offers more than one measure (F7).
possible reasons for differences between the results of different sets of studies (F9)

- Explain (in general terms)
how to minimize bias when deciding whether to include a study in a meta-analysis (F6).
how to minimize bias when appraising the quality of a study (F7).
how the combinability of studies can be appraised (F8).
- State the disadvantages (if any) of
simply pooling study results, as if only one large study had been done (F3).
combining study results by "vote counting" (F3).
combining *P* values from separate studies (F3).
calculating an average of rate ratios (F3).
using effect sizes (F4).
data dredging (9).
comparing subsets of individuals in a meta-analysis (F10).
using the findings of a specific trial or set of trials rather than the overall findings of a meta-analysis of trials (F11).
using the overall findings of a meta-analysis of trials rather than the findings of a specific trial or set of trials (F11).
- Explain (in general terms)
how the results of separate significance tests can be combined (Note F2–2).
how the confidence interval of a summary measure should be interpreted (F9, F11).
how a modifying effect can be investigated in regression analysis (F10).
how a confounding effect can be controlled in regression analysis (F10).
- Explain the following models:
a fixed-effect model (F3).
a fixed-effects model (F8).
a random-effects model (F8 and Note F8–5).
a regression model (F8).
a regression model with the logarithm of the rate ratio as the dependent variable (F8).
- Explain
how to interpret a goodness-of-fit test with a low *P* value (Note F2–1).
how to interpret a heterogeneity test with a low *P* value (F8).
the advantages (if any) of the Mantel-Haenszel procedure over multiple logistic regression analysis (F4).
the advantages (if any) of multiple logistic regression analysis over the Mantel-Haenszel procedure (F4).
why the results in the control groups of different studies should be compared (F8).

what should be done if the results of a meta-analysis would be appreciably modified by the exclusion of one study (F9).
how a meta-analysis can throw light on causation (F11).
what should be looked for in the report of a meta-analysis of trials before a decision is made to apply the treatment to a specific patient (F11).
why a meta-analysis that clearly shows the effectiveness of a treatment does not necessarily rule out the need for new trials of the treatment (F11).

- If you got a low score for *Question F10–5,* try again.

Section G

Putting Study Findings to Use

"That's very important," the King said, turning to the jury. They were just beginning to write this down on their slates, when the White Rabbit interrupted: "*Un*important, your Majesty means, of course," he said. . . .

"*Un*important, of course, I meant," the King hastily said, and went on to himself in an undertone, "important—unimportant—unimportant—important—" as if he were trying to decide which word sounded best.

(Carroll, 1865)

Unit G1

Introduction

The results of epidemiological studies may find practical application in both individual and community health care, as we saw in Unit A17. They may motivate people to alter their own or their family's lifestyles; they may lead to modifications in the preventive or curative care given to patients by physicians, nurses, and others; and they may trigger decisions by public health workers, administrators, and other policymakers with respect to health care at the local, regional, national, or international level.

In clinical care, epidemiological results are commonly used when decisions are made about the performance of screening or diagnostic tests, when test results are interpreted, and when decisions are made about treatment and prognosis. At a community level they may find expression in decisions about screening and prevention programs, programs for the management of common diseases or risk factors, programs for high-risk groups, and so forth.

A number of questions should be asked before deciding to apply study results in practice. The following pages deal with these questions.

Exercise G1

Question G1–1

A magazine published by a highly respected newspaper featured a six-page cover article that dismissed the notion that passive smoking is hazardous; it was entitled "Smoke without fire: Passive smoking—the myth and the reality" (Note G1). The article reported interviews with a number of health professionals who

said that passive smoking was not harmful, and it mentioned that a recent study by the Channing Laboratory of Harvard had not confirmed the hazard. The introduction noted that "comprehensive research in the prestigious medical periodical, the *British Medical Journal*, proves that there is no scientific basis for claims [that passive smoking is] an enemy of the people." To what extent should the reader be influenced by the interviews with health professionals?

Question G1–2

To what extent should the reader be influenced by the reported study by the Channing Laboratory?

Question G1–3

To what extent should the reader be influenced by the reference to the paper in the *British Medical Journal*? This actually referred to a meta-analysis of studies of passive smoking and lung cancer.

Note

G1. This magazine article is described by Siegel-Itzkovich (2000). The meta-analysis is by Copas and Shi (2000). Readers' electronic responses to both these articles can be found on the Internet in the *British Medical Journal*'s archives (www.bmj.com).

■■■■■■■■■■ Unit G2

Are the Results Accurately Known?

If epidemiological findings are to be applied in practice, the obvious first requirement is that these findings should be accurately known.

Reports in the media (press, radio, television, or Internet) should be treated with caution; they cannot always be relied on. "Journalism," it has been said, "is an activity with no scientific methodology" (De Semir, 1996), and this, together with an over-concern with immediacy and novelty and with circulation figures, ratings, or numbers of Web-site hits, can lead to the publication of information that is not completely correct.

Credibility is enhanced if the source of the information is an expert or a trustworthy committee or official agency. In answer to *Question G1–1*, then, the interviews with health professionals should clearly render the report more convincing. But this can be a Catch-22 situation, because the professionals interviewed may not have a sufficient degree of expertise or may not have been selected impartially or may have been reported incorrectly.

In the case of the magazine article under consideration, credence is open to question. The article provoked a furor. Critics pointed out that the selection of people for interviewing was unbalanced, and all eight of them were smokers; and a prominent cardiologist, whom the report had quoted as saying, "Years of work have been destroyed by the new evidence," denied that he had said this or that he had even been interviewed.

The mention of an unidentifiable study (*Question G1–2*) does not make the report more credible. In fact, one reader who tried to trace this "recent study by the Channing Laboratory" wrote to say "Since I couldn't find such a study and the journalist couldn't remember her source, I asked [the] Head of the Channing Laboratory, who replied, 'I am not aware of what article is being referred to. . . . We have published a great deal on passive smoking, and in every case I can remember the results have been associated with some health effects.'"

Mention of a study that can be traced and verified, on the other hand (*Question G1–3*), inspires confidence, particularly if reference is made to a meta-analysis and not to a single isolated study.

But there can be no guarantee that the study has been reported correctly. In this case it was grossly misrepresented. It was based on a meta-analysis of 37 studies, which showed that the risk of lung cancer in nonsmoking women was 24% higher if the woman's spouse or partner smoked. The authors of the meta-analysis appraised the possible effect of publication bias (the nonpublication of studies with negative or inconclusive results) on this finding. "We do not know," they said, "how many unpublished studies have been carried out," and they cited evidence suggesting that the number is "unlikely to be large." But they calculated that if only 60% of studies had been published—that is, if the 37 studies were supplemented by 23 hypothetical unpublished ones—the excess risk might fall from 24% to 15% (but remain statistically significant). This is what the magazine reported as "no scientific basis" for the harmfulness of passive smoking.

When epidemiological findings are gleaned from the media or by hearsay or from some other second-hand source, it is generally prudent to locate and read the original study report before deciding to apply the findings in practice, unless there is good reason to trust the source. This is particularly true when reporting might be influenced by vested interests or political considerations.

Similarly, it may be wise to read a full study report rather than to rely on an abstract, before considering practical application. In these days of easy computer access to literature abstracts (using MEDLINE or other databases) there is, unfortunately, a temptation to use abstracts as substitutes for full reports.

Exercise G2

Question G2–1

A case-control study (Langman et al., 2000) that compared the general practice records of 12,174 cancer cases and 34,934 controls indicated that treatment with aspirin and other anti-inflammatory drugs "may protect against" cancer of the esophagus (odds ratio, 0.61), stomach (0.51), colon (0.76), and rectum (0.75).

These effects were significant, and dose–response relationships were found. Assuming that the study has no methodological faults and that all relevant confounders were well controlled, would you consider applying its findings in practice?

Question G2–2

A meta-analysis may be an especially useful basis for decisions. A search was conducted for meta-analyses and systematic reviews on the treatment of asthma, and they were subjected to a critical appraisal of their quality. This appraisal was based on the way in which studies were sought, the avoidance of bias in the selection of studies, the use of defined criteria in appraising the validity of the studies, and other criteria (Jadad et al., 2000). Can you guess what percentage of the meta-analyses and reviews (over half of which were published in 1989–1999) had serious flaws (a score of 1 to 3 on a 7-point scale)?—about 25%, about 50%, or about 75%?

Question G2–3

An overview of review articles on the health effects of passive smoking that appeared in the medical literature over a 17-year period revealed that 63% concluded that passive smoking was harmful, and 37% that it was not (Barnes and Bero, 1998). The studies' verdicts were not significantly related to the quality of the review, as assessed by a blind evaluation similar to that described in *Question G2–2*. Nor were there significant differences between the conclusions of reviews that dealt with different health outcomes, or between those published in journals that submitted or did not submit papers to peer review, or between papers published in different years. Only one variable was very strongly associated with the direction of the conclusion. Can you guess what it was?

■■■■■■■■■■ Unit G3

Validity of the Findings

If epidemiological findings are to be applied in practice, the obvious next requirement (once these findings are accurately known) is that the validity of the study or studies should not be in doubt. This refers in particular to internal validity (see Unit B4): Were the study methods sound? Is the information they yielded accurate? And are the inferences drawn with respect to the study population well-founded?

A good part of this book has been devoted to these issues, and you should have little difficulty in appraising study validity. It should be easy to recognize a study's

main weak points with respect to sampling, selection of control groups, operational definitions of variables, methods of data-collection, the control of confounding, etc., and to detect questionable inferences, especially with regard to causal processes. This is harder for people without basic epidemiological know-how—which is, of course, one of the reasons why it is so important for all health workers to have some training in epidemiology. There are no simple shortcuts. Reliance on the reputation of the researchers, the sponsoring agency, or the journal in which the results are published can be misleading. Nor is it enough to know what techniques were used, without considering the details of their use. A large sample (though generally better than a small one) does not guarantee accurate results. Strict random sampling is a positive feature, but a so-called random sample chosen without using random numbers (or an equivalent method) can be a negative feature. The use of controls may be laudable; but badly chosen ones can be unhelpful or misleading. Matching is often helpful; but unnecessary matching may obscure associations. Statistical tests are usually a good thing; but they may mar a study if they are not called for or are misinterpreted. Confidence intervals are usually a plus; but they can be misleading if there is bias or confounding.

However certain we are that a study is valid, however, it is unwise to rely on it if it stands alone. Different studies of the same topic often produce different information, as a result of chance variation, differences in the methods or circumstances of the studies, or differences between study populations.

In answer to *Question G2–1,* then, it would be ill-advised to act on the study's findings unless they replicate those of previous studies or are confirmed by subsequent studies. It is of interest that this study found associations in the opposite direction for cancers of the pancreas (odds ratio, 1.49) and prostate (1.33); "these increased risks," say the authors, "could be due to chance or to undetected biases and warrant further investigation." But this also applies to the decreased risks for cancers of the esophagus, stomach, colon, and rectum.

Finding other studies in order to obtain a fuller picture of what is known is not always easy. If meta-analyses have been done, they are thus particularly useful. But it is as important to appraise the validity of a meta-analysis as it is to appraise the validity of a single study. The evaluation cited in *Question G2–2* found that no fewer than 80% of the meta-analyses and reviews that were appraised had serious flaws.

In answer to *Question G2–3,* the one variable that was strongly associated with the direction of the conclusions reached by the review articles on the effects of passive smoking was ("and the winner is . . .") affiliation to the tobacco industry. Almost all (94%) of the reviews whose authors were funded by or associated with the tobacco industry reported that passive smoking was not harmful, as compared with 13% of other reviews. The odds ratio expressing this association was 88 (95% confidence interval, 16 to 476; $P < .001$). The meta-meta-analysis of asthma treatment (*Question G2–2*) included six reviews that were funded by industry, and five of these yielded conclusions favoring the interventions related to the sponsoring companies. The moral is obvious, and it

should be applied to single studies as well as meta-analyses. Investigators are not necessarily fraudulent, but they may perform lesser misdemeanors: "Inventing data would clearly be wrong; suppression of inconvenient results would be less than honest. Yet they need not think too badly of themselves if they gloss over the study's methodological shortcomings, optimise the statistical analysis, cite published work selectively . . ." (Lancet, 1995). Look for the "funding" and "competing interests" statements that some journals append to papers.

Exercise G3

Question G3–1

A case-control study in the Punjab (India) revealed that circumcision in the neonatal period was associated with an increased risk of the subsequent onset of neonatal tetanus (an endemic disease in this area). The odds ratio was 3.1. The odds ratio was not raised (1.1) if antimicrobial agents (usually antibiotics, sometimes antiseptics) were applied to the wound, and it was especially high (4.2) if these substances were not applied (cow-dung was one of the substances commonly used). The estimated proportion of neonatal tetanus in boys in the study area that was attributable to circumcision was 24% (Bennett et al., 1999). Do you think that early circumcision should be discouraged? Should the routine application of antimicrobial agents to circumcision wounds be advocated?

Question G3–2

You wish to use a screening test for diabetes, which a large study has shown to be positive in 75% of diabetics. Can you assume that the test will have a sensitivity of 75%? What would you need to know to calculate the predictive value of a positive test?

Question G3–3

In deciding how to treat a patient, a clinician wishes to use the findings of a clinical trial that has demonstrated that a treatment is effective and safe. However, the criteria for including and excluding cases from this trial were such that this particular patient would not have been included in the trial. Is use of the findings justified?

Question G3–4

A clinician finds an up-to-date meta-analysis that shows that a particular treatment is effective and safe. The studies that were reviewed dealt with patients who differed in age, sex, and the severity of the disease, but the findings were not grossly heterogeneous. When considering application of the results in the care of a specific patient, should the clinician utilize the overall summary findings, or the findings in a specific study or subgroup where the patients' characteristics resemble those of the patient under care?

■■■■■■■■■■ Unit G4

Relevance of the Findings

However valid the epidemiological findings may be, their practical application in health care can be helpful only if the findings are generalizable to the specific individual, group, or community in which we are interested, and if the topic is relevant to a health problem of this individual, group, or community; that is, if it relates to a real or potential problem that is important enough (taking account of competing problems) to warrant action. In community health care, the latter judgment may be based on impressions or (preferably) on an epidemiological appraisal (needs assessment, community diagnosis).

The results cited in *Question G3–1* certainly justify both the discouragement of circumcision and the use of antimicrobial agents in the Punjab study area. (If these objectives are hard to achieve, routine active immunization of expectant mothers may be advocated, to permit the transfer of antibodies to their unborn babies.) But the findings have no relevance in populations where neonatal tetanus is rare or circumcisions are uncommon. In other populations where neonatal tetanus and circumcisions are common, the importance of these findings will depend on (among other things) the way in which circumcision wounds are treated there, and on the relative importance of circumcision and umbilical wounds as sources of tetanus infection in that population.

The sensitivity of a screening test (*Question G3–2*) may vary in different populations, and the sensitivity of the test for diabetes cited in this question has been stated to range from 21–75% (U.S. Preventive Services Task Force, undated). A reported sensitivity is not necessarily applicable in a population other than that in which it was determined. To calculate the predictive value of a positive test, we would need to know (or assume that we know) sensitivity, specificity, and the prevalence of diabetes in the group or population in which the test is to be used (for the formula, see Note C10). If there is doubt about sensitivity, specificity, or prevalence, the effect of different assumptions can be tested (this is a sensitivity analysis: see Unit F7).

With respect to a clinical trial (*Question G3–3*), the following advice has been offered to clinicians: "Rather than slavishly asking: 'Would my patient satisfy the eligibility criteria for the trial?' and rejecting its usefulness if they didn't exactly fit every one of them, we'd suggest bringing in some of your knowledge of human biology and clinical experience, turning the question around and asking: 'Is my patient so different from those in the trial that its results cannot help me make my treatment decision?'" (Sackett et al., 1997).

In answer to *Question G3–4,* opinions on the relative value of a broad meta-analysis or a single trial or subset differ. On the one hand, "When treating Ms Jones, the clinician may want to focus on the single trial or subset of trials conducted in patients most like Ms Jones" (Goodman, 1991), as the summary measure of effect may be "only a rough answer to a rough question about the average effectiveness . . . for a broad class of patients" (Simon, 1991). On the other

hand, as samples get smaller the random error gets bigger, and use of the broad picture may therefore be preferable, even if this provides an apparently less specific indication of the results to be expected in a particular patient. One expert stated, "Knowing the pitfalls, or the variations and the errors that there are—the random noises—even if in a particular subset the treatment did not seem to be particularly beneficial, and I've got a patient who belongs to that subset, but on average the benefit was 25%, I would say I would use the average figure rather than what I saw in that particular subset" (Yusuf, 1987b). "An overview allows a look at the forest through the trees," in the words of Furberg and Morgan (1987).

Exercise G4

Question G4–1

In a randomized controlled trial in Australia, the "Prevent-a-Bite" program, which aims to instill precautionary behavior among children around dogs in order to reduce the incidence of bites, produced striking results. Children aged 7–8 were given a half-hour lesson by a dog handler. After 7–10 days, a dog was tethered in the school grounds; only 9% of the children in experimental schools patted the dog, and did so carefully, whereas 79% of children in control schools patted the dog without hesitation ($P < .0001$) (Chapman et al., 2000). What extra information would help you to decide whether this program should be instituted in some other community where dog bites are a major cause of injury to children?

Question G4–2

Which of the following statements provides the most forceful argument for routine screening for cervical cancer, using Pap smears (U.S. Preventive Services Task Force, undated)? Assume that the statements are correct, although some have reservations that are not mentioned.

1. The sensitivity of Pap smears for the detection of cancer and dysplasia is 55–80%.
2. Their specificity is 90–99%.
3. Pap tests at 3-year intervals reduce the cumulative incidence of invasive cervical cancer by 91%.
4. Case-control studies (comparing women with and without cervical cancer) have shown a strong negative association between the disease and a history of screening.
5. Cervical cancer screening programs reduce cervical mortality rates by 20–60%.

Question G4–3

Should decisions on the use of a new treatment or preventive procedure be based on the risk ratios or the risk differences observed in controlled trials?

Question G4–4

A case-control study in southern Brazil, where incidence rates for cancers of the mouth, pharynx, and larynx are among the highest in the world, showed an odds ratio (controlling for numerous confounders) of 2.45 (95% confidence interval, 1.9 to 3.3) for the association with use of a wood stove. This led to the conclusion that approximately 42% of the incidence of these cancers in this region is attributable to wood stoves (Pintos et al., 1998). How would this attributable risk influence a decision on whether to establish a program aimed at reducing the use of wood stoves in some other population? Assume that the odds ratio is about the same (2.45) in this other population.

Question G4–5

In northern Italy, a large case-control study of thyroid cancer demonstrated a strong and significant association with poor diet, defined as a high intake of refined cereals and a low intake of vegetables and fruit. The odds ratio was 81 in men and 33 in women, after controlling for age, education, a history of benign thyroid disease, radiotherapy, and residence in endemic goitrous areas. The attributable fraction in the population was 41%. The investigators concluded that "intervention is likely to be relevant on a public health scale . . . ; some modification in the diet [is] likely to avoid [about 300] deaths per year in Italy (Fioretti et al., 1999). On the assumption that the association is causal, would you expect similar effects in another country? Would you call this attributable fraction a preventable fraction?

■■■■■■■■■■ Unit G5

Expected Effects

So far we have considered the importance of accurate knowledge of the findings, their validity, and their relevance. We must also take account of the effects (harmful as well as beneficial) to be expected if the epidemiologic findings are applied in practice.

Long-term effects are generally more important than short-term ones. A decision on whether to institute the "Prevent-a-Bite" program (*Question G4–1*), for example, would obviously be easier if we knew whether the change in behavior persists (the investigators suggest that "booster" interventions may be needed) and, more important, whether children exposed to the program sustain fewer dog bites in the long run.

If the issue is a decision on the introduction of a screening program, an effect on the population's health (as in statements 3 and 5 of *Question G4–2*) is more

important than success in identifying previously unknown cases or in bringing them under treatment. Statement 5 obviously provides the most cogent argument for screening. (This statement is actually based only on the observation that mortality declined in a number of countries after the implementation of screening programs; for ethical reasons, there have been no controlled trials.)

Similarly, if a decision has to be made on whether to identify and give special care to persons at high risk, information on the capacity to identify such people is less important than information on the consequent effects on health status.

In answer to *Question G4–3,* both the risk ratios and the risk differences observed in controlled trials can be useful guides in decision making, but the risk difference is generally more helpful. For the clinician concerned with individual patients, the difference (the *absolute risk reduction*) summarizes the procedure's expected effect on the patient's risk of death, disease, complications, side effects, etc. Some clinicians like to express the expected reduction as a percentage of the patient's initial risk; this *relative risk reduction* is, of course, the same as the preventable fraction in those exposed to a protective factor. For the decision-maker interested in the introduction of the treatment on a large scale, the rate difference can provide an estimate of the number of people in a population of a given size who are likely to remain alive or well, recover from illness, and so forth, because of the procedure. If the difference in annual incidence is 1 per 1,000 when people exposed and not exposed to a preventive factor are compared, the expected number of cases avoided in a year, in a population of 200,000, is 200.

All the measures of impact described in Unit E11 (attributable, prevented, and preventable fractions) may be helpful guides. If a risk factor is modifiable, the attributable fraction in the population (the fraction of the incidence or mortality that is attributable to exposure to the factor) is also the preventable fraction, and may be an important consideration when deciding whether to institute a program. Because the fraction is influenced by the prevalence of the exposure in the population (see Note E12), the attributable fraction in one population (*Question G4–4*) is not necessarily valid in another.

The attributable fractions in two populations may also differ because the causal association differs in its strength, as a result of differences in the prevalence of factors that modify the effect of the causal factor, or for other reasons. The possibility of differences between populations in the odds ratio expressing the association between poor diet and thyroid cancer (*Question G4–5*) is supported by the large difference between the odds ratios observed in men and women. The prevalence of "poor diet" will also vary, so that the attributable fraction in another population is hard to predict.

In this context it is difficult to refer to a preventable fraction, as the achievement of appreciable changes in a population's diet is far from easy. In fact, in all instances the estimation of an expected effect should be tempered by a realization that practical constraints may prevent the full realization of projected benefits.

Exercise G5

Question G5–1

If the sensitivity of a screening test is 90% and its specificity is 80%, how many screening tests must be performed, and how many people with positive results must be subjected to more intensive investigation, to identify one case? Since the answers will obviously depend on the prevalence of the disease in the population, assume that this is 1%. Clue: Construct a table like Table C10–1.

Question G5–2

In a large randomized study that showed the effectiveness of screening for colorectal cancer in people aged 54–75 (fecal occult-blood tests every 2 years), the risk of dying of colorectal cancer during a 10-year follow-up was lower by 1.42 per 1,000 in the screened group than in the control group (Kronborg et al., 1996). How many people have to undergo screening to avoid one death from colorectal cancer during a 10-year period? (The method of calculation was explained in Note E6–2.)

Question G5–3

If a randomized control trial shows that the rate of the desired result is higher by 4 per 100 in the treatment group, how many people need to be treated to produce one desired effect? (Parenthetically: If the rate of an adverse effect is higher by 4 per 100 in the treatment group, how many people need to be treated to produce one harmful effect?)

Question G5–4

This is the last question in this book ("O frabjous day! Callooh! Callay! He chortled in his joy"; Carroll, 1872). If the rate of a given disease is higher by 3.3 per 1,000 in smokers than in nonsmokers, and it is assumed that this difference is attributable to smoking, how many people must become nonsmokers to prevent one case?

■■■■■■■■■■ Unit G6

Feasibility and Cost

We have thus far considered the importance of accurate knowledge of the findings, their validity, their relevance, and the expected effects of the application. The missing element, and an essential one, is an appraisal of feasibility and cost.

It is always necessary to ask such questions as whether the treatment or intervention under consideration is likely to be acceptable to the patient or public, whether the required trained and interested personnel, facilities, money, and other resources will be available, and whether the costs are justified by the likely effects on health (cost-effectiveness) or by the likely economic benefits (cost-benefit analysis).

Exercise G5 relates to one limited aspect of the appraisal of feasibility and cost—namely estimating the number of people who will have to undergo the contemplated procedure, alter their lifestyles, etc. This number can be helpful in the appraisal of a program's costs in terms of manpower, time, effort, and anxiety, as well as money.

In answer to *Question G5–1*, it is obvious from Table G6, which was constructed to meet the specified requirements (sensitivity 90%, specificity 80%, prevalence 1%), that 1,000 screening tests will identify nine cases. The number of tests required to identify one case is therefore, 1,000/9, or 111. These 1,000 tests will yield 207 positive results, and the number of more intensive examinations required to identify one case is therefore 207/9, or 23. Depending on the size of the population and the cost of a screening test and a more intensive examination, the total cost of he operation can be estimated.

The number of persons who need to undergo screening to avoid one death from colorectal cancer (*Question G5–2*) is 1/0.00142, or 704.

Similarly in *Question G5–3*, the number needed in the treatment group to produce one desired event (e.g., to avoid one death) is 1/0.04, or 25. Because the data are derived from a treatment trial, this would be referred to as the "number needed to treat" (*NNT*). If the rate of a harmful effect is higher by 4 per 100 in the treatment group, the number required in the control group to avoid one such effect is also 25. Equivalently, the number needed in the treatment group to harm one patient is 25; this can be called the "number needed to treat to harm one patient" (*NNTH*).

In *Question G5–4*, the number who need to become nonsmokers to prevent one case is 1/0.0033, or 303.

Note that if the rate difference is based on person-time, these "numbers needed" must also relate to person-time. A rate difference of 4 per 100 person-years would indicate that 25 person-years of treatment are required to prevent one

Table G6. Expected Results of 1,000 Screening Tests: (Sensitivity 90%, Specificity 80%, Prevalence 1%)

	Disease		
Test Result	Absent	Present	Total
Positive	198	9	207
Negative	792	1	793
Total	990	10	1,000

case, or (more simply) that 25 people must be treated for 1 year to avoid one case.

■■■■■■■■■■ Unit G7

Test Yourself (G)

1. Find a report of a recent study showing the effect of a health care procedure or program; then decide whether the findings should be applied in practice, either in clinical care or in the health care of a specific community in which you are interested.
2. Ask yourself whether, in making the above decision, you took due account of
   - the accuracy with which you knew the findings (G2).
   - the validity of the findings (G3).
   - the relevance of the findings (G4).
   - the expected effects (G5).
   - feasibility and cost (G6).

> "Would you tell me, please, which way I ought to go from here?"
> "That depends a good deal on where you want to get to," said the Cat.
> "I don't much care where—" said Alice.
> "Then it doesn't matter which way you go," said the Cat.
>
> *(Carroll, 1865)*

References

Polonius: What do you read, my lord?
Hamlet: Words, words, words.

(Shakespeare, 1603)

Abramson JH. Meta-analysis: a review of pros and cons. *Public Health Reviews* 18: 1–47, 1990/91.

Abramson JH. Community-oriented primary care—strategy, approaches and practice: a review. *Public Health Reviews* 16: 35–98, 1988.

Abramson JH. Age-standardization in epidemiological data. *International Journal of Epidemiology* 24: 238–239, 1995.

Abramson JH, Abramson ZH. *Survey methods in community medicine: epidemiological research, programme evaluation, clinical trials,* 5th edn. Edinburgh: Churchill Livingstone, 1999.

Abramson, JH, Gahlinger PM. *Computer programs for epidemiologists: PEPI version 4.0.* Salt Lake City: Sagebrush Press, 2001.

Abramson JH, Hopp C. The control of cardiovascular risk factors in the elderly. *Preventive Medicine* 5: 32–47, 1976.

Abramson JH, Hopp C, Epstein LM. The epidemiology of varicose veins: a survey in western Jerusalem. *Journal of Epidemiology and Community Health* 35: 213–217, 1981.

Abramson JH, Kark SL, Palti H. The epidemiological basis for community-oriented primary care. In: *Epidemiologie et médecine communautaire* (Lellouche J, ed). Paris: INSERM, 1983, pp. 231–263.

Abramson JH, Sacks MI, Cahana E. Death certificate data as an indication of the presence of certain common diseases at death. *Journal of Chronic Diseases* 24: 417–431, 1971.

Adams PF, Hendershot GE, Marano MA. Current estimates from the National Health Interview Survey, 1996. *Vital and Health Statistics* 10(200), Hyattsville, MD: National Center for Health Statistics, 1999.

Altman DG. *Practical statistics for medical research.* London: Chapman & Hall, 1991.

Altman DG. Clinical trials and meta-analyses. In: *Statistics with confidence,* 2nd edn. (Altman DG, Machin D, Bryant TN, Gardner MJ, eds). BMJ Books, 2000, pp. 120–138.

Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence.* 2nd edn. BMJ Books, 2000.

Amery A, Birkenhager W, Brixko P, Bulpitt C, Clement D, Deruyttere M, de Schaepdryver A, Dollery C, Fagard R, Forette F, Forte J, Hamdy R, Henry JF, Joossens JV, Leonetti G, Lund-Johansen P, O'Malley K, Petrie J, Strasser T, Tuomilehto J, Williams B. Mortality and morbidity results from the European Working Party in High Blood Pressure in the Elderly trial. *Lancet* 1: 1349–1354, 1985.

Andersen JW, Harrington D. Meta-analyses need new publication standards. *Journal of Clinical Oncology* 10: 878–880, 1992.

Anderson S, Auquier WW, Hauck WW, Oakes D, Vandaele W, Weissberg HI. *Statistical methods for comparative studies.* New York: Wiley, 1980.

Andrews G, MacMahon SW, Austin A, Byrne DG. Hypertension: comparison of drug and non-drug treatments. *British Medical Journal* 284: 1523–1526, 1982.

Armitage P, Berry G. *Statistical methods in medical research*, 3rd edn. Oxford: Blackwell, 1994.

Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? *Statistics in Medicine* 6: 351–358, 1987.

Ban R, Peritz E. *Longitudinal study of borderline hypertension.* Paper presented at International Symposium on Hypertension Control in the Community, Tel Aviv, Nov. 1982.

Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. *Journal of the American Medical Association* 279: 1566–1570, 1998.

Bartko JJ. Measures of agreement: a single procedure. *Statistics in Medicine* 13: 737–745, 1994.

Begg CB, Pilote L, McGlave PB. Bone marrow transplantation versus chemotherapy in acute non-lymphocytic leukemia: a meta-analytic review. *European Journal of Cancer and Clinical Oncology* 25: 1519–1523, 1989.

Bennett J, Breen C, Traverso H, Agha SB, Macia J, Boring J. Circumcision and neonatal tetanus: disclosure of risk and its reduction by topical antibiotics. *International Journal of Epidemiology* 28: 263–266, 1999.

Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. *American Journal of Epidemiology* 132: 612–628, 1990.

Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine* 8: 141–151, 1989.

Berry G, Armitage P. Mid-P confidence intervals: a brief review. *The Statistician* 44: 417–423, 1995.

Boden WE. Meta-analysis in clinical trial reporting: has a tool become a weapon? *American Journal of Cardiology* 69: 681–686, 1992.

Boyce WJ, Vessey MP. Rising incidence of fracture of the proximal femur. *Lancet* 1: 150–151, 1985.

Breslow NE, Day NE. *Statistical methods in cancer research, Vol. 1: The analysis of case-control studies.* Lyon: International Agency for Research on Cancer, 1980.

Breslow NE, Day NE. Statistical methods in cancer research, Vol. II: *The design and analysis of cohort studies.* Lyon: International Agency for Research on Cancer, 1987.

Brett M, Barker DJP. The world distribution of gallstones. *International Journal of Epidemiology* 5: 335–341, 1976.

Bross IDJ. Spurious effects from an extraneous variable. *Journal of Chronic Diseases* 19: 637–647, 1966.

Bross IDJ. Pertinency of an extraneous variable. *Journal of Chronic Diseases* 20: 487–495, 1967.

Brownson RC, Petitti DB (eds). *Applied epidemiology: theory to practice.* New York: Oxford University Press.

Brownson RC, Novotny TE, Perry MC. Cigarette smoking and adult leukemia: a meta-analysis. *Archives of Internal Medicine* 153: 469–475, 1993.

Brunekreef B, Fischer P, Remijn B, van der Lende R, Schouten J, Quanjer P. Indoor air pollution and its effect on pulmonary function of adult non-smoking women: III. Passive smoking and pulmonary function. *International Journal of Epidemiology* 14: 227–230, 1985.

Bulpitt CJ, Beilin LJ, Clifton P, Coles EC, Dollery CT, Gear JSS, Harper GS, Johnson BF, Munro-Faure AD. Risk factors for death in treated hypertensive patients: report from the D.H.S.S. Hypertension Care Computing Project. *Lancet* 2: 134–137, 1979.

Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46: 423–429, 1993.

Campbell DT. Factors relevant to the validity of experiments in social settings. In: *Program evaluation in the health fields* (Schulberg HC, Sheldon A, Baker F, eds). New York: Behavioral Publications, 1969, pp. 165–185.

Campbell DT, Stanley JC. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Canner PL. An overview of six clinical trials of aspirin in coronary heart disease. *Statistics in Medicine* 6: 255–263, 1987.

Carroll L. *Alice's adventures in Wonderland.* 1865.

Carroll L. *Through the looking glass (and what Alice found there).* 1872.

Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, Tonascia S, Chalmers TC. A cohort study of summary reports of clinical trials. *JAMA* 263: 1401–1405, 1990.

Chalmers I, Altman DG (eds). *Systematic reviews.* London: BMJ Publishing Group, 1995.

Chalmers I, Haynes B. Reporting, updating and correcting systematic reviews of the effects of health care. In: *Systematic reviews* (Chalmers I, Altman DG, eds). London: BMJ Publishing Group, 1995, pp. 86–95.

Chalmers TC. Problems induced by meta-analyses. *Statistics in Medicine* 10: 971–980, 1991.

Chalmers TC, Levin H, Sacks HR, Reitman D, Berrier J, Ngalingam R. Meta-analysis of clinical trials as a scientific discipline: I. Control of bias and comparison with large cooperative trials. *Statistics in Medicine* 6: 315–325, 1987.

Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized clinical trial. *Controlled Clinical Trials* 2: 31–49, 1981.

Chapman S, Cornwall J, Righetti J, Sung L. Preventing dog bites in children: randomised controlled trial of an educational intervention. *British Medical Journal* 320: 1512–1513, 2000.

Choi BCK, de Guia NA, Walsh P. Look before you leap: stratify before you standardize. *American Journal of Epidemiology* 149: 1087–1096, 1999.

Choi BCK, Pak AWP. Bias, overview. In: *Encyclopedia of biostatistics,* Vol. 1 (Armitage P, Colton T, eds). Chichester: Wiley, 1998, pp. 331–338.

Clark VA, Aneshensel CS, Frerichs RR, Morgan TM. Analysis of non-response in a prospective study of depression in Los Angeles County. *International Journal of Epidemiology* 12: 193–198, 1983.

Cochran WG. *Planning and analysis of observational studies.* New York: Wiley, 1983, p. 20.

Cochrane AL, St Leger AS, Moore F. Health service 'input' and mortality 'output' in developed countries. *Journal of Epidemiology and Community Health* 32: 200–205, 1978.

Cole TJ. The influence of heights on the decline in ventilatory function. *International Journal of Epidemiology* 3: 145–152, 1974.

Collins R, Gray R, Godwin J, Peto R. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. *Statistics in Medicine* 6: 245–250, 1987.

Collins R, Peto R, MacMahon S, Herbert P, Fiebach NH, Eberlein KA, Godwin J, Qizilbash N, Taylor JO, Hennekens CH. Blood pressure, stroke, and coronary heart disease: Part 2, short-term reductions in blood pressure: overview of randomized drug trials in their epidemiological context. *Lancet* 335: 827–838, 1990.

Connell FA, Koepsell TD. Measures of gain in certainty from a diagnostic test. *American Journal of Epidemiology* 121: 744–753, 1985.

Connor E, Mullan F (eds). Community-oriented primary care: new directions for health service delivery. Washington, D.C.: National Academy Press, 1983.

Cook TD, Campbell DT. *Quasi-experimentation: design and analysis issues for field settings.* Chicago: Rand McNally, 1979.

Copas JB, Shi JQ. Reanalysis of epidemiological evidence on lung cancer and passive smoking. *British Medical Journal* 320: 417–418, 2000.

Coreil J, Augustin A, Holt E, Halsey NA. Use of ethnographic research for instrument development in a case-control study of immunization use in Haiti. *International Journal of Epidemiology* 18: S33–S37, 1989.

Cox DR, Oakes D. *Analysis of survival data.* London: Chapman & Hall, 1984.

Dales LD, Ury HK. An improper use of statistical significance testing in studying covariables. *International Journal of Epidemiology* 4: 373–375, 1978.

Daniel WW. *Biostatistics: a foundation for analysis in the health sciences,* 6th edn. New York: Wiley, 1995.

Davies TW, Williams DR, Whitaker RH. Risk factors for undescended testis. *International Journal of Epidemiology* 4: 169–170, 1986.

DeMets DL. Methods for combining randomized clinical trials: strengths and limitations. *Statistics in Medicine* 6: 341–348, 1987.

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont. Clin-

ical epidemiology rounds. *Canadian Medical Association Journal* 129: 429–432, 559–564, 705–710, 832–835, 947–954, 1093–1099, 1983.

Depue RH. Maternal and gestational factors affecting the risk of cryptorchidism and inguinal hernia. *International Journal of Epidemiology* 13: 311–318, 1984.

DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7: 177–188, 1986.

De Semir V. What is newsworthy? *Lancet* 347: 1163–1166, 1996.

Detels R. Epidemiology: the foundation of public health. In: *Oxford textbook of public health,* 3rd edn., Vol. 2: *The methods of public health* (Detels R, Holland WW, McEwen J, Omenn GS, eds). Oxford: Oxford University Press, 1997, pp. 501–506.

Detels R, McEwan J, Beaglehole R, Tanaka H (eds). *Oxford textbook of public health,* 4th edn., Vol. 2: *The methods of public health.* Oxford: Oxford University Press, 2001.

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology* 45: 255–265, 1992.

Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Comparison of MEDLINE searching with a perinatal clinical trials database. *Controlled Clinical Trials* 6: 306–317, 1985.

Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. In: *Systematic reviews* (Chalmers I, Altman DG, eds). London: BMJ Publishing Group, 1995, pp. 17–36.

Ducimitiere P, Richard JL, Pequignot GP, Warnet JM. Varicose veins: a risk factor for atherosclerotic disease in middle-aged men? *International Journal of Epidemiology* 10: 329–335, 1981.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 337: 867–872, 1991.

Ellenberg JH, Nelson KB. Sample selection and the natural history of disease: studies of febrile seizures. *JAMA* 243: 1337–1340, 1980.

Elwood JM, Lee JAH, Walter SD, Mo T, Green AES. Relationship of melanoma and other skin cancer mortality to latitude and ultraviolet radiation in the United States and Canada. *International Journal of Epidemiology* 3: 325–332, 1974.

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* 19: 1707–1728, 2000.

Epstein LM. Validity of a questionnaire for diagnosis of peptic ulcer in an ethnically heterogeneous population. *Journal of Chronic Diseases* 22: 49–55, 1969.

Evans AS, Wells AV, Ramsay F, Drabkin P, Palmer K. Poliomyelitis, rubella and dengue antibody survey in Barbados. A follow-up study. *International Journal of Epidemiology* 8: 235–241, 1979.

Eysenck HJ. An exercise in mega-silliness. *American Psychologist* 35: 517, 1978.

Eysenck HJ. Problems with meta-analysis. In: *Systematic reviews* (Chalmers I, Altman DG, eds). London: BMJ Publishing Group, 1995, pp. 64–74.

Fawzi WW, Chalmers TC, Herrera MG, Mosteller F. Vitamin A supplementation and child mortality: a meta-analysis. *JAMA* 269: 898–903, 1993.

Feinstein AR. *Clinical biostatics.* St. Louis: Mosby, 1977, p. 91.

Feinstein AR. *Clinical epidemiology: the architecture of clinical research.* Philadelphia: W. B. Saunders, 1985.

Felson DT. Bias in meta-analytic research. *Journal of Clinical Epidemiology* 45: 885–892, 1992.

Fioretti F, Tavani A, Gallus S, Franceschi S, Negri E, La Vecchia C. Case control study of thyroid cancer in northern Italy: attributable risk. *International Journal of Epidemiology* 28: 626–630, 1999.

Fisher KJ, Glasgow RE, Terborg JR. Work site smoking cessation: a meta-analysis of long-term quit rates from controlled studies. *Journal of Occupational Medicine* 32: 429–439, 1990.

Fleiss JL. *Statistical methods for rates and proportions,* 2nd edn. New York: Wiley, 1981.

Fleiss JL. Significance tests have a role in epidemiologic research: reactions to A. M. Walker, *American Journal of Public Health* 76: 559–560, 1986a.

Fleiss JL. Dr. Fleiss responds. *American Journal of Public Health* 76: 1033–1034, 1986b.

Fleiss JL. *The design and analysis of clinical experiments.* New York: Wiley, 1986c.

Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the associa-

tion between exposure to environmental tobacco smoke and lung cancer: a critique. *Journal of Clinical Epidemiology* 44: 127–139, 1991.

Forman MR, Graubard BI, Hoffman HJ, Beren R, Harley EE, Bennett P. The Pima Infant Feeding Study: breastfeeding and respiratory infections during the first year of life. *International Journal of Epidemiology* 13: 447–453, 1984.

Frankel A, Gunnell DJ, Peters DJ, Maynard M, Smith GD. Childhood energy intake and adult mortality from cancer: the Boyd Orr cohort study. *British Medical Journal* 316: 499–504, 1998.

Friedman GD. *Primer of epidemiology,* 2nd edn. New York: McGraw-Hill, 1980.

Furberg CD, Morgan TM. Lessons from overviews of cardiovascular trials. *Statistics in Medicine* 6: 295–303, 1987.

Gershman K. Case-control study of which dogs bite (abstract). *American Journal of Epidemiology* 138: 593, 1992.

Gershman KA, Sacks JJ, Wright JC. Which dogs bite? A case-control study of risk factors. *Pediatrics* 93: 913–917, 1994.

Giagnoni E, Secchi MB, Wu SC, et al. Prognostic value of exercise EKG testing in asymptomatic normotensive individuals. *New England Journal of Medicine* 309: 1085–1089, 1983.

Gifford RH, Feinstein AR. A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction. *New England Journal of Medicine* 280: 351–357, 1969.

Gillam S, Miller R. *COPC—a public health experiment in primary care.* London: King's Fund, 1997.

Glass GV, McGaw B, Smith ML. Meta-analysis in social research. Beverly Hills, California: Sage Publications, 1981.

Glass GV, Smith ML. Reply to Eysenck. *American Psychologist* 35: 517–519, 1978.

Godfrey CM, Morgan P. A controlled trial of the theory of acupuncture in musculoskeletal pain. *Journal of Rheumatology* 5: 121–124, 1978.

Gofin J, Kark E, Mainemer N, Kark SL, Abramson JH, Hopp C, Epstein LM. Prevalence of selected health characteristics of women and comparisons with men: a community health survey in Jerusalem. *Israel Journal of Medical Sciences* 17: 145–149, 1981.

Goldbourt U, Kark JD. The epidemiology of coronary heart disease in the ethnically and culturally diverse population of Israel. *Israel Journal of Medical Sciences* 18: 1077–1097, 1982.

Goldstein R. Epidemiological software. In: *Encyclopedia of epidemiologic methods* (Gail MH, Benichou J, eds). New York: Wiley, 2000.

Goodman SN. Have you ever meta-analysis you didn't like? *Annals of Internal Medicine* 114: 244–246, 1991.

Goodwin PJ, Boyd NF. Mammographic parenchymal pattern and breast cancer risk: a critical appraisal of the evidence. *American Journal of Epidemiology* 127: 1097–1107, 1988.

Gray A, Berlin JA, McKinlay JB, Longcope C. An examination of research design effects on the association of testosterone and male aging: results of a meta-analysis. *Journal of Clinical Epidemiology* 44: 671–684, 1991.

Green, M. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. *American Journal of Epidemiology* 117: 98–105, 1983.

Greenhaigh T, Taylor R. How to read a paper: papers that go beyond numbers (qualitative research). *British Medical Journal* 315: 740, 1997.

Greenland S (ed). *Evolution of epidemiologic ideas: annotated readings on concepts and methods.* Chestnut Hill, Massachusetts: Epidemiology Resources, 1987.

Greenland S. Induction versus Popper: substance versus semantics. *International Journal of Epidemiology* 37: 543–548, 1998a.

Greenland S. Meta-analysis. In: *Modern epidemiology,* 2nd edn. (Rothman KJ, Greenland S). Philadelphia: Lippincott-Raven, 1998b, pp. 643–673.

Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Statistics in Medicine* 9: 247–252, 1990.

Gupta PC, Bhonsle RB, Mehta FS, Pindborgh JJ. Mortality experience in relation to tobacco chewing and smoking habits from a 10-year follow-up study in Ernakulam District, Kerala. *International Journal of Epidemiology* 13: 184–187, 1984.

Hammond EC, Selikoff IJ, Seidman H. Asbestos exposure, cigarette smoking and death rates. *Annals of the New York Academy of Sciences* 330: 473–495, 1979.

Hedges LV. Commentary. *Statistics in Medicine* 6: 381–385, 1987.

Hedges LV, Olkin I. *Statistical methods for meta-analysis.* Orlando, Florida: Academic Press, 1985.

Heggenhaugen H, Pedersen D. Beyond quantitative measures: the relevance of anthropology for public health. In: *Oxford textbook of public health,* 3rd edn., Vol 2: *The methods of public health* (Detels R, Holland WW, McEwen J, Omenn GS, eds). New York, Oxford University Press, 1997, pp. 815–828.

Held PH, Yusuf F, Furburg CD. Calcium channel blockers in acute myocardial infarction and unstable angina: an overview. *British Medical Journal* 299: 1187–1192, 1989.

Hennekens CH, Buring JE. *Epidemiology in medicine.* Boston: Little, Brown, 1987.

Hill AB. *Statistical methods in clinical and preventive medicine.* Edinburgh: Livingstone, 1962.

Hill C, Benhamou E. Age-standardization in epidemiological data. *International Journal of Epidemiology* 24: 241–242, 1995.

Hoddinott P, Pill R. Qualitative study of decisions about infant feeding among women in east end of London. *British Medical Journal* 318: 30–34, 1999.

Holme I. Relation of coronary heart disease incidence and total mortality to plasma cholesterol reduction in randomised trials: use of meta-analysis. *British Heart Journal* 69 (Suppl.): S42–47, 1993.

I-Kuei Lin, L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45: 255–268, 1989.

International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine* 126: 36–47, 1997.

Iyengar S, Greenhouse JB. Selection models and the file-drawer problem. *Statistical Science* 3: 109–135, 1988.

Jacobsen R, Bostofte E, Engholm G, Hansen J, Olsen JH, Skakkebaek NE, Moller H. Risk of testicular cancer in men with abnormal semen characteristics: cohort study. *British Medical Journal* 321: 789–92, 2000.

Jadad AR, Moher M, Browman GP, Booker L, Sigouin C, Fuentes M, Stevens R. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *British Medical Journal* 320: 537–540, 2000.

Jenicek M. Meta-analyses in medicine: where we are and where we want to go. *Journal of Clinical Epidemiology* 42: 35–44, 1989.

Jones DR. Meta-analysis of observational epidemiological studies: a review. *Journal of the Royal Society of Medicine* 85: 165–169, 1992.

Kabat GC, Howson CP, Wynder EL. Beer consumption and rectal cancer. *International Journal of Epidemiology* 15: 494–501, 1986.

Kahn HA. The Dorn study of smoking and mortality among U.S. veterans: report on eight and one-half years of observation. In: *Epidemiological approaches to the study of cancer and other chronic diseases* (Haenszel W, ed). National Cancer Institute Monograph 19, 1966.

Kahn HA, Herman JB, Medalie JH, Neufeld HN, Riss E, Goldbourt U. Factors related to diabetes incidence: a multivariate analysis of two years' observation on 10,000 men. The Israel Ischemic Heart Disease Study. *Journal of Chronic Diseases* 23: 617–629, 1971.

Kahn HA, Sempos CT. *Statistical methods in epidemiology.* New York: Oxford University Press, 1989.

Kark S, Kark E. *Promoting community health: from Pholela to Jerusalem.* Johannesburg: Witwatersrand University Press, 1999.

Kark SL. *Epidemiology and community medicine.* New York: Appleton-Century-Crofts, 1974.

Kark SL. *The practice of community-oriented primary health care.* New York: Appleton-Century-Crofts, 1981.

Kark SL, Gofin J, Abramson JH, Makler A, Mainemer N, Kark E, Epstein LM, Hopp C. Prevalence of selected health characteristics of men: a community health survey in Jerusalem. *Israel Journal of Medical Sciences* 15: 732–741, 1979.

Kark SL, Kark E, Abramson JH, Gofin J (eds). *Atencion primaria orientada a la comunidad (APOC).* Barcelona: Ediciones Doyma S.A., 1994.

Kelsey JL, Thompson WD, Evans AS. *Methods in observational epidemiology.* New York: Oxford University Press, 1986.

Khaw K-T, Rose G. Population study of blood pressure and associated factors in St Lucia, West Indies. *International Journal of Epidemiology* 11: 372–377, 1982.

Klein S, Simes J, Blackburn GL. Total parenteral nutrition and cancer clinical trials. *Cancer* 58: 1378–1386, 1986.

Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods.* Belmont, California: Lifetime Learning Publications, 1982.

Kono S, Ikeda M, Tokudome S, Nishizumi M, Kuratsune M. Alcohol and mortality: a cohort study of male Japanese physicians. *International Journal of Epidemiology* 15: 527–532, 1986.

Kreiger N, Gross A, Hunter G. Dietary factors and fracture in postmenopausal women: a case-control study. *International Journal of Epidemiology* 21: 953–958, 1992.

Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomized study of screening for colorectal cancer with faecal-occult blood test. *Lancet* 348: 1467–1471, 1996.

Lancet. Editorial: Shall we nim a horse? *Lancet* 345: 1585–1586, 1995.

Landis JR, Sharp TJ, Kuritz SJ, Koch GG. Mantel-Haenszel methods. In: *Encyclopedia of epidemiologic methods* (Gail MH, Benichou J, eds). Chichester: Wiley, 2000, pp. 499–512.

Landman JT, Dawes RM. Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist* 37: 504–516, 1982.

Langman MJS, Cheng KK, Gilman EA, Lancashire RJ. Effect of anti-inflammatory drugs on overall risk of common cancer: case-control study in general practice research database. *British Medical Journal* 320: 1642–1645, 2000.

Last JM (ed). *A dictionary of epidemiology,* 4th edn. New York: Oxford University Press, 2001.

Last JM, Foege WH. International health. In: *Maxcy-Rosenau public health and preventive medicine* (Last JM, ed), 12th edn. Norwalk, Connecticut: Appleton-Century-Crofts, 1986.

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analyses of therapeutic trials for myocardial infarction. *New England Journal of Medicine* 327: 248–254, 1992.

Lee J. An insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence. *International Journal of Epidemiology* 15: 22–29, 1986.

Lehtonen A, Luutnen S. Serum lipids of very old people without dementia or with different types of senile dementia. In: *8th Scandinavian Congress of Gerontology: congress proceedings.* Tampere: Societas Gerontologica Fennica, 1986, pp. 489–491.

Lellouch J, Rokotovao R. Estimation of risk as a function of risk factors. *International Journal of Epidemiology* 5: 349–352, 1976.

Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* 115: 92–106, 1982.

Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized clinical trials of primary treatment of breast cancer. *Journal of Clinical Oncology* 4: 942–951, 1986.

Lichtenstein MJ, Mulrow CD, Elwood PC. Guidelines for reading case-control studies. *Journal of Chronic Diseases* 40: 893–903, 1987.

Light RJ. Accumulating evidence from independent studies: what we can win and what we can lose. *Statistics in Medicine* 6: 221–228, 1987.

Lilienfeld AM, Lilienfeld DE. *Foundations of epidemiology,* 2nd edn. New York: Oxford University Press, 1980.

Lindquist C. Risk factors in lip cancer. *American Journal of Epidemiology* 109: 521–530, 1979.

Lipscomb JA, Satin KP, Neutra RR. Reported symptom prevalence rates from comparison populations in community-based environmental studies. *Archives of Environmental Health* 47: 263–269, 1992.

Loevinsohn BP. Health education interventions in developing countries: a methodological review of published articles. *International Journal of Epidemiology* 19: 788–794, 1990.

MacMahon B, Pugh TF, Ipsen J. *Epidemiologic methods.* Boston: Little, Brown, 1960.

Maffei FHA, Magaldi C, Pinho SZ, Lastoria S, Pinho W, Yoshida WB, Rollo HA. Varicose veins and

chronic venous insufficiency in Brazil: prevalence among 1755 inhabitants of a country town. *International Journal of Epidemiology* 15: 210–217, 1986.

McMichael AJ, Hetzel BS. An epidemiological study of the mental health of Australian university students. *International Journal of Epidemiology* 3: 125–134, 1974.

McNeil D. *Epidemiological research methods.* New York: Wiley, 1996.

Mainland D. *Elementary medical statistics,* 2nd edn. Philadelphia: W. B. Saunders, 1964.

Mazzuca SA. Does patient education in chronic disease have therapeutic value? *Journal of Chronic Diseases* 35: 521–529, 1983.

Meade TW, Brennan PJ. Determination of who may derive most benefit from aspirin in primary prevention: subgroup results from a randomized controlled trial. *British Medical Journal* 321: 13–17, 2000.

Meijer WS, Schmitz PIM, Jeekel J. Meta-analysis of randomized, controlled clinical trials of antibiotic prophylaxis in biliary tract surgery. *British Journal of Surgery* 77: 283–290, 1990.

Miettinen OS, Cook EF. Confounding: essence and detection. *American Journal of Epidemiology* 114: 593–603, 1991.

Mittlboeck M, Schemper M. Explained variation for logistic regression. *Statistics in Medicine* 14: 1987–1997, 1996.

Morris JN. *Uses of epidemiology,* 3rd edn. Edinburgh: Churchill Livingstone, 1975.

Mulder PGH, Garretsen HFL. Are epidemiological and sociological surveys a proper instrument for detecting true problem drinkers? (The low sensitivity of an alcohol survey in Rotterdam.) *International Journal of Epidemiology* 12: 442–444, 1983.

National Center for Health Statistics. *Health, United States, 1999.* Hyattsville, Maryland: U.S. Department of Health and Human Services, 1999.

National Center for Health Statistics. *Health, United States, 2000.* Hyattsville, Maryland: U.S. Department of Health and Human Services, 2000.

Naylor CD. Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *Canadian Medical Association Journal* 138: 891–895, 1988.

Niswander K, Henson G, Elbourne D, Chalmers I, Redman C, MacFarlane A, Tizard P. Adverse outcome of pregnancy and the quality of obstetric care. *Lancet* 2: 827–831, 1984.

Nutting PA (ed). Community-oriented primary care: from principle to practice. Washington, D.C.: U.S. Department of Health and Human Services. HRSA Publication No. HRS-A-PE 86–1, 1987.

Nyyssonen K, Parviainen MT, Salonen R, Tuomilehto J, Salonen JT. Vitamin C deficiency and risk of myocardial infarction: prospective population study of men from eastern Finland. *British Medical Journal* 314: 634–638, 1997.

O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *Journal of Clinical Epidemiology* 42: 1021–1026, 1989.

Orwin RG. A fail-safe N for effect size. *Journal of Educational Statistics* 8: 157–159, 1983.

Ostfeld AM, Shekelle RB, Klawans H, et al. Epidemiology of stroke in an elderly welfare population. *American Journal of Public Health* 64: 450–458, 1974.

Owen R. Reader bias. *Journal of the American Medical Association* 247: 2533–2534, 1982.

Oxman AD, Guyatt GH. The science of reviewing research. *Annals of the New York Academy of Sciences* 703: 125–131, 1993.

Palti H. Use of control groups in evaluating the effectiveness of community health programs in primary care. *Israel Journal of Medical Sciences* 19: 756–759, 1983.

Palti H, Adler B, Tepper D. An early infant stimulation program in the maternal and child health service—evaluation at 5 years of age—preliminary findings. In: *Stimulation and intervention in infant development* (Tamir D, ed). London: Freund, 1986, pp. 195–201.

Parkkari J, Kannus P, Niemi S, Koskinen S, Palvanen M, Vuori I, Jarvinen M. Childhood deaths and injuries in Finland in 1971–1995. *International Journal of Epidemiology* 29: 516–523, 2000.

Pearson K. *The grammar of science.* London: Scott, 1892.

Petitti DB. *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine.* New York: Oxford University Press, 1994.

Peto R. Why do we need systematic overviews of randomized trials? *Statistics in Medicine* 6: 233–240, 1987a.

Peto R. Discussion. *Statistics in Medicine* 6: 241–244, 1987b.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. *British Journal of Cancer* 34: 585–612, 1976.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples. *British Journal of Cancer* 35: 1–39, 1977.

Philipp R, Evans EJ, Hughes AO, Grisdale SK, Enticott RG, Jephcott AE. Health risks of snorkel swimming in untreated water. *International Journal of Epidemiology* 14: 624–627, 1985.

Pintos J, Franco EL, Kowalski LP, Oliveira BV, Curado MP. Use of wood stoves and cancer of the upper aero-digestive tract: a case-control study. *International Journal of Epidemiology* 27: 936–940, 1998.

Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Statistics in Medicine* 9: 657–671, 1990.

Pope C, Mays C (eds). *Qualitative research in health care,* 2nd edn. London: BMJ Publishing Group, 2000.

Popiela T, Jedrychowski W, Filipek A, Dolzycki E, Kulig J, Olszanecki S. Validity of questionnaire criteria in mass screening for the diagnosis of peptic ulcer. *International Journal of Epidemiology* 5: 251–253, 1976.

Poulter NR, Sever OS. Intervention in high risk groups: blood pressure. In: *Coronary heart disease epidemiology: from aetiology to public health* (Marmot M, Elliott P, eds). Oxford: Oxford University Press, 1992, pp. 325–342.

Preti P, Miotto P. Suicide and unemployment in Italy, 1982–1994. *Journal of Epidemiology and Community Health* 53: 694–701, 1999.

Rhyne R, Bogue R, Kukulka G, Fulmer H. *Community-oriented primary care: health care for the 21st century.* Washington, D.C.: American Public Health Association, 1998.

Ritchie K, Jildea D, Robine J-M. The relationship between age and the prevalence of senile dementia: a meta-analysis of recent data. *International Journal of Epidemiology* 21: 763–769, 1992.

Roberts RS, Spitzer WO, Delmore T, Sackett DL. An empirical demonstration of Berkson's bias. *Journal of Chronic Diseases* 31: 119–128, 1978.

Rona RJ, Chinn S, Florey CduV. Exposure to cigarette smoking and children's growth. *International Journal of Epidemiology* 14: 402–409, 1985.

Rona RJ, Swan AV, Altman DG. Social factors and height of primary schoolchildren in England and Scotland. *Journal of Epidemiology and Community Health* 32: 147–154, 1978.

Roper HP, David TJ. Decline in deaths from choking on food in infancy: an association with change in feeding practice? *Journal of the Royal Society of Medicine* 80: 2–3, 1987.

Rosenthal R. The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 85: 185–193, 1979.

Rothman KJ. Causes. *American Journal of Epidemiology* 104: 587–592, 1976.

Rothman KJ. *Modern epidemiology.* Boston: Little, Brown, 1986.

Rothman KJ (ed). *Causal inference.* Chestnut Hill, Massachusetts: Epidemiology Resources, 1988.

Rothman KJ, Greenland S. *Modern epidemiology.* 2nd edn. Philadelphia: Lippincott-Raven, 1998.

Ruston A, Clayton J, Calman M. Patients' action during their cardiac event: qualitative study exploring differences and modifiable factors. *British Medical Journal* 316: 1060–1064, 1998.

Sackett DL. Bias in analytic research. *Journal of Chronic Diseases* 32: 51–63, 1979.

Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine.* Boston: Little, Brown, 1985.

Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practice and teach EBM.* New York: Churchill Livingstone, 1997.

Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *New England Journal of Medicine* 316: 450–455, 1987.

Schlesselman JJ. *Case-control studies: design, conduct, analysis.* New York: Oxford University Press, 1982.

Schull WJ, Cobb S. The intrafamilial transmission of rheumatoid arthritis—III. The lack of support for a genetic hypothesis. *Journal of Chronic Diseases* 22: 217–222, 1969.

Scrimshaw SCM, Hurtado E. *Rapid assessment procedures for nutrition and primary health care: anthropological approaches to improving program effectiveness.* Los Angeles: UCLA Latin American Center Publications, 1987.

Selvin S. Statistical analysis of epidemiologic data, 2nd edn. New York: Oxford University Press, 1996.

Severson RK, Linet MS. Does cigarette smoking lead to the subsequent development of leukemia? *Archives of Internal Medicine* 153: 425–427, 1993.

Shakespeare W. *Hamlet, prince of Denmark.* 1603.

Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 39 (Suppl. 6): 2772–2782, 1977.

Shapiro S, Slone D, Rosenberg L, Kaufman DW, Stolley PD, Miettinen OS. Oral-contraceptive use in relation to myocardial infarction. *Lancet* 1: 743–747, 1979.

Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute* 69: 349–355, 1982.

Shirlow MJ, Mathers CD. Caffeine consumption and serum cholesterol levels. *International Journal of Epidemiology* 13: 422–427, 1984.

Shirlow MJ, Mathers CD. A study of caffeine consumption and symptoms: indigestion, palpitations, tremor, headache and insomnia. *International Journal of Epidemiology* 14: 239–248, 1985.

Shoukri MM. Agreement, measurement of. In: *Encyclopedia of epidemiologic methods* (Gail MH, Benichou J, eds). Chichester: Wiley, 2000, pp. 35–49.

Shoukri MM, Pause CA. *Statistical methods for health sciences,* 2nd edn. Boca Raton, Florida: CRC Press, 1998.

Siegel S, Castellan NJ Jr. *Nonparametric statistics for the behavioral sciences,* 2nd edn. New York: McGraw-Hill, 1988.

Siegel-Itzkovich J. "Distortion" of passive smoking evidence provokes controversy in Israel. *British Medical Journal* 320: 626, 2000.

Simon R. A decade of progress in statistical methodology for clinical trials. *Statistics in Medicine* 10: 1789–1817, 1991.

Slavin RE. Best-evidence synthesis: an alternative approach to traditional and meta-analytic reviews. *Educational Researcher* 15(9): 5–11, 1986.

Slavin RE. Best-evidence synthesis: why less is more. *Educational Researcher* 16(5): 15–16, 1987.

Smith GD, Harding S, Rosato M. Relation between infants' birth weight and mothers' mortality: prospective observational study. *British Medical Journal* 320: 839–840, 2000.

Smith GD, Phillips AN, Neaton JD. Smoking as "independent" risk factor for suicide: illustration of an artifact from observational epidemiology? *Lancet* 340: 709–712, 1992.

Smith ML, Glass GV. Meta-analysis of psychotherapy outcome studies. *American Psychologist* 35: 752–760, 1977.

Smith PG, Rodriguez LC, Fine PEM. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *International Journal of Epidemiology* 13: 87–93, 1984.

Sommer A. *Epidemiology and statistics for the ophthalmologist.* New York: Oxford University Press, 1980.

Sosenko JM, Gardner LB. Attribute frequency and misclassification bias. *Journal of Chronic Diseases* 40: 203–207, 1987.

Spitzer WO. Meta-meta-analysis: unanswered questions about aggregating data. *Journal of Clinical Epidemiology* 44: 103–107, 1991.

Sprackling ME, Mitchell JRA, Short AH, Watt G. Blood pressure reduction in the elderly: a randomised controlled trial of methyldopa. *British Medical Journal* 283: 1151–1153, 1981.

Stampfer MJ, Goldhaber SZ, Yusuf S, Peto R, Hennekens CH. Intravenous streptokinase for acute myocardial infarction. *New England Journal of Medicine* 308: 593–594, 1982.

Stark CR, Mantel N. Effects of maternal age and birth order on the risk of mongolism and leukemia. *Journal of the National Cancer Institute* 37: 687–698, 1966.

Stellman SD, Garfinkel L. Artificial sweetener use and one-year weight change among women. *Preventive Medicine* 15: 195–202, 1986.

Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical regression projects. *British Medical Journal* 315: 640–645, 1997.

Stewart AL, Ware JE Jr, Brook R, Davies-Avery A. *Conceptualization and measurement of health for adults in the Health Insurance Study, Vol. II: Physical health in terms of functioning.* Santa Monica, California: Rand Corporation, 1978.

Stewart AW, Jackson RT, Ford MA, Beaglehole R. Underestimation of relative weight by use of self-reported height and weight. *American Journal of Epidemiology* 125: 122–126, 1987.

Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* 341: 418–422, 1993.

Stouffer SA, Suchman EA, De Vinney LC, Star SA, Williams RM Jr. *The American soldier: adjustment during army life,* Vol. 1. Princeton, New Jersey: Princeton University Press, 1949.

Sukwa TY, Bulsara MK, Wurapa FK. The relationship between morbidity and intensity of *Schistosoma mansoni* infection in a rural Zambian community. *International Journal of Epidemiology* 15: 248–251, 1986.

Susser M. *Causal thinking in the health sciences: concepts and strategies in epidemiology.* New York: Oxford University Press, 1973.

Susser M. The logic of Sir Karl Popper and the practice of epidemiology. *American Journal of Epidemiology* 124: 711–718, 1986.

Susser M. Falsification, verification, and causal inference in epidemiology: reconsiderations in the light of Sir Karl Popper's philosophy. In: *Epidemiology, health, & society* (Susser M, ed). New York: Oxford University Press, 1987. [Also in: Rothman KJ (3d). *Causal inference.* Chestnut Hill, Massachusetts: Epidemiology Resources, 1988, pp. 33–57.]

Susser M. Rational science versus a system of logic. In: *Causal inference* (Rothman KJ, ed). Chestnut Hill, Massachusetts: Epidemiology Resources, 1988, pp. 189–199.

Susser M. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *American Journal of Public Health* 86: 674–677, 1996.

Ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. *Journal of Clinical Epidemiology* 43: 1191–1199, 1990.

Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 338: 1127–1130, 1991.

Tillett HE. Statistical analysis of case-control studies of communicable diseases. *International Journal of Epidemiology* 15: 126–133, 1986.

Tuomilehto J, Morelos S, Yason J, Guzman SV, Geizerova H. Trends in cardiovascular disease mortality in the Philippines. *International Journal of Epidemiology* 13: 168–176, 1984.

Turner JA, Herron L, Deyo RA. Meta-analysis of the results of lumbar spine fusion. *Acta Orthopaedica Scandinavica* 64 (Suppl. 251): 120–122, 1993.

Udjo EO. *Additional evidence regarding fertility and mortality trends in South Africa and implications for population projections.* Mimeo. Pretoria: Statistics South Africa, 1998. On Internet: http://www.statssa.gov.za/census96/HTML/Metadata/Docs/fertility_udjo.htm

U.S. Preventive Services Task Force. *Guide to clinical preventive services,* 2nd edn. Washington, D.C.: U.S. Department of Health and Human Services, undated.

Walker AM. Reporting the results of epidemiologic studies. *American Journal of Public Health* 76: 556–558, 1986.

Weddell JM, McDougall A. Road traffic accidents in Sharjah. *International Journal of Epidemiology* 10: 155–159, 1981.

Weinberg CR. Toward a clearer definition of confounding. *American Journal of Epidemiology* 137: 1–8, 1993.

Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 10: 1665–1677, 1991.

WHO Expert Committee on Diabetes Mellitus. Second report. WHO Technical Report Series No. 646, Geneva: WHO, 1980.

Wolf FM. *Meta-analysis: quantitative methods for research synthesis.* Beverly Hills, California: Sage Publications, 1986.

Wortman PM, Yeaton WH. Synthesis of results in controlled clinical trials of coronary bypass graft surgery. In: *Evaluation studies review annual,* Vol. 8 (LIght RJ, ed). Beverly Hills, California: Sage Publications, 1983.

Yusuf S. Obtaining medically meaningful answers from an overview of randomized clinical trials. *Statistics in Medicine* 6: 281–286, 1987a.

Yusuf S. Discussion. *Statistics in Medicine* 6: 287–294, 1987b.

Yusuf S, Held P, Furburg CD. Update of effects of calcium antagonists in myocardial infarction or angina in light of the Second Danish Verapamil Infarction Trial (DAVIT-II) and othe recent studies. *American Journal of Cardiology* 67: 1295–1297, 1991.

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 27: 335–371, 1985.

Yusuf S, Simon R, Ellenberg S. (eds). Proceedings of "Methodologic Issues in Overviews of Randomized Clinical Trials." *Statistics in Medicine* 6: 217–409, 1987.

Zelen M. Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors. *Journal of Clinical Oncology* 1: 164–169, 1983.

Zoloth S, Michaels D, Lacher M, Nagin D, Drucker E. Asbestos disease screening by nonspecialists: results of an evaluation. *American Journal of Public Health* 76: 1392–1395, 1986.

Index

"Are there many crabs here?" said Alice. "Crabs, and all sorts of things," said the Sheep: "plenty of choice, only make up your mind."

(Carroll, 1872)