Second Edition

Epidemiological Studies

A Practical Guide

Alan J. Silman and Gary J. Macfarlane

CAMBRIDGE

more information - www.cambridge.org/0521810973

This page intentionally left blank

Epidemiological Studies

Second Edition

Following on in the footsteps of its acclaimed and popular predecessor, this new edition builds on the successful features that engaged readers of the first edition: it explains the nuts and bolts of epidemiology and serves as a handbook for those who wish to do epidemiology; it uses relevant exercises and examples, taken from real life, to illustrate how to set up a study; it aims to help produce valid results that will satisfy grant bodies, ethical committees and journal editors; ultimately it bridges the gap between theory and practice. By making the subject so easily accessible, it will be an excellent introduction for anyone who is training in epidemiology and public health, and for all those involved in medical research. This edition includes numerous improvements and several new chapters which will further enhance its appeal and usefulness.

Epidemiological Studies

A Practical Guide

Second Edition

Alan J. Silman
The Medical School, University of Manchester
and
Gary J. Macfarlane
The Medical School, University of Manchester

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published in printed format 2002

ISBN 0-511-02979-9 eBook (Adobe Reader)
ISBN 0-521-81097-3 hardback
ISBN 0-521-00939-1 paperback

First published 1995
Second edition 2002

Contents

PART III STUDYING ASSOCIATIONS BETWEEN RISK FACTORS AND DISEASE

PART IV SELECTION OF POPULATIONS AND SAMPLES TO STUDY

Scope of this volume

A quick browse through the epidemiology section of medical libraries and bookshops reveals that most volumes with 'epidemiology' in their title are either theoretical texts covering the concepts of the discipline, with varying degrees of biostatistical complexity, or are reviews of the available 'epidemiological' data on one disease or a group of related diseases. The problem with the latter is that such reviews, often out of necessity, present the available published data without critically reviewing the studies that lead to their generation. In this volume an attempt has been made to bridge the gap between concept and result and to provide a practical guide to epidemiological practice. It is therefore written as a handbook to those who wish to do epidemiology, rather than restricting its aim to those who wish to understand the discipline. The underlying theme, however, has been how to complete a study without compromising validity (in its broadest sense). The hope therefore is that sufficient guidance is provided to produce work that will satisfy grant bodies, ethical committees and journal editors and reviewers.

This volume has taken a deliberate 'holistic' view of epidemiological method. The practical issues, for example, of approaching potential study populations, maximising response and questionnaire design, are similarly independent of the epidemiological study design chosen.

The volume is divided into a number of parts. The first part provides the uninitiated with a list of subjects that are relevant for the application of epidemiological method and follows this with an outline of the main problem areas in undertaking epidemiological research. The remainder of the book aims to cover these issues systematically in completing an epidemiological study.

The second part addresses the options available for measuring the occurrence of disease and methodological considerations in comparing occurrence between populations and over time.

The third part focuses on the choice of an appropriate design to address the questions posed. As such, this part acts also as an introduction to the underlying concepts behind the varying options for choice of study. The objective is that this part provides sufficient material to enable the investigator to choose and justify the optimal method(s) for study.

The fourth part covers the problems inherent in selecting populations for study. The relative merits of the different choices available are discussed together with the practical issues of how to recruit subjects, including, for example, suggestions for letters of invitation. In this part as in the sections on analysis in Part VI, the separate requirements of different study approaches, for example case control and cohort studies, are considered individually.

The fifth part addresses the problems of data collection within epidemiological studies with the objective of selecting the most effective survey method to achieve the goals targeted. Practical issues covered include the assessment of validity and reproducibility of survey methods, suggestions for design of survey instruments and a detailed account of how to maximise participation, possibly the greatest practical problem in conducting a survey. These issues are also considered collectively in the section on conducting a pilot study.

The sixth part covers analysis and interpretation of the data collected. The first focus is on the preparation of data for analysis, including the important but often ignored steps of checking for data errors and ensuring a clean set of data for subsequent analysis. Chapters 16 and 17, on introductory epidemiological data analysis, are not meant to be a substitute for the necessary detailed information on this topic that is required to be obtained from statistics textbooks and the manuals accompanying analytical software packages. Their aim is to provide the necessary formulae with simple worked examples to permit the investigator to understand how measures of occurrence and effect are calculated and to undertake these on a hand calculator. These chapters assume only a limited knowledge of statistics such as an understanding of summary measures (e.g. the mean), and measures of dispersion (e.g. standard deviation). It will also be necessary to understand the concept behind statistical significance testing and particularly the use and limitations of the confidence interval. The subsequent chapters cover the topics of confounding and bias, with the goal that the investigator

will be able to assess the impact, if any, of these phenomena on the results obtained.

The final part covers two important practical areas: how to ensure that a proposed study is ethically sound and how to minimise the costs of undertaking studies.

The text is liberally illustrated with examples from many areas of medicine in an attempt to clarify the points made. Many of the examples are drawn from the real life experience of the authors and their colleagues, others are based on reports of research, both published and unpublished, with the remainder unapologetically invented to illustrate a point where necessary.

Finally, where appropriate, exercises are set at the end of a chapter, with model solutions provided at the end of the book.

Acknowledgements

From first edition

The genesis of this volume lies in the organisation of a teaching programme in epidemiology to those entering public health in the southeast of England. There was a clear need for those students to receive sufficient training to undertake simple epidemiological studies themselves. With no appropriate practical guide available, a course was developed to achieve this goal. With suitable refinements resulting from continuing feedback from the participants, the material developed for this course formed the basis of the current volume. I therefore acknowledge the contribution made by successive cohorts of attenders. The material has grown and altered in further discussions with clinical colleagues in both primary and hospital care who have themselves desired to dip their toes into the fascinating waters of epidemiological science.

I am, however, particularly grateful to my close colleagues Peter Croft, Deborah Symmons and Paul Brennan for their frank and constructive criticisms as the work has progressed; any errors or confusions that remain are mine. My secretary Cath Barrow has worked tirelessly and cheerfully and has coped with substantial rewrites without complaint.

Second edition

The second edition has further benefited from feedback of both staff and postgraduate students at the School of Epidemiology and Health Sciences, University of Manchester. In addition, Lesley Jordan has ably and willingly co-ordinated the administrative work involved in producing this revised edition.

Part I

Introduction

1

Scope of epidemiological enquiry and overview of main problem areas

1.1 What questions can epidemiology answer?

Epidemiology can be defined as the study of the distribution of disease and its determinants in human populations. In other words, it provides the answers to questions on how much disease there is, who gets it and what specific factors put individuals at risk. The epidemiology section in a medical textbook chapter on a particular disease will provide data on these aspects. There is an alternative and broader view of epidemiology, which is that it is a methodology to obtain answers about diseases from their study in humans. This broader definition allows a substantially greater scope for the kinds of question that can be addressed both by those studying the health of populations and by those whose main focus is the study of disease in patient groups. The list in Table 1.1 represents the vast array of topics that epidemiologists would consider as relevant to their discipline.

1.1a Disease definition

Most diseases lack a clear diagnostic test that totally discriminates between disease and normality, though infectious disease and trauma are two obvious exceptions. Most often the diagnosis is based on clinical opinion, with the latter based on experience, prejudice or arbitrary rules. In the absence of a standardised definition of disease, results from aetiological, prognostic or therapeutic studies cannot be directly compared. The development of disease criteria is a separate topic in itself which requires a careful epidemiological study of the occurrence of specific features in cases determined by a notional gold standard, such as the expert clinician's diagnosis. These features are then compared with those from an appropriate group of non-cases and the level of agreement evaluated.

Table 1.1. Questions relevant for epidemiological enquiry

Disease definition	What characteristics or combination of characteristics best discriminate disease from non-disease?
Disease occurrence	What is the rate of development of new cases in a population? What is the proportion of current disease within a population? What are the influences of age, sex, time and geography on the above?
Disease causation	What are the risk factors for disease development and what are their relative strengths with respect to an individual and a population?
Disease outcome	What is the outcome following disease onset and what are the risk factors, including their relative strengths, for a poor outcome?
Disease management	What is the relative effectiveness of proposed therapeutic interventions? (Included within this are health service research questions related to the relative effectiveness of proposed health service delivery options.)
Disease prevention	What is the relative effectiveness of proposed preventive strategies including screening?

Example 1.i

Prior to starting a research programme on migrainous headaches, a neurologist realised that it was necessary to derive criteria that (i) were easy to apply, (ii) could distinguish migraine from other causes of headaches and (iii) would be accepted by the neurological community.

1.1b Disease occurrence

This is the classical focus of epidemiology and the available approaches to measure occurrence are discussed in Chapter 2. Data on occurrence are of interest in their own right, but are also relevant both to the clinician, in weighing up different diagnostic likelihoods in the face of the same evidence, and to those providing health services. A more detailed study will uncover differences in occurrence between sexes and across age groups, over time and between different geographical populations. Indeed, the age and sex effects on disease occurrence are normally so strong that it is absolutely fundamental to gather such data in order to compare disease occurrence both between

populations and within the same population over time. These issues are discussed in Chapter 3. In addition, marked differences between occurrence in different population groups may provide aetiological clues for further enquiry.

1.1c Disease causation

Similarly, the use of epidemiology to unravel causative mechanisms is one of its major roles. It is, however, too simplistic for most chronic diseases, to consider their influence on disease risk as present or absent. It is the strength of any disease association with possible risk factor variables that is of more interest.

Example 1.ii

In planning a study on whether workers exposed to organic dusts were at increased risk of various lung diseases, the investigators aimed to discover (i) whether or not there was an increased risk, (ii) the level of any increase for the major lung disorders considered and (iii) how these risks compared with those from smoking in the same population.

Risk and association

It is appropriate, at this stage, to clarify the meaning of the terms 'risk' and 'association'. In common use, association indicates a relationship with no indication as to the underlying path. As an example, there is an association between an individual's height and his/her weight, although there are a number of possible paths: (i) the taller an individual the greater will be the weight; (ii) (unlikely) the heavier an individual the greater will be the height; or (iii) there are common factors, for example genetic, that influence both height and weight. By contrast, risk implies that the pathway is known (or worthy of investigation). Thus in the example above, the question can be addressed whether height is a risk factor for (being over) weight. In practice, epidemiological investigations aim to uncover associations that, using other information, are translated into risks.

1.1d Disease outcome

Investigations concerning the frequency and prediction of specific disease outcomes in patient populations may be considered as the clinical epidemiological parallels of studies of disease occurrence and causation in normal

populations. Thus the *population* epidemiologist may wish to ascertain the incidence of, and risk factors for, angina in a stated population; whereas the *clinical* epidemiologist may wish to ascertain the incidence of, and risk factors for, subsequent myocardial infarction and sudden death in patients with angina. The methodological concepts, however, are identical, as will be discussed in later chapters.

1.1e Disease management and disease prevention

The use of the clinical trial to evaluate the effectiveness of a particular therapeutic intervention is well established in medicine. Epidemiologically, the clinical trial can be considered as an experimental study where the investigator has intervened to alter the 'exposure', e.g. management, in order to measure the effect on disease occurrence or outcome. The term 'intervention study' describes this broad activity. A further aim of this type of study is to determine whether a link between a suspected risk factor and disease is causative rather than simply an association.

Example 1.iii

In order to examine the possibility that dietary folate deficiency during pregnancy was a causative factor for neural tube defects, an intervention study was carried out on high-risk pregnant women who were randomly allocated to receive folate supplementation or placebo.

Conversely, intervention trial concepts can be applied to health service delivery to answer questions such as whether policy A is likely to be more effective than policy B in reducing waiting lists. Health service research questions such as this require the same rigorous epidemiological approach. In most developed countries with increasing economic pressure to contain health service costs, there is considerable demand (and funding) for epidemiologists to apply their expertise in this area.

An extension of the above is the use of the intervention trial to assess the effectiveness of a population-wide preventive strategy. Population-based primary prevention trials can indeed be considered as clinical trials on a massive scale. Screening for preclinical disease is a widely practised preventive strategy and the evaluation of screening as a tool for reducing morbidity/mortality can be considered under the same heading as above.

1.2 What are the major issues in conducting epidemiological research?

Much of the above seems straightforward, and indeed part of the attraction of epidemiology is its accessibility to the potential investigator. Compared with other approaches to studying biomedical issues, epidemiology often does not require expensive or highly technical equipment and superficially, at least, its language and concepts are those of everyday 'medical speak' that do not require the initiation into a new language as does molecular biology or immunology, for example. There are, however, distinct epidemiological concerns, both for the first-time investigator and for the expert reviewing the work of others, stemming in a large part from the basic tenet that epidemiology deals with 'free living', and frequently healthy, human populations. The consequences of this are: (i) methods of study have to be simple and non-invasive; (ii) subjects, as compared with laboratory animals, may choose to participate or not, or even withdraw from participation during a study; and (iii) the experimental approach where the investigator modifies the conditions to study a specific factor is fraught with difficulties and, as a result, experimental studies are infrequent in epidemiological research. In addition, since many important diseases are relatively rare, studies often need to be large in scope, long in duration or both, with consequences both for the resources required and for the patience and longevity of the investigator.

There are a substantial number of problems to be considered in undertaking any epidemiological study. These are listed in Table 1.2, which provides the framework for the rest of the volume, and are discussed in outline below.

1.2a Study design

The first demand is to clearly frame and thereafter focus on the specific questions posed. In the following two chapters, the various options for studies of disease occurrence and causation are outlined. A decision has to be made about the choice of study design that can best answer the question posed, taking into account the often conflicting demands of scientific validity and practicality.

1.2b Population selection

The subjects to be studied have to be defined both in terms of the group(s) from which they are to be selected and, in selecting individuals, the inclusion

Table 1.2. Major problem areas for epidemiological research

Study design	What is the question posed – what type of study can best answer the question and is most practicable?
Population selection	Who should be studied? How many should be studied?
Information gathering	How should the information be obtained? Is the information obtained correct? Is the method used to obtain the information consistent?
Analysis	How should the data gathered be prepared for analysis? What are the appropriate analytical methods?
Interpretation of results	Can any associations observed be explained by confounding? Are the results explained by bias? Are the results generalisable?
Logistics	Is the research ethical? Is the research affordable?

and exclusion rules. Specific problems arise in comparative studies when it is necessary to recruit two or more groups based on their disease or on their risk factor status. Problems in population selection are one of the major reasons for a study's conclusions being invalid. A specific difficulty is that of sample size. Cost, time, or other practical considerations may limit the availability of subjects for study. A scientific approach to sample size estimation is given for the different study design options later on in the book. Non-response or loss to follow-up can reduce the number of subjects available for analysis and an adequately large study at the onset may prove too small by the end.

1.2c Information quality

This major issue relates to the quality of the data obtained. There is a particular problem when the approach requires a subject to recall past symptoms or exposures. The most appropriate method for obtaining information must be selected. This might, for example, be a choice between interview and self-administered questionnaire. Other sources of information such as data collected prior to the study, often for another purpose such as the medical record, may be available. The classical approach is to consider the quality of

information obtained under the headings of: (i) *validity*, i.e. does the measurement give the true answer, and (ii) *repeatability*, i.e. is the same answer obtained from the same person when repeated measures are made?

1.2d Data handling and analysis

The time spent on this activity is frequently longer than that spent on the data collection itself. In particular, there is a need to ensure that the data analysed are complete and error-free. The next problem is to choose the appropriate method of analysis.

1.2e Interpreting the results

The first issue is that of confounding. Put simply, it is often difficult in human populations to distinguish between attributes that frequently occur together. Thus, in studies to determine the effect of cigarette smoking on the risk for a particular disease, a positive association may be observed that does not relate to the direct impact of smoking on risk but reflects the joint association between a *confounder*, such as alcohol consumption, which is linked to both cigarette smoking and the disease under study. One of the major advances in the practice of epidemiology in the past decade has been the simultaneous development of user-friendly software and accessible hardware that permit the analysis of the impact of any potential confounder in a way that manual methods of statistical analysis could not achieve.

The second issue is whether the results obtained could be explained by bias, either in the selection of subjects, in the subjects who chose to participate, or in the gathering of information.

The third issue is whether the results are generalisable. A study has been conducted amongst university students examining the relationship between coffee consumption and migraine headaches. Students with migraine were more than twice as likely to consume, on average, more than two cups of coffee per day. Is the association generalisable, outside the study population?

1.2f Logistical issues

Two important areas to be addressed are those of ethics and cost. Studying free-living individuals imposes ethical constraints on the investigators, and the need for cost containment is self-evident. Indeed, these issues have to be considered early as they are likely to influence the final study design.

Part II

Measuring the occurrence of disease

2

Which measure of disease occurrence?

Measuring disease occurrence is the basic activity of epidemiology, and the following section aims to provide the background to choosing the appropriate measure of disease occurrence. The term 'disease' can also be taken in this context to describe any personal attribute. Thus, the concepts described apply equally well to assessing the occurrence of an exposure such as cigarette smoking, the development of a particular disability, or the frequency of a medical intervention such as blood transfusion or hip replacement in a population. The first step is to distinguish between incidence and prevalence. These two terms are frequently confused and a simple distinction is to consider incidence as measuring disease *onsets* and prevalence as measuring disease *states*.

2.1 Incidence

The *incidence* of a disease is the number of new onsets in a population within a specified period of time. It will be noted that this measure does not make any reference to the size of population studied and therefore to compare the incidence between a large city and a small village does not make sense. In order to overcome this problem the *incidence rate* is usually calculated. This divides the incidence by the size of the population to which it applies. Thereafter this is usually multiplied by a constant (e.g. 1000, 10 000, 100 000, etc.) for ease of interpretation.

Example 2.i

There were 570 new cases of salmonella food poisoning during 1999 in a city with a population of 8 million.

$$\text{Incidence} = 570 \text{ cases per year}$$

$$\text{Incidence rate} = \frac{570}{8 \times 10^6} = \frac{\text{number of events in time period}}{\text{number in the population}}$$

$$= 0.0000713 \text{ per year}$$

Incidence rate per 100 000 persons per year

$$= 0.0000713 \times 100\,000$$

$$= 7.13$$

Therefore the estimated annual incidence rate (based on 1999 data) is 7.13 per 100 000 persons.

In the example above the denominator population is the average population during the year of interest and represents an approximation to the number of persons and the time for which they were 'at risk' of developing salmonella food-poisoning. Individuals, for example, who move out of the city or die will no longer be at risk, while those who move into the city will be newly at risk. It is therefore more accurate and may practically be possible to determine the time at risk for each individual subject and combine them to form at total measure of *person–time*, most commonly *person–years* at risk. This is obtained from adding up all the person–time data as the denominator. (At its simplest, if two individuals are observed, one for three years and the other for five years, then they have contributed in total eight person–years of observation.)

Example 2.ii

In a study of the incidence of upper respiratory tract infection at a local school during the course of a year, there were 743 children registered at some point during the school year. However, since some children joined/left during the school year the total person–years of follow-up was only 695 person–years.

An example of the calculation of an incidence rate is given in Example 16.i. Since the measure is theoretically a measure of density of events, it is sometimes also referred to as *incidence density*.

2.1a Approaches to measuring incidence

There are a number of different approaches to measuring incidence, and these are shown in Table 2.1. *First-ever incidence* is restricted to the inclusion

Table 2.1. Measures of disease incidence

Incidence	Records disease *onsets*
First-ever incidence	Records only first ever episode
Episode incidence	Records all episodes
Cumulative incidence	Records all onsets up to a certain time point

of only those subjects who present with their first disease episode from a particular pathology during a particular time period. It might, however, be of greater concern to record all episodes, '*the episode incidence*', independent of whether it is the first occurrence or a recurrence.

Example 2.iii

To gain an idea of the age distribution of onset of coronary artery disease, the cardiovascular epidemiologist might wish to study the incidence of first myocardial infarction. By contrast, the public health physician might wish to know the episode incidence in order to provide an appropriate level of acute care facilities.

There is the assumption that it is always possible to distinguish between a recurrent episode of an existing pathological process and a new episode from a separate pathological process. This is frequently not the case.

Example 2.iv

An investigator considered the options for estimating the incidence of low back pain. The problem was that this 'syndrome' is associated with multiple episodes in a single individual during a lifetime, which might result from single or multiple causes. The decision was made to consider first-episode incidence within a time period of one year, ignoring previous history.

An alternative measure of occurrence is *cumulative incidence*. This technically is a *risk* rather than a *rate* and is expressed as a proportion. A *risk* is the combined effect of *rates* operating over a specific time period. This could be calculated for a lifetime; for example the lifetime cumulative incidence of cancer in males is 0.25. This approach is used in follow-up studies to compare the subsequent development of a disease in groups determined by their exposure.

Table 2.2. Measures of disease prevalence

Prevalence	Records disease *states*
Point prevalence	Records all with disease state at a (notional) point in time
Period prevalence	Records all with disease state at some time during a stated period of time
Cumulative prevalence	Records all with disease state at some time up to a certain time point

Example 2.v

In the investigation of an outbreak of gastroenteritis after a banquet, an investigator calculated the cumulative risks, during the arbitrary follow-up time of the study, of developing infection in those who did and did not eat each of the particular foods available.

2.2 Prevalence

The prevalence of disease is the number of individuals in a population with a disease or other personal attribute. Prevalence rates are calculated in an analogous way to incidence rates and again often expressed per multiple (e.g. 1000, 10 000 . . .) of the population. Technically it is a *proportion* and not a *rate* since there is no time element involved. However, the term *prevalence rate* is commonly used:

Example 2.vi

A questionnaire was distributed to the workforce of a large industrial company on a particular working day. Of the 1534 workers, 178 reported headaches on the survey day.

Prevalence = 178 cases

$$\text{Prevalance rate (proportion)} = \frac{178}{1534}$$

$$= 0.12 \text{ or}$$

12 per 100 workers

There are also a number of different approaches to measuring prevalence and these are shown in Table 2.2.

2.2a Approaches to measuring prevalence

Point prevalence records all those with a disease at a notional point in time. In reality the disease status of currently living individuals is assessed at varying points in time, but those who are positive are all assumed to be in the same disease state on 'prevalence day'. Thus, a point prevalence estimate should be in the form of: the proportion of the population with disease on, for example, 1 January 1994, e.g. 35 per 1000 of the population. *Period prevalence* takes account of the common situation that diseases vary within an individual in their clinical manifestations and that an individual sufferer may not be *in state* at a single point in time. The most suitable approach is to describe the occurrence at any time within a defined *period*, typically but not exclusively a calendar year. Appropriate examples include migraine and sleep disturbance, reflecting the fact that within a nominated period an individual may be suffering, but not necessarily at an arbitrary point in time.

Cumulative prevalence extends this to include all those who have been in a disease state during their lives or between two specific time points, for example cumulative prevalence between ages 45 and 64. This concept is useful for those disorders that are variable in their natural history because this measure will 'capture' those with a single resolved episode some years previously as well as those with continuing disease. There is a certain similarity between cumulative incidence and cumulative prevalence, as conceptually they should be recording the same individuals. The only real distinction is that cumulative incidence permits the inclusion of individuals who developed the disease and subsequently died, whereas cumulative prevalence tends to imply a backward (*retrospective*) examination of those currently alive at a certain age. Thus, there is a distinction between the cumulative incidence of a cancer at a particular site by the age of (say) 65, from the cumulative prevalence in a survey of current 65-year-olds of all those that have ever had peptic ulceration and survived to age 65.

2.3 Choice of measure

By implication, from the above paragraphs, it is the nature of the disease itself that determines the appropriate choice of measure. A summary of the issues is given in Table 2.3. The choice of measure has important practical implications because the methodological approaches vary considerably (see

Table 2.3. Appropriate measures of disease occurrence

Disease characteristics	Examples	Appropriate measures[a]
A		
Clearly defined onset	Acute appendicitis	New incidence
Single episode	Colon cancer	Cumulative incidence
Terminated by death, spontaneous resolution or therapy	Major trauma	
B		
Clearly defined onset	Myocardial infarction	Episode incidence
Possible multiple, but infrequent episodes	Influenza	Cumulative incidence
Short duration	Fracture	New incidence
Episodes terminated by death, spontaneous resolution or therapy		
C		
Clearly defined onset	Insulin-dependent diabetes	New incidence
Chronic relatively stable disease state or requiring long-term therapy	Renal failure	Point prevalence Cumulative incidence
D		
Clearly defined onset	Rheumatoid arthritis	New incidence
Single or multiple episodes	Peptic ulceration	Cumulative incidence
Variable duration		
E		
Ill-defined onset	Hypertension	Point prevalence
Chronic relatively stable disease state or requiring long-term therapy	Hip osteoarthritis Deafness	
F		
Ill-defined onset	Asthma	Period prevalence
Multiple episodes with periods of disease absence	Migraine	Episode incidence Cumulative prevalence

Note:

[a] Optimal measure listed first.

Chapter 4). The table can only offer guidance and the decision is not always obvious. Knowledge of the natural history and variation of presentation of disease is important.

Example 2.vii

In a study to determine the incidence of myocardial infarction, the decision had to be made whether to include those for whom an electrocardiograph (ECG) indicated a definite previous infarction even though there was no recall of any clinical event. A study design was chosen that included a baseline ECG, a follow-up ECG after five years and a clinical record review during the same five-year period. The ECG data were used to calculate a five-year episode incidence.

The nature of the question and the objectives of the study also influence the choice.

Example 2.viii

In disorders such as severe mental deficiency, the healthcare planner would need knowledge of the current point prevalence as a priority. By contrast, the perinatal epidemiologist would focus on the incidence during the first year of life.

3

Comparing rates: between and within populations

3.1 Introduction

The previous chapter has outlined types of rates and their measurement. For the public health specialist concerned with the provision of services, measurement of disease occurrence may be an end in itself. For the epidemiologist, however, measurement of rates will commonly be the beginning of a process whose aim is to understand the aetiology of a disease. In order to formulate hypotheses, the rate of a disease under study in a population may be compared with the rate in other populations, or in the same population at different time points. Those populations (or population-groups) with particularly high or low rates can be identified and features of these populations determined in order to formulate hypotheses on the influence of disease to be tested in a formal epidemiological study. If the rates are increasing (or decreasing) one may wish to determine what factors are responsible for such a change.

Whatever the comparison being made, either between populations, between sub-groups of a larger population, or in one population over a period of time, it is important that comparisons are made on a like-for-like basis.

Example 3.i

It was thought, on anecdotal evidence, that Sunnytown had a much higher incidence rate of stroke than its near neighbour Drizzletown. A marked difference in the incidence of stroke would warrant further investigation. However, since the incidence of stroke is strongly age-related it was thought that the difference in incidence may simply be a reflection that the residents of Sunnytown were, on average, older. Interpretations of the difference in rates could only be made after taking into account differences in the age-structure between the populations.

Similarly, increasing incidence rates of stroke within a single population may be a consequence of the population ageing or it may reflect a real increase in rates. If the rates of disease change within a population are found to be real, then further investigation of the pattern of change can also provide clues to the possible reasons.

3.2 Standardisation

Standardisation, as a general procedure for the control of variables that may confound an observed relationship, is discussed in Section 18.4c. In this section age, the most common such factor (and one on which data is most readily available) in comparing rates between or within populations will be considered. The principles outlined can apply to other factors such as gender and social class. There are two methods of standardisation: direct and indirect. They are equally applicable to incidence and mortality rates, and prevalence.

3.2a Direct standardisation

Let us assume that the incidence rates of stroke have been measured in Drizzletown (see Example 3.i), and one wishes to compare rates with Sunnytown taking into account a slightly different age structure between the two populations. (Examples in this chapter relate to age but could be considered for any attribute on which information was available.) Direct standardisation (for age) calculates what the overall disease rate would have been in Drizzletown if the age structure had been the same as in Sunnytown. Thus the directly standardised rate in Drizzletown represents the results of applying the age-specific rates in Drizzletown to a population with the same age-structure as Sunnytown. This latter rate is an 'artificial rate' in that it has not actually been observed. Alternatively, rates in both Sunnytown and Drizzletown could be standardised to the age-structure of a third population (e.g. the age-structure of the county in which the towns are situated). Does it matter which population is chosen as the standard from which the weightings are derived? The weightings will determine the relative contribution of individual age-specific rates to the overall standardised rate and therefore the choice will influence the resulting standardised rates. In terms of interpretation it makes sense to choose a standard population with an age structure

that is close to the populations being compared. For example, if cancer incidence rates are being compared across several European countries, a suitable standard population could be the age structure of all European populations together. It is also desirable to use a population structure to which other investigators have access, thus further facilitating a comparison of rates. One such standard, which has been proposed and widely used, is the World Standard Population (Table 3.1). Directly standardised rates have the advantage that they can be compared with any other rate which has been directly standardised to the same population. A disadvantage, however, is that problems can arise when study populations do not have sufficient numbers of cases to calculate robust age-specific rates.

Example 3.ii

A small seaside town appeared to have a very high prevalence of blindness compared with the rest of the region in which it was situated. However, the town was a popular location for elderly people in retirement. After direct standardisation to the region's population, the prevalence of blindness was still 25% higher suggesting that there was an influence that could not be explained by age.

3.2b Indirect standardisation

In contrast to direct standardisation, which involves the use of population weights, indirect standardisation involves the use of a set of age-specific rates from a standard population. These age-specific rates are then applied to the age structure of the study population to obtain the 'expected' number of cases of disease if these rates had applied. This 'expected' number of cases is then compared to the actual number of cases 'observed' in the study population. The ratio of observed/expected numbers of cases, multiplied by the crude rate in the study population is an indirectly standardised rate and can be compared with the crude rate in the standard population.

Example 3.iii

There were 35 cases of cataract in Oldtown and 23 in Youngtown. If Youngtown's age-specific rates of cataract were applied to the age-structure of Oldtown, 34.4 cases would have been expected in Oldtown. Thus the difference in the number of cases between the towns can be

Table 3.1. The world standard population

Age-group (years)	Population (*n*)
0–4	12 000
5–9	10 000
10–14	9000
15–19	9000
20–24	8000
25–29	8000
30–34	6000
35–39	6000
40–44	6000
45–49	6000
50–54	5000
55–59	4000
60–64	4000
65–69	3000
70–74	2000
75–79	1000
80–84	500
85+	500
All ages	100 000

explained by differences in the age of the residents, rather than by different age-specific rates of cataract.

More commonly, the results are expressed as simply the ratio of observed/expected numbers of cases, usually multiplied by 100, to provide a standardised incidence (or mortality or prevalence) ratio. A standardised incidence ratio (SIR) of 100 implies the incidence rates in the study and standard population are the same, having taken account of differences in the age-structure. An SIR of greater or lower than 100 implies that the rates in the study population are higher or lower than the standard population respectively. Given that in each comparison the 'weight' applied in indirect standardisation is the population structure of the study population, then technically the only valid comparison is between the study and standard populations used to derive the SIR, i.e. two SIRs should not be compared.

Table 3.2. Standardised mortality ratios for selected occupational groups in 1980 (males)[a]

	SMR
Farmers	83
Coal-face workers	143
Chemical workers	104
Welders	124
Plumbers	103
Shoemakers	154
Cotton spinners	133
Brewers	145
Publicans	157
Medical practitioners	79

Note:
[a] The standard population is defined as all working males.

Example 3.iv

Table 3.2 shows a comparison of mortality rates in different occupational groups. Since the age structure of workers in these groups is likely to be different, the comparison has therefore been made using standardised mortality ratios with the standard population as all working persons within the population of interest. It should be noted therefore that each comparison is of mortality in a specific occupational group against the whole working population. Comparisons cannot be made between working groups directly. Thus one *cannot* deduce that the SMR in publicans compared to farmers is 157/83.

3.3 Comparison of rates over time

The above has considered the comparison of rates between populations and between subgroups within a population. A special case of such a comparison is of monitoring rates in a single population over time. Changing crude rates within a population could be a reflection of changes in age-specific rates or it could be a reflection of a population getting older with time. Thus it is important to make comparisons with respect to a defined 'standard population' either by direct or indirect age standardisation. An example

Table 3.3. Cancer mortality in Good records county 1960–1990

Year	Number of cases	Crude rate (per 100 000 person–years)	SMR[a]
1960	6672	270.5	100
1965	7858	296.7	109
1970	8391	326.8	120
1975	10 693	419.3	143
1980	11 576	463.8	154
1985	12 359	492.6	159
1990	13 015	539.4	160

Notes:
[a] Standard Year 1960: SMR = 100.

using the latter method is shown in Table 3.3. Here, information is given on cancer mortality in a population for selected years between 1960 and 1990 in Goodrecords county with 1960 defined as the 'standard population'. Thus the age-specific rates in 1960 are applied to the age structure of the population in subsequent years in order to calculate standardised mortality rates (SMR). It can be seen from the tables that, although the number of cancer deaths and the crude rate have almost doubled over the 30-year period, the increase in the SMR is more modest (from 100 to 160). Thus although there has been an increase in crude mortality over the study period, some of this increase can be attributed to a changing age structure (ageing) of the population.

An alternative approach to comparing rates would have been to directly standardise them. The standard population could be from one of the years under study in Goodrecords county or to an 'external' population such as the entire country or to an imaginary population such as the World Standard Population.

3.3a Age-specific rates

The evaluation of changes in rates of disease in a population, in addition to standardisation for age, should include examination of changes in age-specific rates over time. The pattern of age-specific rates can give some indication of possible aetiology.

Figure 3.1 Cancer of the lung and pleura (age-specific mortality rates per 100 000 person–years vs. median year of death).

	Age-group (years)
A	25–29
B	30–34
C	35–39
D	40–44
E	45–49
F	50–54
G	55–59
H	60–64
I	65–69
J	70–74
K	75–79
L	80–84

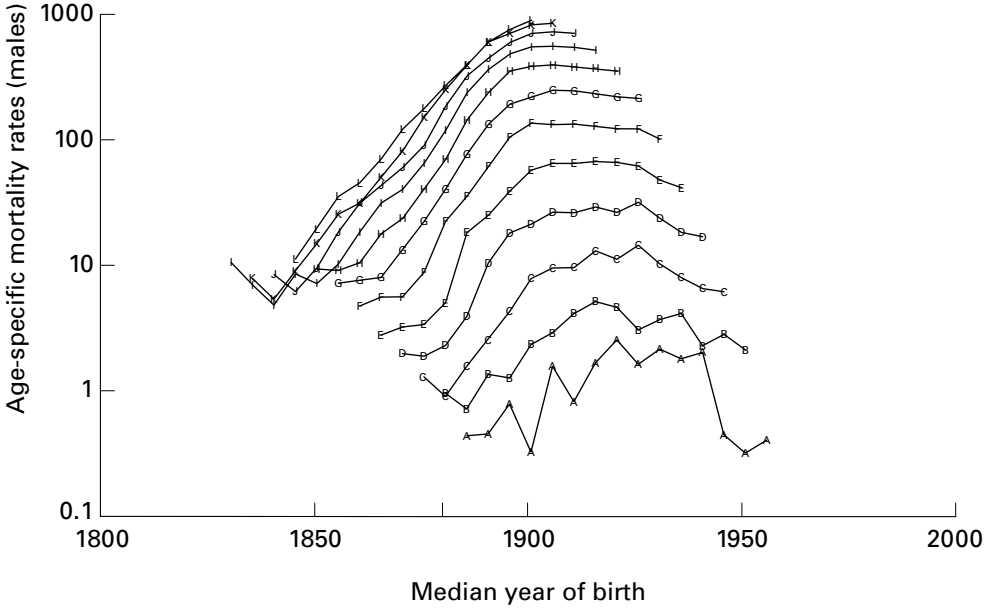

Figure 3.2 Cancer of the lung and pleura (age-specific mortality rates per 100 000 person–years vs. median year of birth).

	Age-group (years)
A	25–29
B	30–34
C	35–39
D	40–44
E	45–49
F	50–54
G	55–59
H	60–64
I	65–69
J	70–74
K	75–79
L	80–84

Example 3.v

A decrease in mortality rates from bladder cancer was observed in Industrialand. On examination of cancer mortality rates within five year age-groups the decrease was noted to have occurred at all ages at the same calendar period. Such effects are known as 'time-period effects'. It was believed that they may have resulted from an improvement in treatment or from an artefactual cause such as a change in registration procedure or coding rules. When these reasons were discounted the most likely explanation was thought to be a prior chemical leak into the local river from a large factory. Everyone, irrespective of age, would have been exposed to this at the same calendar time period.

Changes in rates may show a time-period effect (see Example 3.v). Alternatively, the change in rates may manifest as a 'cohort-effect'. Such an effect is associated with different risks of disease between generations and are a consequence of changes in long-term habits such as diet. Figure 3.1 shows age-specific mortality rates of lung-cancer between 1910 and 1985. The mortality rates are plotted according to the year in which death occurred. Mortality rates are shown to initially increase across this time period. Thereafter, rates have decreased: at younger ages the rates began to decrease earlier than at older ages. In fact, in the oldest age-groups rates were still increasing at the end of the study period. Evaluation of these trends is made easier if, instead of a death being 'attributed' to the time period in which it occurred, it is 'attributed' to the birth cohort to which the dead person belonged. By plotting age-specific rates according to period of birth rather than to date of disease onset/death, one can evaluate whether there is a systematic change in rates according to birth cohort. Figure 3.2 plots the same data as Fig. 3.1 except that mortality rates are plotted according to the birth cohort to which they refer. Thus, for example, persons aged 80–84 in 1911–15 are plotted against the birth cohort centred on 1831 (i.e. persons aged 82 years in 1913 will have been born in 1831). When plotted as such, the data demonstrate a clear 'birth-cohort effect' in changing rates. Mortality rates have increased at all ages for persons born across periods between 1850 and 1900. Thereafter, mortality rates have stabilised and, in the most recent birth cohorts, rates show a fall at all ages.

There are statistical methodologies available (age-period–cohort models) to more formally evaluate the role of period and cohort effects on changing rates, but these are beyond the scope of the current text.

Part III

Studying associations between risk factors and disease

4

Which type of study?

Having settled on a study hypothesis and/or the required measure of disease occurrence, the subsequent decision is which type of study is appropriate. The decision will be based not only on methodological but also on practical considerations. For example, the most appropriate study may be too expensive or take too long to provide an answer. In such circumstances a compromise will require to be made – to undertake a study which can be conducted within the budget and time available and which delivers information which is suitable for answering a hypothesis or provides a useful measure of disease occurrence.

4.1 The ecologic study

The simplest type of study is an ecologic study (also called a correlation study). In this type of study information is collected not on individuals but on groups of people. The unit on which measurement is made may be for example schools, towns, countries, etc.

Example 4.i

Investigators wished to study the hypothesis that the risk of melanoma skin cancer was related to the amount of exposure to ultra-violet rays. They therefore gathered information on the incidence of melanoma skin cancer in 11 regions or countries in the northern hemisphere to determine whether this was related to the geographical latitude of the country (see Fig. 4.i).

The advantage of this type of study is that it is generally inexpensive and quick to conduct. This is especially true if the data on disease and exposure is available from routine sources as in Example 4.i. Even if, for example, the information on level of exposure is not available, the effort to collect this on

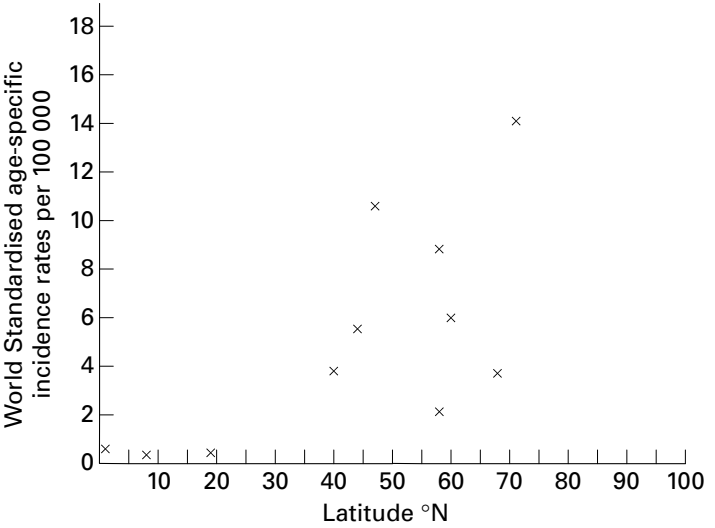

Figure 4.i Incidence rates of melanoma skin cancer.

the aggregated units will generally be less than collecting exposure information on a much larger number of individuals.

The outcome of such a study will be to conclude only that the study either supports or does not support a hypothesis about a relationship between exposure a and disease b. It may also provide some information on the potential type and strength of any relationship which is found. However, it does have serious weaknesses. There is very rarely information on factors which could potentially confound an observed relationship between exposure and disease (see Chapter 18 for a discussion of confounding), and this is the greatest drawback to such studies (Example 4.ii).

Example 4.ii

A study has reported a (negative) correlation, at country level, between the average number of cars owned by families and the level of dental caries. However, the researchers can think of no biological hypothesis which links the two. Instead, they conclude that it is likely that the level of dental caries is influenced by one or more other factors which, if they are a reflection of socio-economic status would, on a country level, show an association with car ownership.

Conversely, there may be no association observed at the aggregate level when, in fact, there is a relationship at the individual level. This may again

be a result of confounding or it may be a result of the exposure measure used being a poor proxy for the actual exposure of interest (Example 4.iii).

Example 4.iii

Investigators wished to examine the relationship between alcohol consumption and mortality from cardiovascular disease in European countries. Information on per capita consumption of alcohol was not available from individual countries. There was however available data on government receipts from excise duty on alcohol. Although in theory this may have been a suitable proxy to measure national consumption, differences in methods of collecting the data with time and between country, and the variable quality of the available data meant that it was not suitable for use.

Given that an association is demonstrated in ecologic studies, and even if information is available on potential confounding factors, it does not guarantee that the relationship holds at the level of the individual. For example, there may be a positive association between per capita fat consumption in a country and the incidence of a disease, but *within* each country it is possible that a higher level of fat consumption results in a lower risk of disease. Drawing inappropriate conclusions about the relationship between an exposure and disease in individuals from ecologic studies has been termed the *ecologic fallacy*.

4.2 The migrant study

In animal studies investigators have the possibility of directly controlling the environment, personal behaviour (e.g. diet) and even genetic factors which are hypothesised as being important in disease occurrence. Rarely will there be such an opportunity for epidemiologists studying free-living populations. Individuals are free to decide on their diet and to choose, for example, whether they smoke, drink alcohol or take regular exercise. Intervening to effect such changes in such lifestyle factors is difficult, even within controlled studies. Therefore the studying of migrants, moving from one country to another, provides an interesting natural experiment which can provide the epidemiologist with important information on the influences on disease occurrence. This is particularly true when considering the relative roles of genetic and environmental factors on disease. The process of migration

involves no change to the genetic influence on individuals' risk of disease but it will almost certainly result in changes to some environmental factors such as climate and diet.

Let us assume that large groups of individuals have migrated from country A to country B and that these countries have low and high rates of disease X, respectively. If the migrants manifest high rates of disease in their new country, it may be assumed that some environmental factors are important in determining disease risk. In contrast if the group maintains a low risk of disease X, there are several possibilities, including (a) genetic factors are the principal influence on disease risk, (b) although environmental factors are important, the migrants have maintained a lifestyle that is very similar to their country of origin, (c) lifestyle factors are important but the migrants were exposed to them at too old an age for them to have an important influence on disease risk.

Further information may be obtained by studying not only the migrants themselves but successive generations of migrants in the new home country. However, in such subjects there may be changes not only in environmental factors but also genetic factors.

Example 4.iv

Breast cancer incidence in the United States is relatively high in comparison to other countries. Investigators examined the incidence of breast cancer in recent migrants to the United States. Those who had come from countries with lower incidence rates, exhibited incidence rates that were higher than their country of origin, but not as high as those of the United States. Those migrants who had come from countries with higher incidence rates still experienced rates that were higher than those in the whole United States population, but not as high in comparison to their country of origin. It was concluded therefore that the 'new environment' in the United States did alter the risks of migrants developing breast cancer. Factors which could contribute to this may be changes in diet, use of hormonal preparations, etc.

In any study of migrants, however, one concern is how representative the migrants are of their original country. Intuitively, it seems likely that those who are ill or have poor levels of general health would be less likely to migrate. They may also differ in other individual ways (e.g. personality) and their lifestyle. Such factors complicate the direct comparison between disease rates in their old and new home countries. The studies can also be

difficult to conduct. Effective ways of identifying the migrants in their new home country would be necessary and following them prospectively to determine disease onset. It obviously becomes easier if the information is collected on a routine basis (e.g. death information or cancer occurrence) with information on their country of birth.

Therefore, ecologic and migrant studies can provide important initial clues to the type of factors which may be implicated in the aetiology of disease. However, they will not provide definitive confirmation or rejection of hypotheses and they are unlikely to be the only type of study undertaken. Instead it would be necessary to proceed to studies which undertake data collection on individuals. There are three main types of study for investigating disease/risk factor associations:

(a) Cross-sectional
(b) Case-control
(c) Cohort (either retrospective or prospective).

These types of study vary in the criteria on which participants are selected (e.g. disease state, exposure state, or neither) the timing of collecting information on exposure and disease. They also vary considerably in the time taken to conduct and their resource implications.

4.3 The cross-sectional study

The cross-sectional study collects information about current disease state and/or current exposure status. In some instances information may be collected about disease over a period of time, e.g. the past month or year. It is the method therefore to determine the point or period prevalence of a disease or attribute, and serial cross-sectional studies can be used in certain circumstances to measure incidence.

Example 4.v

A questionnaire was sent to 500 individuals aged over 65 years, selected from attenders at a single medical clinic. They were asked about whether they had fallen in the home during the past month. In addition, information was collected on general health, current medication and aspects about the layout of their home (e.g. floor coverings). The investigators wished to determine the 1-month period prevalence of falls and whether there was any relationship between falling and the risk factors about which information was collected.

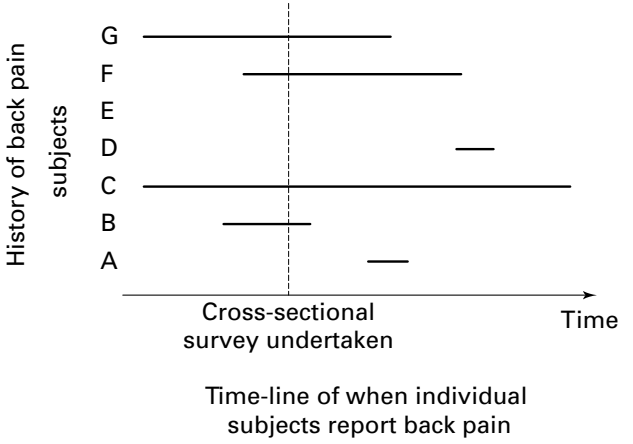

Figure 4.1 Back pain.

The principal advantages of the cross-sectional study are normally that it is relatively inexpensive and can be conducted within a short timescale. Further, given that information is collected about disease state and exposures currently, the problem of recalling past events is less than if subjects were asked about exposures and disease state in the medium or distant past (Example 4.v).

The principal disadvantage is the lack of information on temporality (Example 4.vi). In addition, a cross-sectional survey will preferentially identify chronic cases (and survivors). Those with only short-lived symptoms or disease, or who have died shortly after disease onset, are less likely to be 'in disease state' at the time of survey. In Fig. 4.1, of the seven back pain subjects on which data is presented, four subjects (B,C,F and G) reported back pain at the time of survey. All these subjects have more long-standing symptoms in comparison to the other subjects who were back pain free at the time of survey. This must be borne in mind when interpreting relationships with risk factors.

Example 4.vi

A cross-sectional survey was conducted amongst adult residents of a single town to determine whether mental disorder was more common amongst lower social classes. The study found that mental disorder was twice as common in the lowest social class in comparison to the highest social class. In interpreting these results, the investigators considered that this observation was consistent with the hypothesis that some aspect of the circumstances and lifestyle of those of low social class predisposed them to the onset of a mental disorder. However, they

also realised that the consequences of having a mental disorder may, through changing or loss of employment, result in their being classified as low social class. The study therefore, while demonstrating an association, has not established the temporal order of events.

In addition to measuring prevalence and determining the association of risk factors with disease, cross-sectional studies may be the method through which different study designs are with achieved. For example, a cross-sectional study may be conducted to identify cases of a disease and controls for a case-control study. A cohort study may consist of a series of cross-sectional studies. A series of cross-sectional studies may also be used to determine time-trends in the occurrence of disease, and when investigating possible clustering of disease in time and geographical location. They may also be the method through which information is collected on participants in a migrant study.

4.4 The case-control study

A further design to investigate the relationships between potential risk factors and disease is the case-control study. In this type of study cases of disease X are identified (cases) together with a sample of those without disease (controls), The cases and controls are then compared with respect to their exposure to risk factors. Exposure information may relate to current and/or past exposures. In situations where information cannot be obtained from the cases (e.g. very severe illness or death), it may be appropriate to identify a *proxy respondent* (Example 4.vii) for the case.

Example 4.vii

All individuals admitted to a local coronary care unit with myocardial infarction were enrolled as cases in a case-control study and information gathered on their diet (including alcohol) during the previous year. The authors realised that only subjects who survived the myocardial infarction could be enrolled. Therefore, they also approached the relatives of individuals who were resident in the catchment area of the hospital and had died with an MI before reaching hospital. Similar information to that from surviving cases was gathered from the relatives of dead cases.

The case-control approach is particularly suitable when the disease is rare, and when the aim is to investigate the effect of many exposures on one

disease. Subjects will be required to recall past events ('exposures') and there will be an issue of firstly whether they will be able to recall such exposures accurately and secondly whether recall is similar in cases and controls. The issue of recall bias is discussed in Section 19.4. For some exposures, concerns about recall may be able to be overcome by the use of documentary records (Example 4.viii).

Example 4.viii

A case-control study amongst adults which gathered information relating to hospitalisations prior to the age of 16, was concerned about poor recall. In particular, the study investigators were worried that cases may be more likely to remember hospitalisations which occurred than controls. Therefore, in a subsample of both cases and controls, medical records were examined to determine whether there was any evidence for differences in recall between cases and controls.

Particularly if the cases used in a study are of new onset and subjects are being asked about past exposures, there is not the same concern about temporality of events as in cross-sectional studies. In many situations it will be clear that the exposures reported predated disease onset. However, if care is not taken there is still the possibility that exposures measured in a case-control study will be a consequence of disease (Example 4.ix).

Example 4.ix

A university research department carried out a case-control study of stomach cancer, examining the hypothesis that frequent fruit and vegetable consumption was protective for the occurrence of disease. Cases were subjects recently diagnosed with stomach cancer and controls were free of the disease and had the same distribution of ages, gender and area of residence as the cases. All subjects were asked about their fruit and vegetable consumption during the past year, and the analysis of results showed the surprising result that individuals with stomach cancer had consumed almost double the amount of fruit and vegetables as the controls. Further investigation revealed that since the cases had been ill for a considerable time before diagnosis most had changed to a more healthy diet because of concerns about their health.

The crucial aspects in any such study will be the definition and identification of cases and controls and this is discussed in detail in Chapter 8. In order to ensure the comparability of information from cases and controls study pro-

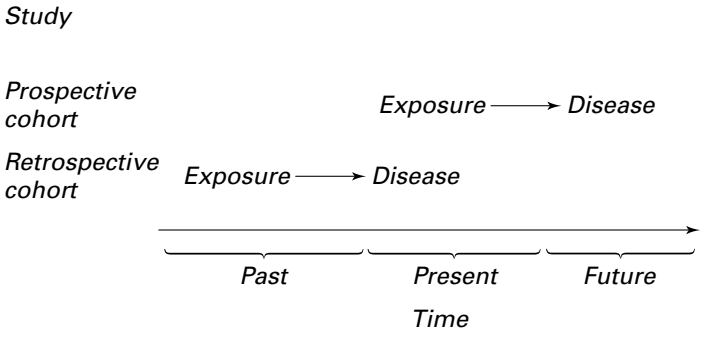

Figure 4.2 Measuring disease and exposure in cohort studies.

cedures should ensure that the collection of information is as similar as possible. This will include the information being collected by the same techniques (e.g. interviewer, questionnaire) in the same setting (e.g. hospital, home) and at the same calendar time. Where possible, if the study involves face-to-face interview, it should be ensured that the interviewer is either not aware of the case-control status of the subject or at least is not aware of the specific hypotheses under investigation. This reduces the possibility of observer bias, which is discussed further in Section 19.4.

In most situations the case-control study will be able to be conducted in the short to medium term. The rate-limiting step in study conduct will usually be the identification of controls, and for particularly rare diseases it may be necessary to have recruitment on a national or international basis, in order to identify enough cases within a reasonable time period.

4.5 The cohort study

A cohort study involves one or more groups of subjects, defined by their exposure status, being followed through time to identify an outcome of interest (usually disease onset). The aim is to determine whether initial exposure status influences risk of subsequent disease. Two particular types of cohort study are *the prospective cohort study* and the *retrospective cohort study*. The only difference between these approaches is with respect to the timing of collecting exposure and disease information (Fig. 4.2).

In the prospective approach cohort(s) are identified by their exposure status presently and are followed up to determine any future disease onset.

The retrospective approach identifies the exposure status of cohort(s) in the past and in a parallel sense they are 'followed-up' until the present time, when their disease status is determined. The latter approach will undoubtedly be quicker and less expensive but may not always be appropriate.

A retrospective study will:

- rely on there being sufficient information available on past exposure status and on being able to determine current disease status;
- involve identification of cohort subjects on national (disease) registers or individually tracing subjects;

and may need to consider:

- whether information on changes in exposure status is available;
- what information is available on other risk factors for the disease which are potentially associated with exposure status i.e. confounding factors (see Chapter 18).

The major advantages of a prospective study are that it can be determined which exposure is measured and how; if and when change in exposure status is measured; procedures to allow future identification and tracing can be implemented; the nature of outcome measures can be determined. Consequently, however, the time-scale of the study will be considerably longer and for some outcomes such as cancer and death, may be as long as 20–30 years.

Overall, the cohort approach is suitable when the disease outcome is common, and is particularly suited to determine the effects of exposure on a variety of disease outcomes. Given exposure status when measured will refer to 'current status' and this is measured prior to disease onset, the issues of temporality of disease/exposure, problems with recall and particularly recall bias will not, in general, affect cohort studies. For this reason, in particular prospective studies, they are considered the most powerful methodology in epidemiologists' armoury.

In some instances it is beneficial to combine the latter two approaches: namely a case-control study nested within a cohort study (known as a *nested case-control study*). Assessment of exposure, or one aspect of exposure, may be very time-consuming and costly and instead of undertaking measurement on everyone in a cohort, it may be more prudent to wait to determine which subjects in a cohort develop the disease under study (i.e. the cases for the nested case-control study). Thereafter, a control group of subjects could

Table 4.1. Choice of strategy

Consideration	Study choice
Disease rare	Case-control study
Investigate multiple exposures on single disease	Case-control study
Investigate multiple outcomes	Cohort study
Accurate assessment of exposure status	Prospective cohort study
Anxiety about stability of exposure status	Prospective cohort study

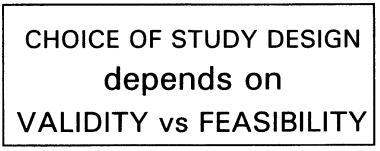

Figure 4.3

be selected amongst those from the original cohort who had not developed the disease. Examples of this may include collecting serum on all members of the cohort but only analysing the 'cases' and selected 'controls'. If there is no efficiency or financial gain in delaying determination of exposure status, then the prospective cohort design should be used.

4.6 Choice of study design

It is apparent that many questions can be addressed by a number of different study designs and thus the decision has to be made as to the most appropriate design to answer a particular question. Figure 4.3 indicates that the choice is often between validity, i.e. obtaining the most accurate answer, and feasibility, i.e. obtaining an answer. Common sense dictates that one without the other negates the value of undertaking the study and also that the two are not mutually exclusive. The decision is normally based on an appraisal of both scientific and logistic considerations. An overview is provided in Table 4.1. There are a number of broad considerations:

- Ecologic and migrant studies are primarily used to generate hypotheses about the aetiology of disease. If appropriate information is routinely collected, they can be conducted quickly and at low cost.
- Cross-sectional studies generally are able to determine only *associations* between risk factor and disease. They can also be the method through which other types of study are conducted.
- The cohort approach allows identification of multiple disease outcomes from a single exposure, whereas the case-control approach allows identification of multiple exposures associated with a single disease entity.
- The lack of quality control of data from a retrospective cohort study, particularly on exposure status, would support a prospective approach. Similarly, data may be sufficient for the primary exposure of interest, but may be lacking on possible confounders (see Chapter 18) that need to be considered.
- The prospective cohort approach, in theory, also permits setting up systems to notify change in exposure status during the follow-up period, an option that may be lacking in a retrospectively derived cohort with only 'point' data on exposure.
- Prospective cohort studies suffer from the problems of potential and unknown loss-to-follow-up rates: it is increasingly difficult to track down individuals after a time interval. Assessment of disease status may then be impossible from within the study.
- Cohort studies are substantially more expensive than the smaller case-control approach. The rarer the disease the more impracticable the cohort approach becomes. Studies that involve population screening to derive either current or future cases are more expensive than those that can utilise an existing morbidity recording system, such as a population-based cancer register.
- Time is relevant in so far as public health questions that require an immediate answer, for example regarding risks from current occupational exposure, might not be able to wait for the 10 years it might take for a prospective study to reach an answer.
- The availability of data may dictate the choice available.

4.6a Examples of choice of study design

Below are a number of examples of studies, with suggested designs, as a guide to how a choice is made.

Example 4.x

The hypothesis was raised that working in a particular occupation led to an increased risk of colon cancer. Excellent occupational records existed going back over a number of years. The most obvious approach to the study was to use such records to establish a retrospective cohort and determine who had developed cancer subsequently. This would give results far quicker than mounting a prospective study. The latter strategy would, however, have permitted the additional collection, at 'baseline', of other potential confounding factors such as diet, smoking and family history.

Example 4.xi

The possibility was raised that a particular serum marker early in pregnancy might predict the future development of a relatively rare congenital anomaly. It was decided to mount a 'nested' case-control study storing serum from a large number of pregnant women at their first ante-natal visit. Subsequently, the sera from all babies born with this anomaly would then be assayed together with a random sample of sera from normal infants. This approach made efficient use of an expensive assay technique.

Example 4.xii

A gynaecologist wished to examine the effectiveness of cervical cancer screening in reducing the incidence of invasive carcinoma. She approached this by undertaking a case-control study comparing the history of cervical smears (the exposure) in a series of women presenting with invasive carcinoma (the cases) to that in a series of normal women from the population (the controls). The hypothesis was that the cases would be less likely to have been screened. In this example, a case control study was being used to examine the effect of a therapeutic or preventive intervention. Although such questions are more ideally addressed by a randomised prospective trial, as suggested earlier the case-control study can give a quick answer for relatively rare events, particularly for studies in which for ethical or other reasons randomisation may be difficult.

Example 4.xiii

A psychiatrist wished to examine the influence of strong family support on reducing the suicide rate after hospital discharge with acute schizophrenia. He felt that the measurement of family support had to be prospective and opted for a prospective cohort study. The option of a case-control study – by comparing the family support (the exposure) in a series of schizophrenic patients who had committed suicide since hospital discharge (the cases) with that in a series of schizophrenic patients who had survived (the controls) – was dropped given the problems in retrospectively assessing family support in the face of a recent suicide.

Example 4.xiv

A cardiologist wished to examine the influence of the severity of angina on the subsequent risk of first myocardial infarction. He decided to undertake a prospective cohort study following up two groups: an 'exposed' group with anginal attacks occurring on walking on the level and an 'unexposed' group whose attacks occurred only on walking up an incline. In this example, the prospective cohort approach was used to assess the influence of specific risk factors on disease prognosis, rather than on disease development per se.

Example 4.xv

A general practitioner wished to examine the effect of salt intake on hypertension. In a pilot study, she reached the conclusion that a randomised trial was impracticable as the compliance in subjects allocated to a low-salt diet was likely to be low, and further, a proportion of subjects in the normal diet group, at their own instigation, attempted to reduce salt. She therefore conducted a cross-sectional population survey and found that there was a modest (positive) relationship between salt intake and blood pressure levels. This provided some support for a relationship but because of concerns about methodological aspects of the cross-sectional study she thereafter proceeded to conduct a prospective cohort study.

All the patients had their baseline salt intake assessed during a 24-hour urine collection. She then followed-up the patients to determine who could be classified as hypertensive in the next five years.

Example 4.xvi

The hypothesis was proposed that there was an increased risk of myocardial infarction following the death of a spouse. A retrospective case-control study was a potential strategy given the relative ease of verifying the exposure. However, this exposure was likely to be rare. It was then necessary to undertake a more expensive and time-consuming prospective cohort study.

Example 4.xvii

A psychiatrist wished to examine the relationship between anxiety and depression and persistent headache. He firstly conducted a cross-sectional survey in which he requested information on whether respondents commonly experienced headaches, and used an instrument to determine levels of anxiety and depression. This showed a strong relationship. He hypothesised that this may be because having headaches often may lead to anxiety and depression, rather than the converse. He therefore conducted a prospective cohort study, measuring levels of anxiety and depression amongst people who did not report headaches and followed them over 1 year to determine who developed symptoms. Having high levels of anxiety and/or depression resulted in an increased risk of reporting headache at follow-up.

5

Which measure of association?

Much epidemiological endeavour is directed towards attempting to discover the aetiology (i.e. the causes) of particular diseases, with a view to prevention. Chapter 4 discussed the options available in terms of epidemiological study design. In practice, most diseases do not have a single identifiable cause, such as infection with a specific micro-organism or exposure to a particular toxin or physical trauma. By contrast, it appears that most diseases are multifactorial and represent the effects of a combination of genetic, constitutional and environmental factors. Thus, most exposures investigated are neither *sufficient*, i.e. they will not cause disease on their own, nor *necessary*, i.e. the disease can occur in the absence of that exposure. A simple example is that although smoking is common and is a risk factor for the development of lung cancer, not all individuals who develop lung cancer smoke and not all smokers develop lung cancer. Epidemiological studies, and in particular their subsequent analysis, are therefore aimed at quantifying the level of increased risk when exposed to a particular factor. The *effect measure* which can be obtained to quantify the strength of the association, varies according to the type of study conducted. These are outlined in Table 5.1.

5.1 Relative risks

5.1a Risk ratios

The first measure is a ratio of the prevalence of disease (or cumulative incidence) in two population groups. This is therefore a ratio of risks and is referred to as a *risk ratio*. Such a measure can be derived, for example, from a cross-sectional study.

Table 5.1

Study design	Selection subjects by status	Information collected on exposure	Information collected on disease
Cross-sectional	No	Current	Current
Case-control	Disease	Past	Current
Cohort			
– prospective	Exposure	Current	Future
– retrospective	Exposure	Past	Current

Example 5.i

A cross-sectional study in a small town collected information from individuals on recent gastrointestinal symptoms and the source of their household water supply. Amongst those with source A, the prevalence (risk) of diarrhoea was 0.11 while the prevalence amongst individuals with source B was 0.04. The risk ratio of 2.8 (i.e. 0.11/0.04) gives an estimate of the magnitude of increased risk associated with source A.

When comparing the risks of disease in population A against population B (i.e. risk in A/risk in B), values above 1 indicate a greater risk in A, values below 1 indicate a greater risk in B, while a value of 1 implies equal risk in both groups.

5.1b Rate ratios

A cohort study measuring incidence (density) rates allows the rates in two population groups to be compared directly in an analogous manner to the above. Given that the comparison is of two rates, this ratio is referred to as a *rate ratio*.

Example 5.ii

A prospective cohort study at a single medical centre identified all strokes which occurred amongst registered patients. The incidence rates amongst those who were and who were not diabetic were 4.7 and 1.6, respectively, per 1000 person–years of follow-up. The rate ratio of 2.9 (i.e. 4.7/1.5) indicates that that those with diabetes had almost three times the risk of a stroke in comparison to those who were not diabetic.

Analogous to the interpretation of risk ratios when comparing the rate of disease in population A against population B rate ratio (i.e. rate in A/rate in B), values above 1 indicate a higher rate in A, values below 1 indicate a higher rate in B, while a value of 1 implies equal rates in both groups.

In many texts the term relative risk is used as an overall term to describe the ratio of occurrence of events in two populations and would include risk ratios and rate ratios. Although superficially acceptable, the difference between risk and rate ratios is more than semantic, because the statistical methods associated with their calculation, and for example an associated confidence interval, differ.

5.2 Odds ratios

Relative risks derived from cross-sectional and cohort studies are intuitively attractive measures to estimate since they reflect what the epidemiologist generally wishes to know. What is the change in rate (risk) of disease associated with being exposed to a particular factor? They are also easily communicated to a wider audience, e.g. 'persons who drink more than 20 units of alcohol per day double their risk of disease X'.

In a case-control study, two groups of subjects are selected according to disease status – a group with disease and a sample of those without disease. In such circumstances it will, in general, not be possible to determine the rate (or risk) of disease amongst subjects exposed to a particular factor in comparison to those not exposed. Thus the rate ratio (or risk ratio) cannot be calculated.

Instead one can calculate the *odds* of exposure amongst the cases and compare this to the odds of exposure amongst the controls. The odds of exposure in a group is simply the number exposed in a group divided by the number not exposed. If the odds of exposure amongst subjects with disease (cases) is determined and similarly the odds of exposure amongst subjects without disease (controls), an *odds ratio* can be calculated (odds of exposure amongst cases/odds of exposure amongst controls). If this ratio is above 1, it implies that cases are more likely to be exposed to a particular factor than controls, and if the ratio is less than 1, the opposite is true. If the ratio is close or equal to 1 it implies that the odds of exposure are very similar in the two groups.

Example 5.iii

Amongst 200 patients with acute appendicitis, 40 had a recent history of urinary tract infection (UTI). The odds of exposure (to UTI) were therefore 40/160 or 0.25. By contrast in 120 normal individuals, only 12 had a recent history of UTI: an odds of exposure of 12/108 or 0.11. The odds ratio of 2.25 indicates that the odds of those with appendicitis of having a recent UTI was more than twice those without appendicitis.

In many circumstances the odds ratio will be a good approximation to the relative risk.

5.3 Attributable risks

An alternative approach to assessing the risk from an exposure may be derived from the difference between the risk in an exposed population and that in an unexposed population. This is referred to appropriately as the risk or rate difference. This difference may be of greater value to those in public health and clinical medicine and represents the absolute risk that can be ascribed to the exposure. For this reason, this difference is frequently termed *attributable risk*.

Example 5.iv

Using the same data from Example 5.ii above, subtracting the rate of strokes in non-diabetics from that in diabetics gave a rate difference of 3.1/1000 person–years of observation. This represents the actual rate of stroke in diabetics that can be explained as a result of their being diabetic, i.e. in excess of their background risk.

The public health physician may, in addition, wish to estimate the proportionate contribution made by the exposure variable in explaining all the cases in a given population. This depends both on the risk associated with the exposure and the proportion of the population exposed; this measure is known as the *population attributable risk fraction*. The approach to the calculation of this is given in Chapter 17 (Example 17.x).

Example 5.v

Using the data from the above example and assuming a 2% frequency of diabetes in the total population, we find that the population attributable risk fraction for diabetes in causing

stroke is itself only 4%. Thus, if diabetes were 'eliminated', then there would be a reduction of only 4% in the number of strokes observed in the population.

5.4 Precision of measures in association

As studies are normally undertaken on samples, any measure of association derived from the samples is subject to error in its ability to describe the effect in the populations from which the samples were derived. Readers will be aware, for example, that an estimate of the error around a sample mean is obtained by calculating the *standard error* (*SE*), with the conventional practice being to present the range produced by adding and subtracting $2 \times$ SE (or, more precisely, $1.96 \times$ SE) to the mean obtained from the sample. This range is referred to as the *95% confidence interval* and the basis for its calculation may be found in any standard textbook on medical statistics. Confidence intervals can also be calculated for the measures of association discussed above. The actual formulae used to calculate the 95% confidence intervals for the epidemiological measures of association and notes on their interpretation are given in Chapter 17.

5.5 Categorisation of exposures

Epidemiological enquiry attempts to assess the association between exposure to a risk factor and a disease. For the sake of simplicity most epidemiological textbooks imply that exposures are dichotomous (have only two forms, e.g. yes/no), whereas in practice the questions of interest relate to the size of the risk with increasing exposure. The available choices for considering exposures are shown in Table 5.2. Most exposure variables in practice are not easily dichotomised and indeed adopting a $+/-$ split does not permit analysis of the crucial 'dose–response' effect. Some exposure variables are, by their nature, ranked or ordinal, such as activity level of occupation in the table. In that instance, it would be of interest to compare the risk of disease in each of the three active groups separately from the sedentary group.

Many exposure data collected are of the continuous type, i.e. an individual can take any value. The exposure can then be classified according to some predetermined 'sensible' categories, as for smoking in the example given in Table 5.2. For some variables there do not appear to be logical (in a biological sense)

Table 5.2. Categorisation of exposures

Type	Exposure example	Appropriate categories
A. Dichotomous (yes/no)	Smoking	Ever Never
B. Ranked	Occupational activity	Sedentary Mildly active Moderately active Very active
C. Continuous		
Strata defined on 'biological' basis	Smoking	Never smoked Smoked ≤1 year >1 and ≤5 years >5 and ≤10 years >10 years
Defined on statistical basis	Body mass index	Bottom third Middle third Upper third
Exposure treated as continuous	Age Serum cholesterol	Actual age Actual serum cholesterol values

divisions and thus the values are ordered and split into strata, with equal numbers of individuals, typically thirds (tertiles), fourths (quartiles) or fifths (quintiles). This approach of forcing equal numbers in each category is also advantageous from the statistical viewpoint as it ensures reasonable numbers in each exposure group. Alternatively the actual values themselves can be considered and risks calculated, for example, per one year increase of age, or per 1 mmol/L increase in serum cholesterol. In many ways such an approach is better as it 'uses all the data', but most statistical methods of generating these risks would assume that the risk is linear, which may not be the case. As an example, the relationship between diastolic blood pressure and stroke is exponential.

Part IV

Selection of populations and samples to study

6

Studies of disease occurrence. I: Identification of the population

This chapter reviews the options for population selection in undertaking investigations aimed at estimating the occurrence of a disease in a target population. The same principles apply, however, if the object of the study is to investigate the occurrence of a risk factor, such as cigarette smoking, or other human attribute. The first requirement is to identify the target population to which the occurrence estimate will apply.

Example 6.i

In England and Wales, data are routinely collected on some infectious diseases and on (nearly) all cancers based on national notification systems. Incidence rates are calculated using national population data derived from the census as the denominator. As censuses occur 10 years apart, adjustment to the denominator has to be made to take account of births, deaths and migrations. For the sake of convenience, disease events occurring in any single calendar year are related to the presumed population alive in the middle of that year (30 June).

In practice, most epidemiological surveys use a variety of sub-national populations (Table 6.1). These may be defined by geopolitical boundaries such as a town or electoral district. Alternatively, a true geographical boundary may be used derived either from a map or, where possible, based on postcodes. Another approach is to use populations registered for other purposes, such as those registered with a general practitioner or other health scheme. Some epidemiological studies target particular age groups such as children or the elderly. In such groups other possibilities are available.

Example 6.ii

In a study to determine the prevalence at birth of congenital dislocation of the hip, the investigators used local birth registers to identify their target population.

Table 6.1. Target general population groups for measuring disease occurrence

1.	*Geopolitical populations*
	Town/city/village population register
	Electoral/local government district
2.	*True geographical population based on map of dwellings*
3.	*Other populations*
	Health registers e.g. general practice
	health insurance
4.	*Age restricted populations*
	Birth registers
	Pension registers

The choice of the appropriate target population is determined by four factors: representativeness, access required, population data accuracy and the size. These are discussed below and issues of sample size are considered in Section 6.4.

6.1 Representativeness

An epidemiological survey of the prevalence of a specific disease conducted in a suburb of a large city can give an estimate of the prevalence only in that suburb. Although it is relatively easy to allow for age and sex differences between the sample studied and a wider population (see Chapter 18), other differences due to socio-economic and related factors might suggest that such estimates cannot be applied more widely, even to the neighbouring city itself. In theory, most epidemiological investigations have as their unstated aim the derivation of 'national' estimates. Although governmental and related institutions often attempt national sample surveys for specific purposes, this approach is rarely practical or available to other investigators. In practice, therefore, for most diseases, published estimates of incidence and prevalence are based on local surveys. This potential lack of representativeness is never an easy dilemma to solve. The most practical approach, if the aim is to present a 'national' view, is to choose a population that is not

extreme, for example in its socio-economic distribution, and to collect the data on those aspects that will either allow formal statistical adjustment of the results or enable the 'reader' to interpret the results. National census data, for example, might give a clue as to the study population's ranking in relation to some key variables, such as unemployment or ethnic mix. Ultimately, however, the estimates derived can strictly be applied only to the population from whom the sample surveyed was selected. Alternatively, the same methodology can be used in a number of geographically and socially disparate populations.

Example 6.iii

A major study was carried out in the United Kingdom to examine the epidemiology of ischaemic heart disease in middle-aged males. The investigators used identical methods to study population samples in their sample of UK towns.

6.2 Access

As discussed below, there are two approaches to measuring disease occurrence. These are: (i) the *catchment population* approach, which counts the number of ascertained cases (based on normal clinical referral and diagnosis) and relates that *numerator* to the *denominator* or catchment population served by those clinical facilities; and (ii) the *population survey* approach, where each member of the study sample is individually investigated to accurately classify their disease status. The choice of the target population is determined by which of these approaches is to be used. Thus, in the catchment population approach it is only necessary to have available the actual numbers in the target population from which the cases were ascertained, and knowledge of their individual identities is not required. Conversely, for a population survey, it is necessary to have a list of individuals and their addresses/telephone numbers, in order that contact can be made. In countries such as the United Kingdom, the only population registers available to investigators are those from general practice and those from registered electors. The general practice option is available, as, in theory, every individual in the entire population is registered with one general practitioner to receive primary care. These registers are increasingly computerised and are available either within the practice or from the local Health Authority – which holds

such lists for all general practices within its area. General practice lists also provide ages whereas electoral registers only list those eligible to vote and do not identify individual ages. In other countries, true geopolitical population listings are available providing age and other key variables. There is, however, a considerable advantage in using a register derived for health purposes in that it provides the subject and the investigator with a strong motive for recruitment and participation. In some countries enrolment to receive healthcare services, for example in the USA with the Health Maintenance Organizations (HMO), is not universal and those registered are not representative of any geographical group.

6.2a Obtaining access

The existence of a population register does not imply that the investigator will be able to access it for research purposes. Government-held registers are often unavailable to researchers, and identifiable data from the UK census is not available for 100 years! Access to general practice registers in the UK, and other western European countries with similar systems, has been relatively easily obtained in the past, but increasing difficulties due to perceived ethical and confidentiality problems are being encountered (see Chapter 20). There is also an increase in the number of commercial organisations who sell mailing lists. Apart from the cost, however, such lists are not derived from population samples and are compiled from a number of sources.

6.3 Population data accuracy

Errors in the denominator population can lead to important errors in the disease estimates obtained. In the catchment population approach, the denominator population is normally obtained from population census estimates, which might be inaccurate owing both to initial errors in their compilation and estimates of population change since the date of the survey due to births, deaths and migration. Calculations are often undertaken to obtain projected population sizes between censuses, but such data may not be available for the chosen study catchment population or for the time period of interest. In reality, for rare disease, it may not matter if there is a 10% error in the denominator population, although it is important to remember that errors are normally greater in young adults (who have high mobility) the

very elderly (for obvious reasons) and also in inner cities and in ethnic minority groups.

In inner-city populations, there is up to 30% mobility each year, i.e. for every notional 100 persons living at a stated address within that area at the start of a year, only 70 still reside at that same address at the end of the year. A further problem is that those who are mobile are likely to be at different risks of disease from those who are more static.

In the population survey approach, the denominator can normally be corrected because knowledge of errors, due to deaths and migration, frequently comes to light during the course of the investigation. However, a register that is very inaccurate can lead both to an initial underestimate of the population size, which it is necessary to survey as well as contributing significantly to costs.

6.4 Study size

Once the target population has been selected, it is necessary to calculate whether it is large enough to provide robust estimates of disease occurrence. Further, if the aim is also to derive robust estimates in each (say) 10-year age group by sex then the number in the target population in each stratum needs to be sufficiently large. The determination of sample size is dependent on two, fairly obvious, aspects: (i) the approximate expected occurrence – the rarer the disease, the larger the population; and (ii) the *precision* required by the investigator, i.e. the acceptable error in the sample estimate as a measure of the true population occurrence. Statistically the calculation of sample size is based on the estimated standard error and the degree of confidence required (normally 95%) that, given a result, the 'true' figure will lie within the derived interval. An example of sample sizes for some typical situations is shown in Table 6.2.

6.4a Sampling from a population

In many surveys it will not be necessary or feasible to study everyone in the target population. In such instances it will only be necessary to study a sample. The basic unit in a study (e.g. a person) is referred to as the *sampling unit*, while the *sampling frame* consists of all the relevant sampling units (e.g. a list of all adults aged 18–65 years resident in town A). Those subjects

Table 6.2. Approximate sample size requirements for population surveys

Expected frequency (%)	Precision required (within ±) (%)	Approximate sample size[a]
1	0.5	1520
	0.2	9500
2	1	750
	0.5	3010
5	2	460
	1	1820
10	3	380
	2	860
	1	3460
20	5	250
	3	680
	1	6140

Note:

[a] These figures assume the population sampling frame is very large and are based on a 95% confidence that the true frequency will lie within the precision limits given.

selected for participation in the study are referred to as the *study sample*. In order that the sample is likely to be representative of the target population from which it is drawn, selection nearly always involves some element of randomness, i.e. the chance of any subject included in the sampling frame (or within a subgroup of the sampling frame) being selected is equal.

The most common method employed is *simple random sampling*. In this method each of the sampling units in the sampling frame would have the same probability of being selected as part of the study sample. This could be achieved by allocating each unit a number and then selecting a sample of a predetermined size using random number tables or a computer-program random number generator. A variant of this is *systematic sampling* where, for example, every twentieth person on a list is selected. In such circumstances each person still has the same probability of being selected for the sample. As discussed previously, in estimating the prevalence of a disease in a population one may wish to have a predetermined level of precision in each 10-year age-group. This can be achieved by *stratified random sampling*. The sampling

frame would be divided into strata defined by age-groups, and thereafter one of the previous methods could be used to select a sample within each 10-year age-group.

In some studies it may be necessary to interview subjects or make a clinical diagnosis. It is unlikely to be feasible to interview or examine all subjects in the study sample.

Example 6.iv

A study wished to examine the prevalence of asthma (defined by clinical diagnosis) but it would have been extremely time consuming and resource intensive to examine all subjects in a large study sample. Most subjects would not have asthma. Instead a multi-stage sampling procedure was employed. The study sample was selected using simple random sampling. Each of the study sample received a postal 'screening' questionnaire. Selection of subjects for the second (clinical) stage of the study then depended on responses to the first-stage questionnaire. This method allowed the prevalence of asthma in the target population to be estimated with efficient use of resources.

Finally *multistage (cluster) sampling* is commonly used. The clusters may, for example, be towns, schools or general practices. The first stage is to select a cluster sample and thereafter from each of the selected clusters to sample individual units (e.g. residents, schoolchildren, patients).

Example 6.v

A study was designed to estimate the prevalence of dental decay amongst 7-year-old children in Rose county. Firstly a sample of schools was randomly selected from all those within the county. Thereafter, from the 7-year-old children attending these schools, a further simple random sample of individual children was selected.

7

Studies of disease occurrence. II: Assessing disease status in study populations

7.1 Approaches to measuring incidence

There are four approaches to measuring disease incidence in a target population (Fig. 7.1): (i) a single population survey identifying those with disease or disease onset during a defined time period; (ii) duplicate population surveys in which two studies are undertaken, separated by an interval of time to permit identification of those cases of disease that develop between the two surveys; (iii) a retrospective review of diagnosed cases attending clinical facilities serving the defined target or 'catchment' population; and (iv) a prospective registration system for recording all new diagnosed cases attending clinical facilities as in (iii) above.

The choice of appropriate strategy is guided by a number of considerations as outlined in the flowchart (Fig. 7.2). The first question is whether all individuals with the disease in question are likely to attend a physician and be diagnosed. If this is not so, for example in self-limiting acute viral illnesses or minor trauma, then a *population survey approach* is essential. Even if the substantial majority of individuals with the disease are likely to seek medical attention, it is necessary to determine whether the *catchment population approach* is appropriate. In inner cities, for instance, there is a wide choice of clinical facilities and it may be impossible to identify those facilities that will cover all the individuals, from the target population, who may develop the disease. It will thus be necessary to use a population survey strategy.

Example 7.i

A study aimed at investigating the incidence of myocardial infarction in a target population. Preliminary enquiries revealed that many such episodes resulted in admission to hospitals outside the immediate geographical area, including, for example, hospitals local to where the

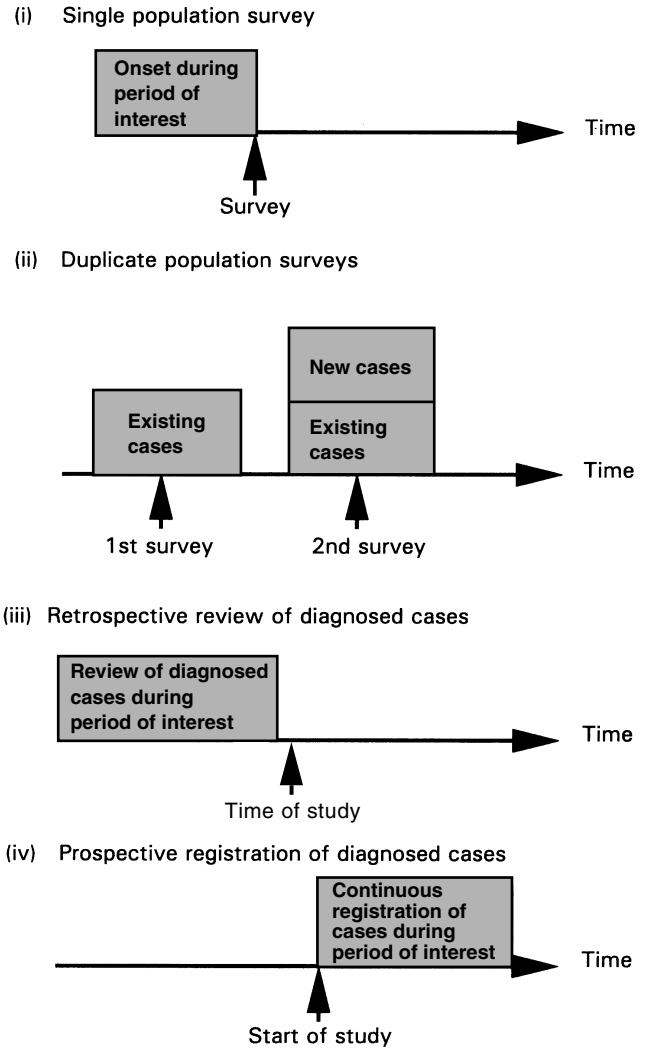

Figure 7.1 Approaches to measuring disease incidence.

cases were employed. Surveying local hospitals was therefore abandoned as a strategy for assessing incidence.

By contrast, in the ideal situation of a single central facility that provides the only health care to a population, the catchment population approach, using available records, is feasible.

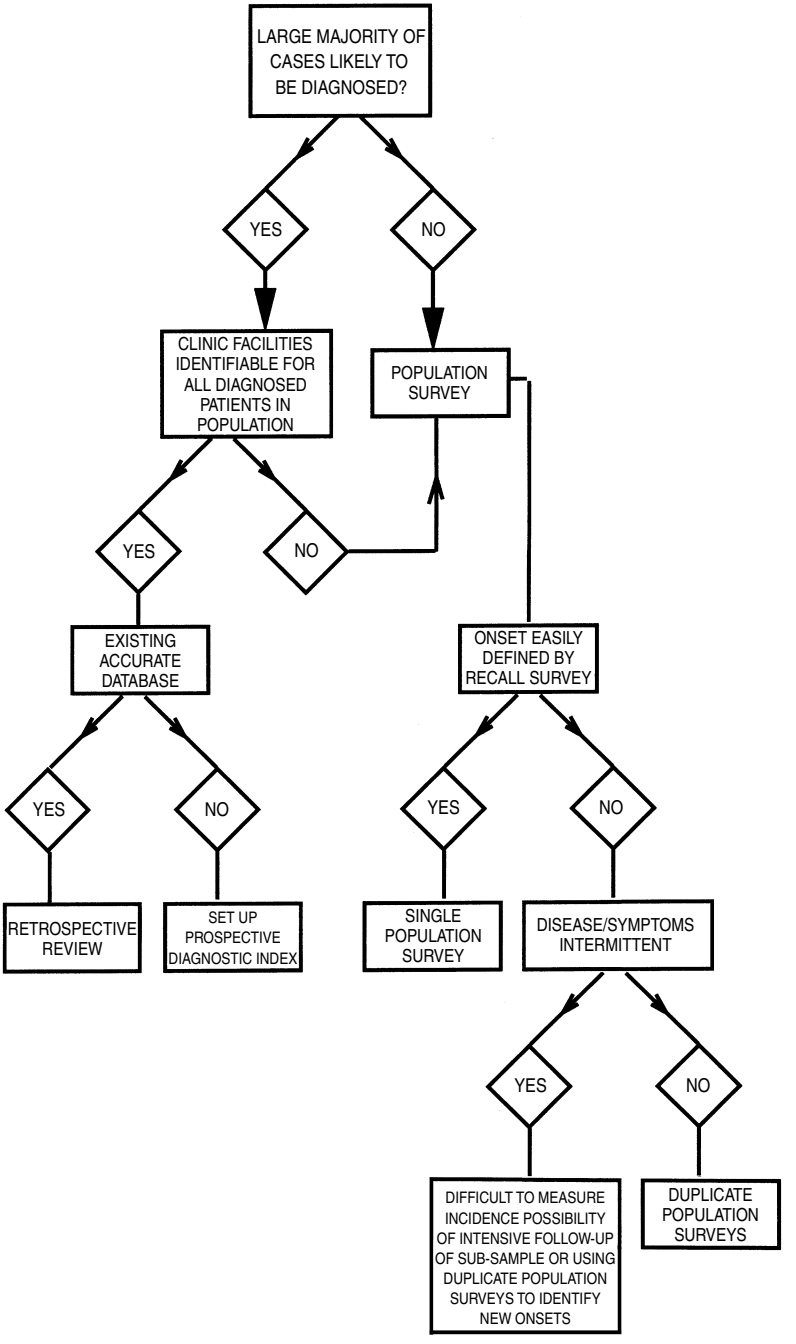

Figure 7.2 Choosing the appropriate strategy.

Example 7.ii

The Rochester Epidemiology Project is based on the necessity that all inhabitants from Olmsted County, Minnesota, USA, seek their medical care either from the Mayo Clinic or a linked medical practice. It is possible to identify Olmsted County residents from the Mayo's diagnostic database, which should include all the diagnosed cases that had arisen within that target population. As a consequence, there have been a large number of incidence estimates of a large variety of diseases from that source.

In situations with centralised facilities, but where there is not an existing database that will permit easy recall of patients, then it will be necessary to establish a prospective system. In this approach, the investigator specifically recruits all relevant physicians to notify any new case with the specific disorder under investigation to the study. A study using the method would require to assess the completeness of ease notification.

If a population survey is required, a single survey will normally suffice if the disease onset can be easily recalled and defined by the subjects surveyed. Alternatively, if onset is unlikely to be recalled, duplicate cross-sectional surveys have to be undertaken. For diseases or symptoms of an episodic nature even such duplicate surveys may not be able to measure incidence, since some episodes occurring entirely between surveys may be missed.

7.2 Use of diagnosed cases: retrospective review or prospective notification?

As indicated above, the presence of an existing diagnostic database will permit the use of the retrospective approach to ascertain cases diagnosed in the past. The problem is that only rarely are such databases accurate or cover all relevant departments. Thus a gastroenterology department within a hospital may keep an index of all new cases of peptic ulcer, but such an index would exclude cases attending a geriatric or general medical unit, for example. In the United Kingdom, Scandinavia and some other countries, there are population-based databases of all hospital in-patient admissions. Thus, provided that the disorder leads to an in-patient stay, such as acute appendicitis or femoral neck fracture, these data sources can provide fairly accurate incidence estimates. The diagnostic accuracy of these databases is often questionable because those recording the data do not have a vested

Table 7.1. Retrospective review or prospective notification

Attribute	Retrospective review	Prospective notification
Existing diagnostic database	Required	Not essential
Physician compliance	Not essential	Required
Disease very rare Examination of time trends	Acceptable if long-term data available	Long-term study required
Diagnostic verification	May be impossible to measure	Easier
Under ascertainment	May be impossible to measure	May be high
Expense	Small	Medium

interest in the data quality, at least as far as the specific question is concerned. The other problem is that it may not be possible to identify all the health facilities serving a catchment population. Specifically, private hospitals and similar institutions do not have a statutory requirement to provide data but in some areas, for some disorders, may be a source of an important proportion of cases.

7.2a Physician compliance

Other factors to be taken into consideration are shown in Table 7.1. With an existing database it is not necessary to ensure physician compliance, though it might be necessary to gain permission to access the data and the all-important case records. It will also be necessary to assess the completeness of the database both in terms of identifying cases of disease and of researching information. One of the problems with the prospective notification is that this normally requires contacting all relevant physicians and asking for their help in identifying and notifying cases. If the disease is rare, or the notification procedure complex, it may be impossible even to approach completeness of ascertainment. Constant reminders are required and the most successful schemes, which are clerically time-consuming, involve circulating all relevant physicians at least monthly or even weekly for common diseases. In some cases it may be necessary to have study personnel dedicated to the task of identifying eligible clinic patients.

Example 7.iii

In a study aimed at ascertaining all new cases of Crohn's disease in a target population, the investigators were aware that no diagnostic database existed. They wrote monthly to all local general physicians, gastroenterologists, paediatricians and geriatricians asking for notification of new cases diagnosed in the previous month.

Compliance may be enhanced by offering a small financial inducement for every patient notified, although the value of this tactic is unclear!

7.2b Problems with rare diseases

If the disease is very rare, the retrospective approach, provided that the database goes back far enough, will permit an answer to be obtained in a much shorter time than the prospective approach, which might require five years of case ascertainment to derive reasonably robust estimates. Similarly, if time trends are of interest, only the retrospective approach can yield an answer in a reasonable time frame.

7.3 Defining cases with the catchment population approach

There are inevitably two issues in defining the numerator for the catchment population approach to incidence studies. These can be considered as case ascertainment and case verification (although the principles are also relevant to the population survey approach). There is a need to ensure that all cases have been ascertained and also that all cases ascertained are verified to have the disease.

7.3a Case ascertainment

Under-ascertainment is a distinct possibility in the catchment population approaches, both retrospective and prospective. The major causes are, in the former, diagnostic misclassification and, in the latter, non-compliance by the participating physicians in addition to those possibilities already mentioned of non-referral or referral to another clinical facility. One approach to checking for under-ascertainment in the retrospective approach is to review a sample of medical records of individuals with a related diagnosis.

Example 7.iv

A surgeon wished to determine the incidence of acute appendicitis in the local population. In addition to gathering from the target hospitals all those listed as having that diagnosis, he reviewed a random sample of records from those with acute abdominal pain to determine whether any additional cases of acute appendicitis had come to light.

In the prospective approach one useful strategy, if there are enough cases, is to compare the rates of notification from different physicians. Obviously, depending on their practices and interests, there will be an inevitable variation in the number of cases reported, but significant deficiencies may come to light.

Example 7.v

In response to a local survey of eight urologists for notifications of cases to determine the incidence of renal colic, five of the surgeons notified an average of ten cases and the remaining three only notified two cases each. Discreet enquiries suggested that the last three surgeons had not fully complied with the request.

7.3b Case verification

Diagnostic verification involves setting up a series of rules to ensure that the cases notified actually have the disease. This will inevitably involve, at the minimum, a review of medical records to establish the basis for the diagnostic assignment and might also require the investigator to interview, examine or even investigate the subjects identified to prove the diagnosis. The advantage of the prospective approach is that because the data collection is contemporaneous, the investigator can both provide the inclusion rules to the participating physicians and check, if necessary, with the patient. For patients registered some time ago such diagnostic information is very difficult to obtain.

Example 7.vi

In a prospective study of the incidence of acute appendicitis, the registration form sought details of findings at surgery and from histology in addition to the clinical diagnosis.

Finally, it is apparent that any prospective system is more costly than a recall system that relies on trawling through existing databases, although even the costs of the former are substantially more modest than those involved in a population survey.

Table 7.2. Single or duplicate population surveys for measuring incidence

Attribute	Example	Survey approach
Onset easy to recall and accurately date	Food poisoning outbreak Myocardial infarction	Single retrospective survey
Onset easy to recall but difficult to date	Acute low back pain	Duplicate retrospective surveys with short interval or intensive follow-up of population sample
Onset not recallable by subject	Hypertension	Duplicate surveys, interval not important

7.4 Use of cross-sectional population surveys to assess incidence

7.4a Single or duplicate surveys?

Incidence estimates may be derived either from a single or from duplicate population surveys. In situations (Table 7.2) where the disease has a clear onset, easily recalled by the subject, a single cross-sectional survey will permit an estimate of the incidence. With disorders which have a major impact such as myocardial infarction, then retrospective recall of the event in a single cross-sectional survey might provide an accurate estimate of the occurrence, even over a fairly prolonged period. With disorders such as gastroenteritis following a point source outbreak, an enquiry fairly close in time to the event is necessary to ascertain cumulative incidence.

Some disorders, though characterised by a clearly defined onset, present as multiple episodes of short duration, for example, epileptic fits pain. The problem in such disorders is that in cross-sectional surveys recall by the individual of the date of onset of the last episode may be poor. It is thus necessary to undertake duplicate surveys, relatively close in time, to derive accurate estimates of the number of episodes that occurred in the interval.

Finally, for those disorders where the date of onset is difficult or impossible to define, such as hypertension, it may be possible to derive an estimate of incidence based on the number who change their disease status between two cross-sectional surveys. This is calculated as an incidence density rate, which is the number of new cases developing divided by the sum of the individual

years of follow-up between the two studies. The interval between surveys could be long.

7.4b Left censorship

One major problem with any cross-sectional (survey) approach to measure incidence is referred to as *left censorship*. This problem results from retrospective enquiries, which, by their nature, exclude those who were in the target population and had a disease incidence event during the period of interest that was not ascertained, due perhaps to death, migration or even full recovery. In circumstances where the length of an episode could be very short, for example low back pain, unless the follow-up interval is also very short, follow-up surveys may miss new onsets. Conversely, intensive follow-up of subjects enquiring about specific symptoms may lead to subjects reporting even very minor symptoms.

Example 7.vii

A retrospective survey aimed to determine the incidence of myocardial infarction in the previous five years. This method would miss detection of those individuals who had developed an acute infarction and had died as a result. Such subjects clearly would not be able to be surveyed, seriously underestimating the true incidence.

Other more subtle examples of this bias might reflect those who have developed the disease and moved subsequently to a different location, perhaps because of loss of job or seeking a different climate, and similarly they would not be ascertained.

7.5 Approaches to measuring prevalence

As with studying incidence, there are a number of approaches to measuring prevalence:

- retrospective review of diagnosed cases,
- prospective recording of current diagnosed cases,
- population survey.

In a similar manner there are the alternatives of (i) review of diagnosed cases from a catchment population, or (ii) a population survey. With preva-

lence, the more usual approach is to undertake a special survey, but there are circumstances where it can be acceptable to use the cheaper and quicker catchment population method.

7.6 Catchment population methods for measuring prevalence

There are a number of requirements before such an approach can be used for deriving prevalence data:

- diagnostic register available,
- substantial majority of cases likely to be diagnosed,
- registers available to cover all individuals in catchment population,
- vital status (dead/alive) known on prevalence day,
- residence known on prevalence day,
- disease status known on prevalence day.

There must be a database available covering the disorder in question; the large majority of cases that arise within a population would be expected to be diagnosed and therefore incorporated into the database, and the clinic facilities surveyed should cover the catchment population. In addition, because (the normally quoted point) prevalence requires knowledge that an individual is in 'disease state' on a specific day (prevalence day) there are other assumptions that need to be met. These are: (i) each of the cases on the register is still a resident of the catchment population on prevalence day (as opposed to the date of onset); (ii) as part of the above each of the cases is known to be alive on prevalence day; (iii) each of the cases is still in 'disease state' on prevalence day. Thus, there are only a small number of conditions that would fit these requirements.

Example 7.viii

One good example is Type I (insulin-dependent) diabetes in childhood. It is reasonable to assume that all cases that arise will be diagnosed in a recognised clinical facility, and the relevant institutions for a particular target population should be identifiable. As the condition is (to all intents and purposes) irreversible, disease presence at diagnosis is equivalent to disease presence on prevalence day in those still alive. Given that virtually all children will probably be followed up in a clinic it should be possible to trace the current residence of all the cases. Thus the point prevalence at an arbitrary date can be estimated.

7.6a Prospective measurement of prevalence

This approach is reasonable for assessing the period prevalence of conditions that are fairly common and occur in episodes that normally require clinical attendance at some time during an arbitrary period.

Example 7.ix

A good example here is asthma. It is unlikely that all cases of asthma will attend hospital. However, it may be reasonable to assume that, in a given year, if there is any significant morbidity from asthma, then the sufferer will need to seek attention at least at the primary-care level. Thus, to determine period prevalence, a system can be instituted for the target primary-care physicians to document, prospectively, all attendances with asthma during the period of (say) one year, to estimate the one-year period prevalence.

In the United Kingdom, such data as these have been collected routinely for all consultations by a number of general practitioners during National Morbidity Surveys. In addition, the computerised software used by some general practitioners results in the automatic recording of the reason for all consultations.

7.7 Population surveys

Prevalence day does not require to be the same day for each study subject. Population surveys may be conducted over a period of several months. On a postal questionnaire it will probably relate to the time around when you fill in the questionnaire. Nevertheless in presenting results a notional 'prevalence day' is assumed in calculating point prevalence rates.

The most frequently used approach to deriving prevalence is the population survey, where a target population is identified and an attempt is made to classify the disease state (present or absent) of every individual on 'prevalence day'.

Example 7.x

A study attempted to estimate the point prevalence of shoulder pain. The survey, which involved home interview took six months to complete. During the survey it became apparent that the symptoms were intermittent and that ascertaining disease status at a single time point was hazardous.

In the above example it may be appropriate to consider a prevalence period, i.e. whether symptoms have been present in a specific period such as the past month. This allows calculation of a period prevalence rate.

7.7a Approaches to population surveys

The approach to be taken depends on the information required to classify disease accurately. A hierarchy of strategies for population prevalence surveys of increasing complexity and cost is as follows:

- postal survey,
- postal screening survey with telephone or postal follow-up,
- postal screening survey with interview or examination follow-up,
- interview survey,
- interview screening survey with examination,
- interview screening survey with investigation,
- investigation survey.

Thus, at the lowest level a postal survey may provide all the necessary information and has been used, for example, in surveys of diseases whose diagnosis is essentially based on history alone, such as angina and chronic bronchitis. If the questions are prone to misinterpretation or error, an interview may be required either by telephone (a frequent approach if there is a good coverage of the population by telephone), or in person either at the subject's home or on special premises. Detailed aspects of the relative advantages of a questionnaire over an interview for obtaining epidemiological information are given in Chapter 10. It may be necessary to undertake a restricted physical examination, for example blood pressure measurement. Obviously direct contact is required when investigations such as blood, urine or radiological testing are required. A cost-effective approach is often to undertake a two-stage screening procedure if the nature of the disease makes it appropriate. The first stage is thus a postal or telephone interview to all, with follow-up of subjects depending on the results of the initial screen.

7.8 Other (indirect) measures

It will be appreciated that when measuring incidence, determining the date of disease 'onset' may only be an approximation. For example, the date of onset of cancer recorded will usually refer to a date of diagnosis, even though

symptoms are likely to have been present some time previously. Indeed the true onset of the cancer will predate even symptom onset.

In some circumstances it may not be possible to measure incidence and some proxy measure may be used. If the disease under investigation has very poor survival then mortality may be a good approximation to disease onset.

Example 7.xi

In an international study of lung cancer, some countries involved did not routinely measure lung cancer incidence. However, since the survival rate from lung cancer is very poor (approximately 3% at 5 years) it was considered reasonable to use available mortality data.

In other circumstances it may be possible to use health service data to approximate disease state or onset. This may not capture all persons in the disease state or with the disease onset of interest but it is likely to reflect those of a certain severity or perceived severity. Investigators will, however, have to consider the quality of the data recording and whether it is in a form suitable for the purposes of the study.

Example 7.xii

Investigators wished to determine the prevalence of hypothyroidism (an under-active thyroid) in a small town. They did not have the resources to investigate all the residents. Instead they obtained permission to scrutinise the medical records of the single general practice in the town. It was possible to determine from the computerised and paper medical records the number of residents who had been formally diagnosed during the previous year.

There will be occasions where the recorded data of the condition of interest may not be suitable for use for a variety of reasons, e.g. not closely enough related to incidence, poor quality of the data etc.

Example 7.xiii

A company had requested local investigators to determine how common back pain was amongst their workforce. Analysis of the company absenteeism records showed that back pain was by far the most common reason self-reported by workers for short-term absences. However, in many cases the records were incomplete and there was no cause given. The inves-

tigators thought that the quality of the records, the inconsistent way in which the information was recorded and the fact that many people with back pain may not have taken time off work, meant that they had to undertake a special enquiry for the purpose.

Overall, the decision would need to be made on the basis of specific needs relating to an individual project where actual incidence data were not available as to whether an alternative (proxy) measure may be suitable.

8

Studies of disease causation. I: Selection of subjects for case-control studies

8.1 Recruitment of cases

There are two major issues to be addressed. These are (i) what type of cases to include – all cases or only newly occurring cases, and (ii) whether attempts are made to recruit all eligible patients from a target population. These two issues are interlinked. A flow chart for considering the options is shown in Fig. 8.1. A major factor underlying the choice is the rarity of the disease. It is clearly necessary for the investigator to have an idea of the likely incidence and/or prevalence of the disorder in order to make the appropriate choice. The approach to calculation of the number of cases required is described in Section 8.5.

8.1a Incident or prevalent cases?

The choice exists either to recruit only new or *incident* cases or to include all those with current 'active' disease or *prevalent* cases. One benefit of the former is that the cases are not selected for disease severity, whereas selection of prevalent cases will exclude those whose disorder has become inactive or totally remitted. In addition, there is the paradoxical problem that if a disease, such as some forms of cancer, is associated with a rapid mortality in an important proportion of cases, then selecting prevalent cases will *skew* recruitment to those surviving. It is entirely plausible that the association between a possible risk factor and a disease may be different in those who survive from those who die. Further, in attempting to investigate aetiological factors for a disease, in subjects who have had the disease for several months or years it is difficult to distinguish risk factor exposure prior to and after onset of the disease.

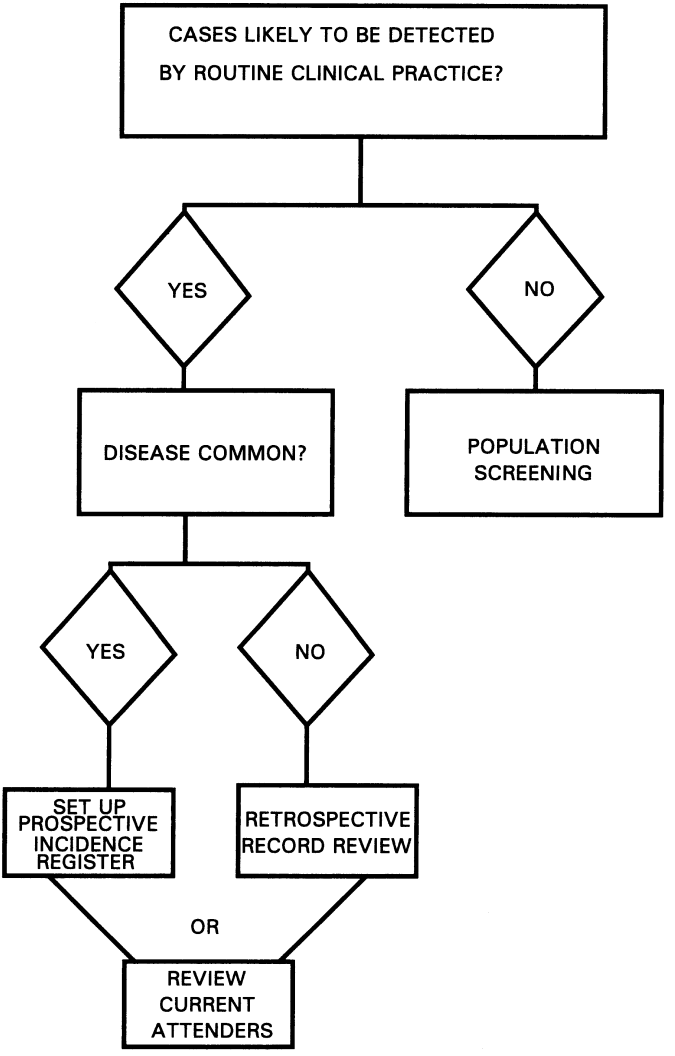

Figure 8.1 Strategies for case selection.

Example 8.i

A case-control study investigated the relationship between cigarette smoking and stroke, taking as cases those who had recovered sufficiently to be discharged from hospital. If there were, amongst people who had a stroke, an increased risk of sudden death in smokers, then restricting study to surviving cases would bias the estimate obtained of the risk from smoking.

It may be possible to identify cases that have remitted or died by a trawl of medical records or from a contemporary disease register. Indeed, it may be possible to use proxies to obtain data, for example by interviewing surviving relatives of deceased cases. (Consideration has to be given also whether to adopt a similar strategy for the controls.) As suggested in Fig. 8.1, if contemporary registers of new cases do not exist, it may prove impossible to prospectively recruit a sufficient number of new cases within a reasonable time. The inclusion of all available prevalent cases is the only option. This situation is particularly likely for rare disorders.

8.1b Population-based or 'hospital-based series'

Ideally, the cases selected, whether incident or prevalent, should be either all or, depending on the number required, a random sample of all cases from a target geographical population. In that way one can be sure that there has been no *selection bias*. In some situations this may not be possible or practicable. If the disease is not likely to be detected by routine clinical practice, that is, a substantial proportion of cases are asymptomatic, then population screening may be the only available option.

Example 8.ii

A study was undertaken to examine occupational influences on osteoarthritis of the hip. Given that a substantial proportion of cases are pain free, at least at any single point in time, the decision was taken to screen radiologically a random population sample to ascertain cases rather than rely on identifying symptomatic individuals who were referred to hospital. Another advantage of this approach was that the controls could be chosen from amongst the disease-free individuals from the same population, ensuring their comparability.

For many rare diseases, population screening is too expensive, and may not be necessary if cases are likely to come to clinical attention anyway. In such circumstances, it may be possible to ascertain most or all incident cases

within a population if (i) the investigator can identify all the clinical facilities serving the target population, (ii) there are a few cases, resident in the study area, who use residential facilities elsewhere, and (iii) it is possible to identify (and therefore exclude) cases that live outside the study area. The approach used is identical to that for measuring incidence (see Section 7.1).

Example 8.iii

For a proposed case-control study of acute appendicitis, an investigator chose a local government boundary as the population base. The operating theatre records and surgical admissions were reviewed both from the two local hospitals serving that population as well as from the hospitals' four adjacent areas to which some patients, from within the target area, might attend. Only cases who normally resided within the target area were eligible for study.

A frequent and convenient source of cases is the patient population of one or more physicians with an interest in a particular disease. There are considerable advantages in terms of accuracy of diagnosis and high participation, but the question should be posed as to whether the cases seen in their practice are representative of the population at large.

8.1c Recruitment of diagnosed cases: using existing databases

It is often appropriate to recruit cases, either incident or prevalent, from a number of physician colleagues because it is rare for one physician to have a sufficiently large number of cases 'on the books'. Computerised databases will often permit the identification of patients attending other physicians and indeed other hospitals; permission may readily be granted for access both to the database and to the patients themselves.

Example 8.iv

To recruit cases for a case-control study of genetic factors in pre-eclampsia, the investigators used the local hospitals' obstetrics databases to identify all pregnancies in the previous two years with a positive entry for pre-eclampsia.

There are two problems that frequently occur with this path: (i) in most countries routine hospital databases include diagnostic data on in-patients only, and thus disorders that do not routinely result in an in-patient admission are missed (a more recent development in the UK and in other countries is that

there are a number of commercially available databases for primary-care physicians designed for continuous morbidity recording); (ii) hospital and primary-care databases are set up for wider purposes than the demands of the proposed study. Thus, there will have been little obligation for those who enter the data to be particularly aware of the needs of a study for what may amount to be a tiny fraction of all cases entered. The reliance on these routine databases may also be problematic because there may be considerable diagnostic errors. Diagnostic accuracy is potentially a problem in primary care, and for rare disorders, such a source may provide an insufficient number of cases compared with the enriched patient population of a hospital specialist.

8.1d Recruitment of diagnosed cases: *ad hoc* recruitment from colleagues

Particularly in rare diseases it is necessary to set up a special system for recruiting cases from interested colleagues. This is always easier in theory than in practice. Colleagues invariably overestimate the number of cases they have available and their compliance with prospective notification is variable. It is, however, preferable to rely on prospective recruitment, i.e. to ask for notifications of all previous attenders. Similarly the recruitment phase should be restricted. Compliance is enhanced by the knowledge that this is only to be done for a fixed period, (say) three months. The notification process should be as simple as possible with clear eligibility criteria provided. A suitable letter for sending to colleagues is shown in Fig. 8.2 together with an example of a patient notification form (Fig. 8.3). It may be a worthwhile investment to offer a prize to those notifying the greatest number of potential cases or a payment per patient recruited to the study.

8.1e Case verification

The principles are the same as those discussed in Section 7.3b in relation to population prevalence and incidence surveys. It is necessary to state very clearly the rules used that allowed potential cases to be considered as cases. It may be difficult, particularly if the cases were diagnosed at some stage in the past, to verify the diagnosis with absolute certainty. Contemporary medical records are not normally sufficiently comprehensive. A hierarchy of desirability in case verification is shown in Table 8.1. Thus clinical/pathological/radiological or similar confirmation using standard criteria, depending on the disorder, is the gold standard. Ideally, the investigator should have

Disease Research Clinic
St Anywhere's Hospital

Dear Colleague

I am planning a case control study to examine the hypothesis that exposure to pets may be linked to the subsequent development of Von Sturmmer's Syndrome. Unfortunately, as you know, this is a very rare syndrome and I am writing to all the neurologists locally to help in case recruitment. I would be grateful if you would be willing to participate.

I am looking for patients with Von Sturmmer's Syndrome. They should have:

- **a positive serological test**
- **a disease duration of less than 2 years.**

Any patient who agreed to take part would be interviewed by my research nurse, Mary Jones, about pet ownership and other aspects of lifestyle. The interview would take place at the patient's home.

I enclose a copy of our enrolment form, which I would be grateful if you could complete and return for all eligible patients. The most practical approach is to arrange for the nurse in charge to make sure that there is a supply of these forms in your clinic room.

The study has the approval of St Anywhere's Ethical Committee and, of course, on the interview forms there will be no personal identification information. The study will be fully explained to eligible patients prior to requesting their written consent to participate. I also enclose an outline protocol, a copy of our interview schedule, a photocopy of the ethical approval, and a patient information sheet.

I look forward to hearing from you.

Yours sincerely

Figure 8.2 Example of a letter inviting recruitment of cases.

access to the actual diagnostic data, but for independent validation a contemporary report is normally acceptable. Standardisation of diagnostic verification is very difficult in practice and thus if all the histology slides, X-rays or even physical findings, etc., can be reviewed by a single expert then so much the better.

Example 8.v

In a case-control study of liver cirrhosis, it was decided to use histology as the defining criterion for acceptance as a case. The microscope slides were gathered from all potential cases and were reviewed by a single expert histopathologist.

For many disorders, pathological or similar confirmation is impossible and sets of clinical criteria have been formulated with the aim of standardising

Table 8.1. Hierarchical list of possibilities for case verification

Most ideal	Diagnostic confirmation using agreed criteria by investigator, including physical examination and results of investigations where relevant
	Diagnostic confirmation using agreed criteria by investigator based on contemporary records
	Diagnostic confirmation using agreed criteria by recruiting physician based on contemporary records
	Diagnostic status based on clinical opinion as judged by experienced physician
Least ideal	Self-reported diagnosis by subject

VON STURMMER'S SYNDROME STUDY

Patient Name .....

Address .....

..... or patient identification label

Date of Birth .....

Hospital Number .....

This patient has had Von Sturmmer's Syndrome for ..... years and has a positive serology test. He/she is willing to be contacted for an interview at home.

Telephone number .....

Signed ..... Date .....

Figure 8.3 Example of a patient registration form.

disease classification internationally. This approach is of particular relevance in psychiatry, but is also the main approach used in a number of organic disorders. Again, ideally, the investigator should clinically review all the cases. This may be impracticable, or the patients may be currently in remission and thus reliance has to be placed on contemporary clinical records. These again may provide insufficient data for the purposes of applying a standard disease classification scheme. Ultimately, the investigator has two choices: either exclude those cases for whom diagnosis may not be proven, or include them but undertake, at the analysis stage, a separate analysis to determine the influence of including unconfirmed cases. The choice of the above depends on the number of cases available and the proportion with missing data.

Finally, although self-reported diagnosis is listed in Table 8.1 as the least ideal, in certain circumstances self-report is the only option available. This is true, for example, when studying symptoms, e.g. pain. Only the individual subject knows whether they have a headache or not!

8.1f Exclusion criteria

In addition to specifying rules for which cases are to be included, it is necessary to state at the beginning the rules for those cases to be excluded. Examples of such criteria are:

- disease status not verified,
- emigrated,
- significant physical/psychiatric co-morbidity,
- language problems.

Besides a lack of diagnostic data, another reason for a case being excluded could be the unavailability of the patient, for geographical, medical, psychological or other reasons. It is important for the reader of the final report to have a feel for what these exclusion rules are, again because the results can only be extrapolated to those who are studied. If the study method involves an interview or completion of a questionnaire, familiarity with the native language is required and those who are illiterate may need to be excluded. Often for reasons of *homogeneity*, studies may exclude those who do not belong to the predominant ethnic group, particularly if ethnicity is related to the risk of exposure. Sometimes unexpected exclusion criteria are forced on the investigator. In a collaborative study that one of the authors was undertaking in a Mediterranean country, many of the young males were

unavailable for study because of military service duties – this was not an exclusion criterion that had been planned!

8.2 Recruitment of controls

This is probably the most vexed issue in case control study design. In broad terms, the controls should be representative of the population from which the cases have arisen. In other words, if a subject chosen as a control had become a case, then they should have had the same chance of being included in the study as did the cases. For population-based case-control studies, when all cases in a defined population area are selected, the source for the controls is obvious. It is a greater problem deciding from where to recruit controls when cases have been selected from a hospital or clinic.

8.2a Strategies for control selection: population-based

A flow chart for determining the most appropriate control group is shown in Fig. 8.4. The optimal situation, as stated above, is where a whole population has been screened to detect the cases and a complete list or sampling frame exists. Then the only requirement is to take a true random sample of those who do not meet the criteria for being a case.

Example 8.vi

A general practitioner used a standard questionnaire and screened his female registered population, aged 45–64 years, to detect those with probable depression. He then undertook a case-control study comparing socio-economic variables in the cases detected with an equal number of randomly selected women aged 45–64 years from his practice who had a normal depression score on screening.

Frequently, a population sample of cases is based on ascertainment from the relevant clinical facility or facilities, but a population register does not exist for the conceptual catchment population from which such cases arose. Thus, in the United Kingdom, there does not exist a list of residents in a particular geographical area that includes their age. An alternative approach to be used in these circumstances is to take another population-based listing approximately covering the population concerned, such as the list of individuals registered to receive medical care at one or more general practices. In

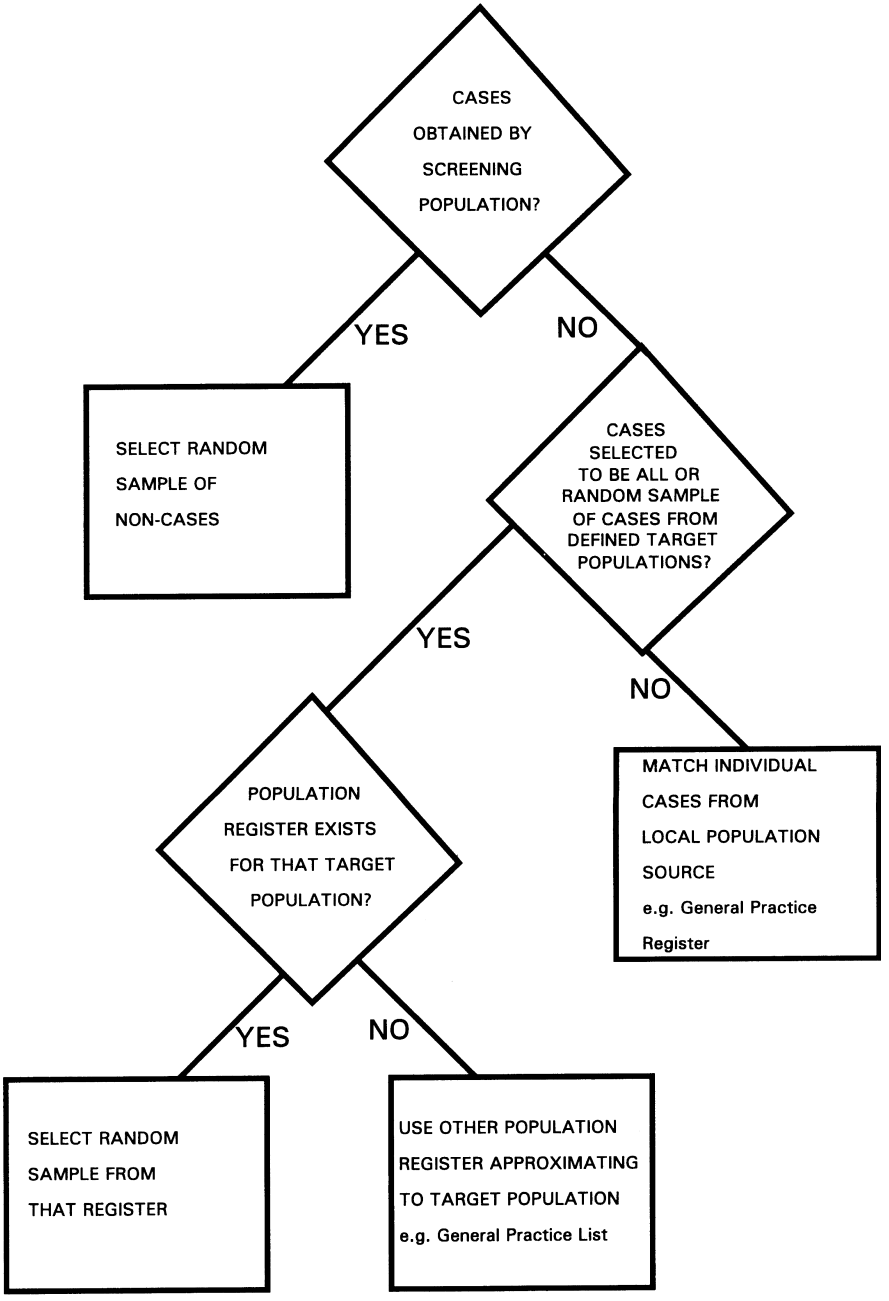

Figure 8.4 Ideal options for selecting controls.

this example, the controls should be as closely representative as possible of the population from which the cases were selected. Thus, the general practices selected should be the same as those from which the cases arose. If the cases are selected from a very wide area, i.e. no single target population is identifiable, selecting controls individually from the general practice list of each case is appropriate.

Example 8.vii

In Example 8.iii above on acute appendicitis, the general practitioner for each case of appendicitis ascertained was asked to provide, as a control, the name of the next three individuals on the age–sex register of the same sex and age (within 1 year).

8.2b Other sources of population controls

The registers of individuals listed above may not be available for all populations and access may be denied to the investigator. There are other sources of population controls.

(i) Population registers without ages

In the United Kingdom, local electoral registers (lists of eligible voters) are available to the public. They are reasonably up to date but exclude those under 18 and those who do not register or are not eligible to vote in any elections (e.g. Non-European Union citizens). They cover perhaps 90–95% of the population. Their major drawback is the absence of information on age, but they can be useful.

Example 8.viii

In a case-control study, an investigator undertook the role of assessing reproductive factors in the aetiology of rheumatoid arthritis; the cases were women with the disease attending a number of hospitals in the London area. The controls were selected from random samples of women chosen from three local electoral registers. Electoral-register responders who were outside the age range of the cases were subsequently excluded from the analysis.

(ii) Telephone recruitment

In populations with good telephone coverage, random-digit dialling within certain telephone areas can be a cost-effective method of obtaining controls. It is often a two-stage process, with the initial call identifying the presence

within the household of a suitable individual (say as regards age and sex), and a follow-up call to obtain participation. This is a very popular method in the USA and a number of commercial companies undertake the background calling. The main problem with this method generally is that telephone coverage may not be high enough to obtain reasonably representative samples. An additional and more recent problem is that this approach is being increasingly used for 'cold sales calling' by life insurance and similar companies with the result that there is an increased rate of not getting past the first sentence and hence the study having a low participation rate.

(iii) Neighbour controls

The final option is to use local street maps to obtain a near neighbour. This approach will at least ensure that the controls are from the same socio-geographical population as the cases. If the street layout is appropriate, rules can be established.

Example 8.ix

In a case-control study of home factors (lighting, ill-fitting carpets, etc.) related to fractured neck of femur in elderly females (above the age of 75), the investigator identified the address of all eligible cases from hospital registers who were non-institutionalised at the time of injury. For each case a researcher visited the household on each side and all female residents aged over 75 were eligible for inclusion as controls.

8.2c Strategies for non-population-based control selection

In some situations the strategies detailed above will not be appropriate. Firstly, the population base from which the cases arose may be impossible to determine. This may be true when cases are recruited from hospital. The selection factors that resulted in an individual with a particular disorder attending and being diagnosed by a particular hospital are complex. Secondly, response rates from randomly chosen population controls are often very low. There is a greater personal incentive for cases to participate than controls. There are, however, a number of choices for controls that address these problems.

(i) Disease controls

One valuable approach is therefore to use cases attending the same clinical facilities with a different disease. The problem is which disease to choose? If

one is conducting a study of myocardial infarction and chooses as potential controls patients attending the accident and emergency department, this is likely to result in a very different control population (e.g. in terms of age) from the cases under study. The fundamental premise that the controls should have been eligible to be cases (had they had a myocardial infarction) may also have been lost. Further in terms of 'exposures' the choice of such controls is likely to be unsuitable.

Example 8.x

A hospital-based case-control study of myocardial infarction was conducted using patients admitted to the hospital with myocardial infarction as 'cases', while age- and sex-matched 'controls' were selected from those attending the accident and emergency department on the same day (without myocardial infarction). The investigators were surprised to find that this study showed that cases consumed much less alcohol than controls, despite the fact that most other studies showed that heavy alcohol consumption was a risk factor for myocardial infarction. The result was explained by the fact that many of the conditions presenting in the accident and emergency department were related to heavy alcohol consumption.

This is a particular problem if one relies on choosing controls from a single disease group. It is therefore preferable to choose several conditions, which the investigator has no *a priori* reason to believe has an association with the major risk factors under study, and to select controls from these groups. Using such methods will ensure that one disease group will not dominate the control group and have undue influence on the results.

Example 8.xi

In a case-control study of rheumatoid arthritis, women with other musculoskeletal conditions attending the same clinics were chosen as controls. Both cases and controls were approached in the same manner ('we are interested in looking at the relationship between arthritis and hormonal/gynaecological factors') and the presumption was that if the controls had developed rheumatoid arthritis instead they would also have been ascertained as cases.

(ii) Family/friend controls

This again can be a useful strategy for two reasons: (i) knowledge of the case will aid recruitment and encourage compliance; (ii) relatives and friends are likely to be from a socio-economically similar population. As household contacts will share the same environment, it is appropriate to choose a non-

household relative. A non-blood relative, if appropriate, can be useful if genetic factors may be of relevance. The typical strategy for friend (US 'buddy') controls is to ask the case to supply the names of three friends of the same age and sex, from whom the investigator then selects one at random for study. One potential problem with either friends or family is one of *over-matching*, that is, the case and controls are too similar with regard to the exposure under investigation. For example, smokers tend to have friends who smoke and the same applies to alcohol consumption, aspects of diet, level of physical activity and other lifestyle factors. An unexpected problem with this approach is the poor response. Many cases are reluctant to involve friends or relatives and tend to select unrepresentative super-healthy and cooperative individuals. One final example of control selection illustrates the difficulties.

Example 8.xii

In a case-control study of women with a chronic disease, the investigator proposed to study, as controls, sisters-in-law of the cases on the basis that, unlike friend controls, the cases did not have to make the selection, only provide the names. These middle-aged women should have been very likely to have at least one sister-in-law as the term encompasses wife/wives of the brother(s), sister(s) of the husband, and indeed wife/wives of the husband's brother(s)! In the event, the investigator was able to obtain data on sisters-in-law for only 20% of the cases – family feuds, unknown whereabouts and the like were more common than expected!

8.3 One or two control groups?

If no single control group is ideal, it is a reasonable and indeed prudent approach to recruit two very dissimilar control groups, with different selection factors, in the hope that the comparison of the case group with either will reveal the same answer as regards association with the risk factor under study. In these circumstances it would be very difficult to argue that control selection had explained the results obtained. Differences in results using two comparison groups indicate that differences in the selection methods employed do affect the results and conclusions. In the absence of an obviously superior control group, interpretation of the results can be difficult. However, the absence of a second control group may result in inappropriate confidence being placed from the results on the single control group.

Example 8.xiii

In one study of a rare connective-tissue disease, cases were recruited from the national membership list of a patient self-help group. Two control groups were selected. The first was from the members of a self-help group, with a disorder thought to be unrelated to the exposure being investigated. A second group was chosen from primary-care registers. The aim was to distinguish factors associated with disease from those associated with self-help group membership.

8.4 Matching

This refers to the process by which the investigator attempts to ensure that the cases and controls are similar with respect to those other variables that are risk factors for the disease and are associated with the exposure under study, so-called *confounding* variables (see Chapter 18).

Example 8.xiv

In a case-control study investigating the risk of smoking for chronic lung cancer the cases had both higher smoking rates and lower consumption of fruit and vegetables than the controls. It is difficult to disentangle whether there was a true effect of smoking, independent of level of fruit and vegetable consumption.

In this example diet is a potential confounder for the effect of smoking. There are two possible approaches to confounding variables. The first is to match controls so that they have the same exposure to the confounding factors as the cases (this can either be done on an individual case-control basis or on a group basis). However, collecting information on diet from potential controls would involve a substantial amount of work. The alternative is to recruit controls, to measure the confounding factor of interest, and to deal with this issue in the analysis stage (see Section 18.4).

8.4a Problems with individual matching

There are also a number of specific practical problems with matching individual cases to one or more controls:

(i) It is subsequently impossible to examine the influence of any of the matching variables on disease risk.

(ii) There is a danger of overmatching, i.e. controls are selected to be so similar to the cases in respect of their risk of exposure that they end up with an almost identical exposure frequency.
(iii) One can only match if data on matching variables are available, in advance, to the investigator. Apart from age and sex (and perhaps proxy measures of social class), this is rarely the situation. Thus, to match for smoking status would require a prior survey to determine the smoking status of the potential controls!
(iv) It may be impossible to find a perfect match if a number of different variables are being controlled for at the same time. The search for controls then becomes expensive and time-consuming.
(v) Matching may also be inefficient: if no controls can be identified for a case or the controls chosen refuse to participate, the case also has to be excluded from the analysis.

There are, however, some situations where matching may be of value, particularly when the cases are recruited from diverse populations, each with the possibility of unknown confounders. This is useful because analysis can take account of only those potential confounders for which data have been collected. Although confounding issues can be dealt with at the analysis stage one is relying on having a reasonable proportion of both cases and controls exposed to the confounding factor of interest. If, for example, in a study of lung cancer, approximately half of the controls but very few of the cases report regular fruit and vegetable consumption, then dealing with the possible confounding effects of diet at the analysis stage will be very inefficient.

There may be specific situations where matched-pair designs are appropriate. One classical example is in disorders whose aetiology is presumed to lie in the influence of specific environmental susceptibility factors on a particular genetic background. In this instance, a case-control study examining environmental factors in disease-discordant identical twins is a clear way of matching for the genetic factors.

8.4b Frequency matching

An approach which retains the advantages of matching but which overcomes some of the problems associated with individual matching is 'frequency' matching. This ensures that overall the groups are similar even if individual matched pairs cannot be identified.

Example 8.xv

In a case-control study on the effect of various environmental exposures on the risk of a relatively rare congenital malformation, the investigators recruited cases from the obstetric registers of four large maternity units. The investigators wished to ensure that the cases and controls were similar in relation to maternal age, parity and calendar year of birth, each of which might be a potential confounder. They used the facilities afforded by the obstetric databases to group their non-affected births by these variables and take random samples within each group in proportion to the number of cases from each group. This ensured that the distributions of these key variables were similar between cases and controls. This also proved a substantially easier option than attempting individual matching.

8.5 Study size

The final consideration in setting up a case-control study is deciding on the appropriate sample size. This is a crucial step to ensure that the study has a reasonable expectation of being able to produce a useful answer. If the calculations suggest that more cases or resources are needed than are available to the investigator, the study should not proceed. Conversely (and very rarely) it may be that the number needed to be studied is smaller than expected and that expense and effort can be saved.

The major effects on sample size are the expected frequency of exposure in the controls (which requires either prior knowledge or a pilot study to determine) and the size of the effect that the investigator would not want to miss. First, it seems intuitively likely that the rarer the exposure the larger the number of cases required. Indeed, the case-control approach is inefficient in the face of a rare exposure. Secondly, if the investigator wishes to be able to detect an increased risk of, say, 1.5-fold, this would require a larger study than if only large increases in risk were being sought.

There are many simple-to-use software programs available that make sample size calculation easy. As an example, the data in Table 8.2 were obtained from the 'EPI INFO' package and cover a broad range of exposure frequencies and risks. The table shows the substantial influence of the control exposure frequency on sample size. The figures also show how much easier it is to detect large increases in risk than small ones. If exposure is being stratified into a number of different levels, a greater number of individuals will need to be studied. Similarly, if interest is in the interaction

Table 8.2. Sample size for case-control studies

Estimated proportion of exposure in controls	Number of cases required for minimum odds ratio to be detected				
	1.5	2	3	4	5
0.05	1774	559	199	116	81
0.10	957	307	112	67	48
0.20	562	186	72	45	33
0.30	446	153	62	40	31
0.40	407	144	61	41	32

Note:
Assume power 80%, 95% confidence level, and one control per case.

Table 8.3. Sample size for case-control studies

Estimated proportion of controls reporting exposure	Power of study				
	70%	75%	80%	85%	90%
0.10	246	274	307	347	402
0.15	181	201	225	255	295
0.20	150	166	186	211	244

Note:
Odds ratio to detect 2.0, 95% confidence level, and one control per case.

between two risk factors then a large increase in the required sample size is inevitable.

The figures obtained should be taken as a guide rather than the exact number to be recruited. Although software programs will provide seemingly 'accurate' sample size requirements (e.g. 429 cases), it should be remembered that uncertainty in the parameter's input (e.g. proportion of exposure in controls) will be reflected in uncertainty of the sample size required. For this reason it is often best, rather than carrying out a single sample size calculation, to conduct calculations under several assumptions.

In a study (with the sample size calculations shown in Table 8.3) the 'best guess' for the proportion of controls exposed is 0.15. It is therefore decided

Table 8.4. Influence on sample size of increasing the number of controls per case

Number of controls per case	Required number of:		
	Cases	Controls	Total
1:1	307	307	614
2:1	223	446	669
3:1	194	582	776
4:1	180	720	900
5:1	171	855	1026
6:1	166	996	1162

Notes:
Estimated proportion of controls exposed 0.10.
Odds ratio to detect 2.0.
Power 80%.

to conduct a study with 255 subjects to give 85% power to detect an odds ratio of 2.0. If, however, the proportion exposed is only 0.10 the study will still have power of 70–75%, while if the proportion exposed is 0.20 power will be greater than 90%.

The previous calculations assume that equal numbers of cases and controls will be obtained: this is the most efficient approach. However, when the number of cases is limited (i.e. when the disease is rare) it is possible to compensate by increasing the number of controls (Table 8.4). As shown in this example, there is little gain beyond having four times as many controls as cases.

Finally, a word of warning: these figures assume that the numbers stated are the numbers actually studied. Ineligible and unconfirmed cases, together with deaths, wrong addresses, refusers and other non-respondents, will need to be taken into consideration. Thus, for example, with a typical final participation rate of 60%, the numbers stated have to be multiplied by approximately 1.7 to provide the number of cases that need to be approached.

9

Studies of disease causation. II: Selection of subjects for cohort (longitudinal) studies

The central issue in setting up a cohort study is the selection of the population(s) to be studied according to their exposure status. As discussed in Chapter 4, the issues are the same whether the aim of the study is (i) to determine the development of a disease in those exposed or not exposed to a suspected particular disease risk factor at baseline, or (ii) to determine the development of a specific outcome in individuals with or without the presence of a disease state at baseline.

The main questions that influence the approach to be used for population selection are:

(i) Should the study ascertain the exposure status from historical data and the outcome currently (the retrospective cohort approach), or should exposure status be ascertained currently with the outcome determined by future follow-up?
(ii) Should the exposure be considered as just present or absent, or at multiple levels?
(iii) Should the different exposure groups studied be derived from a single target population?
(iv) How many subjects should be studied?

9.1 Retrospective or prospective study cohorts?

The choice of retrospective or prospective cohort design was discussed in Chapter 4. At a practical level, the choice is determined by the availability of high-quality exposure data on the one hand, and the practicality and costs of a prospective study on the other. If a retrospective study is being considered the following questions need to be addressed:

(i) Are data available on a sufficiently large number of individuals?
(ii) Is the quality of the exposure data sufficiently high to permit accurate allocation to different exposure category groups?
(iii) If the exposure changes over time, does the information available allow the investigator to assess this?
(iv) Are there any data on potential confounding variables?
(v) Are there sufficient data available from follow-up to determine the disease status of the population both in terms of being able to ascertain completely the presence of disease and to verify the diagnosis in those ascertained?

Two examples illustrating the use of retrospective cohorts may be of help here.

Example 9.i

In a study designed to investigate the relationship between body weight at age 20 and subsequent risk of non-insulin-dependent diabetes, an investigator discovered the availability of army medical records giving details of body weight in male conscripts in the late 1950s. In a pilot study using general practice and other record systems, he was able to trace around 40% of the original cohort, whose diabetic status was then easily assessed. The investigator decided that the study was practicable. He was concerned about this loss to follow-up given that those who had developed diabetes might have died. Clearly, this would need to be taken into consideration when interpreting and extrapolating the results.

Example 9.ii

A paediatrician wished to investigate the relationship between weight gain in the first year of life and subsequent intellectual development. She used records from community child health clinics to identify a group of children born 10 years previously on whom there were data available on birthweight and weight around the first birthday, and was able to link these to subsequent school records.

By contrast, recruitment of a prospective cohort will require knowledge of current exposure status. Depending on the question to be asked, the decision is between using existing data sources or setting up a special survey to determine exposure status. The former is cheaper and quicker. However, the latter permits the investigator to document exposure accurately and to collect data on other important variables. One advantage of the prospective approach is, even if existing data sources are being used, the investigator can influence the collection of the exposure data.

Example 9.iii

An occupational-health physician wished to investigate the association between certain industrial exposures and the risk of skin rashes in an industry. The quality of the exposure data available from the employment medical records was poor and the employees themselves were unaware of their exposure history in sufficient detail. He therefore set up a prospective cohort study and provided the personnel department with a specially designed pro forma on which they could record both current and future exposures in a structured way.

Example 9.iv

A proposed investigation on risk factors for falling in the elderly required the investigators to recruit a population sample that were then surveyed by trained interviewers on possible risk factors for subsequent falls. The interviewers also undertook a limited examination of visual acuity and postural stability.

Clearly, in Example 9.iv there was no alternative to the rather expensive and lengthy recruitment procedure, given the exposure data required.

9.2 How should exposure be categorised?

The categorisation of exposure depends on the main question(s) to be addressed, as discussed in Chapter 4. This decision will influence the choice of populations to be studied and the size of the study (see below).

Example 9.v

The hypothesis was proposed that the serum level of a particular enzyme was a marker for a subsequent increased risk of ovarian cancer. It was decided to compare the cancer risk in those with raised level (above an arbitrary level) with those with a normal level. It was therefore necessary to screen a large population to obtain a sufficient number with a high level for follow-up.

Example 9.vi

In an epidemiological study of the influence of height on the subsequent risk of back pain, the investigator considered that it would be appropriate to consider the risk by dividing the population into five equal groups, or quintiles, by height. The risk of back pain could then be assessed separately in each of the groups. In planning the study, she needed to ensure there would be a sufficient number in each of the five groups to permit valid comparisons.

In such an approach, the groups do not have to be equal in size but may be divided into biologically more coherent groups. The advantage of having equal-sized groups is that it is often statistically more efficient, i.e. the comparisons would have greater power. The decision about the approach to be used in considering the exposure can be (and frequently is) left until the analysis stage.

9.2a Studying the effect of multiple exposures

For the sake of simplicity, many introductory epidemiological texts imply that cohort studies are best applied to single exposures, and preferably those that can be categorised into present or absent. In practice, the relatively recent ready availability of the necessary computer hardware and software has permitted multiple exposures to be studied simultaneously and their separate risks assessed.

Example 9.vii

An investigation was planned to determine the risk factors for premature death in elderly women admitted to hospital with a fractured neck of femur. Multiple exposure data were gathered at baseline assessment, including mental test scores, indices of body mass, presence of related disorders and haemoglobin level. Thus, this cohort study did not consist of a single exposed population, but, rather, a single study population from whom the various exposures could be ascertained for subsequent analysis.

9.2b Ascertainment of exposure status

Given the above, the most obvious strategy is to select a 'whole population' and investigate the exposure status of every individual, perhaps by a screening survey. The population is then divided into two or more exposure groups depending on their exposure status at screening. The advantage of this approach is that both the 'exposed' and the 'non-exposed' groups are truly representative of the exposed and non-exposed individuals from the population under study. The anxiety is that when the exposed and non-exposed groups are selected from different 'parent' populations, it may be impossible to separate the influence on disease risk of differences in other disease-related variables in the two populations. This strategy, however, is unavoidable if the exposure is relatively rare and the investigator needs to use an 'enriched' sample to obtain sufficient numbers of exposed individuals.

Example 9.viii

A gastroenterologist wished to investigate the hypothesis that being a vegetarian was associated with a reduced risk of developing large-bowel cancer. In order to obtain sufficient numbers, he recruited his 'exposed' cohort from the membership of local vegetarian societies and compared their risk with populations drawn from general practice who had been surveyed about their meat-eating habits. He obtained data on other variables such as alcohol consumption in both groups as he recognised that members of the vegetarian societies might have differed in other ways in regard to their risk of developing bowel cancer.

In the above example, the investigator attempted to collect data on all other variables (possible confounders, see Chapter 18) that could be adjusted for in the analysis, with the hope that if a protective effect was seen in the vegetarians it could be determined whether this was due to having a vegetarian diet. The problem is that there may be unknown variables associated with being a member of vegetarian societies that are thus impossible to measure. In these circumstances, rather than not do the study, the best option is to do what is reasonable and practicable, but note in the report that it is impossible to exclude the possibility of a selection effect.

9.3 Study size

In all epidemiological studies, recruitment of sufficient sample sizes for study is of crucial importance. As with case-control studies, the sample size is determined by the statistical power considered appropriate to detect stated increases in risk, traditionally at the 95% confidence level. The determinants of sample size are thus the increase in risk to be detected, the background risk in the unexposed or lowest exposed group, and the frequency of exposure. The calculations are readily done with widely available software packages. Table 9.1 provides some typical sample-size calculations under a variety of assumptions. The first half of the table assumes either that the population can be split equally into two, based on their exposure status, or that the investigator aims to recruit an enriched exposed cohort and an equal-sized unexposed group. The second half of the table is based on the desire to investigate a rare exposure (approximately 10%) of the population. The figures are based on the cumulative occurrence of the disease outcome by the end of follow-up. In an analogous fashion to Table 8.3 it is often

Table 9.1. Sample-size calculations for cohort studies

Frequency of exposure (%)	Cumulative disease risk in unexposed (%)	Risk ratio to be detected	Number of subjects to be recruited	
			Exposed	Unexposed
50	10	3	71	71
		2	219	219
		1.5	725	725
	5	3	160	160
		2	475	475
		1.5	1550	1550
	1	3	865	865
		2	2510	2510
		1.5	8145	8145
	0.1	3	8800	8800
		2	25000	25000
		1.5	82000	82000
10	10	3	35	305
		2	110	990
		1.5	380	3410
	5	3	75	665
		2	235	2115
		1.5	810	7260
	1	3	400	3580
		2	1240	11170
		1.5	4230	38060
	0.1	3	4000	36400
		2	12500	113000
		1.5	43000	385000

prudent not to rely on a single power calculation. Having established the order of magnitude of certain parameters (e.g. cumulative disease risk in unexposed), it is useful to examine the effects on power of variations around this estimate (Table 9.2).

Manifestly, the rarer the disease the greater the number that need to be studied. Similarly, to uncover a largish effect (risk ratio >3) it requires a substantially smaller sample size than it does to detect a smaller ratio of (say) 1.5.

Table 9.2. Sample-size calculations for cohort studies

Cumulative disease risk in unexposed (%)	Power of study (%)				
	70	75	80	85	90
12	142	158	176	199	230
10	176	196	219	247	286
8	227	253	282	319	369
6	313	347	389	440	509

Notes:
The number of cases required is shown assuming: ratio of exposed : unexposed, risk ratio to be detected = 2.0, 95% confidence interval, frequency of exposure 50%.

If the exposure is rare, the number of exposed subjects that need to be studied can be reduced by increasing the number of the non-exposed. Conversely, using an enriched exposure sample, to give equal-sized exposed and non-exposed cohorts, can clearly reduce the overall size of the study. The decision is therefore a compromise between cost and the difficulty in obtaining a sufficiently large exposed group.

The sample size can be reduced by continuing the follow-up for longer if disease risk remains constant or increases over time. Thus, if the disease risk is a uniform 1% per annum, the longer the period of study the greater the number of cases that develop and thus the smaller the numbers to be followed up. The balance between long follow-up of a relatively small cohort compared with short follow-up of a relatively large cohort depends on the cost implications, the requirement to obtain a quick result, the problems with loss to follow-up with increasing time, and the pattern of disease risk amongst other factors. For example, a short follow-up may also be inappropriate if the latency between exposure and disease is long. This is, for example, particularly true in relation to cancers, where too short a study may fail to detect a real effect. It also must be remembered that the sample size calculations assume no loss to follow-up and they will need to be weighted accordingly to take account of the estimated losses.

Part V

Information from epidemiological surveys

Conceptually, the logical next step, having selected the populations, is to obtain information from the members of the populations chosen that is needed to answer the questions posed. The information normally required is about disease status and/or (current or past) exposure to the factors under investigation. In the previous section, the possible approaches to recruiting study populations for the various types of epidemiological study were discussed. It is appropriate in this section to consider the problems of information-gathering across the spectrum of study designs as a whole. The issues are the same whether a case-control design is used, where recruitment requires accurate information on disease status and the 'body' of the study requires information on possible linked exposures, or a cohort design is used, where the reverse is the case.

Four practical problem areas are addressed in the chapters in this section in relation to the obtaining of information. These are:

(i) What are the best approaches to obtain information?
(ii) How valid (accurate) is the information obtained?
(iii) How reproducible (consistent) is the information?
(iv) How best to maximise participation in a study and/or to effectively follow-up subjects over time?

Having made decisions about how to obtain valid and reliable information and to achieve high participation and follow-up rates, it would be prudent, prior to embarking on a large (and costly) study, to test some of your assumptions and methods. The final chapter in this section is therefore about conducting a pilot study.

10

Collecting information

Epidemiological information comes from a variety of sources. These may be conveniently divided into those that are available from previously documented data and those that require the gathering of information specifically for the task in hand.

Examples of the former include disease registers, medical records, occupational records and related data sources. New information gathered may also require data from clinical examination, blood, urine and related tests. The most frequent approach is to use a questionnaire to obtain data direct from the subject. It may be necessary to obtain data about the subject from a proxy respondent if, for example the subject is a child, is too ill, is dead or is otherwise incapable of answering. The questionnaire poses a number of methodological challenges. This chapter is therefore focused on the methodological aspects of questionnaire development and design.

10.1 Interview or subject completing questionnaire?

The two most frequently adopted tools used in information-gathering epidemiological surveys are the interview-administered and the self-completed questionnaire. The clear advantage of the latter is that it can normally be distributed by post, thereby covering a large population, at little cost, in a short period of time. There are, however, other considerations in deciding between these two modalities (Table 10.1).

(i) The response rate may be higher in the interview approach, which necessitates a direct contact (including by telephone) to the subject. The main reason for non-participation is apathy rather than antipathy, and it is the experience of many that direct personal contact frequently

Table 10.1. Interview or subject completing questionnaire?

	Interview	Subject
Cost	High	Low
Response rate	May be higher	May be lower
Completion of questions	High	May be lower
Complexity of questions	Can be high	Should be minimised
Interviewer bias	May be present	Not relevant
Interviewer variability	May be present	Not relevant
Sensitive questions	May be difficult	May be easier
Total study duration	Slower	Rapid
Recall	May be different	

ensures a greater response. This advantage is lost, however, if the initial request for interview is sent by post, requiring a positive response.

(ii) The completeness of the answers can be more assured by the interview approach. In self-completed questionnaires, questions may be ignored either deliberately or by accident. A useful ploy to minimise this in postal questionnaires is to add at the bottom something to the effect of, 'Can you please check that you have filled in all the relevant questions before returning this questionnaire.' Sometimes questionnaires can be fairly complex in structure, for example, 'if you answer "no" please go to question 7, otherwise go to question 11'. Such requirements, though simple, can result in confusion. The best answer is as far as possible to avoid these internal routings, otherwise an interview-administered instrument may be a better approach.

(iii) A related issue is that complex questions in self-completed questionnaires may be misinterpreted and answered inappropriately. By contrast, the interviewer can guide the answers and explain those that are more complex. Some areas, such as details of medical history and occupational exposure, are difficult to encapsulate in simple, easily answered questions, whereas the trained interviewer can separate the relevant from the irrelevant and obtain a more useful product.

(iv) By contrast, some issues, such as those relating to sexual and family relationships, may be better addressed by a postal questionnaire, where

the subject does not have to confront the interviewer. Typically, adolescents, for example, may be more willing to fill in an anonymous questionnaire about various aspects of their lifestyles than relate them to 'an adult in authority'.

(v) Interviewers are also subject to bias, particularly if they are aware of the hypothesis under investigation. They may, for example, probe more deeply in investigating cases than controls for a particular premorbid exposure.

(vi) A related issue is that in studies involving more than one interviewer, there is scope for variability due to differences in the approach used by the interviewer and in the interpretation of any replies. Some of this may be overcome by training and standardised methods of administering the questions. In large surveys, the interviewers are frequently given a scripted text to read to ensure that they are presenting the material in a similar manner. It is difficult to exclude the possibility of interaction between the subject and the interviewer, with subjects responding differently to different personality types, however standardised the delivery of the questions.

(vii) Recall by the subject may be different when given the same question by an interviewer and in a self-completed questionnaire. The setting for completion is different between the two. In the interview situation there may be no time for the subject to consider his or her answer and/or to change it after an initial response. By contrast, given the greater time for completion of a postal questionnaire, a motivated subject may enquire of close family members and look for other supporting information for previous events. It is perhaps not surprising that in many situations the same question, administered by an interviewer, produces only half the rate of positives compared with a self-completed questionnaire.

Summary

There is no single best approach. Financial and time constraints may exclude an interview-administered survey. If that is not the situation, the decision has to be made taking account of the points above. Ultimately, it is the nature of the information requested that will determine which of these imperfect approaches is the preferred option.

10.1a Interview-administered questionnaires by telephone

The use of the telephone is an increasingly attractive compromise between the face-to-face interview and the postal questionnaire. It is cheaper, saving in time and travel costs. Many subjects, reluctant either to attend for an interview or to invite a stranger into their home, may be content to answer questions over the telephone. The telephone option also permits the initial contact and the information gathering in a single act. By contrast, the interview done face to face requires an initial contact either by telephone or post, with the likelihood of a higher non-response. However, there are a number of disadvantages compared with the postal questionnaire and these are:

- population selection is restricted to those with a telephone, leading to under-representation of lower social groups in some countries;
- telephone numbers may not be listed in directories;
- contact will need to be made outside normal office hours to maximise participation;
- multiple attempts may be necessary to 'capture' a given subject;
- a direct refusal has to be accepted, whereas in a postal survey a second or subsequent mailing may be issued to initial non-responders.

10.2 How to formulate a questionnaire

This is one of the most difficult tasks in survey research and there can be few investigators, despite the strenuous attempts at piloting and pretesting their questionnaire, who would not change their instrument if they were to repeat the study. The most common mistake is to ask too many and too detailed questions that are never analysed. The authors' experience, of sending out a questionnaire for comment from colleagues, is to receive numerous suggestions for additional topics ('Then while you are asking about *x*, why not ask about *y* syndrome?'). The computer files of many epidemiologists are littered with answers to questions never to be analysed!

10.2a Open or closed questions?

An important issue in survey research is the choice between open and closed questions. Unlike social science, most epidemiological surveys adopt the closed question approach as being the most efficient approach for data handling and analysis. A typical example of the contrast is shown in Table 10.2. In

Table 10.2. Open and closed questions

Open

What sports and other physical activities do you undertake each week on a regular basis (at least 30 minutes)?

.....

.....

Closed

For each of the following sports tick the box if you regularly spend more than 30 minutes each week in that activity.

Walking	☐
Jogging	☐
Cycling	☐
Swimming	☐
Racket sports	☐

Table 10.3. Open and closed questions?

	Open	Closed
Subject recall	Reduced	Enhanced
Accuracy of response	Easier to express complex situations	Difficult to investigate complex situations
Coverage	May pick up unanticipated situations	Will miss areas not anticipated
Size of questionnaire	May need fewer lines of text	May need many pages of text
Analysis	More complex	Simpler

the first situation, the subject is being asked to consider what activities he or she undertakes that would fit in with this description and to write them down. In the alternative formulation, the subject is presented with a list of options to make a decision (yes or no) about each one. Not surprisingly, the decision about the better approach is not absolutely clear (Table 10.3).

Subject recall may be enhanced by the closed approach. Thus, in the example given in Table 10.2, walking and cycling may not be considered by

some subjects as 'physical activities' and thus are ignored. By contrast, the open question can detect an enormous number of activities, which would be impossible in the close approach without an overlong survey form, much of which would be left blank. The open question allows the subject to address and describe complex issues which the investigator had not considered: thus, if the subject is a football referee, he or she may be reluctant to tick against *playing* football, yet the investigator may wish to detect this activity. Results from open questions are more difficult to handle analytically. The design of the database to handle the answers will need to take account of the material received. Similarly, the material will need to be carefully reviewed to determine how it fits in with the planned analysis. The most useful approach, in the absence of any previously available off-the-shelf questionnaire, is to use an open approach in developing and piloting the questionnaire and then to use closed questions formulated on the basis of the answers received. A modification of the closed approach which may be appropriate is to provide a list of options which includes 'Other'. If respondents tick this box they can be invited to 'Please specify'.

10.2b Questionnaire design

The rules of questionnaire design are straightforward. Questions should:

- be simple,
- cover one dimension,
- have a comprehensive choice of answers, and
- have a mutually exclusive choice of answers.

To expand on these: first, the questions should be simple and cover one dimension only. It is better to ask separate questions, for example:

1. Have you ever smoked?

Yes No

If yes,

2a. What age did you start smoking?

Age . . . . . .

2b. Do you regularly (at least one cigarette per day) smoke now?

Yes No

2c. If you have stopped smoking, what age did you last regularly smoke?

Age . . . . . .

rather than the more complex question:

1. If you have ever regularly smoked (at least one cigarette per day), how many years was this for?

The former approach is simpler to complete and does not require the subject to perform calculations with the risk of error or lack of precision.

Secondly, in any closed question with a choice of answers, the choice should cover all possible answers including 'don't know/can't remember' which is different from 'no'. The choice of answers should be mutually exclusive and it should not be possible for a respondent to tick more than one choice unless this is specifically requested.

Some examples of poor question design are shown in Table 10.4. Amongst other items, they display the problems that subjects may have in recall and the problem of lack of mutual exclusivity. Thus in the second question the 'it varies' option may be the most accurate answer for all but the teetotal, ruling out any useful information emerging.

There are numerous research studies discussing colour, size, layout and length of questionnaires produced for self-completion. Some are inevitably more 'user-friendly' than others, enhancing response. Certainly, desktop publishing techniques have enhanced the appearance of questionnaires, but it is not clear whether these aspects substantially improve either the response rate or the quality of the answers: ultimately it is the quality of the questions that is the crucial factor. Some design features seem useful. A front page just with a study title and a few lines of instructions seems to be preferred in so far as there is a perception that the answers are kept 'under cover'. The print should be large enough and sufficiently widely spaced to be read with ease. Complex words should be avoided, as should long sentences. Shading may be used to highlight where the answers should be inserted. Instructions should always be given on what to do with the questionnaire after completion, including provision of prepaid envelopes for return.

Table 10.4. Examples of poor question design

Question	Problem
What was your height at age 25? _____	Subject may have difficulty in recalling height. No option for 'can't remember', and thus subject may be forced into inaccurate response
How many days a week do you drink alcohol? Never ☐ 1–2 ☐ 3–4 ☐ 5–6 ☐ Every day ☐ It varies ☐	The 'It varies' option is difficult to follow. As no subject is likely to have an identical pattern every week, then expect a substantial proportion completing this non-informative option
What is your current occupation – if not currently employed, please put your most recent job? _____ _____	The question is too vague concerning the amount of detail required. The questions need to be very specific, requesting name of organisation/firm and exact job title
Do you take any of the following drugs or medicines? Pain killers ☐ Anti-dyspepsia ☐ Diuretics ☐	Subject may not understand medical terms. Better to ask an open question to list names of current medication
Have you ever taken the oral contraceptive pill? yes/no If yes, how long did you take it for? _____ years	A superficially simple question that is full of problems. Does taking one tablet count as 'ever taken'? More problematically, it is very difficult – if not impossible – to aggregate cumulative consumption over a lifetime, taking into account breaks for pregnancies, etc.
How many times a week do you eat cheese? _____ times	One of the problems with this type of question is asking about a theoretical time period; far better to ask about a real time period, e.g. last week

11

Obtaining valid information

11.1 Introduction

The information obtained in any study should be valid, i.e. the 'truth'. The limitations imposed on studying 'free-living' populations, as opposed to studying volunteers or laboratory animals, is that indirect methods frequently have to be used to obtain the data required, often by interview or questionnaire. The answers obtained may not represent the true state of the individual.

Example 11.i

A questionnaire is widely accepted as a simple means of screening a population for the presence of angina (the Rose Angina Questionnaire). It relies on self-reports of exertional chest pain. Clearly, there will be errors in its use in classifying coronary artery disease, but the alternative of coronary artery angiography is not appropriate for an epidemiological study.

Example 11.ii

As part of a study, it was necessary to investigate the dietary intake of vitamin D. The most valid approach would be for the subjects to weigh all food consumed during the period of study and provide duplicate portions for biochemical analysis. The approach chosen was 24-hour recall, where the subject recalls his or her total dietary intake during the previous 24 hours. This approach was substantially more acceptable and less costly.

The issue, at a practical level, is therefore not which is the valid method, but what is the size of the error in adopting the feasible option, and, by implication, does an error of this size alter the interpretation of the data obtained?

Table 11.1. Classical assessment of validity

	'The truth'	
Test result	+ve	−ve
+ve	TP	FP
−ve	FN	TN
	TP + FN	FP + TN

Notes:
TP, true positives, correctly identified;
FN, false negatives;
FP, false positives;
TN, true negatives, correctly identified.

$$\text{Sensitivity} = \text{Proportion of persons with disease correctly identified as disease positive} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \text{Proportion of persons without disease correctly identified as disease negative} = \frac{TN}{TN + FP}$$

11.2 Sensitivity and specificity

The reader will probably be aware of the application of validity tests in assessing the performance of diagnostic and screening tests. In this situation, the objective is to evaluate the ability of a test to distinguish correctly between true disease-positive and true disease-negative individuals. It is conventional to evaluate this performance in terms of the two indicators as shown in Table 11.1 *sensitivity* and *specificity*.

Sensitivity is the proportion of subjects who truly have disease, who are identified by the test as 'disease positive'. Specificity is the proportion of subjects who truly do not have disease, who are identified by the test as 'disease negative'. These proportions are often multiplied by 100 and expressed as a percentage.

When the result of a diagnostic test is dichotomous (positive or negative), the investigator simply needs to calculate the sensitivity and specific-

ity of the test and decide whether they are satisfactory for the purposes intended. Often, however, the result of a test may be a category or a reading on a continuous scale. In these cases the investigator will need to evaluate the specificity and sensitivity at each (or a sample) of possible cut-off definitions for positive and negative. Thereafter a choice will need to be made of the cut-off that optimises sensitivity and specificity for the purposes intended.

Example 11.iii

A study measures the concentration of substance X in the urine as a screening test for disease Y. In practice, during a population study all those screened have a concentration between 6 and 92 mg/l. There is an overlap between the distributions of concentrations between those with and without the disease. At the extremes, if the cut-off for disease is taken as a score of ≥ 0 then the sensitivity of this cut-off will be 1, i.e. everyone with the disease will be labelled as disease positive. It is clearly not a useful cut-off; however, since the specificity is 0, i.e. no-one without the disease is labelled as disease negative. Conversely if the cut-off is >95, the disease sensitivity is 0 and the specificity 100. The challenge will be to find a cut-off value between these extremes that satisfactorily optimises the values of sensitivity and specificity.

There is no 'magic figure' for either sensitivity or specificity above which a test can be said to perform satisfactorily. If the test result is a score then the investigator, by making the cut-off less stringent, will be able to increase the sensitivity. However, specificity is also likely to fall. One approach to choosing a cut-off is to plot a graph of sensitivity v. 1 − specificity – a receiver operating characteristic (ROC) curve (Fig. 11.1).

The diagonal (dotted) line in Fig. 11.1 represents the results from a hypothetical test that was no better than random at distinguishing positive from negative. The more discriminatory a test, the steeper the initial upper portion of the curve, and the greater the area under the curve. One possible 'optimal' cut-off is to choose the value corresponding to the shoulder of the curve (the point nearest the top left-hand corner of the graph).

The most suitable cut-off point will depend on its intended use. If screening for cancer one may wish to choose a cut-off other than the so-called 'optimal' point. By increasing the sensitivity of the test, the number of false negatives would be reduced. This would likely be at the expense of decreasing the specificity, i.e. the number of false positives would rise.

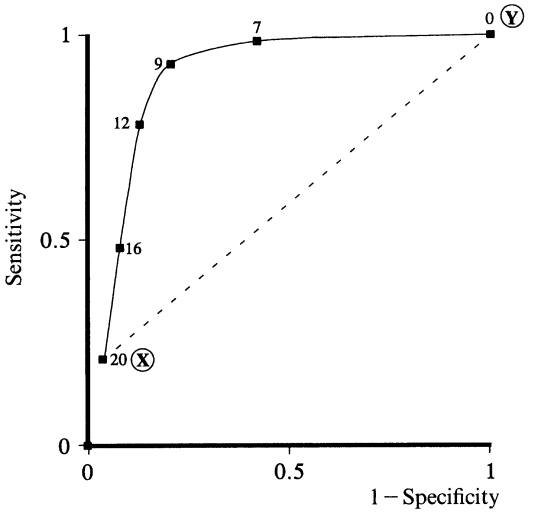

Figure 11.1 Example of ROC curve: results of an interview questionnaire to determine presence of clinical depression. The 20-item questionnaire yields a score between 0 and 20. The values for sensitivity and specificity are calculated from a sample of 200 individuals, all of whom were interviewed with the use of the questionnaire and, additionally, clinically evaluated by a psychiatrist.

At cut-off point X (equivalent to a cut-off of 20), very few individuals score positive, hence detecting few cases (sensitivity 0.2) but also few false positives (specificity 0.95). At cut-off point Y (equivalent to a cut-off of 0), all individuals are positive and hence the entire population is detected as a case (sensitivity = 1) with all true negatives classified as positive (specificity = 0).

The results from five of the potential cut-offs are shown. At a score of 9 or more, the interview score would appear to be at its most discriminatory. A higher cut-off (e.g. 12) would miss too many true cases. A higher cut-off would result in a steep drop in sensitivity with only a small improvement in specificity, and vice versa for a lower cut-off.

Example 11.iv

In a prospective large-cohort study of the possible long-term hazards from the use of the oral contraceptive pill, the aim was not to miss any possible morbid event. In seeking, therefore, to maximise their sensitivity by the ascertainment of all subsequent cases with a stroke, the investigators included a very broad range of neurological signs and symptoms that would be considered as positive.

Alternatively, in other circumstances high specificity will be important.

Example 11.v

In a case-control study of hypertension, the decision was made to restrict recruitment of cases to those individuals who had an untreated diastolic blood pressure of greater than 100 mmHg, sustained over three readings taken at weekly intervals. The desire was to have maximal specificity and not include individuals with a transient rise in blood pressure due to anxiety.

The above discussion has focused on determining the validity of disease classification. However, the same principles apply to the classification of exposure in epidemiological studies. If it is appropriate to consider the exposure as dichotomous (exposed/not exposed) and the results from a gold standard are available, then the sensitivity and specificity of the method used can be determined as above.

Example 11.vi

In a large prospective study aimed at determining whether a particular trace-element deficiency was linked to the development of stomach cancer, deficiency was defined on the basis of a simple 'dipstick' test on a random urine sample. The validity of this approach had been previously investigated by comparison with the more cumbersome and costly collection and analysis of 24-hour urine samples collected from 25 volunteers.

11.3 Validity for variables that are not dichotomous

More typically, particularly in relation to exposure, a state may not be dichotomous, and then the question at issue is not whether the survey method got the right answer, but how close the value obtained was to the true value. The consequence of any impaired validity is misclassification of either disease or exposure status. The investigator therefore needs an estimate of the extent of any misclassification. One approach is to grade misclassification, based on the comparison of the answers to a survey with the 'true' grade.

Example 11.vii

The validity of self-reported levels of alcohol consumption was assessed by comparison with presumed 'true' data derived from a sample of spouse interviews. In this example, one can distinguish minor misclassification (one grade out) from more major misclassifications (Fig. 11.2). The weighting given to the misclassification depends on the investigator.

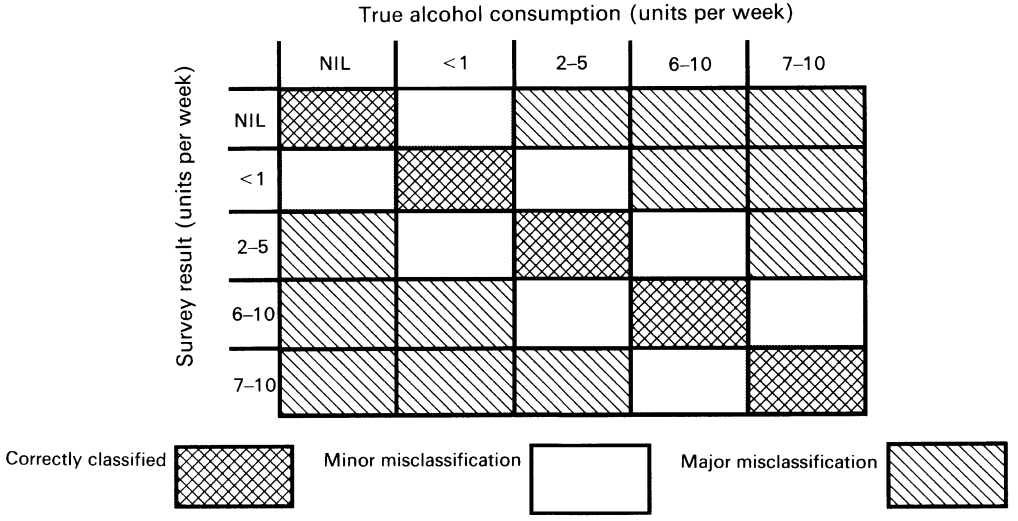

Figure 11.2 Assessment of validity in more complex situations.

11.4 Possible approaches for independent validation

In the situations described in Examples 11.vi and 11.vii, the assumption is made that there is an independent gold standard measurement that will give the truth and allow an evaluation of validity and the size of the potential problem of misclassification. In reality, this is frequently neither available nor practical. Survey methods are often based on uncorroborated subject recall, and validation of the answers appears elusive. There are, however, a number of approaches that can be used to attempt to measure validity:

- validate in other studies;
- contemporary documented data, e.g. medical records, pregnancy/birth records, occupational/personnel files;
- reports from spouse or close household contact or other close contact;
- in-depth investigation of random sample of respondents.

Note the following:

(i) It is always worthwhile to use a method that has been validated in other studies of similar populations. Thus, the questionnaire to detect angina, mentioned in Example 11.i, has been validated against clinical

diagnosis. No further validation is necessary unless it is thought that the characteristics of the questionnaire are such that there may be differences in sensitivity between populations, perhaps due to variation in language. There are many other examples where standard questionnaires have been validated against an independent gold standard, for example on medication use or dietary intake. Unfortunately, secular, social and geographical differences between populations mean that it may not be possible to assume that if a survey method has been shown to be valid in one group, it is necessarily valid in another.

(ii) Where possible, the investigator should obtain contemporary collected information to validate recalled responses by using such sources as medical and occupational records.

Example 11.viii

A self-complete questionnaire was used to estimate previous oral contraceptive use. It was possible to validate the responses by checking against the respondent's general practice and family planning clinical records. Indeed, it was only necessary to study a small sample, at the start, to confirm the validity of the method used. It was also necessary to examine the records of a sample of those who denied ever taking oral contraceptives to confirm the validity of the negative response.

(iii) In the absence of contemporary records, it is often possible to corroborate recalled information against that obtained from a household or other close informant. Again, this can be restricted to investigation of a sample. The problem in this situation is that when the two sources give conflicting answers, it cannot necessarily be assumed that one is inherently more accurate than the other, although, intuitively, the interpretation of such conflicts will depend on the topic investigated.

(iv) If no other sources are available then the investigator may have to find an alternative method of investigation for validating results from a sample, which might require an invasive or semi-invasive procedure. Some examples of this are shown in Table 11.2. Thus, urine cotinine is a sensitive test for cigarette smoking. A seven-day weighed-diet survey is probably the gold standard against which other dietary survey methods should be compared. In postal questionnaires using self-recording of height and weight, these measurements may be validated

Table 11.2. Examples of validation approaches for samples from surveys

Survey variable	Validation approach
Cigarette smoking	Urine cotinine Blood carboxyhaemoglobin
Dietary intake based on food frequency	Seven-day weighed diary record
Self-reported height/weight	Standardised measurement
Self-reported disease status	Evaluation by experienced clinician

against the results obtained from true measurements. In surveys designed to detect clinical conditions, self-reported diagnoses, for example of psoriasis, may be checked with evaluation by 'an experienced physician'. The hope is that, first, the survey method will be shown to be valid and, secondly, it will be necessary to undertake the validation in only a sample.

11.5 Misclassification

The final consideration, in relation to validity, is that it is worth remembering that the problem is one of misclassification. An imperfectly valid method will misclassify individuals, as regards either their disease or their exposure status. Perhaps surprisingly, misclassification need not be a disaster. If, in comparative studies (such as case-control or cohort investigations), the misclassification is *random* then the effect will be to make it more difficult to discover a real difference against this background 'noise'. Thus, if the study detects a difference (say) in exposure between cases and controls, then the fact that there is random misclassification could not explain the result, which still stands. Indeed that true association between an exposure and disease state will be greater than that observed in the study.

Example 11.ix

A case-control study of middle-aged women examined the influence of self-reported weight at age 25 on current bone density. It was found that there was a substantial positive association between current bone density and recalled weight. The investigators, although anxious

about the dubious validity of recalled weight from many years previously, thought it unlikely that those with low bone density were more likely to underestimate their recalled weight than those with high bone density.

Non-random misclassification is more serious and is based on there being differences in the validity of the survey method in its application between groups. Here the effect on the study results cannot be predicted, and depends on the direction of misclassification. Hence it may be possible to explain an observed association by non-random misclassification. These concepts are discussed further in the discussion of bias in Chapter 19.

12

Repeatability

12.1 Introduction

The second consideration in obtaining information from or about individuals is the difficulty in obtaining consistency of the results. A number of terms are used in describing this phenomenon. They are often used interchangeably, although some authors have attempted to distinguish between them.

(i) Repeatability: strictly, given the same set of circumstances, how consistent are the results obtained?

(ii) Reproducibility: best considered as an alternative expression for repeatability.

(iii) Reliability: normally reserved to determine consistency between observers or between techniques. If consistent results can be explained by true subject change, this is therefore not due to poor reliability of the measurement.

(iv) Agreement: typically reserved for consistency between observers.

Common sense determines that there may be a variety of reasons for failure to obtain consistent or reproducible results. First, there may be true subject variation. This is particularly relevant in physical measurements such as blood pressure, heart rate and similar variables that, in normal individuals, vary with time, even over short periods. The greater concern, however, is that of a lack of consistency in the result obtained in the absence of true subject change. This lack of consistency may be due either to a single observer being inconsistent in the method of measurement or to inconsistencies between multiple observers. If this occurs to any major extent, data obtained from studies may be difficult to interpret because apparent differences between subjects might reflect differences in measurement.

12.1a Variation within and between observers

In many studies it may be impossible to use a single observer to obtain all the necessary data and there is therefore a need to use multiple observers. However, a single observer may prove to be just as inconsistent over time as multiple observers. In theory, with rigorous training and standardisation of approach, it should be possible to measure and minimise inconsistencies.

Example 12.i

A study involved the use of a stadiometer to measure height. In pilot studies of a single observer undertaking repeated measurements on the same subjects, some inconsistency was noted. After extra training this inconsistency disappeared.

Example 12.ii

In the same study as 9.1, a second observer had difficulty in obtaining acceptably close results to the first observer when assessing the same test subject. Again, training removed this inconsistency. As a quality control against 'drift', the two observers were checked every three months during the study by duplicate measures on six, randomly selected subjects.

There is, however, the potential problem of subject–observer interaction such that even in response to an interview, different subjects respond differently to minor differences between observers: this requires very close attention during training.

It is not always necessary to formally test consistency within observers.

Example 12.iii

In an interview-based study that involved a quality of life assessment, a pilot study suggested that certain questions were answered differently depending on the sex of the interviewer. The answer to this was to use interviewers of one gender only.

If between-observer reproducibility is good, it can be assumed that within-observer reproducibility is also satisfactory. Clearly, the converse is not the case.

Example 12.iv

In a study assessing perceived severity of eczema on a 0–4 scale, the investigator was concerned about inconsistency in assessment within an observer. It was not sensible for the

observer to undertake duplicate assessments on the same patients because it would be impossible not to be influenced by the results of the first assessment, given the difficulty in ensuring blindness of assessment. However, duplicate assessments by different observers were acceptably close. It was thus unlikely that there was a serious inconsistency within an observer.

12.1b Bias and agreement

These are different concepts, but both may explain inconsistencies or lack of reproducibility in results between observers. Bias occurs if one observer systematically scores in a different direction from another observer, whereas poor agreement can also occur when the lack of precision by a single observer is random.

The possibility of bias is relatively easily assessed from differences in the distribution of measures obtained. Thus, (i) for continuous variables it is relatively simple to compare the results from two observers by using a paired *t* test or similar approach, and (ii) for categorical variables the frequencies of those scored in the different categories can be compared with a chi-squared or similar test (see also Section 12.2a below).

The assessment of agreement is discussed below in detail.

12.2 Study designs to measure repeatability

In any study in which multiple observers are to be used, it is necessary to measure their agreement before the start of information gathering. The same principles apply when there is a concern about a lack of consistency within a single observer. The essence of any study to investigate this phenomenon is to obtain multiple observations on the same subjects by the different observers (or replicate measures by the same observer) done sufficiently closely in time to reduce the likelihood of true subject change. An ideal approach is to enlist some subject volunteers who are willing to be assessed in a single session by multiple observers. It is important to take account of any *order* effect, i.e. where there is a systematic difference in measurement response with increasing number of assessments. One strategy to allow for this is to use the so-called 'Latin square' design (Table 12.1). In this example, five subjects are assessed by five observers in a predetermined order. With

Table 12.1. 'Latin square' design for a study of repeatability: five subjects (1–5) and five observers (A–E) giving the order in which the observers assess the subjects

	Subject				
Observer	1	2	3	4	5
A	1st	2nd	3rd	4th	5th
B	5th	1st	2nd	3rd	4th
C	4th	5th	1st	2nd	3rd
D	3rd	4th	5th	1st	2nd
E	2nd	3rd	4th	5th	1st

this kind of design it is relatively simple statistically, by using an analysis-of-variance approach, to separate the variation between different observers from that due to order and, of course, that due to the subjects themselves. A similar approach may be used when testing for reproducibility within an observer. One problem, however, particularly in assessing interview schedules, is that both the subject and the observer may remember the response. In such circumstances the replicate interviews need to be spaced in time, but not to such an extent that the true state of the subject has changed. The particular details of individual studies will determine what is an appropriate interval.

12.2a Analysis of repeatability

The analytical approach is different for measures that are categorical and those that are continuous.

For categorical measures, the kappa (κ) statistic is the appropriate measure of agreement. It is a measure of level of agreement in excess of that which would be expected by chance. It may be calculated for multiple observers and across measures which are dichotomous or with multiple categories of answers. For the purposes of illustration, the simplest example is of two observers measuring a series of subjects who can be either positive or negative for a characteristic.

The kappa statistic is calculated as follows.

Measurements by observer B	Measurements by observer A		
	Positive	Negative	Total
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=N$

Proportion that A scored positive $= \dfrac{a+c}{N}$.

Proportion that B scored positive $= \dfrac{a+b}{N}$.

Therefore, by chance alone it would be expected that the proportion of subjects that would be scored positive by both observers $= \left[\dfrac{a+c}{N} \cdot \dfrac{a+b}{N}\right]$.

Proportion that A scored negative $= \dfrac{b+d}{N}$.

Proportion that B scored negative $= \dfrac{c+d}{N}$.

Therefore, by chance alone it would be expected that the proportion of subjects that would be scored negative by both observers $= \left[\dfrac{b+d}{N} \cdot \dfrac{c+d}{N}\right]$.

Therefore, total expected proportion of agreement $= \left[\dfrac{a+c}{N} \cdot \dfrac{a+b}{N}\right] + \left[\dfrac{b+d}{N} \cdot \dfrac{c+d}{N}\right] = P_e$.

Maximum proportion of agreement in excess of chance $= 1 - P_e$

Total observed proportion of agreement $= \dfrac{a+d}{N} = P_o$

Therefore, proportion of observed agreement in excess of chance $= P_o - P_e$

The observed agreement in excess of chance, expressed as a proportion of the maximum possible agreement in excess of chance (kappa) is:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Table 12.2. Interpretation of kappa

Value	Strength of agreement
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

Example 12.v

	Observer A		
Observer B	Positive	Negative	Total
Positive	57	13	70
Negative	16	34	50
Total	73	47	120

$$\text{Observed agreement} = \frac{57+34}{120} = 0.76;$$

$$\text{Expected agreement} = \left[\frac{73}{120}\cdot\frac{70}{120}\right] + \left[\frac{47}{120}\cdot\frac{50}{120}\right].$$

$$= 0.52$$

$$\kappa = \frac{0.76-0.52}{1-0.52} = 0.50.$$

The use of kappa is important, as the often-used proportion of total agreement does not allow for the fact that some agreement is due to chance. The interpretation of kappa values is subjective, but as a guide Table 12.2 may be useful. Mathematically kappa can range from -1 to $+1$. Values below zero suggest negative agreement, not normally of relevance unless circumstances are bizarre. Values close to zero suggest that the level of agreement is close to that expected by chance.

Bias can be assessed by examining the marginal totals. In Example 12.v, the proportions scored positive by the two observers were similar (70/120 vs. 73/120), excluding any serious systematic bias even though the agreement is only moderate.

Example 12.vi

Two observers are grading X-rays as disease positive or negative for evidence of lung disease. Even if they were scoring the X-rays randomly they would be expected to agree the status in half the X-rays (proportion agreement 0.5). The kappa statistic measures agreement in excess of that expected by chance, i.e. the agreement in excess of 0.5 expressed as a proportion of the maximum excess (0.5).

For continuous measures, the simplest initial approach is to determine for each individual subject the absolute level of disagreement between observers, as in Table 12.3. A number of measures of agreement can then be obtained. First, calculation of the mean disagreement and the standard deviation around that mean can give an estimate of the range of disagreements. Secondly, calculation of the standard error of the mean disagreement can be used to provide a 95% confidence range for the likely mean disagreement. Finally, the closer the mean disagreement is to zero, the less likely is the presence of systematic bias.

Table 12.3. Assessment of repeatability between two observers for continuous measure

	Weight (kg) measured by:		
Subject number	Observer A	Observer B	Difference, d
1	64.2	64.6	−0.4
2	71.3	71.0	+0.3
3	80.4	84.2	−3.8
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
40	66.2	65.4	+0.8
Mean	70.6	71.0	−0.4

Notes:

Mean difference $\bar{d} = -0.4$ kg.

Level of disagreement

If standard deviation = 0.25

95% range of disagreements (observer A–B) $= -0.4 \pm (1.96 \times 0.25)$

$= -0.9$ to $+0.1$.

Thus, the disagreement between these two observers for 95% of subjects will be between −0.9 and +0.1 kg

Level of mean disagreement

SE mean difference $= 0.25/\sqrt{40}$

$= 0.04$

95% confidence limits for mean disagreement (observer A–B) $= -0.4 \pm (1.96 \times 0.04)$

$= -0.48$ to -0.32 kg.

Thus, it is 95% likely that the mean difference between weights recorded by observers A and B will be 0.32 and 0.48 kg, with observer B recording the heavier weights.

13

Maximising participation

13.1 Introduction

At a practical level, the greatest challenge is to maximise participation. In all studies there is the issue of ensuring a high response rate, and in prospective studies there is the additional problem of maintaining participation over the period of observation. The consequence of a significant loss of recruitment is to introduce the potential for non-response bias, the assessment of which is discussed in Chapter 15. In this chapter, strategies are discussed to ensure the greatest participation in surveys, including the specific problem of follow-up studies.

13.2 Reasons for non-participation

There are a number of reasons why subjects refuse to participate in epidemiological research, whether this involves completion of a mailed questionnaire or attendance at a clinical facility for a moderately invasive procedure such as a blood test or X-ray. The major reasons are:

- lack of interest or perceived personal relevance,
- inconvenience,
- avoidance of discomfort,
- financial cost,
- anxiety about health,
- antipathy to research.

In practice it is often difficult and perhaps unethical to seek an explanation for non-participation. These various reasons for non-participation are, however, discussed below with suitable strategies for their reduction.

> *' we have therefore chosen to study individuals from the general population chosen at random from the electoral register. It is only by receiving answers from a large broad section of the population that we can understand why some people develop this disorder ... '*

Figure 13.1

13.2a Lack of interest or perceived personal relevance

Many subjects do not participate because they do not see any personal gain. Those with a disease are more likely to enrol to seek a greater understanding than those who perceive themselves to be healthy. In general, males, and those who are younger, are less likely to participate for the 'good of humanity' than females and those in older age groups. Differential response in a case-control study can be a major problem, and the use of controls with another disease, or friends or family of the cases, might generate a higher response than recruiting random population controls. This has to be balanced against the scientific advantage of the latter approach (see Section 8.2a). In reality, it is difficult to solve this problem.

In sending out letters of invitation, it is often valuable to stress that there is a need to study normal people in order to understand the abnormal: healthy subjects may not realise that they may contribute useful information. It is useful to state clearly that this is the goal of the study and why a population recruitment policy has been followed (Fig. 13.1).

It is often a more productive strategy to use registers constructed from healthcare sources to approach general population samples. If the letterhead also includes the address of the local practice and is signed (on behalf of the study) by the general practitioner, there is a greater perceived relevance and sense of involvement, as well as approval from a trusted source. Suitable phrasing is shown in Fig. 13.2.

13.2b Inconvenience

This may reflect an unwillingness to spend time completing a form, a reticence to receive a home visit or a reluctance to attend a survey centre. Clearly, depending on the nature of the survey, much can be done to reduce the level of inconvenience to the subjects targeted (Table 13.1). Diaries of symptoms

Table 13.1. Strategies to reduce inconvenience

Type of survey	Strategies
Postal survey	Clear instructions Keep questions simple Keep number of questions to a minimum Avoid embarrassing/sensitive questions if possible Provide stamped or reply paid envelope
Home visit	Offer choice of timings including outside work hours Keep visit time to a minimum and emphasise shortness of visit Be sensitive to needs of elderly living alone, mothers of young children
Survey-centre visit	Offer choice of timings including outside working hours Provide transport costs and parking facilities Reduce waiting time to a minimum Give prior information on likely duration of visit Provide telephone number for cancellations/rebookings

The Health Centre
High Street
Newtown

Dear Mr Smith

re: Health Survey

I am collaborating with Newtown University to look at how common some kinds of ill health are in my patient population. I am writing to all males, aged 46 – 64, to ask them if they would be willing to .....

Figure 13.2

or of variable exposures, such as diet, are useful but they are tedious to complete and response is disappointing. In letters sent prior to home or survey-centre visit, it is useful to cover the issues mentioned in the table. The maximum effort should be extended to make access easy and without cost, and to make the visit as short as possible, particularly when attendance at a survey centre is required. Poor organisation rapidly becomes known in local-based studies, hindering later recruitment.

13.2c Avoidance of discomfort

This is obvious and is not often a feature of epidemiological surveys. However, detailed questions on sensitive issues, or long surveys requiring considerable recall, may prove too onerous for some, particularly the elderly.

13.2d Financial cost

As mentioned in Table 13.1, the survey should be prepared to reimburse the costs of travel and parking. The timing of visits should be done to reduce the need for time lost from work or the need to obtain extra child care. Sensitivity shown in these issues is often amply rewarded. It is becoming more common for epidemiological surveys to reimburse subjects directly for their time.

Example 13.i

An occupational study conducted by the authors included a group of young male army recruits. It was of concern that the non-participation rate may be high in this group, and therefore they were offered £5 on return of a completed questionnaire.

13.2e Anxiety about health

The monitoring of those who do respond to health surveys frequently reveals a close relationship between the speed of response and perceived health needs. Rapid responders frequently over-represent those individuals with relevant symptoms who, depending on the nature of the enquiry, use their response as an entrée into the healthcare system, however tenuous the links. Conversely, there is inevitably a proportion who will not respond to any survey likely to reveal undiagnosed morbidity.

Example 13.ii

In a hypertension screening survey, a sample of the population who failed to attend for screening were carefully interviewed by telephone in an attempt to discover the reasons. A proportion of the cases admitted that they would 'rather not know' if their blood pressure was raised.

It may be possible to overcome such anxieties in advance, where relevant by highlighting (i) the rarity of the disorder, (ii) the benign nature, (iii) the considerable advantage of early detection. If the survey is not going to reveal any unsuspected morbidity then this should be made clear to reduce any suspicion of a 'hidden agenda'.

13.2f Antipathy to research

Finally, there is the rare problem that there will always be one or two strongly worded letters or telephone calls in response to a request for participation. To the novice these can be unsettling and time-consuming to deal with. The reality is that, provided that the project is ethically sound (see Chapter 20), these inevitable, but infrequent, events can be easily managed, without compromising the research, by a polite letter expressing regret at any distress.

13.3 Maximising participation in follow-up

13.3a Introduction

The crucial activity of a cohort-based study is the follow-up of the individuals recruited after their baseline investigation. There is a requirement to keep track of the study population. There are a number of reasons for following up the study cohorts, the major ones being to determine their vital status (dead or alive) as well as ascertaining their disease status.

Example 13.iii

In a study determining the future risk of a relapse following chemotherapy for a particular malignancy, the investigator sent out a postcard every six months to determine whether the individual still lived at his or her baseline address. She was concerned to identify those who had moved or died (from whatever cause) and who were not continuing in the follow-up.

It may also be important to determine whether there has been a change in the exposure status since baseline assessment. The necessity to obtain this information depends on whether the study is examining the risk of 'ever exposed' or of length or other measure of exposure dose. The investigator should also consider whether those previously unexposed have become exposed.

Example 13.iv

In a cohort study to determine the relationship between medical students' alcohol intake and their subsequent success (or otherwise!) in the final examinations, the investigator surveyed the students at the start of the course about their alcohol consumption and then resurveyed the cohort at the start of the second year. He proposed to examine separately those whose status had remained unchanged from those whose consumption had increased or decreased.

Table 13.2. Options for obtaining follow-up information on disease status

1. Use existing disease notification registers
   (i) National/Sub-National Disease Registers
   (ii) Hospital in-patient records
   (iii) General practice computerised morbidity records
2. Continuous monitoring by notification forms held by subject or general practitioner
3. Cross-sectional surveys
   (i) General practice record review
   (ii) Hospital record review
   (iii) Subject questionnaire/interview/examination surveys

The follow-up can also be used to determine change in status for other potentially important variables that might affect the analysis (see Chapter 18). Thus, in the example above, if the investigator was concerned that poor diet might contribute to low examination success and had collected dietary data at baseline, it would be important to collect similar data at the follow-up. If, as is likely with medical students, increased expenditure on alcohol was accompanied by a decreased expenditure on a balanced diet, then changes in diet would need to be examined in analysing the main effect of alcohol.

13.3b Ascertaining disease status

There are a number of strategies for ascertaining disease status during follow-up (Table 13.2).

Using existing registers

If the disease or event is one for which an existing register is likely to ascertain all cases, data collection is simplified. This might be achieved using national cause-of-death registers, national morbidity registration systems (particularly for cancer), or hospital in-patient diagnostic or primary care registers. The success of these approaches relies on the likely completeness and accuracy of ascertainment. It may be necessary to undertake pilot validation studies to confirm these assumptions.

Example 13.v

In an occupational cancer study, the investigators used the national cancer registry to ascertain the cases that developed. In the UK, they were able to take advantage of the system where they could 'tag' all their study population at the National Health Service Central Register (NHSCR) at the start of the study. They then received notification of cancers as and when they were registered.

It is possible to use a similar approach when population-based diagnostic registers for hospital in-patients exist.

Example 13.vi

In a Swedish study, an epidemiologist was able to obtain questionnaire data on marijuana use by army recruits. He was then able to link the national identity numbers to hospital in-patient databases to obtain subsequent information on admissions with schizophrenia.

Such linked registers are not as readily available in most countries as in Scandinavia and, in the current climate of data protection, are becoming more difficult to access. Increasingly, though, general practitioners, who provide the primary care needs for a defined population, are developing computerised diagnostic registers resulting from all patient encounters. Diagnostic accuracy (and between observer repeatability) is clearly more of a problem in a primary-care setting, but if the 'disease' to be identified is a simple one to define, this source may be invaluable.

Example 13.vii

In a general-practice-based study of low back pain, a baseline population survey of psychological and other risk factors was followed by the continuous notification by the general practitioners of all new consultations with low back pain. This notification was achieved 'electronically' because the reason for all consultations in these general practices was routinely logged on a computerised register.

Developing specific information systems

In most circumstances, such registers are insufficient or are unavailable, and a specific system for disease ascertainment needs to be introduced. In theory, the most appropriate tactic is to supply the subjects or their general practitioners with a special form to post back if a specified outcome occurs.

Example 13.viii

In a study, an attempt was made to identify subsequent strokes in general-practice patients who had had a baseline blood pressure measurement. A boldly coloured special card was designed to fit in the general-practice record folder, with a message to alert the general practitioner to notify the study if that particular patient suffered a stroke. This tactic, though fine in theory, failed in practice for a number of reasons: lost cards, failure to recall procedure, loss of interest, etc.

Others have tried giving study subjects postcards to post back if they develop the disease under question. Apart from very short-term investigations, for example an outbreak of food poisoning, it is likely that such cards will be lost or forgotten.

Thus, in many prospective surveys, it is necessary for the investigator to contact the study population directly during the study to obtain the necessary outcome data. This will require a cross-sectional survey using a postal questionnaire, interview or examination as appropriate. The timing of the survey must be sufficiently close to the baseline in order not to miss events. Multiple surveys may be necessary.

Example 13.ix

In a large prospective study of fracture risk, the participants were sent a simple annual questionnaire enquiring about fractures in the previous 12 months. Other studies had suggested that recall of fractures over a 12-month period would be adequate.

As part of such regular survey questionnaires, the investigator can also use the opportunity to check the current whereabouts of the subject, which becomes more difficult the greater the interval, and to obtain data on changes in other variables since baseline. Further, if the questionnaire needs to be followed by (say) confirmatory documentation from medical records, it makes sense to obtain all necessary information and permissions at the same time (Fig. 13.3).

13.3c Minimising loss to follow-up

The final issue in the design is ensuring that losses to follow-up are minimised, both to reduce the likelihood of bias and to maintain sufficient numbers so that the pre-study sample-size calculations remain valid.

<Mr J Smith> <04261>
<10 Hill Street>
<Hightown>
<HI17 3XZ> <Today's date>

1. Has your name or address changed since <date of last follow-up>?
YES/NO

If YES, what is your correct name/address?

.....
.....
.....

2. Have you acquired any new pets since <date of last follow-up>?
YES/NO

If YES, what pets have you acquired?

Dog ☐
Cat ☐
Bird ☐
Other ☐

3. Have you attended a hospital or clinic with any of the following problems since <date of last follow-up>?

Cough
Phlegm
Shortness of breath YES/NO

If YES, can you give us the name of the hospital/clinic and (if possible) the name of the doctor you saw?

Hospital/Clinic.....
.....

Name of doctor .....

4. Have you attended your general practitioner with any of the following problems since <date of last follow-up>?

Cough
Phlegm
Shortness of breath YES/NO

If YES, can you give us the name of your general practitioner and (if possible) the name of the doctor you saw?

General Practitioner .....
.....

Name of doctor .....
Can we approach the doctors/clinics listed above for further details about your attendance? YES/NO

....
Signature Date

Figure 13.3 Example of follow-up letter during prospective study.

The following strategies can minimise losses.

(i) Use disease registers that do not require direct contact with subjects.

(ii) Keep follow-up period to a minimum.

(iii) Encourage continuing participation by regular feedback and other contacts, for example newsletters, birthday or Christmas cards. Participants in studies like to believe that their contribution is valued, and a lay summary of progress is helpful.

(iv) Minimise the obligations and inconvenience to participants. Answering a questionnaire by post regularly is likely to be more acceptable than repeated requests to attend for examination or blood taking.

(v) Collect information at baseline that offers alternative approaches to contact if a subject moves or dies. Suggestions include names and addresses of non-household relatives, neighbours or employers. In the UK (as in other countries), there are systems for locating individuals who have moved. In the UK this is based on the NHSCR, which keeps information on the local health authority of the general practitioner (GP) for each GP-registered patient. Thus, individuals who have changed GP can theoretically be traced.

(vi) Provide postcards at baseline to participants in the study, asking for them to be posted back in case of relocation (although, as mentioned earlier, this may not be a foolproof method).

Clearly, if funds are no problem, household enquiries can be conducted locally. Indeed, in some countries commercial organisations exist for this purpose.

14

Conducting a pilot study

14.1 Aims

Having decided on appropriate study methods to test your hypothesis, identified the study population, designed the study questionnaire and finalised the protocol for study conduct, it is essential to undertake a pilot study prior to embarking upon the principal study. The main study may well involve considerable resources in terms of human time and finances and it would be regrettable (to say the least) to reach the end of your endeavours only to find, for example, that there was an important omission in the questionnaire or that your methods resulted in a poor response. Despite the fact that pilot studies are commonly conducted it is the authors' experience, firstly that students are often not clear about the aims of such an exercise and, secondly it is rare that the information which a pilot study provides is fully utilised. It is not possible to be comprehensive regarding the possible aims of a pilot study, because these will vary between studies, but most commonly it will relate to one or more of the following:

- Participation rates (or subject identification rate)
- Data/record availability/quality
- Study instruments
  - comprehensibility
  - completion
  - validity/repeatability of instruments
- Sample size information

It is also important to emphasise that there may not be one single pilot study but a series of pilot studies as study methods are being developed and tested.

14.1a Participation rates

As discussed previously it is desirable to achieve high participation rates in a study in order to reduce the possibility of non-participation bias. A pilot study will give an estimate of the likely participation rates with the chosen protocol. If participation rates are low then additional or alternative methods will need to be considered. For example if after two mailings of a questionnaire the participation rates is only 35%, then one will need to consider, e.g. an additional mailing, mailing a shortened questionnaire and/or contacting subjects by telephone or door-knocking. A small pilot study may also afford the opportunity to conduct a short interview with non-participants to evaluate to what extent they differ either with respect to the disease of interest or risk factors for its development, and to understand why they did not participate. Reasons may include that they did not think the study topic was relevant to them, they felt too unwell, or that the questionnaire was too long.

Case-control studies will be concerned with evaluating the rate of case recruitment to the study. Once study commences, the number of cases recruited, e.g. through out-patient clinics, are rarely as high as originally envisaged, particularly once refusals to participate and exclusion criteria are considered. A pilot study will give a realistic estimate of recruitment rate.

14.1b Data/record availability

Some studies may require the availability of historical record data, for example occupational or medical records. Such records may be unavailable to the researcher or their quality may be disappointingly low.

Example 14.i

A study relied on occupational personnel records detailing exposures to a number of chemicals by the workforce. A pilot study of 20 individuals revealed many gaps and thus the records could not be used as a reliable source of chemical exposure

14.1c Study instruments

Instruments used to obtain data include questionnaires, interview schedules and examination procedures. There are often a variety of reasons for pilot-testing such instruments. The first is acceptability and comprehensibility –

did the study subjects find the methods acceptable, for example, did they understand the questions being asked? To assess this, one could give pilot study participants an opportunity to list at the end of the survey any problems which they had in its completion, either understanding the question or in providing a response. There may be questions, however, where the participant misunderstood what was being asked, and this may only be detected by interviewing some of the participants. A pilot study will also allow completeness of returns to be assessed, e.g. did respondents omit the last page of the questionnaire because they failed to notice it, were there problems with internal routings, was an obvious response category missed or were responses not mutually exclusive? The use of open questions in a pilot study may also allow these to be changed into closed questions in the main study, on the basis of the most common responses given.

The study questionnaire may include some questions/instruments which have not been validated previously within the study setting. This may therefore involve requesting a sub-sample of participants to complete the questionnaire on more than one occasion (repeatability) and evaluating responses against an external standard if possible (validity). Full details of assessing validity have been given in Chapter 11.

Having conducted a pilot study, there are often time constraints requiring the study investigator to move quickly on to the main study. Although this is often reality, if insufficient time is taken to consider the results of a pilot study, then the effort and advantage involved in its conduct may be wasted. It is important, even though there may be small numbers of subjects involved, to undertake an exploratory data analysis. Only by doing this will problems such as high levels of missing data, missing questions from an instrument or unusual response patterns be detected.

14.1d Other methods of collecting information

If the study involves an interviewer-administered questionnaire or if an examination is conducted on all (or a subsample) of respondents, then a goal of a pilot study will be to examine intra- and inter-observer repeatability. Details of this assessment are given in Chapter 12. Lack of repeatability will necessitate further training. Indeed evaluation of repeatability should not be considered necessary only at the pilot stages of a study – it should form part of the protocol for the main study also. It is, for example, possible that exam-

iners may develop habits during the course of the study which result in systematically different measurements or observation from each other.

14.1e Sample size requirements

Chapters 8 and 9 have discussed the factors which determine the sample size required in conducting case-control and cohort studies respectively. It may be that not all of the required pieces of information are known, e.g. the prevalence of disease or the exposure of interest. In this case, estimates of these can be obtained from the pilot study before finalising the size of the main study.

In summary, the pilot study (or pilot studies) is an essential component of preparing to conduct an epidemiological study. Ample time and resources invested at this stage will be rewarded with a greatly enhanced chance of a successfully conducted main study.

Part VI

Analysis and interpretation of epidemiological data

15

Preparation of survey data for statistical analysis

15.1 Introduction

Preceding sections have discussed, in some detail, the necessary steps in the design of epidemiological studies to collect the data appropriate to answer the question(s) posed. The next stage is to turn the data into information. Thus, the real task is to assimilate the large number of data items, be they numerical or categorical, to generate the epidemiological measures targeted, such as estimates of disease occurrence or disease risk. Although statistical analysis is a necessary, and conceptually perhaps the most difficult, part of the task, most of the work in the analysis of any epidemiological survey is in the preparation of the data. Indeed, the availability of modern statistical software means that the stage of analysis often requires much less time than that required for the preparation of the data. Although reviewers and readers of epidemiological reports may comment on the dexterity of a statistical analysis, the painstaking and laborious data checking is often taken for granted. This lack of emphasis is also true for those planning studies. In many grant applications, resources are requested for data collection and for statistical and computing expertise, but no provision has been made for the more resource-consuming tasks of data preparation.

15.1a Use of computers

It is appropriate to discuss here the use of computers for data preparation. As discussed in Chapter 10, the design of questionnaire and other data forms needs to take account of the fact of computer analysis, and if the computer software likely to be used is considered during the planning of data collection, an enormous amount of work can be saved subsequently. Computers have three main tasks in epidemiological analysis, these being those of (i) a

Table 15.1. Stages in preparing data for analysis

1.	Checking completeness and accuracy of data
2.	Linkage by subject of data from multiple sources
3.	Development of data coding schedule
4.	Development of computer database
5.	Procedure for data entry
6.	Checking for errors and inaccuracies on entered data
7.	Consideration of problem of missing data
8.	Re-coding entered data for planned analyses
9.	Storage of dataset and database

database, (ii) statistical analysis and (iii) graphics for display. As far as data preparation is concerned, a suitable database package is a prerequisite not only for storing the data but also for checking for errors and modifying the data for particular analytical problems. Suitable and widely available database packages particularly for survey analysis, include SPSS PC and Access. The former has the advantage that the same package can be used for both data preparation and statistical analysis, but it is not a problem to prepare the data with one package and analyse it with one or more other different analytical packages. Most database packages are very easy to use and most readers are likely to be familiar with at least one.

15.1b Stages in preparing data for analysis

In Table 15.1 are listed nine separate stages involved in preparing data for analysis, and the remainder of this chapter will discuss these in detail. Some items will not be relevant or may have been considered at the design stage. Thus, if only a single source of data is being analysed, step 2 is irrelevant. Similarly, a survey instrument such as a questionnaire may have been designed to be 'self-coding', with the coding schedule decided at the design stage (see Fig. 15.1), and need not be considered again. The time that each of these individual stages takes clearly depends on the size of the data set and the number of errors found. Irritating delays often emerge when missing or inaccurate data require referral back to the subject or other primary data source. However, it is preferable for this to occur at the preparation stage, rather than during analysis. The actual task of keying data into the computer

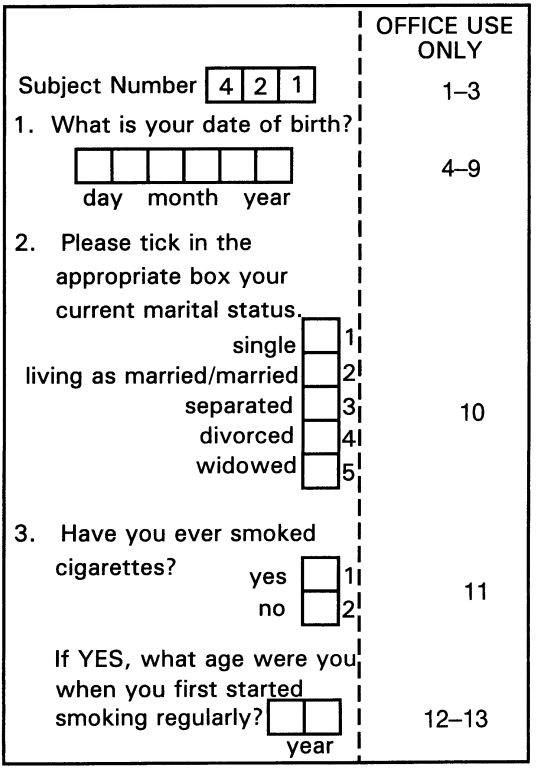

Figure 15.1 Example of pre-coded questionnaire.

can be very rapid, and even in very large studies it may not take up the greatest amount of data preparation time.

15.2 Initial checking for completeness and accuracy

The first stage is a manual review of the data gathered. This may involve a checking of self-completed questionnaires or a review of data forms used to obtain interview or other data. Self-completed questionnaires are prone to a number of errors. Questions may have been missed, multiple answers given when only one was required, inconsistencies emerging such as a subject responding negatively to 'ever pregnant?' but then giving the age of first child. Another problem occurs when the subject ignores the choices given and writes a small (or even a long) note giving a detailed life history relevant

to that particular question. Piloting and field testing may reduce the likelihood of such problems but cannot eliminate them altogether. Decisions have to be made on how to deal with such errors. The counsel of perfection is to contact the subject again for clarification, although this is often not feasible or indeed ethical. Otherwise rules have to be adopted based on common sense.

Example 15.i

In a large population-based survey covering different lifestyle factors, a number of subjects had answered negatively to questions relating to alcohol, tobacco and the consumption of specific pharmaceutical preparations including the oral contraceptive pill, but had then gone on to give a starting date and duration of exposure. The decision was made to convert the initial negative answers to those lifestyle exposures to positives on the assumption that the subject had made an error.

Frequently it is difficult, if not impossible, to work out the 'correct' answer, and the data on that particular question for that subject has to be considered as missing.

Example 15.ii

In a survey on pregnancy-related factors, one subject had ticked 'yes' to the question asking about ever having a successful pregnancy, but crossed out the subsequent question dealing with dates of pregnancies. There may have been reasons for that woman not wishing to provide the data, or the initial 'yes' might have been wrong. Without further information it is impossible to determine which of these alternatives was correct.

For interviewer-obtained data, the interviewers can be trained to check the interview form carefully, before leaving the subject, in order to pick up omissions, obvious errors and common transcription problems such as inserting the current year instead of year of birth. Often within such a short time, the interviewer can recall the correct answer and at the very worst it is easier to recontact the subject if the time interval is short. When information is being obtained from a self-complete questionnaire, particularly from a physical examination by a trained individual such as a nurse, the researcher should check the data obtained, limiting the interval between completion and review to a minimum, so that it is easier to recontact the subject and request the missing/ambiguous information.

Example 15.iii

In a survey involving home visits, trained nurses were asked to examine the subject for psoriasis (amongst other items of physical examination). In cases of doubt the nurses were asked to describe carefully what they saw, and their findings were discussed soon after the interviews with an experienced clinician, who could then make a judgement as to the likelihood of psoriasis being present.

15.3 Linkage by subject of data from multiple sources

Frequently, survey data are obtained from a number of sources for an individual subject, including, for example, a self-completed questionnaire, data obtained from a physician's records and perhaps the results from a laboratory test. Although it might seem unlikely, it is disturbing how often it proves difficult to link these items to a single subject. Confidentiality frequently precludes using names on study record forms, and subject identification numbers often get transposed or entered incorrectly. Two strategies may be used to minimise such errors. The first is to include both a date of birth and a unique study number on all sources of data. In most studies, the relative uniqueness of a date of birth will ensure that items that should be linked are, and those that should not are separated. Such an approach would fail if errors were made in entering birth dates. An alternative approach is to obtain preprinted self-adhesive labels in multiple copies, printed with either the subject's study number alone or, if the information is available from a register before the survey, in combination with date of birth. The labels are then attached to all documents to be used for obtaining data. Mailmerge word processing programs may also be used for this task.

15.4 Development of a data coding schedule

It is not always possible for the data collected to be in the appropriate format required for data analysis. It is therefore necessary to modify the data before entering it on to the computer database. This modification is referred to as *data coding*, the 'coding' implying that a simple 'language' is developed to facilitate data entry and analysis. Data that are already in numerical form such as year of birth or number of children, can be entered directly and do not require coding. Other simple categorical variables such as sex may be

modified for ease of data entry. Typical codes might be 'm' for male and 'f' for female. The layout of the typical computer keyboard with a separate number keypad accessed via the 'Num Lock' key means that data entry is often physically easier if only numbers are used. Thus, one can easily substitute '1' for male and '2' for female. Multicategory answers can be coded similarly, e.g. '1' for never smoker, '2' for ex-smoker and '3' for current smoker. Decisions can also be made about missing or inappropriate answers. If an answer is missing, one option is to leave that field blank for that subject; alternatively a standard number such as '9' can be used. A separate code may be used for questions that are inapplicable. It also might be scientifically important to distinguish between individuals who, after consideration, answer 'don't know' from answers that are missing.

Example 15.iv

In a survey of women's health, a series of questions were asked about the age at last menstrual period. A coding scheme was devised such that if an age was given this was entered as a two-digit number. For other possible answers the following codes were applied: questions unanswered, code '00'; if a woman was still menstruating, code '88'; if a woman had recorded a surgical menopause (i.e. hysterectomy), code '77'; if a woman had indicated that she was menopausal but could not recall when she had her last period, code '99'. The advantage of this scheme is that it allows separation of these very different groups of women.

The task can be made simpler if the data-recording instrument, such as a questionnaire, is precoded at the design stage, as in Fig. 15.1. Using this approach, the answers can be entered directly. In this example, a subject ticking 'single' for question 2 would be automatically coded and could be directly entered onto the database as a '1'. The numbers under 'Office Use Only' refer to the place on the database where the answers are to be entered. The use of such instruments does not reduce the need for careful checking before data entry.

This approach may not be appropriate if the coding scheme is complex. It is also sometimes necessary to design the coding scheme once the answers have been reviewed. Figure 15.2 shows an example of a coding scheme that was used to apply codes after data collection. In particular, the subject was asked about current medication in question 7, but the researcher was interested only in certain categories of drugs as indicated in the coding schedule.

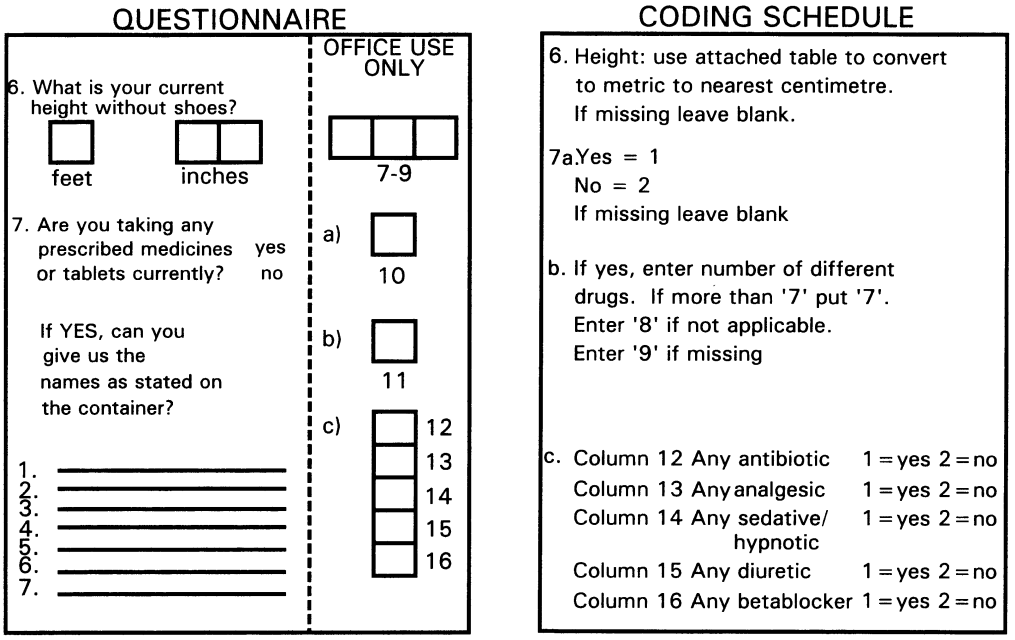

Figure 15.2 Example of questionnaire and accompanying coding schedule.

It was thought easier to ask an 'open' question about drug use rather than rely on the subject accurately ascribing their own medications to a particular class of drug. Note also in this example that the researcher converted heights from imperial to metric before data entry. This is not strictly necessary or indeed desirable, as a computer routine can be used after data entry to achieve such a conversion, minimising the possibility of human error. In this example the coding schedule to question 7.b permitted a separation between those not taking any prescribed medications (code '8') and those who were, but could not remember the number of different preparations (code '9').

15.5 Development of a computer database

This is probably the easiest task and simply creates in a computer file the coded version of the data forms used.

Table 15.2. Examples of common range and consistency checks

Range	
Age:	within expected range
Adult height:	between 1.45 m and 1.95 m
Adult weight:	between 35 kg and 110 kg
Age at menopause:	between 40 and 55
No. of cigarettes/day:	below 60
Consistency checks	
Males not pregnant	
Non-smokers giving number smoked per day	
Age at menopause given before age at birth of youngest child	
Unemployed giving details of current occupation	

15.5a Subject identification number

It is customary always to commence with the subject's study identification number. Study numbers can be 'customised' to facilitate the conduct of a study.

Example 15.v

In a survey involving a number of different interviewers, each interviewer was given a separate series of subject numbers, e.g. interviewer no. 1: 0–999, interviewer no. 2: 1000–1999, etc. Each interviewer may only have planned to survey 150 subjects but this numbering system prevented duplication errors as well as permitting easier *post hoc* identification of particular interviewers for any subgroup analysis.

15.5b In-built range and consistency checks

One advantage of using a computer database is that it can check data as they are being entered and prevent the entry of 'illegal' data. Thus, packages such as SPSS Data Entry allow the investigator to state what ranges are permitted. In Table 15.2 examples are shown of typical range checks used. Thus the investigator decides *a priori* that any weight of under 35 kg or over 110 kg is likely to be erroneous and wishes the computer to reject weights outside this range being entered. Although such errors may be detected in the initial manual data checking, with large surveys it provides an additional and more

watertight procedure. In a similar manner an investigator will set up the database to prevent the entry of any answers that are clearly inconsistent with other data, for example pregnancy details being recorded for males, or an age of an event given which is older than the subject's current age. Other examples of prior consistency checks are given in the table. Such checks are substantially more difficult to carry out manually than range checks because they involve examining simultaneously a number of different variables, and can realistically be achieved only by a computer.

15.6 Procedure for data entry

The next step is to enter the data, either: (i) indirectly from the coded data forms; or (ii) directly from the uncoded data, with the data entry clerk coding simultaneously. If the coding schedule is simple this is an easy matter, otherwise it is better to separate the tasks of coding and entry. Data entry is a skilled task and, given its repetitive and boring nature, mistakes are inevitable for the expert as well as the novice. Data sets can also be entered by professional agencies, who by their speed and accuracy can work out to be a cheaper option as well. 'Double data entry' is also to be encouraged, although as the name suggests it doubles the effort. With this approach, after entering the data, they are then re-entered using a specially written simple program that highlights inconsistencies between the two entries. This procedure substantially reduces the likelihood of data entry errors. For example, if the probability that entry clerk A makes an error for question 1 of study subject is 1/100, and entry clerk B independently has the same error rate, then the probability with double entry that an error is undetected is 1/10 000 ($1/100 \times 1/100$).

If detection of errors is postponed until after the statistical analysis has started, this can lead to substantial difficulties, delays and errors in interpretation.

15.6a Immediate data entry

A further, and likely to be an increasingly popular, option for interview surveys is to enter replies directly onto a computer carried by the interviewer, omitting the 'paper' stage. This has the advantages of minimising transcription errors and increasing speed. Conversely, though, the lack of a hard-copy

record may be discomforting to some. It is also easier on hard copy to scribble some marginal notes for later consultation, whereas entry at the time of interview means a rapid decision has to be made as to the single best answer.

15.6b Optical scanning

The procedure of keying in data is being eliminated altogether in many studies, through the use of optical scanning machines. These 'read' the questionnaires and enter the data directly on to a database. However, operator intervention is necessary for queries and errors, which this process can generate (see above).

As the flexibility of the approach increases and the cost decreases, it is likely to become commonplace in many studies.

15.7 Checking for errors in entered data

A number of strategies can be employed to check for errors in data entry. The stringency with which this task is done perhaps best characterises the rigorous investigator who would rather delay the interesting part of the analysis until he or she is entirely satisfied that the database is free of errors.

(i) In a similar manner, as mentioned above, range and consistency checks can be carried out after data entry to check for obvious errors. A greater stringency can be used after data entry because the emphasis is not on rejecting rogue values but checking for extreme, and hence possibly suspect, values that could be checked against the original data.

Example 15.vi

From Table 15.2, the investigator permitted data entry of all weights between 35 and 110 kg. As an extra check, after data entry, all subjects with weights outside the range 50–95 kg had their weights checked against the primary data source.

(ii) A small number of subjects (20 would be a reasonable number) randomly selected should have their entered data checked against the original data forms. This is particularly vital if a double data entry system has not been used and is a useful back-up. If errors are detected there may be no alternative but to check the entire database, depending on the number of errors detected and the size of the task.

15.8 Missing data

One of the major problems in the analysis and interpretation of epidemiological surveys is that of missing data. This might arise for a number of reasons. In self-completed questionnaires, and even interview-administered surveys, data may be missing because of:

(i) poor recall by the subject,
(ii) a question not understood,
(iii) the lack of a suitable category for an unexpected answer,
(iv) genuine error in missing out a question,
(v) concerns about confidentiality.

Thus, in addition to the problems posed by total non-response by a subject, important data may be missing from some of those who do participate, hindering the interpretation of the study as a whole. If important items are missing, such as variables that are essential in disease or exposure classification, that subject may need to be excluded completely. A more frequent problem is the absence of more minor items. This results in different totals being used in the analysis as well as opening up the possibility of bias: for example, in a survey of women's health, those for whom no age of menopause is available may be more likely to have an irregular menstrual history than responders to that question.

There is little, however, that can be done at the stage of data preparation. It is important, as mentioned earlier, to distinguish in the design of the study instrument between, for example, those who are unable to recall and those for whom no suitable category is available. If missing data are identified soon after the time of the survey, the subject can be contacted or a proxy source of information used to fill in the gaps. It may be possible to take an average value, particularly in relation to items with fixed ranges and for which the consequences from error are small.

Example 15.vii

In a long-term outcome study following a femoral neck fracture, 30% of patients could not recall the month of their fracture. They were each assumed to have fractured in June of the recalled year of fracture (i.e. mid-year), thereby allowing person-months of follow-up to be calculated for each subject, with the error for an individual subject being limited to six months.

In preparing the data for analysis, therefore, the possible reasons for data absence should be entered if likely to be of relevance. It is then possible to undertake some preliminary analyses comparing the frequencies of other important variables between those with and those without missing data.

Example 15.viii

In a case-control study of cervical cancer, 20% of those who were interviewed declined to answer the question about the number of sexual partners. These non-responders were not, however, different from women who answered this question in relation to other variables of sexual activity gathered, such as age at first intercourse. In presenting their data on the number of partners as a risk factor, the investigators commented that they had an incomplete data set to address this question but from the data available there was no reason to believe that the results from those answering the question could not be extrapolated to the study population as a whole.

15.9 Recoding of entered data

Figure 15.2 showed an example of data collected on height in imperial units, which were to be converted into metric units. This and similar conversions are most efficiently achieved by recoding the data before the main analysis.

Other common examples include:

(i) Converting ages to calendar years and vice versa: thus, for a study conducted in 2001 one can create a variable 'AGE' (age now) which is calculated from 'BIRTHYR' (year of birth).

(ii) Converting continuous data to stratified data for analysis by exposure levels: thus, for a study on smoking, the number of cigarettes smoked on average per day was recoded into <5, 5–14, 15–29 and ≥ 30 because the investigator felt that biologically it would be more useful to treat smoking in this way.

(iii) Combining data from different variables: thus, an investigator had collected recalled data on height at entry to military service ('HTARMY') and had measured current height ('HTCURR') in a population of elderly men, and wished to examine height loss in relation to blood pressure. A new variable was therefore created 'HTLOSS' from 'HTARMY' – 'HTCURR'.

Table 15.3. Items to be stored

1. Study protocol
2. Primary data (questionnaires, lab. reports, etc.)
3. Coding schedule
4. Coding sheet, if used
5. Disc copy of entered data, with clear file names

Clearly these are only a few examples of what is a very useful and easily applied technique. The consequence is that for the purposes of data entry it is best to include raw data where possible and to restrict calculations and modification to post-entry analysis.

15.10 Storage of data and data set

The final aspect of data preparation to be considered is a simple, but frequently overlooked, issue: that is, the establishment of an archive so that either the same investigator or others in the future can refer to the data set to undertake repeat or different analyses. This is good scientific practice, and a minimum list of items to be archived is shown in Table 15.3. To this could be added data files containing modified data, together with the programming instructions used to make the modifications. In practical terms, if the principal person involved in the study were 'run over by a bus', would the data be in a sufficiently acceptable state for another investigator to work on (and presumably provide recognition, even if posthumous, to the original researcher!)? Such an exercise is also sound scientific practice and provides a useful measure against scientific fraud.

16

Introductory data analysis: descriptive epidemiology

16.1 Introduction

The purpose of this chapter is to outline the approaches used in calculating the tools of descriptive epidemiology, i.e. describing the risks and rates of disease. It is not intended to replace statistical textbooks on this subject but will provide an introduction to the information required for measures of disease occurrence (or death) and their calculation. When computer packages are readily available to carry out such calculations, one may ask why it is necessary to be aware of such detail. In fact, their availability as something of a 'black-box' makes understanding of the basic methods used even more important: readers are encouraged to work through the examples shown using a hand calculator. The ability to conduct such simple calculations will permit the reader, for example, to check their work or further explore data published by others. The authors have seen many examples of results presented which, on making some simple calculations, are clearly shown to be wrong. It is assumed that the reader has a basic knowledge of statistics and is aware of the use of summary measures including means and proportions, simple measures of variability including variance and standard deviation, and understands the conceptual basis for (i) making inferences about populations from studying samples and (ii) is familiar with the assessment of standard errors and confidence intervals.

16.2 Incidence rates

The calculation of an incidence rate (incidence density) is simple, and requires (i) the number of incident events and (ii) calculation of the total person–years at risk.

Example 16.i Calculation of incidence rate (density)

IR = incidence rate
x = number of incident events;
npyr = number of person–years at risk;
IR = x/npyr.

621 individuals followed up for a total of 1480 person–years, of whom 40 developed the disease.

IR = 40/1480
= 0.02703/pyr,
= 27.03/1000 person–years at risk.

In this example, the rate was expressed as rate per 1000 person–years to give a number that is convenient to handle. The exact denominator chosen will depend on the rarity of the event.

16.2a Confidence interval around an incidence rate

The Poisson distribution can be used to calculate limits around an incidence rate. This is based on the assumption that cases occur randomly in time in relation to each other. If the incidence estimate is based on more than approximately 75 cases, the normal approximation to the Poisson distribution provides a simple formula for obtaining the confidence interval.

Computationally, in practice, the 95% confidence interval using the Poisson distribution is obtained by looking up the answers in standard tables. Table 16.1 gives an abridged example from *Geigy's Scientific Tables.*

Example 16.ii Calculation of confidence interval around incidence rate

(Data are taken from Example 16.i)

(a) Using Poisson distribution:

From Table 16.1 for $x = 40$,

lower 95% limit = 28.58;
upper 95% limit = 54.74.

Therefore 95% confidence limits are:

28.58/1480 and 54.74/1480
= 0.0193/pyr and 0.0370/pyr,
= 19.3–37.0/1000 person–years at risk.

(b) Using normal approximation to Poisson:

x_u = upper 95% confidence limit to number of events;

x_l = lower 95% confidence limit to number of events.

$x_u = x + 1.96\sqrt{x}$;

$x_l = x - 1.96\sqrt{x}$.

For $x = 40$,

$$x_u = 40 + 1.96\sqrt{40}$$
$$= 40 + 1.96 \times 6.32$$
$$= 40 + 12.40$$
$$= 52.40.$$

$$x_l = 40 - 1.96\sqrt{40}$$
$$= 40 - 1.96 \times 6.32,$$
$$= 40 - 12.40,$$
$$= 27.61$$

Therefore 95% confidence limits are:

27.61/1480 and 52.40/1480

= 0.0187/pyr and 0.0354/pyr

= 18.7–35.4/1000 person–years at risk.

The points to note are, first, that even for 40 cases the two methods give very similar answers, and, secondly, both methods calculate the confidence interval for the number of events, which then have to be divided by the person–years in the denominator to derive the actual rates.

16.3 Prevalence (proportions)

Prevalence is estimated as a proportion of the population at risk and is very simple to calculate. All calculations in this section also apply to cumulative incidence or prevalence, both of which are also expressed as a proportion.

Example 16.iii Calculation of prevalence proportion

n = number in denominator;

x = number with disease;

p = proportion with disease;

$p = x/n$.

621 individuals examined, of whom 40 had disease

$p = 40/621$,

= 0.0644 (or 64.4 per 1000 persons)

Table 16.1. Exact confidence intervals for number of incident events based on Poisson distribution

Number of observed events	Lower 95% limit	Upper 95% limit
0	0	3.69
1	0.03	5.57
2	0.24	7.22
5	1.62	11.67
10	4.80	18.39
15	8.40	24.74
20	12.22	30.89
25	16.18	36.91
30	20.24	42.83
40	28.58	54.47
50	37.11	65.92
75	58.99	94.01
100	81.36	121.63

16.3a Confidence interval around a prevalence (proportion)

As with incidence rates, there are two approaches to calculating confidence intervals around a proportion (whether applying to prevalence or cumulative incidence or prevalence). The first, for use with small samples, where the denominator is (say) less than 100, relies on using published tables that provide the confidence interval for every possible observed proportion of events for each denominator up to 100. In the example below (16.iv) is shown a small part of the table for the denominator of 50.

For a larger sample size, again the normal approximation is used, which computationally is still simple (Example 16.v).

Example 16.iv Calculation of confidence interval around prevalence proportion or cumulative incidence: small sample size ($n < 100$)

Use statistical tables for exact confidence limits for a binomial proportion.

50 individuals, of whom 23 had disease:
$p = 0.46$.

Extract of statistical table

		95% confidence limits	
n	*x*	lower	upper
50	22	0.300	0.588
	23	0.318	0.607
	24	0.337	0.262
	⋮	⋮	⋮

From table, for $n = 50$, $x = 23$,
the confidence interval for the proportion of individuals with disease is 0.318–0.607.

Example 16.v Calculation of confidence interval around prevalence proportion or cumulative incidence: large sample size ($n > 75$)

Use normal approximation to binomial.

$$p_l \text{ (lower confidence interval limit)} = p - 1.96\sqrt{\frac{p(1-p)}{n}};$$

$$p_u \text{ (upper confidence interval limit)} = p + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Example: Using data from Example 16.iv,

$$p_l = 0.0644 - 1.96\sqrt{\frac{0.0644 \times 0.9356}{621}},$$

$$= 0.0644 - 1.96\sqrt{0.000097},$$

$$= 0.0644 - 1.96 \times 0.00985,$$

$$= 0.0644 - 0.0193,$$

$$= 0.0451.$$

$$P_u = 0.0644 + 1.96\sqrt{\frac{0.0644 \times 0.9356}{621}}$$

$$= 0.0644 + 1.96\sqrt{0.000097},$$

$$= 0.0644 + 1.96 \times 0.00985,$$

$$= 0.0644 + 0.0193,$$

$$= 0.0837$$

95% confidence interval for the proportion of individuals with disease = 0.0451–0.0831,
= 45.1–83.1 per 1000 persons

16.4 Crude, age-specific and standardised rates

The previous sections have demonstrated the calculation of rates overall in a defined population. These are known as crude rates. The same principle applies to calculating rates in population sub-groups. This is most commonly of interest for different age groups (Example 16.vi). In this example it can be seen that the incidence rate of the disease of interest increases with older age.

A frequent scenario in epidemiological investigations involves the comparison of rates of disease (whether prevalence, incidence or mortality) between populations or indeed within a population at different points in time. Using crude rates can lead to erroneous conclusions. For example Town A has a mortality rate of 5 per 100 000 person-years at risk (pyr) and Town B a mortality rate of 10 per 100 000 pyr. Does that mean a person in Town B has double the risk of dying compared to a person in Town A? Overall yes, but the difference in mortality rate in the two towns may be explained, for example, by the fact that the residents of Town B are older. As can be seen from Example 16.vii although Lowtown has increased mortality overall (i.e. crude rate) in comparison to Hightown, the mortality rate in every age group is, in fact, lower than in Hightown. The excess crude mortality rate in Lowtown is due to a much higher proportion of its residents being in the oldest age-group. It is important therefore in comparing rates between populations to take account of such factors which may differ between towns and have an important influence on the outcome of interest (see Chapter 18 for a fuller discussion of such *confounding factors*). With respect to population data such information is often restricted to age and gender.

It is important therefore to examine strata-specific rates in comparing populations with different age structures. In addition there are methods for producing a summary measure taking account of age differences (i.e. an age-standardised rate). However, particularly in relation to populations with very different age structures it should be emphasised that such measures should be calculated in addition to examining age-specific rates rather than as an alternative. The two methods of standardisation are *direct* and *indirect standardisation.*

Example 16.vi Age-specific rates

Index (i)	Age group years	Number of incident events (x_i)	Person–years at risk ($npyr_i$)	Incidence rate (per 10000 person–years) (r_i)
1	0–15	2	4432	4.51
2	16–34	6	3978	15.08
3	35–64	13	5396	24.09
4	65+	12	2159	55.58
	All ages	33	15965	20.67

Incidence rate in age-group $i = r_i = x_i / npyr_i$

All ages (crude) incidence rate: $R = \sum_i x_i \Big/ \sum_i n\ pyr_i$

Example 16.vii Comparing rates

	Hightown			Lowtown		
Age group (years)	No. of incident events	Person–years at risk	Incidence rate per 10000	No. of incident events	Person–years at risk	Incidence rate per 10000
0–15	10	9415	10.6	3	4103	7.3
16–34	18	8346	21.6	6	3765	15.9
35–64	20	6215	32.2	12	4192	28.6
65+	22	2196	100.2	73	7426	98.3
All ages	70	26172	26.7	94	19486	48.2

16.4a Direct standardisation

A directly age-standardised rate is the theoretical rate which would have been observed in the population under study if the age-structure of the population was that of a defined reference population. This reference population may be real or hypothetical. For example if comparing rates between six countries, the age-structure of one of the populations may be designated as the reference. Alternatively there are reference population age structures which have been proposed for different parts of the world. The calculations involved in direct age-standardisation are shown in Example 16.viii.

Thus in Example 16.viii the higher crude incidence rate of disease X in Lowtown is explained by the different age structures of the two towns. If Lowtown had the same age structure then its crude rate of disease would be lower than that in Hightown.

16.4b Indirect standardisation

The second method to compare rates between different populations, allowing for differences in age structure, is referred to as indirect standardisation. It involves applying a set of age-specific rates from the 'standard' population to the age-structure of the second population under study, to determine the 'expected' number of cases if such rates had applied in that population. This 'expected' number of cases is then compared with the actual number of cases 'observed' in the population under study. The ratio of observed/expected number of cases (often multiplied by 100) is called, when considering incidence rates, the Standardised Incidence Ratio (SIR) – with corresponding terms for Mortality and Prevalence. The calculations involved in indirect age-standardisation are shown in Example 16.ix. By definition the SIR of a population defined as the 'standard population' will be 100. In Example 16.ix Lowtown has an SIR of 93.7 in comparison to an SIR of 100 in Hightown (the 'standard population'). Thus the number of incident cases in Lowtown was only 93.7% of that expected if the age-specific rates of Hightown had operated.

It is important to note that if making several comparisons using indirect standardisation, each time applying age-specific incidence rates from a 'standard' population (say town A) to other populations (say town B and town C), then the weights used in the standardisation procedure are those of the population being compared to the standard. In such circumstances each population under study can only be compared to the standard and not directly to each other, i.e. the SIRs compare town B vs. town A and town C vs. town A but cannot be used directly to compare town B vs. town C. A common usage of standard ratios is in comparing rates amongst different population subgroups, e.g. occupational groups.

Example 16.viii Direct age-standardisation

Using data from Example 16.vii let us denote Hightown as the 'standard' population and Lowtown as the 'study' population.

Age group		Standard population (Hightown)		Study population (Lowtown)	
Index (i)	Years	Person–years at risk (pyr_i)	Proportion (pyr_i/PYR*)	Incidence rate (per 10000) ($r_i \times 10000$)	Weighted rate (per 10000) ($r_i \times 10000 \times pyr_i/PYR$)
1	0–15	9415	0.36	7.3	2.628
2	16–34	8346	0.32	15.9	5.088
3	35–64	6215	0.24	28.6	6.864
4	65+	2196	0.08	98.3	7.864
Total		26172	1		22.444

Directly age-standardised rate: $\text{DASR} = \sum_i \frac{r_i \times \text{py}r_i}{PYR}$

$= 22.4$ per 10000

Thus the DASR in Lowtown is lower than the incidence rate in Hightown (the standard population).

*PYR = total person–years at risk.

Example 16.ix Indirect age-standardisation

Again using data from Example 16.vi let us denote Hightown as the 'standard' population and Lowtown as the 'study' population.

Age group		Study population (Lowtown)		Standard population (Hightown)	
Index (i)	Years	Person–years at risk (pyr_i)	Number of incident events (c_i)	Incidence rate (per 10000) ($r_i \times 10000$)	'Expected cases' (ec_i)*
1	0–15	4103	3	10.6	4.3
2	16–34	3765	6	21.6	8.1
3	35–64	4192	12	32.2	13.5
4	65+	7426	73	100.2	74.4
Total			94		100.3

* $ec_i = r_i \times pyr_i$

$$\text{Standardised incidence rate: SIR} = \frac{\sum_i c_i}{\sum_i r_i \times pyr_i} \times 100$$

$$= \sum_i c_i \Big/ \sum_i ec_i \times 100$$

$$= 93.7$$

By definition, the 'standard' population (Hightown) has an SIR of 100, in comparison with the study population (Lowtown) has an SIR of 93.7.

17

Introductory data analysis: analytical epidemiology

17.1 Introduction

The purpose of this chapter is to outline the major simple analytical approaches to answering epidemiological questions with the use of data generated by the study designs described in previous chapters. Traditionally, analysis is first undertaken to examine the main effect of the factors under study. This is followed by consideration of whether any observed major effects can be explained by their association with other variables. This issue of confounding is dealt with in Chapter 18. In practice, relatively little mathematical calculation is done by hand calculator because many easy-to-use epidemiological programs exist for personal computers that permit a rapid and accurate approach to statistical analysis. Most statistical packages either include their own database or standard database files can be imported in a format that they can analyse. In this chapter, however, formulae are presented for the major measures of effect together with simple worked examples. Indeed, when data are available in tabulated form, as opposed to raw data files, it is frequently a simple task to calculate the important measures by hand. The formulae presented will permit the reader, for example, to check or further explore data published by others.

It is not the aim of this chapter to review all available statistical tests: reference should be made to appropriate statistical textbooks. Further, it is assumed that the reader has a basic knowledge of statistics and is aware of the use of summary measures including means and proportions, simple measures of variability including variance and standard deviation, and understands the conceptual basis for (i) making inferences about populations from studying samples and is familiar with the assessment of standard

errors and confidence intervals, and (ii) making comparisons between samples by using significance testing.

Sections are included on the analysis of case-control and cohort studies. There is no specific section on the analysis of cross-sectional studies – the appropriate methods will be found in the other sections. When a cross-sectional study is measuring the relative risk (risk ratio) of disease associated with an exposure then the methods outlined in Section 17.4b for prevalence/cumulative incidence data are appropriate. Alternatively, if the cross-sectional study is being used as a method for identifying diseased and non-diseased individuals (with further data collection thereafter on exposures) then the relationship between disease and exposure will be analysed using methods for case-control studies.

17.1a Statistical packages

There are now a large number of statistical packages available for epidemiological analysis, many of which incorporate a database package to permit data entry, checking and modification. Some packages are designed specifically for the epidemiologist, such as 'EPI-INFO' and 'EGRET'. Others, such as 'SPSS' (Statistical Package for Social Sciences) and 'SAS' (Statistical Analysis Software), have wider applications in survey research generally. Still others are focused more on the needs of biomedical research, such as 'BMDP' (BioMedical Data Program) and 'GLIM' (General Linear Interactive Models). Many programs now incorporate excellent graphical features to plot and print data in publishable form; examples are 'STAT GRAPHICS' and 'STATA'. All major analytical software is constantly being upgraded and most are available in the user-convenient Windows format including PC versions. The choice of package is frequently dictated by its availability within an institution. For the novice, it makes sense to work with a package for which there are other individuals available locally with experience who can sort out problems. The manuals that accompany most packages are a cross between a technical manual on using the program and a textbook of analytical methods. It thus makes sense to peruse the manuals of potential packages, not only to consider the scope of the applications, but also to gauge their ease of use given the experience (or lack) of the investigator.

17.2 Effect measurement, interval assessment and significance testing

Analysis of any study can reveal three items of potential value: an assessment of the major effect, the precision of that estimate, and its statistical significance.

Example 17.i

In a case-control study examining the influences of working in the dyeing industry for the development of bladder carcinoma, the following results were obtained: odds ratio for ever working with dyes, 3.2; 95% confidence interval 1.2–8.4; $0.01 < p < 0.05$.

The results give us different pieces of information. The odds ratio provides the best single estimate of the effect, assuming an unbiased study. The 95% confidence interval gives a range for the precision of that estimate and shows, in this example, that the data are consistent with a true effect that might only be marginal (20% increased) or indeed very large (over six-fold). The *p* value suggests that there is between 1% and 5% probability of observing by chance the odds ratio of 3.2, when the truth is that there is no increased risk (true odds ratio = 1). Hence the conclusion is that working in this industry increases the risk. The *p* value is perhaps the least valuable of all the analytical outcomes because most epidemiological questions do not require a yes/no answer, but are concerned with magnitude. In practice, many will quote the fact that a lower 95% confidence limit greater than unity indicates a real effect, in the belief that this will increase the likelihood of publication. Indeed, studies where the confidence intervals span unity are viewed as problematic by virtue of being of insufficient size. This is unfortunate because a study that yields a 95% confidence interval for an odds ratio of 0.8–1.3 gives as much information as one that yields a 95% confidence interval of 1.1–17.4. The first result implies that if there is an increased risk, it is likely to be small. The second result implies that an increased risk is likely, but its magnitude cannot be accurately estimated.

In all examples below, for simplicity only, 95% confidence intervals are calculated. These are the most frequently presented and accepted in practice. To obtain 90% or 99% intervals it is necessary to substitute the standard normal deviates of 1.58 and 2.64, respectively, for the figure of 1.96 in all the formulae presented.

17.3 Analysis of case-control studies

In a case-control study the aim is to measure the association between a risk factor and a disease by using the odds ratio (strictly, the ratio of odds of exposure in disease persons to that of exposure in the non-diseased). This odds ratio, generally, provides a good estimate of the relative risk. Where the exposure is continuous, for example as blood pressure, it is often preferable to express the risk in terms of an increase in disease risk per unit increase in exposure, for example each 1 or 10 mmHg increase in blood pressure. Such a calculation requires the procedure of logistic regression, which is readily available on most computer packages. Where the exposure is dichotomous, a two-by-two table can be drawn up and the odds ratio calculated easily as the ratio of the 'cross-products'.

Example 17.ii Calculation of odds ratio for dichotomous exposure

	Disease	
Exposure	Case	Control
Present	*a*	*b*
Absent	*c*	*d*

Odds ratio = (*a*:*c*)/(*b*:*d*),
= *ad*/*bc*

42 cases, 18 exposed;
61 controls, 17 exposed.

	Disease	
Exposure	Case	Control
Present	18	17
Absent	24	44
Total	42	61

Odds ratio = $18 \times 44/17 \times 24$,
= 1.94.

An odds ratio of 1 implies that the odds of exposure are the same amongst cases and controls, i.e. there is no relationship between exposure and disease. Odds ratios greater than 1 imply that the odds of cases being exposed is greater than controls (i.e. the exposure is a potential risk factor). While odds ratios less than 1 imply that the odds of cases being exposed is less than controls (i.e. a potential protective factor).

17.3a Calculation of confidence interval for an odds ratio

There are two methods in common use, the 'test'-based method and Woolf's method. They normally give very similar results; problems only arise with small numbers. Most computer programs specify which method is used; some present both. The 'test' method is based on the χ^2 statistic, which will be familiar to many readers as the standard statistical technique for comparing proportions from a contingency table.

Example 17.iii Calculation of confidence interval around odds ratio: (a) use of 'test'-based method

	Disease category		
Exposure	Case	Control	Total
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$a+c$	$b+d$	N

$\chi^2 = \sum (O-E)^2/E$

For a: Observed, $O = a$;
Expected, $E = (a+b)\ (a+c)/N$.

For b: $O = b$;
$E = (a+b)\ (b+d)/N$.

For c: $O = c$;
$E = (a+c)\ (c+d)/N$.

For d: $O = d$;
$E = (c+d)\ (b+d)/N$.

95% confidence limits $= \mathrm{OR}^{(1 \pm 1.96/\chi)}$

Note that a short-cut formula can be used:

$$\chi^2 = \frac{(ad - bc)^2 N}{(a + c)(a + b)(b + d)(c + d)}.$$

Using data from Example 17.ii

	Observed	Expected	$(O–E)^2/E$
a	18	$35 \times 42/103 = 14.27$	0.97
b	17	$35 \times 61/103 = 20.73$	0.67
c	24	$42 \times 68/103 = 27.73$	0.50
d	44	$61 \times 68/103 = 40.27$	0.35
		Total	2.49

$\chi^2 = 2.49$;
$\chi = 1.58$.

Lower 95% confidence limit $= 1.94^{(1-1.96/1.58)}$,
$= 1.94^{-0.24}$,
$= 0.85$.

Upper 95% confidence limit $= 1.94^{(1+1.96/1.58)}$,
$= 1.94^{2.24}$,
$= 4.41$.

Example 17.iv Calculation of confidence interval around odds ratio: (b) use of Woolf's method

$$95\%\ \text{Confidence limits} = \exp\left(\log_e \text{OR} \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

Using data from Example 13.viii,
OR = 1.94;
$\log_e$ OR = 0.663.

$$\text{Lower 95\% confidence limit} = \exp\left(0.663 - 1.96\sqrt{\frac{1}{18} + \frac{1}{17} + \frac{1}{24} + \frac{1}{44}}\right),$$

$= \exp(0.663 - 1.96\sqrt{0.179})$.
$= \exp(0.663 - 1.96 \times 0.423)$,
$= \exp(0.663 - 0.829)$,
$= \exp(-0.166)$,
$= 0.85$.

$$\text{Upper 95\% confidence limit} = \exp\left(0.663 + 1.96\sqrt{\frac{1}{18}+\frac{1}{17}+\frac{1}{24}+\frac{1}{44}}\right),$$
$$= \exp\,(0.663 + 1.96\sqrt{0.179}),$$
$$= \exp\,(0.663 + 1.96 \times 0.423),$$
$$= \exp\,(0.663 + 0.829),$$
$$= \exp\,(1.492),$$
$$= 4.45.$$

(Note: exp is the inverse of the natural logarithm function. The exp function is available on most calculators.)

The examples above illustrate that both methods give virtually identical results. Note that the odds ratio does not lie mathematically in the centre between the two limits, reflecting its logarithmic properties. The consequence of this is that displaying confidence intervals graphically should be done on a log scale, which would then suggest equal distances from the observed odds ratio.

17.3b Calculation of odds ratios with multiple levels of exposure

Frequently, the exposure can be considered after stratifying into a number of different levels. Indeed, dichotomising exposures into an arbitrary yes/no does not use all the data, whereas demonstration of a trend of increasing risk with increasing exposure is valuable evidence of a real effect (dose–response effect).

The major analytical principle is to relate the risk in each exposure stratum to that of one reference stratum, normally that presumed to be at lowest risk, or absence of exposure. This stratum is then given an odds ratio of 1.

The choice of the number of strata to be used is dependent on the numbers available and the biological sense of splitting up the exposure. The best statistical use is made when the strata have equal numbers and thus, in the absence of powerful biological or clinical arguments, the entire cohort is divided up by tertiles, quartiles or quintiles (see Section 5.5). There is no further gain in demonstrating a dose response in going beyond five categories. For other exposures unequal categorisation may be more appropriate, for example for smoking: never smoked, ex-smoker, currently less than 5 cigarettes per day, 5–20 per day and more than 20 per day.

Example 17.v Calculation of odds ratio with multiple levels of exposure

Exposure level	Cases	Controls	Odds ratio
1	a_1	b_1	$\frac{a_1 b_1}{a_1 b_1} = 1$
2	a_2	b_2	$\frac{a_2 b_1}{a_1 b_2}$
3	a_3	b_3	$\frac{a_3 b_1}{a_1 b_3}$
⋮	⋮	⋮	⋮
i	a_i	b_i	$\frac{a_i b_1}{a_1 b_i}$

In a case-control study comparing body mass index (BMI) between affected and non-affected individuals, it was decided to stratify BMI into four levels.

BMI stratum	Cases	Controls	Odds ratio
1	21	30	$\frac{21 \times 30}{21 \times 30} = 1$
2	31	26	$\frac{31 \times 30}{21 \times 26} = 1.70$
3	24	11	$\frac{24 \times 30}{21 \times 11} = 3.12$
4	17	4	$\frac{17 \times 30}{21 \times 4} = 6.07$

Estimate of linear trend

In the Example 17.v, the data suggest that there is a trend of increasing risk with increasing levels of obesity. This can be tested for statistical significance by using the χ^2 test for trend. It is possible to do this relatively easily by hand calculator as shown, but a number of computer programs will also do this calculation.

Example 17.vi Determination of presence of linear trend with increasing exposure

Notation:

x_i arbitrary score given to exposure stratum i;
n_i number of cases and controls in stratum i;
d_i number of cases in stratum i;
N, total number of cases and controls.

Using data layout from Example 17.v:

Exposure stratum	Cases	Controls	x_i	n_i	d_i	$n_i x_i$	$d_i x_i$	$n_i x_i^2$
1	a_1	b_1	1	$a_1 + b_1$	a_1	$1(a_1 + b_1)$	$1(a_1)$	$1^2(a_1 + b_1)$
2	a_2	b_2	2	$a_2 + b_2$	a_2	$2(a_2 + b_2)$	$2(a_2)$	$2^2(a_2 + b_2)$
3	a_3	b_3	3	$a_3 + b_3$	a_3	$3(a_3 + b_3)$	$3(a_3)$	$3^3(a_3 + b_3)$
i	a_i	b_i	i	$a_i + b_i$	a_i	$i(a_i + b_i)$	$i(a_i)$	$i^2(a_i + b_i)$
Total				N	Σd_i	$\Sigma n_i x_i$	$\Sigma d_i x_i$	$\Sigma n_i x_i^2$

$$\chi^2_{\text{trend 1df}} = \frac{N[N(\Sigma d_i x_i) - \Sigma d_i (n_i x_i)]^2}{\Sigma d_i (N - \Sigma d_i)[N(\Sigma n_i x_i)^2 - (\Sigma n_i x_i)^2]}.$$

Using data from Example 17.v:

BMI stratum	x_i	n_i	d_i	$n_i x_i$	$d_i x_i$	$n_i x_i^2$
1	1	51	21	51	21	51
2	2	57	31	114	62	228
3	3	35	24	105	72	315
4	4	21	17	84	68	336
Total		164	93	354	223	930

$$\begin{aligned}
\chi^2_{\text{trend 1df}} &= \frac{164[164(223) - 93(354)]^2}{93(164 - 93) \times [164(930) - (354)^2]}, \\
&= \frac{164(36572 - 32922)^2}{93(71) \times (152520 - 125316)}, \\
&= \frac{164(3650)^2}{93(71) \times (27204)}, \\
&= 12.16.
\end{aligned}$$

Thus, from tables, $p = 0.0005$.

The resultant value for χ^2 has one degree of freedom. Thus, values in excess of 3.84 are said to describe a significant trend, although in practice inspection of the data will yield the same conclusion! Inspection of the data will also permit an idea of trends that are not linear, including the so-called J-shaped curve, where low dose exposure is associated with a higher risk than an intermediate dose, but subsequent increases in exposure are associated with an increase in risk.

17.3c Analysis of matched pairs

The calculation of an odds ratio for matched pairs is different from that for an unmatched analysis. In practice, a study that is designed as matched may frequently be analysed as though it were unmatched because, as a consequence of dropouts and withdrawals, there are incomplete pairs. Therefore, where an unmatched analysis will include all the subjects seen, a matched analysis will include only complete matched sets. For the simplest form of 1:1 matching, the calculation of the odds ratio can be rapidly achieved. It is important to note that the unit included in each cell of the table is a pair rather than an individual. The odds ratio is calculated as the ratio of the discordant pairs. Thus, in simple terms the odds ratio will be greater than unity if there are more pairs with the case exposed and the control not exposed than the reverse.

Example 17.vii Calculation of odds ratios for matched pairs

Taking *p*, *q*, *r* and *s* as numbers of pairs:

	Case	
Control	Exposure present	Exposure absent
Exposure present	*p*	*q*
Exposure absent	*r*	*s*

OR = *r*/*q*.

In a study of 78 matched case-control pairs, in 11 pairs both smoked, in 17 pairs neither smoked, in 36 pairs the case smoked and the control did not smoke, whereas in the remaining 14 pairs the converse was the case.

	Case	
Control	Smoked	Not smoked
Smoked	11	14
Not smoked	36	17

OR = 36/14 = 2.57.

It is clear that it is the exposure-discordant pairs that are informative. In studies where there are multiple controls per case or, as is often the case in practice, the study ends up with a variable number of controls per case, the calculations become rather complex. The conditional logistic regression procedure available on many software packages can undertake this kind of analysis as well as examining exposure, either at multiple levels or as a continuous variable similar to the way that 'normal' or unconditional logistic regression does for unmatched case-control studies.

Example 17.viii Calculation of confidence interval around odds ratio derived from a matched-pair analysis

As with the calculation of a confidence interval around an unmatched derived odds ratio, the appropriate calculation is test-based.

In this instance the 'test' is McNemar's test for matched pairs:

$$\chi^2 = \frac{[(r-q)-1]^2}{r+q}.$$

95% confidence interval = $\mathrm{OR}^{(1 \pm 1.96/\chi)}$

Using data from Example 17.xii,

$$\chi^2 = \frac{[(36-14)-1]^2}{(36+14)}$$

$$= 8.82,$$

$$\chi = 2.97.$$

Lower 95% confidence limit $= 2.57^{(1-1.96/2.97)}$,

$= 2.57^{(0.340)}$,

$= 1.38.$

Upper 95% confidence limit $= 2.57^{(1+1.96/2.97)}$,

$= 2.57^{(1.660)}$,

$= 4.79.$

17.4 Analysis of cohort studies

17.4a Calculation of rate ratio from incidence data

A ratio of incidence (density) rates between two exposure groups provides an estimate of the rate ratio, with the incidence rates being calculated as in Example 17.ix.

Example 17.ix Calculation of rate ratio (RR) estimate from incidence (density) rates

	Person–years at risk	Number with outcome	Incidence rate
Exposed	$npyr_e$	x_e	$x_e/npyr_e$
Non-exposed	$npyr_0$	x_0	$x_0/npyr_0$

$RR = (x_e/npyr_e)/(x_0/npyr_0)$

In a follow-up study comparing the incidence of melanoma in two groups, stratified at baseline by the presence of a significant number of benign naevi, the following results were obtained.

	Person–years	No. of cases
High naevi count	12 100	19
Low naevi count	41 500	20

Incidence in high count group = 15.70/10 000 pyr;
Incidence in low count group = 4.82/10 000 pyr.

RR = 3.26.

The rate difference is also of importance, being the incidence in the exposed minus that in the non-exposed, and gives a 'real' incidence value for the risk of exposure. If data are available on the proportion of individuals in the population who are exposed, it is also possible to calculate the proportion of the total number of cases arising within the population that are due to the exposure.

Example 17.x Calculation of population attributable risk

Incidence in exposed $= I_e$;
Incidence in non-exposed $= I_0$.

Incidence in exposed due to exposure $= I_e - I_0$.

Proportion of incidence in exposed due to exposure $= \dfrac{I_e - I_0}{I_e}$

As $RR = I_e/I_0$, dividing by I_0, $= \dfrac{RR - 1}{RR}$.

If proportion of population exposed $= p_e$, then proportion of all cases in population due to exposure (population attributable risk)

$$= \frac{p_e(RR - 1)}{1 + p_e(RR - 1)}$$

In a prospective study of ever use of oral contraceptives and stroke, a RR of 3.1 was found for ever users. The proportion of ever users in the population studied was 0.6.

Proportion of risk in oral contraceptive users due to oral contraceptive use $= \dfrac{3.1 - 1}{3.1} = 0.68$.

Proportion of cases in female population due to oral contraceptive use

$$= \frac{0.6(3.1 - 1)}{1 + 0.6(3.1 - 1)},$$

$$= 0.56.$$

Thus, in this example, 68% of the stroke risk in oral contraceptive takers is due to their use of oral contraceptives and the remainder is due to their background (population) risk. Further, given the frequency of oral contraceptive use in the female population, it can be estimated that 56% of all the cases that arise can be explained by their use. Alternatively, if these data were true, abolishing oral contraceptive use might have the potential for reducing the number of cases of stroke in a female population by over half.

Calculation of confidence interval around incidence rate ratio

As with the calculation of the confidence interval around an incidence rate, the crucial factor is the number of cases; the number of person–years at risk does not influence the calculation.

Example 17.xi Calculation of confidence interval around incidence rate ratio

An approximation may be found from the following formulae:

$$95\%\ \text{confidence interval} = \exp\left(\log_e \text{RR} \pm 1.96\sqrt{\frac{1}{x_0}+\frac{1}{x_e}}\right).$$

Using data from Example 17.ix,

$$\text{lower 95\% confidence limit} = \exp\left(\log_e \text{RR} \pm 1.96\sqrt{\frac{1}{19}+\frac{1}{20}}\right).$$

$$= \exp\,(1.18 - 1.96\ \sqrt{0.1026}),$$
$$= \exp\,(1.18 - 0.63),$$
$$= \exp\,0.552,$$
$$= 1.74.$$

$$\text{upper 95\% confidence limit} = (1.18 + 0.63),$$
$$= 1.18,$$
$$= 6.11.$$

As with the confidence interval around an odds ratio, the distribution is logarithmic.

17.4b Calculation of risk ratio estimate from prevalence or cumulative incidence data

This, as discussed in Section 17.3, is relevant to a comparison of prevalence proportions or cumulative incidences. The risk ratio is simply the ratio of the two proportions.

Example 17.xii Calculation of risk ratio estimate from prevalence or cumulative incidence data

	Case	Not case
Exposed	a	b
Not exposed	c	d

$$\text{Risk in exposed} = a/(a+b);$$
$$\text{Risk in not exposed} = c/(c+d).$$

$$\text{RR} = [a/(a+b)]/[c/(c+d)];$$

221 infants, of whom 49 weighed under 2000 g at birth, were studied; 18 of the low birth-weight group, and 29 of the normal birth-weight group, had learning difficulties at age 6.

	Learning difficulties		
Birth weight	Yes	No	Total
Low (<2000 g)	18	31	49
Normal (≥2000 g)	29	143	172

Risk of learning difficulty in low birth-weight group = 18/49;
risk of learning difficulty in normal birth-weight group = 29/172.

RR = (18/49)/(29/172)
= 2.18.

Calculation of a confidence interval around a risk ratio

As with other confidence-interval calculations, there is the possibility of using either a 'test'-based approach or a logarithmic transformation: both give very similar results.

Example 17.xiii Calculation of confidence interval for risk ratio: (a) use of test-based method

95% confidence limits = $RR^{(1 \pm 1.96/\chi)}$.

Using data from Example 17.xii,

χ^2 (by short-cut formula from Example 17.iii)
$= 9.00$
$\chi = 3$.

Lower 95% confidence limit $= 2.18^{(1 - 1.96/3)}$
$= 2.18^{(0.347)}$
$= 1.31$.

Upper 95% confidence limit $= 2.18^{(1 + 1.96/3)}$
$= 2.18^{(1.653)}$
$= 3.63$.

17.4c Life-table method for incidence data

One problem with comparing incidence rates, as in Section 17.4a above, is that such a comparison takes no account of the time to an event. A simple

example will suffice to clarify the point. The ultimate cumulative incidence of death in any group of individuals is always 100%; the point of interest is the time course of the deaths. Thus, it is frequently the rate ratio of an event at a particular point in time that is of interest: this is often referred to as the *hazard ratio*. The life-table approach permits a calculation of the probability of an event at a particular time since the start of observation. The probability at the end of the second year (say) for developing an event is equivalent to that probability of developing the event during the first year multiplied by the probability of developing the event in the second year if free from the event at the start of the second year.

Example 17.xiv Use of life-table methods for analysis of incidence data

Time interval	Number entering interval	Number developing event	Probability of event within interval	Probability of no event within interval	Cumulative probability of no event
1	n_1	x_1	x_1/n_1	$1-x_1/n_1=p_1$	p_1
2	$n_1-x_1=n_2$	x_2	x_2/n_2	$1-x_2/n_2=p_2$	p_1p_2
i	$n_{i-1}-x_{i-1}=n_i$	x_i	x_i/n_i	$1-x_i/n_i=p_i$	$p_1p_2\ldots p_i$

One hundred smokers were given intensive health education to give up smoking. At the end of each of the subsequent five years, 5, 10, 8, 7 and 6 subjects, respectively, had restarted the habit.

Time interval	Number entering	Number restarting smoking	Probability of restarting	Probability of remaining non-smoker	Cumulative success
1	100	5	0.05	0.95	0.95
2	95	10	0.105	0.895	0.85
3	85	8	0.094	0.906	0.770
4	77	7	0.091	0.909	0.700
5	70	6	0.086	0.914	0.640

In Example 17.xiv, the cumulative probability of remaining a non-smoker at the end of the five years could have been calculated much more easily by noting that at the end of the fifth year there were 64 of the original 100 subjects who had not relapsed.

However, in practice, it is exceedingly unlikely that all subjects entered at time '0' will have exactly five years of follow-up. Some will die; others will be lost to follow-up for a variety of other reasons. More importantly, no study recruits all individuals at a single point in time. Thus, in considering the disease risk following an occupational exposure, the duration of follow-up will vary, being the interval since a particular subject was first employed to the date that the follow-up of that individual was completed: the censorship date. Some individuals may therefore have had 10 years of follow-up, whereas others, depending on the study design, had perhaps less than one year. Thus, there is frequently incomplete information, a phenomenon known as *right censorship*, i.e. incidence data are missing to the 'right' of the conceptual data line. If the assumption is made that the risk of disease is unrelated to the chances of being 'censored', each individual can be included up until the time they are censored. When this analysis is not undertaken by computer, the approach is to assume that the individual was censored midway through an interval and hence contributed half a person–time of follow-up for that interval. Thus, taking the same data as in Example 17.xiv and adding information on the losses or end of follow-up during each interval produces the result as shown below.

Example 17.xv Use of life-table methods with losses

Time interval	Number entering	Number 'lost' during interval	Number at risk, n_i	Number restarting smoking, x_i	Probability of restarting smoking, x_i/n_i	Probability of remaining non-smoker	Cumulative success
1	100	4	98	5	0.051	0.949	0.949
2	91	3	89.5	10	0.112	0.888	0.842
3	78	2	77	8	0.104	0.896	0.755
4	68	5	65.5	7	0.107	0.893	0.674
5	56	4	54	6	0.111	0.889	0.599

Clearly, the cumulative success for remaining a non-smoker is actually lower than that seen in Example 17.xiv, owing to the censored observations.

Data from these calculations can be used to plot 'survival' (i.e. event-free survival) curves from which the event-free survival can be read off at a particular time of interest (Fig. 17.1). It is easier to undertake this analysis by

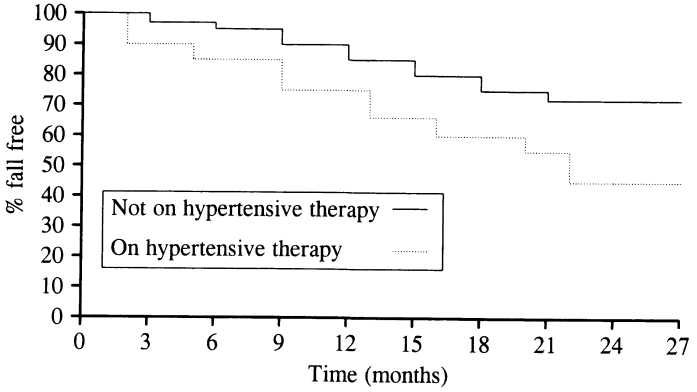

Figure 17.1 Life-table analysis of incidence of falls in the elderly.

computer, which uses the actual date of each event to generate a more accurate curve than can be obtained by the interval approach used above; however, the principle is the same. The curves generated are often known as Kaplan–Meier curves and the survival estimate at a particular point in time is known as the Kaplan–Meier estimate. It is also possible to calculate the confidence interval around the estimates at each time point.

Comparison of survival curves

A comparison of the curves for disease-free survival (say) between two exposure groups can give an estimate of the effect of the exposure. Survival curves can be compared for statistical significance by using the LogRank test, which can be done easily by hand calculator; a suitable layout is shown below.

Example 17.xvi Use of LogRank test to compare survival curves

	Group A		Group B		Combined	
Time interval	At risk	Number of observed events	At risk	Number of observed events	At risk	Number of observed events
1	n_{1A}	o_{1A}	n_{1B}	o_{1B}	$n_{1A} + n_{1B}$	$o_{1A} + o_{1B}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	n_{iA}	o_{iB}	n_{iB}	o_{iB}	$n_{iA} + n_{iB}$	$o_{iA} + o_{iB}$

If the numbers of events in both groups were distributed randomly related only to the number at risk in each group, then, in any given time interval i:

expected number of events in Group A, $e_{i\mathrm{A}} = \dfrac{n_{i\mathrm{A}}}{n_{i\mathrm{A}} + n_{i\mathrm{B}}}(o_{i\mathrm{A}} + o_{i\mathrm{B}})$;

expected number of events in Group B, $e_{i\mathrm{B}} = \dfrac{n_{i\mathrm{B}}}{n_{i\mathrm{A}} + n_{i\mathrm{B}}}(o_{i\mathrm{A}} + o_{i\mathrm{B}})$.

Then for all time intervals considered from 1 to i:

total number of observed events in Group A $= \sum_{1}^{i} o_{i\mathrm{A}}$;

total number of observed events in Group B $= \sum_{1}^{i} o_{i\mathrm{B}}$;

total number of expected events in Group A $= \sum_{1}^{i} e_{i\mathrm{A}}$;

total number of expected events in Group B $= \sum_{1}^{i} e_{i\mathrm{B}}$.

LogRank test gives the value for $\chi^2_{1\mathrm{df}}$ as

$$\frac{(\Sigma o_{i\mathrm{A}} - \Sigma e_{i\mathrm{A}})^2}{\Sigma e_{i\mathrm{A}}} + \frac{(\Sigma o_{i\mathrm{B}} - \Sigma e_{i\mathrm{B}})^2}{\Sigma e_{i\mathrm{B}}}.$$

Using data from Example 17.xv for Group A and adding data for Group B:

	At risk		Observed events		Expected events	
Time period	Group A $n_{i\mathrm{A}}$	Group B $n_{i\mathrm{B}}$	Group A $o_{i\mathrm{A}}$	Group B $o_{i\mathrm{B}}$	Group A $e_{i\mathrm{A}}$	Group B $e_{i\mathrm{B}}$
1	98	97	5	4	4.52[a]	4.48
2	89.5	91	10	6	7.93	8.07
3	77	82	8	5	6.30	6.70
4	65.5	74.5	7	5	5.61	6.39
5	54	66	6	7	5.85	7.15
Total			36	27	30.21	32.79

Note:

[a] Calculated from $\dfrac{98}{97 + 98} \times (4 + 5)$.

$$\chi^2_{1\mathrm{df}} = \frac{(36 - 30.21)^2}{30.21} + \frac{(27 - 32.79)^2}{32.79},$$

$$= 1.11 + 1.02,$$

$$= 2.13.$$

From statistical tables for $p = 0.05$,

$$\chi^2_{1df} = 3.84.$$

Thus we conclude that the curves from the two groups are unlikely to represent different responses.

17.5 Conclusion

The exciting point in any epidemiological study comes when the raw data collected are analysed to reveal the answer to the main question posed, be it an incidence rate, an odds ratio or other measure. In practice, the calculations are most often done by computer, but the investigator needs to be aware of how the results were arrived at. For all studies it is necessary to calculate not only an estimate of the main effect but also a confidence interval around that result.

18

Confounding

18.1 Introduction

Confounding is best introduced by way of a simple example:

Example 18.i

In a prospective cohort study, coal miners who worked at the coal face were found to have twice the incidence of carcinoma of the lung as miners who worked on the surface. It was questioned whether this represented a real increase in risk associated with face work or whether the increased risk could be explained by face workers smoking more than surface workers.

In this example the question is raised as to whether the apparent relationship between the disease and the risk factor is explained (*confounded*) by their joint association with the 'true' risk factor of smoking. Unlike the discussion of bias in Chapter 19, the issue here is not one of impaired validity: if the study had been carefully conducted then the relationship observed was correct. The problem of confounding is therefore one of interpretation. (Some textbooks of epidemiology refer to 'confounding bias', a conjunction that is misleading and unhelpful.)

Further evaluation of this simple example reveals some key points about confounding.

(i) Confounding can only be proved after appropriate analysis; in the example above, confounding due to cigarette smoking is suspected but remains to be demonstrated. As a consequence, the observation that a particular variable is (or is not) a confounder in one study does not answer the question as to whether it is acting similarly in the current investigation.

(ii) Confounding as an explanation of an association is not (or is only extremely rarely) an all-or-nothing phenomenon. The effect of confounding will be to alter the strength of an apparent relationship between two variables (e.g. risk factor and disease). The effect of the confounder may make the apparent relationship disappear. An accurate estimate of the strength of any observed association requires consideration of the effects of all likely important confounders. Occasionally, however, *adjusting* for a confounder can strengthen the observed association, as in the example below (negative confounding).

Example 18.ii

A prospective study comparing pregnancy outcome in smokers and non-smokers suggested that the former resulted in babies of lower birth weight. It was known that smoking varies with maternal age and that the latter also affects birth weight. However, adjusting for maternal age increased the strength of the association.

(iii) The demonstration of confounding cannot be taken as an indication of the direction of a cause-and-effect relationship. Although common sense and previous experience are helpful, it may be impossible to reach a clear conclusion, as the following example shows:

Example 18.iii

A study was undertaken in a Bangladeshi migrant population sample living in London, to determine whether blood pressure was related to duration of living in the UK. A strong association was found, which disappeared after adjusting for the possible confounding effect of age. It was impossible to distinguish between the two alternative hypotheses: (i) that duration in the UK explained an observed hypertensive effect of age, and (ii) that age explained an observed hypertensive effect of duration in the UK.

(iv) Confounding should also be distinguished from one variable acting *on the path* in explaining an association, i.e. an intermediate variable that exists as a consequence of the exposure and therefore should not be adjusted for. As an obvious example, the demonstration that individuals who have a high energy (calorie) intake have an increased risk of developing maturity onset (Type II) diabetes is probably mediated via obesity. It would thus be nonsensical to adjust for body weight because

this would eliminate the relationship, and the real risk from overeating would be masked!

One of the major advances in epidemiological method in the past decade has been the widespread availability of statistical software for personal computers that enables the simultaneous adjustment of an association for multiple potential confounders. Indeed, in the same analysis it is now relatively easy to examine the strength of association, not only of the main variable under study, *the main effect*, but also of a variety of potentially important confounders. In practice, therefore, it is possible to examine a number of variables simultaneously for their *independent* effects and it is not necessary to specify which effect is the main one of interest. This is discussed in detail in Section 18.4d.

It is clear that confounding can be a problem and hinders appropriate interpretations and it is important that it should be minimised where possible and evaluated for its impact where not.

18.2 Minimising confounding in study design

The first step is to think about the potential confounders in a proposed study. A list can be drawn up from the results of previous studies and from common sense. The consequence of drawing up such a list is that it is incumbent on the investigator to attempt to collect these data from the study's subjects, if at all possible. The list should, however, be of manageable size. It is useful for the investigator to adopt the role of a potential reviewer or other critical reader who might ask whether an association revealed could be explained by a confounder that the investigator had not considered.

18.2a Selection of study subjects

Matching

One approach to minimising confounding is to make the groups being compared as similar as possible in respect of potential confounders by using either individual or frequency matching, as discussed in Section 8.4. The problems with this strategy as outlined are, first, that the effects of the matched variables cannot be examined and, secondly, that the matching process is inefficient in time, may be difficult in respect of a large number of

potential confounders and can seriously reduce the numbers available for study.

The unknown confounders

In comparative studies there is often the concern that any difference observed between the groups under study may result from confounding, though the exact nature of the confounding variable(s) remains obscure, preventing their ascertainment. This is a particular problem in case-control studies where cases are not selected from a true population base. These unknown confounders can therefore be controlled by attempting to recruit controls from a similar base. Two examples illustrate the point.

Example 18.iv

A case-control study of a rare disease recruited cases from the register of a specialist unit in a major academic centre. The investigator was unsure what factors, such as socio-economic, educational, etc., determined whether a patient with that disease would be referred to such a unit. The anxiety was that these very factors might confound any observed association between the exposures studied and the disease. To minimise this, it was felt that an appropriate comparison group would be cases with a broad range of other diseases referred to the same unit, but for which there was no evidence that the exposure under investigation was linked.

Example 18.v

In a case-control study of an even rarer disease, after numerous enquiries an investigator was able to reveal 80 cases nationwide. The population base for these cases was unknown and the investigator was concerned that any association uncovered might be explained by unknown confounders relating to geographical and other local factors. It was decided to use friend controls in an attempt to minimise this possibility.

18.3 Conduct of study

There is little that can be done during the data-collection phase to reduce confounding. In studies with sequential recruitment over a long period, it might be appropriate to monitor the subjects to ensure a reasonable balance between the major potential confounders such as age and sex. This is of particular relevance in studies where the target number to be recruited is small. If the study groups end up with substantially different distributions in

Table 18.1. Comparison of potential confounding variables between study groups

	Case, $N = 80$	Controls, $N = 94$
Age: mean (SD)	41.4 (6.2)	42.3 (7.1)
Female (%)	31 (38.8)	40 (42.6)
Current smokers (%)	19 (23.8)	25 (26.6)
Years full-time education: mean (SD)	9.8 (2.1)	10.3 (2.0)

respect of the main potential confounders, then interpretations of effects may be impossible. It sometimes becomes apparent, once a study has moved into the data-collection phase, that an unforeseen confounder may be operating. In such circumstances it might be possible to collect data on this potential confounder, at least in those still to be recruited.

Example 18.vi

In a prospective cohort study comparing the frequency of headache and related minor disorders in secretaries in relation to the number of weekly hours spent in front of a monitor, early in the study the investigators became aware that those who were high users also worked longer hours and were often working under greater pressure. Initially they had not thought to collect these data but the study had to be amended to allow such information to be obtained.

18.4 Analysis

There are four analytical approaches to permit an estimate of confounding in a completed study. The details of the analytical methods involved are beyond the scope of this volume, and adjusting for possible confounders is a major topic in its own right. In the discussion below, the aim is to present the major uses of the different strategies with illustrations and to provide appropriate cautions about their use.

18.4a Baseline comparison of potential confounders

The first stage is to draw up a table comparing the frequency distributions of the potential confounders (or their summary statistics such as means, medians, etc.) to determine whether a serious problem is likely. Thus, in Table 18.1 data are presented from a case-control study comparing the dis-

tribution of some potential confounding variables. The differences are relatively small but the controls are slightly older, more likely to be female, to be current smokers and to have spent more years in full-time education. None of these differences seem large and they are intuitively unlikely to affect the results to any major extent. This reasoning has the disadvantage of being 'ecological' in so far as the two groups may have similar overall age distributions and overall sex distributions but (say) in one group the males might be much older than the females, and vice versa. If the effect of the main variable under study is itself altered by this interaction this could be a problem. None the less this 'baseline' comparison is an important first step, and with luck it might provide some reassurance.

18.4b Stratification

Inspection of raw data may suggest that there are important differences between the groups being compared in relation to the distribution of a potential confounder, to the extent that these differences might explain the observed result. In those instances where it is possible (given the available numbers) and biologically sensible to divide up the subjects into different categories (*strata*) of the possible confounders, a stratified analysis is an appropriate first step. Consider the data provided in Table 18.2. The same data will be used to illustrate the approach to be adopted if the study is (i) a cohort study with the end point being a risk ratio of the cumulative incidences, or (ii) a case-control study with the end point being an odds ratio.

In this example the risk of disease in the exposed population appears to be twice that in the unexposed after crude analysis. However, the same study's data revealed (Table 18.2b) that, overall, males have an approximately 3-fold increased risk compared with females and thus it is pertinent to enquire whether the difference between exposed and non-exposed in Table 18.2a could be explained by differences in the sex distribution between the exposure groups. The relevant data are given in Table 18.2c, which indeed show that the males are heavily *weighted* towards being exposed. As a consequence, analysis by *sex-specific strata* shows that within strata there is probably no increased risk due to exposure. The original observed relationship between exposure and disease was therefore confounded by sex.

It is possible to calculate relatively easily, with a hand calculator, an adjusted relative risk or odds ratio by using the data from stratified two-by-two tables

Table 18.2. Investigation of possible confounding by stratification

(a) *Crude analysis*

Exposure status (number)	Disease: Yes	Disease: No	Cumulative incidence (%)	Risk ratio	Odds ratio
Exposed (500)	15	485	3	2.0	2.0
Non-exposed (2500)	38	2462	1.5		

(b) *Relationship to possible confounder of sex*

Sex stratum (number)	Disease: Yes	Disease: No	Cumulative incidence (%)	Risk ratio	Odds ratio
Male (1000)	31	969	3.1	2.8	2.9
Female (2000)	22	1978	1.1		

(c) *Analysis stratified by sex*

Sex stratum	Exposure status (number)	Disease: Yes	Disease: No	Cumulative incidence (%)	Risk ratio	Odds ratio
Male	Exposed (400)	14	386	3.5	1.2	1.2
	Non-exposed (600)	17	583	2.8		
Female	Exposed (100)	1	99	1.0	0.9	0.9
	Non-exposed (1900)	21	1879	1.1		

(Table 18.3). The principle is to calculate an 'average' or weighted effect estimate across the strata examined. The calculations for an odds ratio are shown in Table 18.3a and for a risk ratio in Table 18.3b. The upper box in both tables shows the calculation for the first stratum and the second box gives the calculation for the *i*th stratum. The summary estimate is then obtained by summation across the *i* strata analysed. The application of this formula to the data in Table 18.2 is given and shows how adjusting for sex virtually abolishes any effect of exposure. The results are summarised in Table 18.4.

The weighted odds ratio is referred to as the *Mantel–Haenszel* estimate after the authors who first described the procedure. It is also necessary to calculate confidence intervals around these estimates, and formulae are readily available. Indeed, many simple statistical packages calculate these relatively easily. Table 18.4 gives the confidence interval results from the data in Table 18.2.

Table 18.3a. Calculation of Mantel–Haenszel summary

(a) *Odds ratio (OR)*

(i) *First stratum of confounder*

	Disease	
Exposure	+ve	−ve
+ve	a_1	b_1
−ve	c_1	d_1

Notes:

$n_1 = a_1 + b_1 + c_1 + d_1$;

$$\text{OR} = \frac{a_1 d_1}{n_1} \Bigg/ \frac{b_1 c_1}{n_1}.$$

(ii) i*th stratum of confounder*

	Disease	
Exposure	+ve	−ve
+ve	a_i	b_i
−ve	c_i	d_i

Notes:

$n_i = a_i + b_i + c_i + d_i$;

$$\text{OR} = \frac{a_i d_i}{n_i} \Bigg/ \frac{b_i c_i}{n_i}.$$

$$\text{Summary OR} = \sum_1^i \frac{a_i d_i}{n_i} \Bigg/ \sum_1^i \frac{b_i c_i}{n_i}.$$

Using data from Table 18.2, summary odds ratio

$$= \left\{ \left(\frac{14 \times 583}{1000} \right) + \left(\frac{1 \times 1879}{2000} \right) \right\} \Bigg/ \left\{ \left(\frac{386 \times 17}{1000} \right) + \left(\frac{21 \times 99}{2000} \right) \right\},$$

$$= (8.16 + 0.94)/(6.56 + 1.04),$$

$$= 9.10/7.60,$$

$$= 1.2.$$

Table 18.3b.

(b) *Risk ratio (RR)*

(i) *First stratum of confounder*

	Disease	
Exposure	+ve	−ve
+ve	a_1	b_1
−ve	c_1	d_1

Note:

$$\text{Relative risk} = \frac{a_1}{a_1 + b_1} \Bigg/ \frac{c_1}{c_1 + d_1}.$$

(ii) i*th stratum of confounder*

	Disease	
Exposure	+ve	−ve
+ve	a_i	b_i
−ve	c_i	d_i

Notes:

$$\text{Relative risk} = \frac{a_i}{a_i + b_i} \Bigg/ \frac{c_i}{c_i + d_i}.$$

$$\text{Summary RR} = \sum_1^i \frac{a_i(c_i + d_i)}{n_i} \Bigg/ \sum_1^i \frac{c_i(a_i + b_i)}{n_i}.$$

Using data from Table 18.2, summary risk ratio

$$= \left\{\left(\frac{14 \times 600}{1000}\right) + \left(\frac{1 \times 1900}{2000}\right)\right\} \Bigg/ \left\{\left(\frac{17 \times 400}{1000}\right) + \left(\frac{21 \times 100}{2000}\right)\right\},$$

$$= (8.40 + 0.95)/(6.80 + 1.05),$$

$$= 9.35/7.85,$$

$$= 1.2.$$

Table 18.4. Calculation of confidence limits from the data in Table 18.2

Crude odds ratio for all strata = 2.00
Mantel–Haenszel weighted OR = 1.20
95% confidence limits: 0.58, 2.44
Crude risk ratio for all strata = 2.00
Mantel–Haenszel weighted OR = 1.20
95% confidence limits: 0.62, 2.30

There are some limitations to the use of this method. First, it may not be appropriate to consider a confounder in a categorical state when its effect may be continuous. Secondly, small numbers in some strata preclude a valid application of the method. Thirdly, for practical reasons, it is often impossible to stratify by more than one variable simultaneously, for example age and sex, whereas such adjustment is normally considered desirable. Fourthly, a summary measure is not appropriate when there is evidence of *heterogeneity* across strata, i.e. the association between exposure and disease is substantially different between strata.

18.4c Standardisation

An alternative, but related, approach to stratification is standardisation. The principle of this has been outlined in Section 3.2. It permits comparison of two or more groups, by weighting for a potential confounder. Age is the most typical variable used for standardisation. Consider the data in Table 18.5, the crude data being the same as in Table 18.2. When stratified by age (Table 18.5b), it is clearly seen that incidence rises with increasing age, and thus a difference in age distribution between the exposed and unexposed groups should be considered to determine whether this might explain their observed difference in incidence. As the data are revealed (Table 18.5c), the differences between the exposure groups within each age band are less striking than the overall difference. A comparison of the age distributions (Table 18.5d) shows that the exposed individuals are weighted towards being older (44% aged over 65 compared with 31%). Standardisation is therefore the process of applying a single or standard set of weights (in this case age-specific proportions) to

Table 18.5. Investigation of possible confounding by standardisation

(a) *Crude analysis*

Exposure status (number)	Disease: Yes	Disease: No	Incidence (%)	RR
Exposed (500)	15	485	3	2.0
Non-exposed (2500)	38	2462	1.5	

(b) *Relationship to possible confounder of age*

Age stratum (number)	Disease: Yes	Disease: No	Incidence (%)
25–44 (1000)	6	994	0.6
45–64 (1000)	14	986	1.4
65+ (1000)	33	967	3.3

(c) *Analysis stratified by age*

	Exposed			Non-exposed			
	Disease			Disease			
Age stratum	Yes	No	Incidence (%)	Yes	No	Incidence (%)	RR
25–44	1	99	1.0	5	895	0.5	2.0
45–64	4	176	2.2	10	810	1.2	1.8
65+	10	210	4.5	23	757	2.9	1.6

(d) *Age distribution of exposed/non-exposed*

	Exposed		Non-exposed	
Age	Number	Proportion	Number	Proportion
25–44	100	0.20	900	0.36
45–64	180	0.36	820	0.33
65+	220	0.44	780	0.31
Total	500	1.00	2500	1.00

(e) *Apply non-exposed age weighting to incidence rates of exposure*

Age	Exposed incidence, I (%)	Non-exposed weight, W	Weighted incidence, IW (%)
25–44	1.0	0.36	0.36
45–64	2.2	0.33	0.73
65+	4.5	0.31	1.40
All	3.0		2.5(ΣIW)

(f) *Summary of analysis*

Ratio of crude incidence = 3/1.5,
= 2.
Ratio of age-adjusted incidence = 2.5/1.5,
= 1.65.

both groups. In Table 18.5e, the weights that are applied are those from the non-exposed group. When these weights are applied to the age-specific incidence figures in the exposed group, this results in an age-standardised incidence in the exposed group of 2.5%, slightly less than the observed incidence of 3%. This incidence is not a 'real' figure in any way, but describes the incidence that would have been obtained in the exposed population if their age distribution had been that of the non-exposed. The consequence of this is to lower the apparent risk ratio from 2 to 1.65, suggesting that exposure is still associated with an increased disease risk, but not as large as suggested by the crude data.

The standardisation can be done the other way round, i.e. applying the weights from the exposed to the non-exposed. When this is done using the **exposed** weights from Table 18.5d, an age standardised rate of 1.8 is obtained for the **unexposed** group. The result of this is to lower the risk ratio from 2 to 1.65 (3/1.8): an identical result. This process can be repeated for other potential confounding variables. Standardisation simultaneously across two or more variables is also possible, for example applying weights specific to age and sex. As with stratification, the limiting factor is the numbers in each level, and standardisation is not normally practicable unless the data sets are very large.

18.4d Multivariate techniques

The increasing availability of computer software has permitted the adjustment of associations simultaneously for a number of potential confounders in order to obtain an accurate estimate of their independent effects. Thus, in one routine the effects of (say) age, sex and socio-economic status as well as the main exposure under investigation can be quantified. All software programs will generate standard error values and/or confidence intervals around the estimates obtained. In addition, it is possible to look for interactions between a confounder and a main effect; for example, is the risk from an exposure different in different age groups?

Logistic regression is the method of choice for examining associations in case-control studies where the outcome is dichotomous: i.e. predicting whether a subject is a case or a control. The output is expressed in terms of *coefficients*, which are interpreted as giving the change in the $\log_e$ of the odds of being a case per unit change in the risk factor considered.

Example 18.vii

In a case-control study of diabetes the coefficient resulting from a logistic regression for age was +0.1225. This was calculated as being equivalent to an increased disease risk of $e^{(0.1225)} = 1.13$ for each increase in age of one year (within the age range studied).

Logistic regression is only one multivariate method of adjustment. Other methods used, particularly in cohort studies, are *Poisson regression* and *Cox's (proportional hazards) regression.* The former is used in studies where the outcome is the development of a rare incident event, and the latter is used in life-table analyses when the time to the development of an event is the outcome of interest. The details and constraints of all these approaches are outside the scope of this volume, and expert help will probably be required. One problem is that the availability of statistical software means that these analyses can be done with only one line of programming. The danger is that incorrect inferences will be made. Many would argue that it would have been better if these sophisticated statistical techniques were not so accessible to the novice!

19

Bias

19.1 Introduction

Bias is the largest source of anxiety for those undertaking an epidemiological study. It can be usefully defined as a systematic deviation from the truth, i.e. the study produces an incorrect conclusion, either about the existence of an association or about its strength. Bias can also exist in cross-sectional prevalence surveys, in which case bias results in a false estimate of the true occurrence of the disease. In studies comparing two populations, such as case-control or cohort studies, the results of bias can be in either direction: it may (i) falsely show the presence of an association or (ii) falsely show the absence of an effect. Clearly, bias is not an 'all or nothing' phenomenon and it will therefore typically falsely express the magnitude of an effect in either direction. Textbooks often oversimplify this.

Bias results from problems in the design or conduct of the study. The importance of this is that such problems cannot be overcome by analysis, in contrast to those due to confounding (see Chapter 18).

Example 19.i

In the same cohort study of coal miners as outlined in the previous chapter (Example 18.i), it was discovered that the coal-face workers also had twice the incidence of chronic bronchitis as the surface workers, based on sickness absence records. It was questioned whether this result was due to bias in so far as coal-face workers with respiratory symptoms might be more likely to have work absences. This possibility could not be assessed from the available data.

The message is that biases have to be considered at the design stage and during conduct of the study and not at the end of the study, although it is incumbent on investigators to consider the action of potential biases in explaining the study results.

19.1a Direction of bias

The direction of any potential bias should always be considered and, indeed, may be helpful. Consider the following example.

Example 19.ii

In a case-control study, where the exposure of interest was alcohol intake, an investigator took, as the source of controls, friends of the cases. At the end of the study, which showed that the cases had an increased exposure to alcohol, criticism was levelled on the basis of the possibility that the source of controls represents a potential bias, i.e. friends may be selectively more similar to their matched case in relation to alcohol consumption. The investigator pointed out that if this were true, it would have acted in a direction to make it more difficult to find a real effect. Thus, if this bias had not been present then the real difference would have been larger.

Such a situation is not uncommon and the point is a valid one: before levelling an accusation either at others or, indeed, yourself about the possibility of bias explaining an observed result, it is incumbent to consider the direction of a possible bias.

19.1b Non-directional misclassification

Another frequent situation relates to a study involving a comparison, where the same level of error occurs in both groups under study.

Example 19.iii

In a case-control study of osteoarthritis of the knee, one of the exposure variables under study was subject-recalled weight at age 25. The study suggested that those with arthritis had a higher average weight at this age, but the study was criticised about the likely errors in recalling weight many years previously. The investigators accepted this, but argued that despite being an inaccurate method of assessing weight, the errors would be similar in both groups. Hence the errors should act in the direction of making it more difficult to detect a real difference, and thus the fact that this study found a difference is likely to represent a real effect. In this case the investigators were making the assumption that the level of inaccuracy was similar in both groups.

A word of caution is required. The above two comments seem to suggest that study designs that are clearly imperfect are acceptable if they come to the 'right' answer. This is not ideal for three reasons:

(i) In both instances the investigator was lucky not to have missed a real association!

(ii) Studies, such as those in Examples 19.ii and 19.iii, should aim not only at assessing whether there is an association, but also to estimate its magnitude. An underestimate of the true effect is not an ideal solution.
(iii) Such studies are **inefficient**, in so far as to observe an effect it is necessary to study more subjects at greater cost.

Clearly it is preferable, if the choice has to be made, to slant the study design to bias against the hypothesis under test, but there are, as stated, important negative consequences.

19.2 Major sources of bias

It is useful to separate biases into two groups: those that result from the subjects studied and those that result from errors in information gathering. Biases that result from the sample of subjects studied mean that assumptions about randomness and representativeness cannot be assumed and the conclusions to be drawn are strictly limited. Biases in information gathering will have effects that depend on whether the data were gathered for the purposes of ascertainment of cases or of exposure. Each of the major sources of bias will be discussed, examples given and strategies suggested for their reduction.

Bias affects the validity of a study. It is important however to differentiate internal validity from external validity. Internal validity refers to whether the results obtained (e.g. in measuring prevalence or studying associations) reflect the 'truth' in the study population. The internal validity of the study can be compromised by selection bias (Section 19.3), information bias (Section 19.4) and confounding (Chapter 18). External validity refers to the extent to which the study results can be generalised to populations other than that included in the study. To a certain extent this is a matter of judgement taking into account relevant factors. A common misconception is that in order to generalise to a wider population, the study population must be 'representative' of that wider population.

Example 19.iv

Early observations of a relationship between smoking and lung cancer came from studying British doctors. Could such observations be generalised outside this population group? It seemed unlikely that the carcinogenic effects of smoking would be altered either by one's

occupation or nationality – therefore it seemed reasonable to extrapolate this result more generally – and subsequent studies worldwide have supported this initial observation.

Example 19.v

In a specialist pain clinic setting it was noted that patients showed high levels of anxiety and depression. This observation was confirmed when these subjects were compared (using a case-control design) with a sample of pain-free subjects attending a population screening programme. Could an observation on this sample be generalised to conclude that pain symptoms were associated with measured levels of anxiety and depression? This generalisation does not seem reasonable since firstly those subjects in the specialist pain clinic are likely to be highly 'selected' on the basis of pain symptoms (e.g. duration, severity, cause). Further it may be particularly those with high levels of anxiety that are more likely to be referred to such clinics. It would be necessary to conduct studies in other population groups before assuming that such a relationship held.

19.3 Selection bias

The first possibility for selection bias comes from the selection of subjects for *participation.*

Example 19.vi

In a case-control study of duodenal ulcer and smoking, an investigator selected the cases from patients attending the clinic of a specialist gastroenterologist. Clearly, such patients are not a random sample of the conceptual population of all patients with duodenal ulcer and are likely to have more severe disease. Thus, any risk factors derived from those cases may represent risk factors for severity rather than ulcer **susceptibility** per se. In addition, the referral practice for that gastroenterologist may be determined by a number of socio-demographic and related variables, each of which might contribute to suggesting a spurious relationship with a putative risk factor.

In this example, the problem is clearly laid out, which results from selecting cases of a disease from a specialist referral population. However, in practice, apart from a few conditions such as cancer, disease population registers do not exist and in the absence of screening at potentially exorbitant cost, selecting cases from a specialist centre is the only practicable approach, particularly for rare diseases. The potential bias can be minimised, however, by

attempting to cover all such specialist units that cover a stated target population and to exclude cases that do not come from the population.

Example 19.vii

In a case-control study of a relatively rare neurological disorder, an investigator defined the target population for the cases (case base) as individuals resident within a defined administrative area. She contacted all neurologists who would be likely to see cases arising from that population and asked them to notify all cases. Cases referred from outside the case base were excluded.

One particular type of selection bias is the so-called *incidence/prevalence* bias. This is the bias that results from restricting recruitment of cases to those who have active disease at an arbitrary point in time (prevalent cases). Such cases will tend to have more chronic disease in comparison to a random sample of all possible cases.

Example 19.viii

In a study of multiple sclerosis, an investigator recruited cases from those who had attended a specialist service within a three-month period. As a consequence of this recruitment method, individuals who had severe disease and had died, or conversely those with mild disease and did not currently attend for treatment, were excluded. The better approach would have been for the recruitment of all cases who had been diagnosed during an interval of time: an *incident cohort*. It might have been possible to do this retrospectively if a register had existed of all new cases attending in the period of interest.

19.3a Non-response bias

The most frequent source of concern to investigators, in any epidemiological study, is that of non-response bias. This applies equally to cross-sectional prevalence surveys, case-control studies and even prospective cohort investigations where the non-response results from both failure to participate at the recruitment stage and losses to follow-up. The principal point at issue is whether those who are studied are selectively different from those who choose not to participate, in respect of the variable under study, such as disease or exposure status. In free-living populations there will always be a proportion of individuals who will decline to complete a questionnaire, be interviewed or to attend for an examination. The likelihood of non-

response is greater in males and in those at the extremes of age, is related to socio-economic status and education and to the perceived benefits from participation. In case-control studies, cases are likely to be much more willing to take part than controls. In cross-sectional studies, those with a disease are more likely to respond than those completely well. However, some health surveys have found that study subjects with positive health behaviours (and consequently better health) were more likely to participate than those who did not. This emphasises that without any information on non-participants, it may be difficult to be certain about the direction of any non-response bias.

Example 19.ix

A random population sample was mailed and invited to attend a special screening clinic for diabetes. The response rate was 42% and over-represented those with a family history and those who had symptoms of thirst and urinary frequency. Thus, this sample could not yield an accurate estimate of disease prevalence.

There is no 'magic' percentage response when bias can be confidently excluded. Conversely, a low response does not indicate that bias is present. Indeed it may be easier to interpret a study with a low response rate but with an analysis of the potential non-response bias in comparison to a study with a high response rate but without any information about non-participants.

Example 19.x

In a prospective cohort study, looking at the development of an otherwise rare disease in two occupationally exposed groups (high and low exposure) 88% were successfully followed up for 10 years. The investigator was concerned that even this relatively small loss to follow-up could be problematic if those who had developed the disease had died or moved away, i.e. development of the disease was associated with an increased likelihood of loss.

Example 19.xi

In a molecular genetic study of a rare disease, an investigator invited both cases and controls to give a blood sample. The response rates were low at 35% in the cases and 27% in the controls. The investigator argued that given the lack of known phenotypic consequence of the genetic variants he was investigating, he could not conceive of a mechanism by which the likelihood of responding was related to the genetic factor under study.

Given the frequent potential for non-response bias, it is appropriate to consider strategies for its control. Clearly, the study should be designed and conducted to minimise the non-response rate, and strategies for this have been described in previous chapters. It is also important to attempt to ensure that non-response is not related to the variable under study, though this might create ethical difficulty.

Example 19.xii

A population-based screening survey for vertebral osteoporosis involved inviting a random elderly population sample to attend for X-ray screening. Pilot studies suggested that the response rate in the target population might be low at around 55%. The investigator was concerned that those who attended might be biased towards those who perceived themselves at higher risk owing to family history, etc. He therefore described the study loosely in terms of a research study of 'spinal health'.

The strategy that normally has to be used is a *post hoc* investigation to attempt to uncover the presence and strength of any non-response bias. The results of such an exercise have to be considered when interpreting the data found. There are a number of possible approaches.

19.3b Approaches to assessing non-response bias

Demographic approach

Data are normally available from the sampling frame on age and sex distribution of responders and non-responders. These should be evaluated in relation to the observed effects of age and sex on the main variable under study.

'Reluctant responders'

In most surveys, subjects who initially do not participate are invited a second or even a third time to take part. A comparison of the distribution of the main variable under study between first-time and subsequent responders may be instructive. Failure to find a difference would provide some support against response being related to outcome measured. In contrast, a finding that the prevalence of the disease decreased with time to response would give rise to concern that the prevalence amongst non-participants was even more extreme.

Alternative data

Depending on the sampling frame, it might be possible to obtain data about the non-responders from an alternative source that might give either direct or indirect data on any difference between responders and non-responders.

Example 19.xiii

In a cardiovascular risk-factor survey based on an occupational group, the investigator was allowed access to the occupational health records of both responders and non-responders. She was able to discover that the non-responders had fewer sickness absences and weighed slightly less at their pre-employment medical examination. These data suggested that there might be differences in other risk factors, including cigarette smoking and blood pressure.

Sample survey

One useful approach is to survey a small sample of the non-responders, possibly using a different approach, for example a home visit or telephone interview, to try to seek out differences. It is impossible to obtain a truly random sample of non-responders in this way because there will always be a small percentage of individuals who refuse any participation. Depending on the nature of the investigation, again either direct or indirect data relating to the major variable under study may be obtained. Thus, in a screening survey to attend for blood-pressure measurement, non-responders may be willing to provide information on body weight and family history by telephone.

With respect to the above approaches to assessing non-response bias, depending on the precise data available, it may be possible to 'adjust' the results at the analysis stage for the possible effects of this bias.

It is not unreasonable to attempt to recalculate what would have been the outcome of the study under a number of realistic assumptions of bias, for example if the non-responders were (say) half as likely to have the disease or symptoms as the responders. In that way, a single study can generate a number of estimates of the main result, each gathered under a different set of assumptions, with the reader left to consider the merits of each estimate.

19.3c Other forms of selection bias

There are many other possible ways by which selection bias may occur, but the above are by far the most important. There may, however, be special situations where surprising biases can arise.

Example 19.xiv

In a case-control study, a general practitioner used her computerised diagnostic index to retrieve information of patients who had consulted with migraine in the previous 24 months. By comparing their records with those of an age-matched random sample, she demonstrated that the condition was linked to the use of the oral contraceptive pill (OCP). Her results, however, could have been explained by surveillance bias in that women on the OCP were reviewed every six months and were asked about their general health. Hence this surveillance was likely to result in an increased detection of migraine in OCP users compared with non-users.

Example 19.xv

In a different case-control study, it was found that oestrogen use was related to the risk of uterine fibroids. The possibility needed to be addressed of protopathic bias, i.e. that early unrecognised disease led to an increased chance of exposure. Thus, in this case the possibility was considered that women with unusual uterine bleeding as a first sign of fibroids, but for whom the diagnosis was not made until a later date, had been started on hormones as a means of controlling the symptom. The cases that subsequently came to light were therefore more likely to have been 'exposed'.

19.4 Information bias

The origin of information bias may lie either within the observer or the subject.

19.4a Observer bias

Comparative studies, involving interviews or clinical examination, rely on the impartiality of the observer. Structured interviews, using standardised wording of introductions to questions and clearly standardised examination techniques, may help. However, if the observer is aware of the disease (or exposure) status of the participant, it may be impossible to rule out a 'nudging' towards the hypothesis being tested. Ideally, the observer should be unaware (be blinded) as to the status of the subject although, particularly in case-control studies, this is frequently impossible to maintain. Timing the duration of the interviews may be useful.

Example 19.xvi

In a case-control study, which suggested a possible link between febrile illness and subsequent early pregnancy loss, it was found that the interviews with the cases were taking about 10 minutes longer than those of the normal pregnant controls. Whilst some of this increase may be ascribed to greater care in introducing the questions to the case mothers, the study co-ordinator was concerned that the interviewer might have probed more closely for history of fever in the cases.

Other problems may arise if, for logistic reasons, the cases are interviewed by one observer (say in hospital), and the controls interviewed by another (at home). A not uncommon source of bias occurs when there is a delay between interviewing cases and interviewing controls. Thus, the cases could (say) be interviewed in winter and the controls interviewed in the summer. These differences may be important if a seasonally affected exposure such as diet is being investigated. Alternatively, in a case-control study, subjects with a disease may have thought at some length about past relevant exposures, in comparison to controls who may have never previously considered these.

19.4b Subject bias

This most typically takes the form of recall bias. For example in a cohort study subjects who know they are exposed to a factor of interest may be more likely to recall disease symptoms than those who are not exposed. It is human nature to attempt to explain the occurrence of an unpleasant event like an illness by a simple exposure. Thus, in clinical practice, many patients invoke a stressful (life) event as the cause of their symptoms, even though the relationship in time is coincidental. Attempts can be made to reduce this by:

(i) making the subjects blind to the hypothesis under test (not always practical);
(ii) using exposures that can be corroborated from other sources such as contemporary records or interviews with household contacts;
(iii) minimising the period of recall to maximise accuracy, although clearly the study may have to involve a longer period of recall;
(iv) selecting a comparison group for whom the potential for recall bias is likely to be similar.

Example 19.xvii

In a case-control study investigating the link between running in early adult life and the later development of osteoarthritis of the hip, the investigators relied on recalled data on the amount of running undertaken by cases and controls. Blinding the subjects to the hypothesis would have proved unrealistic and no corroborating contemporary data were likely to be available. The investigators therefore altered their design to take, as controls, individuals who had hip pain, but who on X-ray were shown to be free from osteoarthritis. The investigators reasoned that both cases and controls might be subject to the same level of recall bias in recalling past activity. They also assumed, perhaps incorrectly, that there was no relation between early adult running and the later development of non-osteoarthritic hip pain.

There is no reason to believe that recall bias results in the false reporting of events, either exposure or disease that had not occurred. The problem typically lies in more accurate recall by (say) cases rather than controls. Recall may be heightened and may be more inaccurate temporally, with an artificial collapsing of the time interval.

Example 19.xviii

A case-control study of inflammatory bowel disease suggested a greater exposure to stressful life events in the six months preceding disease onset in the cases. However, a sample survey of household contacts, in both cases and controls, suggested that although there was no bias in recall of whether or not an event had occurred, the cases (but not the controls) were more likely to have errors in the dating of the event towards a closer time interval to disease onset.

Example 19.xix

In a case-control study of adults with chronic widespread body pain, subjects were asked about events in childhood including hospitalisation and operation. Analysis revealed strong associations between case status and reporting these events. However, because of concern about subject bias (in particular, differential recall between cases and controls), information on these events was also collected from the medical records of cases and controls. Re-analysis of the data revealed no association between the record-derived exposure and case status. Further analysis revealed that cases always self-reported hospitalisations and operations which were documented, while controls 'forgot' about half of all such events.

19.5 Is an unbiased study ever possible?

The above comments are likely to have led to the pessimistic conclusion that it is either impossible or impracticable to perform an unbiased study. Indeed, in epidemiological classes, a favourite exercise is to take a published paper, preferably from a renowned source, and suggest a large number of reasons why the conclusions may be invalid owing to bias in design and conduct of the study! The message from this chapter is more practical and hopeful. It is possible, in both the design and the conduct of a study, within the available resources, to attempt to minimise the potential for bias. The skill of the epidemiologist lies not in conducting the perfect study, but in documenting and assessing the likely impact of its imperfections! However, perhaps the major task at the end of the study is for the investigator to be self-critical and consider carefully what potential for bias remained and how far the data collected permit an estimate of the likely effect of any bias.

Part VII

Other practical issues

20

Ethical issues in epidemiology

20.1 Introduction

The procedures involved in undertaking most epidemiological investigations do not produce either hazard or discomfort to the individual studied. The most commonly collected data arise either from interview or from other written sources. Occasionally a limited physical examination is undertaken. Much less often, there is a requirement to take samples of biological fluid such as blood and urine or to undergo simple investigations such as electrocardiography or plain radiography, but even such investigations are associated with trivial risk. It is therefore reasonable to question whether there are ethical concerns in the design and conduct of epidemiological investigations.

Ethical issues do arise in epidemiological studies for a number of reasons. First, the main focus of most studies is often normal populations and the need to obtain a high response rate from individuals who, unlike a laboratory animal, can refuse to participate without explanation. The study population does not normally 'belong' to the investigator, and the latter frequently requires data that could be considered both confidential and sensitive. Secondly, many epidemiological studies, often as their primary aim, uncover previously unidentified and possibly asymptomatic disease, which presents the problem of whether intervention, not part of the study design, is required. Finally, the method of investigation itself may not be without problems.

This chapter considers the major ethical problems that may arise in epidemiological research and discusses approaches to minimise their impact.

20.2 Ethical approval

In most countries, investigators need to seek approval from either their institutional or another local authority that their proposed study is ethically

sound. The epidemiologist will need to complete the same procedures as a clinical investigator who might, for example, wish to perform a clinical trial with a new drug or to undertake a study involving an invasive procedure such as endoscopy. Ethical approval is rapidly becoming a requirement before grant funding will be awarded, access to patient or population groups permitted or results published. Thus, there are strong practical imperatives for the epidemiologist to seek such approval, in addition to legal and ethical obligations.

20.2a Scientific validity and ethical research

Increasingly, ethical committees include a methodologist (who may be a statistician) to consider the validity of the proposed study design. The rationale for considering the latter is the premise that a research study that is incapable of answering the question it poses is unethical because it wastes resources and investigator time. Further, even in epidemiological enquiries with simple procedures such as a questionnaire, it is unethical for a population to be recruited for a study that is unable to provide an answer. A well-designed study will avoid this problem. In practice the most frequently occurring sins are:

(i) insufficient sample size;
(ii) selection bias;
(iii) information bias, for example using an invalid method of enquiry.

In addition, it could reasonably be considered unethical to undertake a study on too large a study population if the question can be answered with fewer subjects, for all the same reasons stated above.

20.3 Ethical constraints in maximising response rate

One of the greatest sources of conflict is the scientific need to minimise non-response bias, conflicting with the ethical requisite not to bother unwilling subjects unnecessarily. There is certainly much that can be achieved to maximise response rates without problem. It is not unreasonable to send one or even two follow-up mailings by post to those who have not replied to a first mailing, provided that the accompanying letter is not critical of the previous non-response. A useful tactic is to encourage those who do not wish to participate to make a positive affirmation to that effect (see the example letter

> *'…If you prefer not to participate in the study I would be grateful if you could return the blank questionnaire in the envelope provided and we would trouble you no further…'*

Figure 20.1 Suggested wording for identifying 'positive' non-responders.

extract in Fig. 20.1), which would then distinguish them from those non-responding from apathy or forgetfulness, who might be approached again without anxiety.

A decision should be made, preferably at the outset, as to the strategy to be adopted in the face of an unacceptable level of non-response, either to a postal survey or attendance for interview/examination at a survey centre. There are several possible approaches: further mailings, telephone contact or home visit. The first could take the form of a shortened version of the original questionnaire (for example, a single sheet), which includes the most important questions and/or would allow a more detailed assessment of non-response. Telephone contact although cost-efficient does involve a considerable amount of staff time. However, a diplomatic approach asking about the reasons for non-response may frequently result in participation. Home visits might require sending the researcher, without a prior appointment, to the subject's home to obtain the required data. This would be unacceptable if a definite refusal had already been communicated by mail (see above). It may, however, be the only option for those groups with high frequencies of illiteracy and low telephone coverage. A prior letter giving notice of the visit is important. The researcher should have a suitable form of identification. In addition, the author has found it useful in the past to warn the local police in some areas, particularly in studies involving the elderly, that home visits are being undertaken. The elderly, in particular, are properly advised not to admit strangers to their home and the advice to contact the police for approval can be helpful.

20.3a Opting out or opting in

Studies that require face-to-face contact with the subject normally require that an appointment time is agreed. Alternative approaches include (i) sending a specific appointment time and asking the subject to telephone or mail back only if they wish to *opt out*, and (ii) inviting the subject to telephone

or mail back only if they wish to *opt in*. The former is more efficient and raises response rates, although it has been criticised for giving the implication, particularly to the less well educated, that attendance is necessary regardless of problems with transport, work and other constraints. A carefully worded letter should avoid most of these pitfalls.

20.4 Confidentiality and data protection

This is a fairly complex and increasingly important issue that impinges on epidemiological studies. There are three areas of interest:

(i) legal requirements regarding data registers and storage;
(ii) confidentiality of data provided to the study direct by the subject;
(iii) access to data held on the study population by other sources and, in particular, their medical records.

20.4a Legal requirements

Most societies have a legal framework regarding the acquisition and storage of data from the population. There are, however, a number of misconceptions about the legal position. In the United Kingdom, under the Data Protection Act (1998), the requirement is for the investigator to register with the appropriate agency that data are being held. The act provides a set of 'principles', which must be adhered to – but they do not provide details of what is and what is not acceptable. Interpretation of the act differs and could only be judged on a case-by-case basis. Instead, researchers are best advised to consult and follow guidelines issued by relevant bodies such as the Medical Research Council. Normally, clinical researchers must ensure that they have consent to hold or use personal information, and in most clinical research this is practicable. When consent is impracticable (and this would have to be justified), it is suggested that confidential information can be disclosed without consent if (a) it is justified by the importance of the study (b) there is no intention to contact individuals (c) there are no practicable alternatives of equal effectiveness and/or (d) the infringement of confidentiality is kept to a minimum. Although the above discussion is, in some respects, specific to the United Kingdom, a more general point is that, when conducting studies, researchers should be aware of the legal framework and of any recommendations of professional bodies in the country concerned.

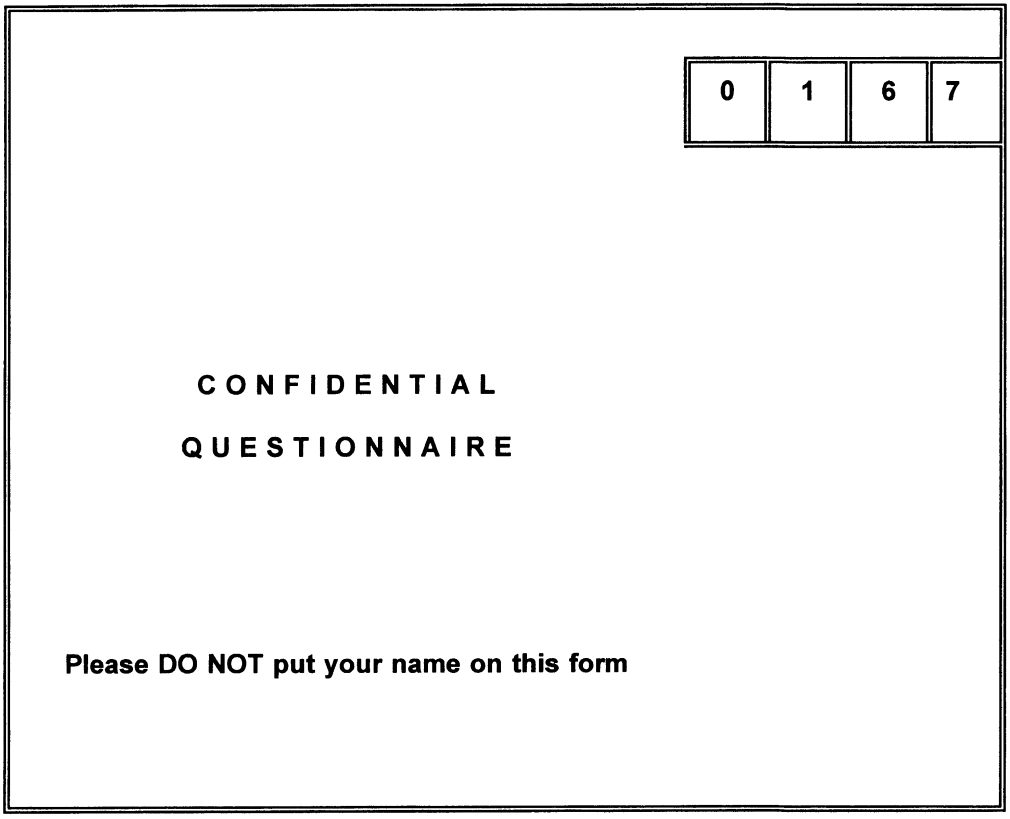

Figure 20.2

20.4b Confidentiality of data collected for the study

There are clear legal and ethical responsibilities on researchers for maintaining confidentiality. The study population should be confident that data provided by them will be kept confidential. There are a number of simple measures that can enhance this confidence.

Figure 20.2 shows a suitable layout for the front sheet for a self-completed questionnaire. It incorporates a study number rather than the subject's name, and, should the form be lost or misplaced, no identification is possible without access to the register of names and study numbers. The simple act of having no data on the front page, although relatively wasteful of space, also engenders a feeling of confidence. A suitable covering letter for such a

'...This information will only be used for the purposes of the research mentioned above. The data gathered will be used for looking at groups of individuals and the results presented such that the information from a single individual cannot be identified separately. Access to the information you provide will be limited to members of the research team...'

Figure 20.3 Extract from a letter covering the issue of confidentiality.

questionnaire should also address the issue of confidentiality, as in the example shown in Fig. 20.3.

The letter emphasises that the research is interested in looking at groups rather than single individuals and that it will not be possible to identify individuals from the way that the results will be presented. Subjects are concerned that data given for research may be used for other purposes and given to other bodies and individuals. The letter should reassure the subject that only those involved in the particular project will have access to the original data.

Ethical committees frequently express concern about the inclusion in questionnaires of data concerning sensitive topics like sexual practices and use of addictive substances. The strategies outlined above should provide sufficient reassurance, but it is not unusual to have questionnaires returned with the study number obliterated! A useful additional tactic is to put the study number on the envelope or elsewhere on the questionnaire to permit identification of the responder.

20.4c Access to confidential data from a third party

In epidemiological studies, it is frequently necessary to obtain data from an individual's medical records, perhaps to verify diagnostic or treatment information given in answer to a questionnaire or to uncover new disease episodes in a prospective study. It is the responsibility of the third party to protect the individual's privacy. If a physician, either in hospital or in general practice, is satisfied by the credentials of the research team then access either to the records themselves, or to specific items of information,

may be permitted without permission of the patient. Thus, in a survey of patients registered with a single general practitioner, the researcher may wish to study a random sample of the records of the non-responders, to determine some information about their health status. In such circumstances, access may be allowed with the normal conditions that the data gathered are to be used for the purposes of aggregation, no individual's medical history will be identifiable from the published or presented results and the research team do not intend to contact the individuals concerned or reveal the findings to them.

An alternative and preferable strategy is to request the permission of the subject in advance, and this is often essential if the subject's recruitment to the study was previously unknown to the physician holding his or her medical records. In such circumstances, a written consent such as that shown in Fig. 20.4 is appropriate. Note that the consent form provides the name of specific doctors who can be approached. It is the author's experience that this usually results in permission being granted. In addition, the obtaining of such permission before requesting access encourages the third-party physician to participate.

Some morbidity data on individuals are obtainable from routine data-collection sources, though separate ethical approval is normally required. Thus, in the United Kingdom, it is possible to 'tag' individual names on the National Health Service Central Register for the purposes of being informed about subsequent death or cancer registration. This register is held by the Office for National Statistics. The provision of such record linkage is of substantial benefit in undertaking very large long-term prospective studies, and it would be a major additional burden, and often impractical, if individual consent were required for such data to be made available. Access to this system requires formal ethical approval from the Office for National Statistics.

20.5 Detection of previously unrecognised disease

Many epidemiological studies have as their primary aim the ascertainment of morbidity, whether previously recognised by the subject or not. The action to be taken depends on the nature of the problem, its natural history and the availability of any effective intervention.

..... It may be valuable for us to obtain information about your past medical history from your medical records. We would be grateful if you could give us permission to contact your doctor to provide us with such information that is relevant. Please could you complete the tear-off slip below and send it back in the envelope provided.

Yours sincerely

DR JOHN SMITH for Study Team

Mrs B Jones
21 Pine Road
Newtown

PLEASE CIRCLE

I agree to you contacting the doctor(s) below to obtain information relevant to your research. **YES NO**

General practitioner
Name _____

Address _____
_____

Hospital doctor
Name _____

Address _____
_____

Signed ..... **Date** .....

Figure 20.4 Extract from a letter seeking consent to obtain medical information.

20.5a Ascertainment of presymptomatic disorders not requiring intervention

There are many instances where such presymptomatic recognition has no practical consequences in so far as no intervention at that stage would alter the prognosis. It might be argued that rather than create anxiety it is perhaps reasonable not to inform the subject. The best approach is to set out in the consent form the possible results of the survey procedure for an individual and that any relevant information about the condition could be conveyed to

Dear Dr Brown

Re: Albert Pink, age 72, 12 Retirement Cottages

Your patient, Mr Pink, agreed to participate in a research study into the occurrence of hip osteoarthritis in retired council gardeners. His X-ray shows evidence of osteoarthritis particularly affecting the left hip. Mr Pink answered negatively to the presence of hip pain. I am not aware that there is either any treatment or alteration to lifestyle that would be of benefit in such cases. Participants in the study were advised that they would not receive any individual results and only those with hip pain, not already under treatment were advised to visit their general practitioner. If you require any further information.....

Figure 20.5 Example of a letter to a general practitioner with the results of a study.

'...We found his blood pressure to be 210/120mmHg, which was sustained despite a 10 minute rest. We advised Mr Scarlet that he should seek medical attention for this and to make an appointment to see you within the next week. If you require any further information you can telephone me on...'

Figure 20.6 Extract from a letter for a positive result from screening.

the subject's general practitioner. A typical letter is shown in Fig. 20.5. Such a step is useful in giving the practitioner a baseline for future developments. It should also be made clear what action, if any, is felt appropriate and what information has been conveyed to the subject.

20.5b Ascertainment of disorders requiring intervention

Some previously unrecognised disorders, for example hypertension, do require intervention, and any population-screening survey would need to incorporate in its protocol the action to be taken after case detection even if the intervention is to be handed over to others. An extract from a suitable letter to be used in such circumstances is shown in Fig. 20.6. The letter

should also explain (as in Fig. 20.5) how the subject was recruited and the purpose of the study.

The letter gives the relevant result and provides a contact should further information be required. In this example, for hypertension, it is reasonable not to give further guidance to the physician about specifics of future action; indeed, it might be resented. By contrast, if the study involves a specialised technique with results requiring interpretation by experts, there may be concern that few general practitioners would feel sufficiently competent to interpret the findings and to plan the appropriate interventions. In such instances, the researcher should provide more guidance in the letter and preferably offer the services of a professional colleague from whom specific advice could be sought, or to whom the patient could be referred.

20.5c Ascertainment of unrelated disorders

It is not infrequent that, during the course of a survey, other health problems become apparent that are not part of the study and hence are unanticipated, but are perceived by the researcher, perhaps a junior doctor or nurse, as requiring some intervention. Two recent examples include a leukaemia discovered on separating white blood cells for immunogenetic analysis, and diabetes that came to light during an interview with a research nurse for other data. The ethical position is clear. First, managing or even advising about management is not part of the research and should be avoided. Secondly, the subject needs to be advised that there may be a problem and that they should seek medical attention. Simultaneously, the subject's general practitioner should be contacted, either by telephone or letter depending on the circumstances, and given such information as is available. Any such episode should be notified to the project leader and any action taken documented. Obviously, medical omniscience is not expected, but the possession of knowledge that is not followed by appropriate action is culpably negligent.

20.5d Distress amongst participating subjects

Other than the unanticipated ascertainment of disorders, most questionnaire and interview-based studies will have few other major ethical issues to consider. However, in some instances, completing a questionnaire or participating in an interview may cause distress to the subject. An example from the authors' recent experience involved conducting a study amongst war vet-

erans. The study involved a self-complete postal questionnaire which asked both about their current health and about their experiences during deployment. Recalling particularly traumatic events could potentially cause distress. In such circumstances a counsellor was available to the study participants. Alternatively if conducting home interviews, the interviewer may need to be qualified, or receive special training to deal with circumstances in which the subject is distressed by the nature of questions.

20.6 Consent

Consent is necessary for all research investigations on humans. Frequently, in epidemiological studies, written consent is not obtained and participation assumes **implied** consent, i.e. the subject participated only because he or she was willing to do so. There is no clear distinction between surveys that can rely totally on implied consent and those that require formal written consent. There are certain circumstances where obtaining written consent is probably mandatory. These include (i) obtaining information from a third party such as medical or occupational records, (ii) where an investigation undertaken as part of the survey may uncover previously unrecognised disease that requires further intervention, and (iii) where the survey procedure involves any risk. By contrast, it may be assumed that a postal questionnaire will be answered and returned only if the subject consents to do so. Similarly, the subject would attend for a screening examination only if that was his or her wish. The problem comes if such participation is not based on accurate or complete information. It is normally the wording of the letter accompanying the questionnaire or screening invitation that provides the necessary information. If the first contact is by telephone or a home visit, without prior warning, then such information would need to be given either verbally or, preferably, in written form. Ethical committees will look very carefully at letters to ensure that the following points are covered:

(i) a clear statement that the study is for research, and that participation is voluntary;
(ii) the exact nature of the study is given, including its purpose and the procedures to be used;
(iii) a clear statement indicating that non-participation will not jeopardise the medical care they are receiving;

(a) *Dear Mrs Green*

I am writing to you following your recent discharge from hospital with an ulcer. Dr Grey has given me permission to contact you about a research project we are undertaking on whether the diet alters the chances of a recurrence. We aim to look in detail at the diets of 150 patients such as yourself and investigate whether there are certain features in the diet that might affect the chances of having another ulcer.

Your participation in the study would require you to carefully weigh and record everything you eat over a period of seven days and to weigh the leftovers on your plate. We would obviously give you full instructions and a set of scales. I do hope you will agree to help in the research which we hope will help in understanding why some patients develop further ulcers.

If you would be interested in participating in this study please complete the form below and our dietician, Mrs Blue, will contact you by telephone to fix up a home visit when she will explain all the details and answer any questions you have.

If you would prefer not to take part in this research, that is no problem and Dr Grey will continue to see you in the clinic as arranged.

Yours sincerely

DR WHITE

Name.....
Address.....
.....

I agree/do not agree (delete where not applicable) to a visit from the research dietician in connection with the study on diet and ulcers.

Signature ..... *Date* .....

(b) *Mrs Green*

I am writing following your recent discharge from hospital with an ulcer. Dr Grey and myself are very keen to do all we can to stop our patients having a recurrence and we believe that diet may be important.

I am writing to ask if you would be willing to provide us with details of your diet which we can then analyse. Our dietician, Mrs Blue, will telephone you next week and will make an arrangement to visit you at home and explain exactly what's involved and answer any questions you have.

Yours sincerely

DR BLACK

Figure 20.7 Examples of (a) a suitable and (b) an unsuitable letter of invitation.

(iv) indicating why they were chosen (e.g. at random);
(v) even if the subject agrees to participate they can, thereafter, withdraw at any time.

Examples of letters that are ethically satisfactory and unsatisfactory are shown in Fig. 20.7. The second letter is unsatisfactory for a number of reasons. It implies that the study is part of the subject's continuing care and does not explicitly give the option of not participating. The research basis is obscured, and the subject may feel it is a necessary part of her management. It is also less than candid about the tasks involved, which are not trivial. There is no formal consent form. It will also be more difficult to say 'no' once the dietitian has telephoned.

21

The costs of epidemiological studies

21.1 Costs versus design

Study design is frequently undertaken without consideration of the cost implications of different available options. Indeed, in many standard texts and taught courses, the student is encouraged to seek the most valid approach to answering a question independently of any cost consequences. Similarly, in refereeing articles submitted to a journal or criticising published studies it is tempting to suggest that alternative and, frequently, considerably more expensive strategies should have been used. Although in theory the study design influences the costs, in practice the resources available will often constrain the methodological choices (Fig. 21.1). Epidemiology should be both the science of the theoretical and the art of the practical – the skill of the epidemiologist lies not in planning the perfect study, but rather in planning the most valid study given practical limitations of cost and time. Two points, however, are clear. First, no study, however cheap, can be of value if the study design is incapable of producing an answer to the question posed (e.g. a valid estimate of a particular effect). Secondly, and conversely, it is possible that the potential benefit to emerge from a study may be substantially less than the costs of undertaking a methodologically sound study. This is particularly true for relatively rare diseases and could lead to a reassessment of the rationale for undertaking the study.

Example 21.i

Researchers were intrigued by the reported negative association between schizophrenia and rheumatoid arthritis. They postulated that the susceptibility gene(s) for the one disorder might be protective against the other and vice versa. They planned to test this by comparing, in a case-control design, the frequency of schizophrenia in first-degree relatives, among rheu-

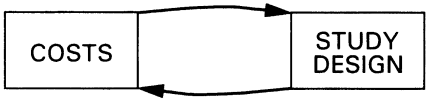

Figure 21.1 Study costs and study design: chicken or egg?

matoid arthritis and control subjects. Some rough sample-size calculations suggested that to prove this negative effect would require an enormous study necessitating substantial interview time, with the study estimated to cost some £250000. The idea was dropped because it was felt that no funding body would consider the question of such value to justify that expenditure. Given the nature of the diseases being studied, a much cheaper study based on postal questionnaires would have been inappropriate.

21.2 Costing an epidemiological study

All epidemiological studies incur costs and it is essential to be able to estimate these in advance, both for obtaining funding from grant-giving bodies and to ensure that sufficient resources are available to complete the task. Population surveys will vary considerably in their cost, but as there are many common elements it is perhaps useful to take an example of a simple survey to provide a model for how costs can be calculated.

The first step is to construct a flowchart of the methodological approach to be used. In this example (Fig. 21.2), the plan is to derive the study population from a primary-care register (stage I), with the objective of assessing the prevalence of respiratory symptoms (cough and wheeze) in middle-aged adults. It is decided to undertake a postal survey in the study sample, with a second mailing to those non-responding after four weeks. A telephone interview will be undertaken of a random sample of non-responders to determine whether they are selectively different from the responders in respect of some key variables such as symptom presence and smoking history. It is also intended to undertake home follow-up of a sample of the questionnaire responders to validate their answers against a more detailed probing interview. In estimating the costs, a pilot study had already been undertaken to determine the likely response to the first mailing (stage II). The advantage of setting out the study design as in the figure is that it permits a straightforward approach to estimating the costs of the different stages. Thus, Table 21.1 shows the resources necessary for each of the stages based on, for

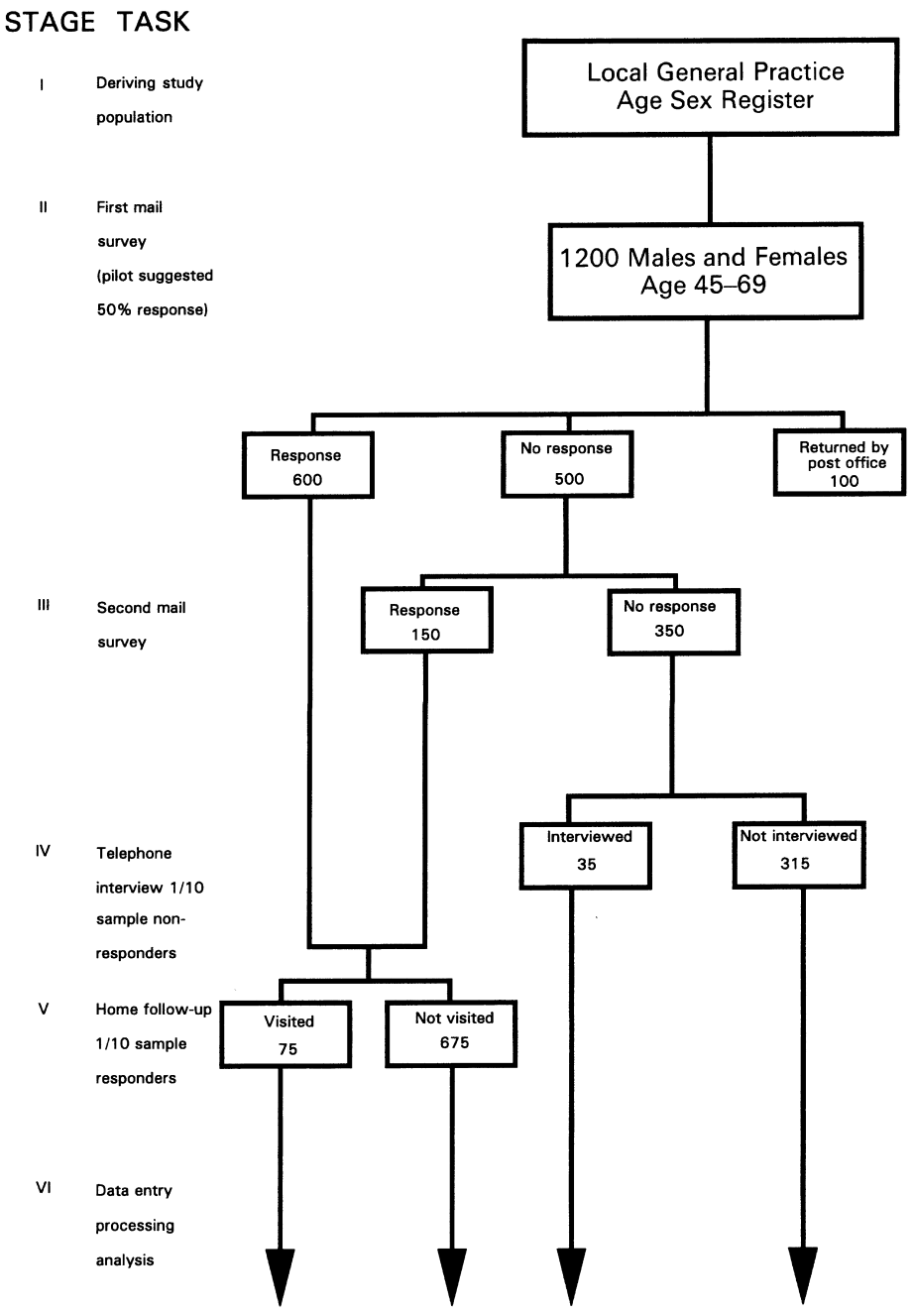

Figure 21.2 Flowchart for a population survey as a basis for estimating costs.

Table 21.1. Hypothetical budget based on study outlined in Figure 21.2

Stage	Staff	Expenses
I	Clerical: 3 weeks FTE[a]	Floppy discs Local travel to general practice
II	Clerical: 6 weeks FTE	Printing questionnaires Envelopes/return envelopes Postage/return postage
III	Clerical: 4 weeks FTE	As for stage II (pro-rata)
IV	Interviewer: 2 weeks FTE	Telephone
V	Interviewer: 12 weeks FTE	Local travel Telephone
VI	Clerical: 8 weeks FTE Data processor: 8 weeks FTE	Computer sundries

Note:

[a] Full-time equivalent, e.g. may need one person half-time over six weeks.

example, sending out 1200 questionnaires in stage II and 500 in stage III. Whereas postage, printing and related costs can be fairly accurately predicted, estimating the personnel time to complete a particular stage is more hazardous and is best based on experience. Some examples of person–time needed to complete some of the tasks associated with large mail surveys are shown in Table 21.2. They should be considered as rough estimates and will obviously vary according to the expertise of the individuals concerned and the computing and other facilities available. One major additional time is if the survey forms are not precoded, i.e. if the answers have to be translated into a computer-readable format. Obviously, the complexity of the coding schedule will have a major influence on the time, which might range from 2 to 15 minutes per form. The times listed should also be considered as the calculated number of hours to complete a specific task rather than the true numbers of employee–hours; the reason is obvious in that most of the tasks are fairly repetitive and it will be necessary to have a number of breaks during the work.

One feature of the resources listed in Table 21.1 is that three personnel tasks have been identified: (i) a clerical one of organising the mailings; (ii)

Table 21.2. Estimates for person–hours needed in relation to specific aspects of mail surveys

Task	Time per 100 subjects
Entering names and addresses from written source onto a computer	3 hours
Generation and printing of letters from mail merge	5 hours[a] 15 min[b]
Preparation of survey forms for mailing, including checking of study numbers, names and addresses on envelope labels, etc.	1 hour 30 min
Recording replies received to survey on computer register	1 hour
Checking for completion errors in typical four-page precoded questionnaire	1 hour 30 min
Entry of coded data from typical four-page precoded questionnaire	5 hours

Notes:

[a] To set up for initial run.

[b] Actual additional time for each 100 letters.

an interviewer-based one (such as a research nurse) to undertake both the telephone interviews and the home visits; and (iii) a data-processing one of coding, entering, checking and undertaking simple analyses of the collected data. This latter task may be split, as shown in the example, with the data entry being considered a clerical task and the other tasks being allocated to a data processor. In a large research unit it may be possible to identify three separate individuals who can undertake these tasks, and who can be freed from other duties to undertake the work. More often, personnel have to be recruited specifically to undertake a survey. In practice, it makes little sense to have three individuals employed on a single short-term survey, and one would opt for a single person who could undertake the complete task. The advantage of the latter approach is that there is a greater sense of both 'ownership' and of the need to maintain high standards of data collection.

Technological advances mean that some traditional tasks associated with large-scale epidemiological studies can now be accomplished by alternative (potentially time-saving) methods. For example, the sampling frame of a study may be able to be downloaded directly on to the computer, or

Table 21.3. Possibilities for cost reduction

Design	Sample size should not be in excess of statistical requirements
	Use samples of study population where relevant for sub-studies
Personnel	Paramedical (metrologists) rather than physicians
	Data entry undertaken by commercial bureau or using automated scanning or similar system
Expenses	Questionnaire size/weight
	Bulk mailing
	Reply-paid envelopes

completed questionnaires may be scanned in to the computer (instead of manual data entry for both tasks). While extremely useful, experience has shown that the saving in time is usually not as great as first envisaged. These new technologies still need person–hours devoted to them, albeit requiring different skills.

21.3 Possibilities for cost containment

There are a number of possibilities for reducing the costs of an epidemiological survey, many of which are fairly obvious (Table 21.3). If a final sample size of 400 is sufficient to answer the questions posed, it is a waste of resources to study more. It is often not necessary to investigate the entire study population in relation to one particular component of the study, and a random sub-sample will suffice. Thus, in the example in Fig. 21.2, it was decided that the answers to the questionnaire required validation by home interview, but that this would only be necessary for a 1 in 10 random sample. The sampling fraction to be used in these circumstances is based on the results of pilot studies or other work suggesting that robust (in a statistical sense) estimates of the effect or variable being measured can be obtained from that proportion of the total study population.

Another approach to cost containment by sampling is related to the analysis of data collected in large prospective cohort studies. As an example, an investigator may wish to take detailed dietary histories in a baseline survey as a prelude to evaluating prospectively the effect of diet on the subsequent risk of disease. The task of translating dietary data into actual nutritional

intakes is large. Thus, the investigator stores the dietary data unanalysed, and, using the nested case-control approach, only considers the dietary data of those who develop the particular disease with that of a random control group from within the cohort.

21.3a Personnel costs

This is the major component of research costs, as Table 21.1 shows. Although, traditionally, surveys requiring clinical evaluation have used a physician, a paramedically trained person such as a nurse or physiotherapist can frequently be trained to be equally or even more reliable and accurate in collecting clinical data, and at lower cost! The disadvantage is the perceived credibility of the data to those external to the survey. There may well be a willingness to accept that a blood-pressure measurement can be undertaken by a non-physician, but a belief may persist that a description of retinal vascular abnormality in a diabetic population, or the determination of the presence or absence of ankle jerks, requires a physician. In such circumstances, if resources do not permit the employment of a physician, it will be necessary to train the survey worker, to show in pilot studies conducted blind that there is good agreement between a physician and the proposed worker, and that this agreement is maintained during the study.

Data entry

One major staff item in any survey is the cost of data entry. Modern technology, with optical scanning equipment as discussed earlier, can automate this task such that survey forms can be read directly by a computer without manual input. The equipment itself is expensive, as is the printing of the survey forms, though for large studies with compliant subjects, such as in the Nurses Health Study in the USA, then such an investment can be efficient in terms of both cost and time. Alternatively, a commercial bureau can be employed to input data both quickly and accurately, with the investigator paying only for the inputting time. This is often preferable to employing an individual in large studies to input the data, both because it is impossible to do this non-stop for more than a short period, thereby resulting in paying for non-inputting time, and also because the job itself is inherently boring and leads to lowering of morale.

21.3b Non-personnel costs

Available approaches include using cheap postal opportunities including bulk mailing rates when available. Printing rather than photocopying survey forms may be cheaper because the cost is proportionally lower as the quantities increase. Others are less obvious. The author has recently conducted a mail survey of some 6000 individuals; the weight of the mail package was just above one charge band, adding £1000 to the postage cost of the study. In retrospect, it would have been simple to redesign the survey forms to fit on fewer pages! If a follow-up mailing is planned, any strategy that will increase the response to the first mailing will reduce costs. Some have argued in favour of using a special mail service that requires the recipient to sign for the arrival of the mailing, such as the Recorded Delivery system in the UK. Though expensive, this has the advantage of accurately identifying those who have left their registered address, hence reducing unnecessary follow-up mailings. It may, however, be detrimental to participation rates, often requiring a special journey to the Post Office to collect the letter.

21.4 Wasting resources

With hindsight, there is the inevitable reaction at the end of a study that, with a better design or more thought, it would have been possible to have used resources more efficiently. The greatest problem is an unrealistic time scale, particularly during the setting-up stage. Obtaining permissions, securing population access and finalising study instruments, such as questionnaires, always seem to take longer than envisaged. The resource problem comes when new staff are specially recruited and have to mark time for some weeks until the survey is ready to run.

21.4a Questionnaire size

Questionnaires printed in bulk, before adequate piloting has revealed the unexpected deficiencies in design, is a misdemeanour that all have committed. In addition, it should be remembered that any changes to the questionnaire after the pilot study, should themselves be tested. The authors know of no epidemiological survey, involving a detailed questionnaire, that used all the data for analysis. It is often discovered retrospectively either that many questions are ill-understood or that the frequency of a 'positive' is so rare as

to preclude useful interpretation. The wastage here is not so much in the fact that a shorter form could have been used, but in the time spent in coding and entering data. Each additional question adds to the cost, though few studies are sufficiently rigorous to exclude such items at the onset.

21.4b Statistical analysis

A well-planned statistical analysis may reduce the almost inevitable torrent of printout that is a feature of most surveys. Statistical time, processing time and computer printout are frequently expended with little thought as to the benefits of each particular stage of the analysis. This is particularly true as access to statistical packages increases and computing time used decreases. Many large computer runs are repeated after modifying the data or excluding some individuals, perhaps those with incomplete information. Prior planning, however, could reduce this wasted effort.

Index

Numbers in italics indicate *tables* or *figures*.