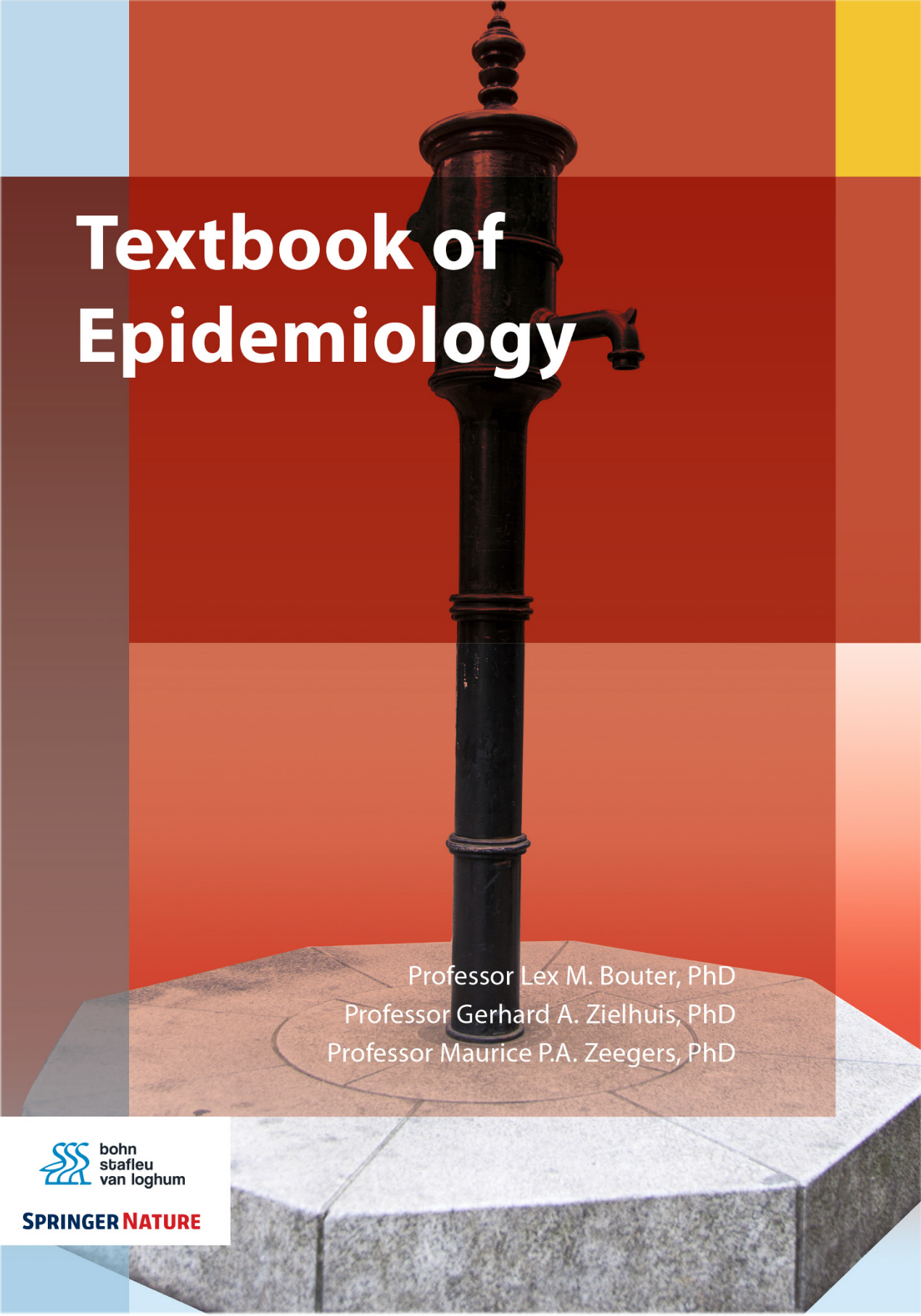
Textbook of Epidemiology



Professor Lex M. Bouter, PhD
Professor Gerhard A. Zielhuis, PhD
Professor Maurice P.A. Zeegers, PhD

bohn stafleu van loghum
SPRINGER NATURE

L.M. Bouter
G.A. Zielhuis
M.P.A. Zeegers

Textbook of Epidemiology

L.M. Bouter
G.A. Zielhuis
M.P.A. Zeegers

Textbook of Epidemiology

bohn stafleu van loghum

Houten 2018

ISBN 978-90-368-1740-0
ISBN 978-90-368-1741-7 (eBook)
DOI 10.1007/978-90-368-1741-7

NUR 870

Basisontwerp omslag: Studio Bassa, Culemborg
Automatische opmaak: Pre Press Media Groep, Zeist

The first edition based on the seventh edition 2016 of the Dutch book, Leerboek epidemiologie with
ISBN. 978-90-368-0561-2

Bohn Stafleu van Loghum
Waldmolen 1
Postbus 246
3990 GA Houten

www.bsl.nl

Foreword

For three decades this book, currently in its 7th edition, has been the most prominent textbook of epidemiology in Dutch speaking countries. It has served as a basis of education in epidemiology of generations of bachelor's, master's and PhD students receiving training in the biomedical sciences, and for practitioners in the medical and allied health professions.

In 2016, the European Epidemiology Federation (EEF), which is the European branch of the International Epidemiological Association (IEA), encouraged the authors to translate the book to English with a view to broaden the audience that can – and certainly will – benefit from its didactical approach. The translation has been carefully worked out by a highly professional team at Springer, in close collaboration with the authors, professors Bouter, Zielhuis and Zeegers.

This *Textbook of Epidemiology* contains the essential concepts of epidemiology presented in an elegant and easy-to-read manner. The chapters are short and suitable to be used both in individual studies or group learning. Teachers of epidemiology will find this book particularly useful as an essential core text for their course. Researchers and other professionals with basic or advanced experience of epidemiology will find the *Textbook of Epidemiology* a useful reference guide on epidemiological methods.

We are grateful to Eline Krijkamp, Sander Kuick, Tinca Polderman, Teun Bousema, Eva Grill, Susan Hahné, Jeannine Hautvast, Bart Kiemeney, Eline Krijkamp, Sander Kuick, Tinca Polderman, Adele Seniori-Costantini, Lau Caspar Thygesen, Hanneke Trines and Sita Vermeulen for their critical reading and helpful comments. We also thank Martien van Dongen who was an author on previous versions of the Dutch book.

We are proud to have the EEF associated with this initiative. It is an important step into the further professionalisation of the federation and making teaching and learning of epidemiology enjoyable and interesting. The EEF warmly recommends the *Textbook of Epidemiology* to everyone who is interested in the field.

Professor Elisabete Weiderpass, M.D., M.Sc. Ph.D.

IEA – International Epidemiological Association – Europe councillor
Chair, European Epidemiological Federation (IEA-EEF)

Table of Contents

Epidemiology

1

Case 1.1 Scurvy and citrus fruits

James Lind, a Scottish ship's doctor, carried out an inventive experiment on scurvy patients in 1747. Scurvy (or scorbutus) is a disease characterized by bleeding gums, internal bleeding, stiff limbs and rough skin, among other things. It was very common in those days among crews of ocean-going ships. Long after Lind made his observations it was found that scurvy is caused primarily by a deficiency of vitamin C (ascorbic acid), which is needed to synthesize collagen, a substance that strengthens blood vessels. James Lind selected ten cases in the same stage of the disease on his ship. Two patients were prescribed cider, two elixir of vitriol, two sea water, two a mixture of nutmeg and a commonly used medicine based on garlic and mustard seed, and two were given two oranges and one lemon a day. Only the patients who were given the orange and lemon treatment experienced a rapid cure. The fact that the mechanism by which scurvy can be cured and prevented using citrus fruit was not known did not present an obstacle to taking effective measures. And yet it took many years for citrus fruit to be introduced as a prophylactic for this dreaded disease on ships.

1.1 What is epidemiology?

1.1.1 Epidemiology is the study of the occurrence of disease in human populations

'Epidemiology' is a term that not many people are familiar with: most will trip over the word when they try to pronounce it for the first time. Many professionals in healthcare and health research have only a vague notion of what it means. This is certainly not the case with the concept of an 'epidemic', which in most people conjures up a picture of a sudden outbreak of an infectious disease such as influenza or an intestinal disorder. It will come as no surprise to learn that the familiar term 'epidemic' and the relatively unfamiliar term 'epidemiology' are closely related. Associating these two concepts can be instructive, but it can also be dangerous, as an overly narrow interpretation of 'epidemic' can easily cause misunderstanding as to what epidemiology really is and what present-day epidemiologists study.

The word 'epidemic' is derived from the Greek words 'epi' (=on) and 'demos' (=people). Epidemics are phenomena (plagues, diseases, health outcomes) that are 'thrown' onto a people (population). Epidemiology is the theory or science ('logos' = doctrine) concerned with studying the frequency with which these phenomena occur in a population. Strictly speaking, we ought to use the term 'human epidemiology' to indicate that we are always concerned with human populations, especially since there is nowadays also a veterinary branch of epidemiology. For the sake of convenience, however, we shall simply use the word 'epidemiology' in this textbook. The term epidemic means a dramatic increase in the extent to which certain diseases or health phenomena occur during a particular period (weeks, months, years, decades). In other words, it is a question of the frequency of disease or disease phenomena at a particular time, which we implicitly compare with what we would expect, namely the normal background level up to that time. Alternatively, we can compare the frequency in a particular geographical area with what we would expect based on the frequency elsewhere. If diseases or disease phenomena remain at a constantly high frequency in a population for a lengthy period that is referred to as an endemic. Malaria, for example, is endemic in many tropical regions, and people travelling to those regions are advised to protect themselves against it with prophylactic medication.

The terms 'epidemic' and 'endemic' can be used in connection with the occurrence of many different types of diseases: not only can infectious diseases be epidemic or endemic, so can chronic diseases, injuries due to accidents and other health problems. The seven cases described in this chapter provide a good cross-section. The fact that we nevertheless tend to associate these terms with acute infectious diseases rather than with other types of disease is due to past

practice, which focused on the study of infectious diseases (see ▶ par. 1.2).

Following on from the above, we can characterize epidemiology as follows:

- Disease or health outcomes are the main objects of epidemiological research (see ▶ par. 1.1.2).
- Epidemiology is the study of disease outcomes in human populations (see ▶ par. 1.1.3).
- Epidemiologists usually look at disease outcome in relation to other phenomena. These are factors that are suspected to have an influence on the development of the disease in question (etiological factors), to give an indication of the presence of the particular disease (diagnostic factors), or to be associated with the course of the disease (prognostic factors) (see ▶ par. 1.1.4).

1.1.2 The object of epidemiological research is disease in humans

The basic variable in epidemiology, and in any epidemiological study, is a disease or health outcome. Epidemiologists are interested in the extent to which diseases occur (i.e. their frequency) among the population. This defines the object of epidemiological research, namely disease outcome in humans. The word 'disease' needs to be interpreted broadly in this context: as already mentioned, it can mean a broad variety of medical conditions, both infectious and non-infectious, acute and chronic, somatic and mental. In this context 'disease outcome' can refer to all sorts of phenomena on a continuum from full health to death from a particular condition. It can also refer to a disability, injury due to trauma, quality of life, or a physiological measure in sport at the highest professional level. As there is no term that encompasses all these aspects, we generally use the phrase 'disease outcome' – or simply 'disease' or 'health outcome' – in this book. ▶ Chapter 2 looks in more detail at the process of measuring various specific aspects of health and disease. For the purpose of epidemiological research we need to define the disease outcome that we are interested in as precisely as possible so as to enable us to detect specific relationships with other variables.

1.1.3 Measuring frequency is the epidemiological research method

Epidemiology is the study of disease outcomes and related variables in human populations. Animal experiments and observations in cell or organ cultures do not therefore fall into the domain of epidemiology. Although the measurements are usually carried out on individuals, the results of epidemiological research always relate to groups of people, i.e. disease frequency. As the disease frequency for a group can be interpreted as the mean disease risk for each member of the group, the knowledge gained from epidemiological research is valuable not only at the group level (for public health) but also at the individual level (in healthcare). Epidemiological researchers calculate and compare disease frequencies in groups of people with different characteristics. They ascertain whether each individual has the disease and then count the number of individuals with the disease in the group as a whole, thus yielding the epidemiological fraction.

$$\frac{\text{Number of cases}}{\text{Total number of persons in the group from which these cases are taken}}$$

This fraction provides the basis for all epidemiological measures of frequency (see ▶ chap. 2). For example, we refer to incidence when counting new cases of a disease in a group at risk of contracting that disease (the at-risk population). Prevalence, on the other hand, relates to the number of existing cases of the disease. Incidence is also a measure of risk: it indicates the average risk an individual belonging to the group in question has of getting the disease.

Health and disease are not equally distributed among the population, and it is this fact that lends epidemiology its raison d'être. The purpose of epidemiology is first and foremost to identify differences in health and disease outcomes between human populations. The distribution pattern of a disease outcome among the population becomes clear when we investigate the differences in frequency between groups of people at different times, in different places and with different individual characteristics. Differences in time can manifest themselves e.g. between seasons or over a series of years or decades. Differences between geographical areas can relate to

e.g. continents, countries, regions of a country, urban versus rural areas, or districts of a town or city. Examples of personal characteristics that can be associated with differences in disease frequency are age, gender, race, genetic predisposition, occupation and specific lifestyle characteristics such as smoking, drinking and sporting activities. Dividing the population up into subpopulations based on time, place or personal characteristics thus gives us an understanding of the distribution of the disease risk, and can enable us to identify risk periods, risk areas or at-risk groups. Identifying the distribution pattern of a disease outcome falls within the domain of descriptive epidemiology. The branch of epidemiology that tries to identify the causal factors underlying the frequency of a disease is known as analytical epidemiology.

1.1.4 Determinants influence disease frequency

In addition to the main variable (disease outcome), epidemiology is concerned particularly with the factors related to the occurrence of a disease. These factors fall into three categories: etiological, diagnostic and prognostic. The term 'determinants' or 'exposures' is used for these factors in epidemiology. Individuals can be exposed to several determinants at the same time or successively. Epidemiologists are interested first and foremost in determinants that are causally responsible or co-responsible for the development of a disease (etiological factors) or that influence the course of the disease (prognostic factors). They may also look for factors that distinguish people who have a particular disease from those who do not have it (diagnostic factors).

The factors that influence the development or course of a disease can be divided into three categories: genetics, lifestyle and environment. Genetic properties and biological characteristics of genetic origin are important determinants, but they are as of yet, more difficult to manipulate. Interventions to influence the risk of disease are more likely to be available with regard to the environmental and lifestyle factors to which people are exposed voluntarily or involuntarily (e.g. diet, smoking, alcohol, drugs, sexual habits, microorganisms, environmental and occupational exposures). Preventive, diagnostic and therapeutic interventions can also be regarded as determinants of disease and prognosis: for example, screening programmes and various medical and paramedical interventions such as dietary restrictions, surgery, pharmaceuticals, radiotherapy or physiotherapy. An entire treatment protocol or a particular facility (e.g. the thrombosis prevention service or the well baby clinic) can also be regarded as an intervention in this context.

Epidemiology still has a lot to contribute to medical science by finding out what determinants are associated with a disease outcome, how strong the link is between each determinant and the disease, and what relative (and possibly interactive) contribution these determinants make to the occurrence of the disease. There are always multiple exposures involved in every disease and every stage of the disease. This can be expressed in symbolic notation as follows:

$$P(O) = f(D_i)$$

This epidemiological function states that the occurrence of the disease outcome (O) is a mathematical function of a series of k determinants (D_i, where i = 1, …, k). In this equation the disease outcome is the dependent variable or outcome variable. The determinants are the independent variables in the formula. Various measures of association can be used to express the strength of the relationship between the occurrence of the determinant and the disease. The most important measures of association will be discussed in ▶ chap. 3.

Although an epidemiological study usually focuses on a single determinant of the disease outcome if we want to gain a good understanding of the relationship between that determinant and the disease we will almost always need to include other determinants in the study design, measure them and take them into account when analysing the results.

The presence of these other determinants can affect the results. We know, for example, that if we want to examine the effect of alcohol consumption on driving behaviour, we need to include the weight of the drivers, as the effect is stronger for fat versus thin people. In other words, the effect of alcohol on driving is modified by weight (effect modification). If we want to study the effect of exercise behaviour

on cardiovascular disease outcomes, we must not forget to adjust for the effect of healthy diet: people who take healthy exercise may also have a healthier diet. Unless we include these extraneous determinants we cannot be sure whether the observed effect of exercise behaviour is actually due to healthy diet (confounding). More information on effect modification and confounding will be given in ▶ chap 4 and ▶ 5.

Case 1.2 Cannabis and depression

In 2015 the United Nations Office on Drugs and Crime estimated the annual prevalence of cannabis use worldwide at around 2.8–4.5%. This frequency varies between countries, with the highest prevalence (>10%) in Canada, the US and Australia and a prevalence between 2.5 and 10% in most European countries. With around 182.5 million users globally, cannabis ranks first among illicit drug use.

Research has linked cannabis use to the development of depression. Most of these studies showed an increased risk of depression among cannabis users after adjusting for other possible determinants of depression (confounding variables). There is also evidence of a dose-response relationship, since the risk of depression increases with higher levels of cannabis use.

Recently a US cohort study was performed to study the effect of reducing cannabis use on depressive symptoms in young female adults. This cohort study included 332 females of different ethnicities between the ages of 18 and 24 who smoked cannabis at least three times in the past three months. At baseline and after three and six months of follow-up the young females filled out the Beck Depression Inventory-II (BDI-II) questionnaire, which measures depression on a continuous scale, with higher scores indicating more severe depressive symptoms.

After controlling for alcohol use – which has been shown to be associated with depression – the results showed a significant reduction in depressive symptoms in participants who had stopped using cannabis. This reduction was more pronounced among initially moderately or severely depressed women compared to minimally depressed women.

This cohort study supports the view that cannabis use may be a cause of depression. Since the research focused only on young female subjects, the results cannot be generalized to males directly. It may be worthwhile to study the existence of effect modification by gender in a mixed population. Before using these insights in the treatment of depressive patients who are also cannabis users the effect of advice to stop using cannabis needs to be studied in a randomized trial (see ▶ chap. 4).

1.1.5 The research question specifies the epidemiological function and is a central element in the empirical cycle

The epidemiological function is the formal expression of the research question. The question whether cannabis use in humans increases the risk of depression (▶ case 1.2) can be represented by the following function:

$$P(\text{depression}) = f(\text{cannabis use})$$

How do epidemiologists arrive at a research question? Research questions do not just arrive unbidden; often accidental observations of e.g. coinciding events (clusters) lead investigators to tackle a particular topic. Researchers' curiosity can also be aroused by results of previous research or reports of other people's results, and they will feel the need to confirm the results (verification), dispute them (contradiction), refute them (falsification) or make them more specific (elaboration). Sometimes they will feel the need to improve on weak points in the design of previous research. Successive studies on similar research questions will gradually increase our understanding of that particular aspect of reality. This is referred to as the empirical cycle. In its simplest form the empirical cycle can be represented as shown in ◘ fig. 1.1.

The investigator starts with a particular theory, a statement or coherent series of statements intended

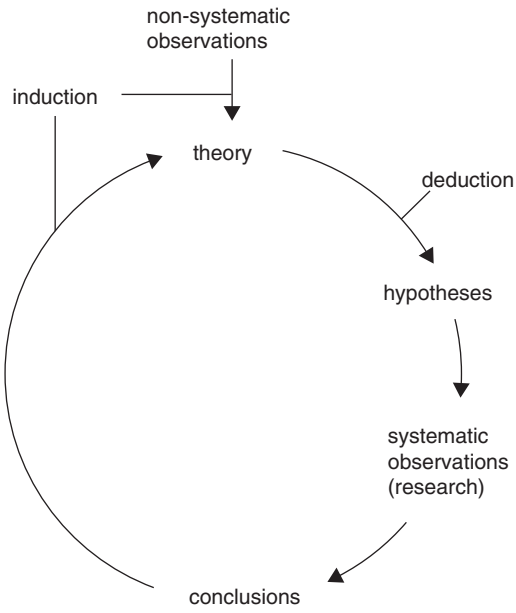


Figure 1.1 The empirical cycle

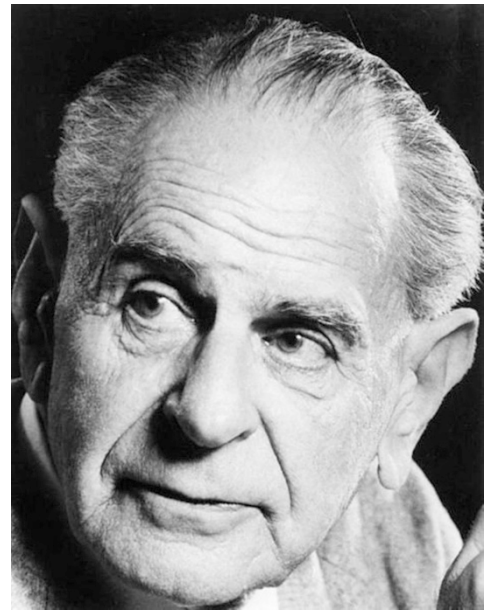


Figure 1.2 Karl Popper (1902–1994), philosopher of science and founding father of the falsifiability theory

to have universal validity. In some cases the theory will have been handed down, but it is usually based partly on observations of actual events, which do not by definition have to be systematic. Sometimes they come from incidental encounters of the researcher. However, usually systematic observations form the basis for theoretical explorations: results of previous research into the same topic conducted by the researcher himself or other researchers. This previous research will by no means always be epidemiological in nature.

Incorporating these systematic and unsystematic observations into a theory is no simple matter; it calls for a good deal of creativity and inventiveness. This process of developing a more abstract, universally valid picture of reality from specific observations is known as induction. According to inductive reasoning, a new theory applies not only to the cases observed but also to all similar cases. To find out whether this is true we carry out new – systematic – observations. If they tally with the expectations, confirmation is obtained that we are on the right track with our theory.

Another way of developing a theory is to try to prove that it is incorrect. This idea was introduced in the twentieth century by the philosopher of science Karl Popper (1902–1994) (fig. 1.2) and now forms a cornerstone of empirical science. It is impossible to confirm anything with certainty, but it is possible to falsify it. Popper therefore stated that every scientific theory must be falsifiable. Falsifiability means that we can try to undermine the theory with e.g. epidemiological research. To do this we first need – taking the theory as our starting point – to formulate hypotheses, statements that can or cannot be refuted based on real-world observations.

This process of going from an abstract theory to one or more hypotheses for testing is known as deduction. In effect, the deduction process corresponds to translating a general research idea into one or more research questions. In contrast to what fig. 1.1 suggests, not every systematic study is designed to test a hypothesis; there is also research of a more exploratory nature. Exploratory research is conducted mainly on new and relatively unexplored problems, with the aim of generating promising hypotheses. The findings from this type of research also contribute to scientific theory.

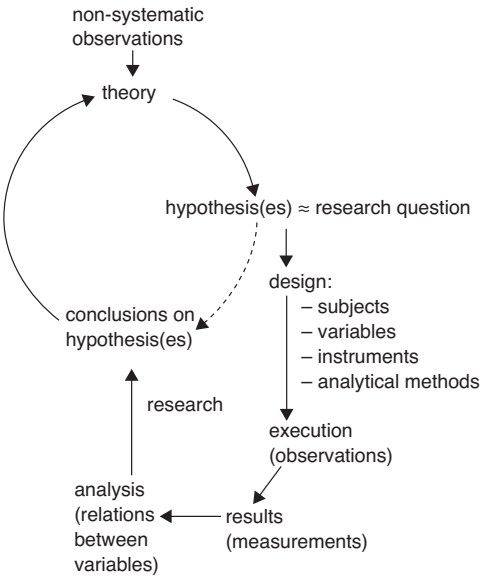


Figure 1.3 Research as an element in the empirical cycle

Figure 1.4 Fokke and Sukke on the empirical cycle

As we have said, empirical research is needed to test a hypothesis or explore a new problem area. The phases of empirical research are as follows:

1. Formulating the research question
2. Drawing up a study design:
   - Selecting the study population (specifying the sampling framework, the sampling procedure, the subpopulations to be compared and the inclusion and exclusion criteria)
   - Selecting the instruments (deriving measurable variables from the concepts in the research question, and if necessary developing measuring tools and criteria for classifying the participants in terms of each factor)
   - Selecting the measuring times (when observations take place)
   - Selecting the statistical analysis techniques
3. Conducting the research (carrying out the observations, collecting data)
4. Analysing the results:
   - The frequencies of the measured values of the relevant variables (univariate analysis, descriptive in nature)
   - The relationships between the variables (bivariate and multivariate analysis, either descriptive or explanatory)
5. Interpreting the results (conclusions).

The empirical cycle can therefore be expanded as shown in fig. 1.3.

The results of a single study can rarely make or break a theory or hypothesis. A theory represents current knowledge on a particular subject at a particular time. Plenty of research is usually needed before a theory can gain a satisfactory empirical basis. In effect each study means a fresh circuit of the empirical cycle, or rather the empirical spiral, working towards an ever better description of reality.

See also fig. 1.4.

Case 1.3 Treating tuberculosis patients with Streptomycin

Although Lind's research (see ▶ case 1.1) could be characterized as an experiment, randomized intervention studies into the efficacy of medical treatment did not come into general use until after World War II. The prototype for clinical trials

1

Table 1.1 Diagnosing Alzheimer's disease using a visuo-spatial associative learning test

test	alzheimer's disease	possible dementia	depression	healthy	total
below threshold	25	17	3	1	46
above threshold	1	26	34	38	99
total	26	43	37	39	145

– an experiment on patients with pulmonary tuberculosis – was published by the British Medical Research Council (MRC) in 1948. The reason for conducting the experiment was that Great Britain had only limited stocks of streptomycin after the war. Its use on patients with very severe forms of tuberculosis was uncontroversial, but what was left of it was completely inadequate to cover all the other types, including pulmonary tuberculosis. The MRC decided to make a virtue of necessity by carrying out an experiment to study the effect of streptomycin on pulmonary tuberculosis patients.

From six clinics, 107 patients were recruited and allocated at random to two groups without asking the respective patients for consent or telling them which group they had been assigned to. Random allocation within subgroups took place based on gender and research centre. Of the patients, 55 were given streptomycin and 52 received nothing.

The result was statistically significant: in the treatment group four of the 55 patients died within six months, against fourteen of the 52 patients in the control group. This trial demonstrated that streptomycin is also effective in treating pulmonary tuberculosis.

Case 1.4 Diagnosing Alzheimer's disease

With new treatments being developed for Alzheimer's disease, diagnosis at an early stage of the condition needs to be improved. Defects in short-term memory have been consistently found in patients with the disease, but as these signs are also found with other conditions such as depression, making a good differential diagnosis is no simple matter. A British study subjected four groups of patients (26 with mild Alzheimer's disease, 43 with possible dementia, 37 with major depression and 39 healthy controls) to a battery of computerized tests. A visuo-spatial associative learning test yielded the results shown in tab. 1.1.

The researchers concluded that this test was able to distinguish patients with Alzheimer's disease from depressed patients and healthy people. Also, the subgroup of patients with possible dementia who had test results below a certain threshold displayed substantial cognitive decline during follow-up. While the test is not perfect (47% of the patients with positive test results did not have Alzheimer's disease, and 4% of the Alzheimer's disease patients had normal test results) and only involved relatively small groups of people, it would seem to be promising for the diagnosis of early Alzheimer's disease.

1.2 Developments in epidemiology

1.2.1 From Hippocrates to death records and infectious diseases to research into smoking and lung cancer

As an epidemiologist *avant la lettre*, Hippocrates (circa 470–430 BC) stressed the importance of describing cases of disease meticulously (fig. 1.5). Hippocrates demonstrated in his book *On Airs, Waters and Places* that health and disease are deter-

Figure 1.5 Hippocrates (circa 470–430 BC), epidemiologist avant la lettre

Figure 1.6 Captain John Graunt (1620–1674), founding father of descriptive epidemiology

mined by all sorts of observable environmental factors. Subsequently, for more than two millennia, ideas on disease remained related to humoral pathology (water, earth, fire and bile), with scarcely any attention being paid to empirical observation. Not until the end of the seventeenth century was the Hippocratic approach readopted, for example by the Italian clinician Bernardino Ramazzini (1633–1714), who stressed the importance of asking patients about their medical history, nutritional habits and working conditions. He even demonstrated the added value of comparing similar cases and looking for common circumstances. The idea of recording, counting and analysing causes of death dates from the same period. The London physician John Graunt (1620–1674) is regarded as the founding father of this type of descriptive epidemiology (fig. 1.6).

The advent of empirical research, introduced by Galileo Galilei (1564–1642), William Harvey (1578–1657) and others, laid the foundation for the further development of epidemiology. The systematic recording and analysis of causes of death yielded a wealth of information, on the basis of which researchers such as Farr (1807–1883) presented impressive and highly relevant reports to the British Minister of Health. In this tradition we also find John Snow (1813–1858), who studied the causes of cholera (see ▶ case 1.5), presenting one of the first clear examples of the epidemiological approach (fig. 1.7).

In the eighteenth and early nineteenth century great progress was made in statistics, whose influence also grew in medical science. An important person in this regard is the French physician Pierre Louis (1787–1872), who introduced the 'numerical method' in medicine and demonstrated with statistics that bloodletting was ineffective and even harmful. Empirical medical research became popular in the first half of the nineteenth century, an era with major developments in biology, pathology and public health and the first classifications of diseases as

Figure 1.7 John Snow (1813–1858), one of the founding fathers of analytical epidemiology

well. The German scientist Rudolf Virchow (1821–1902) played a vital role here, being not only an eminent pathologist but also a great advocate of a vigorous approach to public health problems. The work of the Frenchman Louis Pasteur (1822–1895) and the German Robert Koch (1843–1910) in the field of microorganisms uncovered the causes of many infectious diseases, including the tuberculosis, which was widespread. Thanks to Koch we have a first set of criteria for causality in epidemiological research. Around 1900 epidemiology was virtually synonymous with the epidemiology of infectious diseases, but these became less important with the development of bacteriology. a sense the epidemiological approach was rediscovered in the mid-twentieth century for the study of chronic conditions.[1]

Present-day epidemiology has been highly influenced by the studies into the relationship between smoking and lung cancer carried out around World War II. Following initial reports of a possible link in the late 1930s, three studies published in 1950 showed that smoking is likely to be causally linked to the risk of lung cancer. One of these studies was carried out by Doll and Hill, two researchers who

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine ; Honorary Director of the Statistical Research Unit of the Medical Research Council

Figure 1.8 Doll and Hill's famous article on smoking and lung cancer

1 The Epidemiology Hall of Fame (video). ▶ http://bit.ly/1GetcFu.

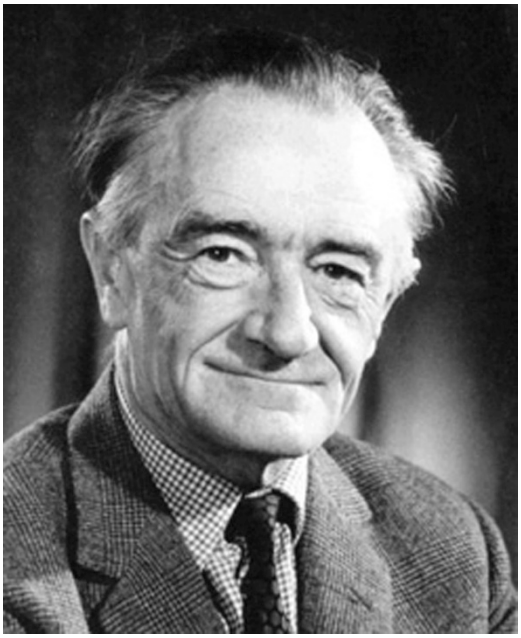


Figure 1.9 Archie Cochrane (1909–1988), who introduced the use of randomized clinical trials in epidemiology

continued their studies into the harmful effects of smoking on health during subsequent decades. In 1964 they published the results of a major study among British physicians that clearly showed a causal relationship (fig. 1.8) (see also ▶ case 4.5 in ▶ chap. 4). It is no coincidence that soon afterwards (in 1965) Hill published a revised set of criteria for causality, which were applicable both to infectious diseases (as were Koch's criteria mentioned above) and non-infectious diseases (see ▶ chap. 6).

From the particularly heated debate about the relationship between smoking and lung cancer in the 1950s and 1960s we can infer that lack of understanding of the exact mechanisms of disease is both a strength and a weakness of the epidemiological approach. This convincing epidemiological example has certainly raised interest in the contribution of epidemiology to the solution of many other public health problems. In addition clinicians became interested in the potential of epidemiology for solving diagnostic, prognostic and treatment problems. This development gained a major boost from the publication in 1972 of *Effectiveness and Efficiency: Random Reflections on Health Services*, in which Archie Cochrane (1909–1988) ardently advocated the systematic use of randomized clinical trials to measure the efficacy of curative and preventive interventions. Present-day epidemiology can be regarded as the combination of classic observational epidemiological methods with experimental study designs rooted in statistical theory developed since the 1930s (fig. 1.9).

Case 1.5 Cholera and drinking water

The London physician John Snow is often named as one of the founding fathers of epidemiology. A cholera epidemic broke out in the London district of Soho in 1848, and 500 residents of a single neighbourhood died of the disease within ten days. Snow started surveying the cholera cases systematically. He suspected that there was a connection with drinking water, as none of the monks at a nearby monastery – who drank no water, only beer – had cholera. Snow then compared the water supply provided by the Southwark and Vauxhall Company with that provided by the Lambeth Company. He concluded that the victims must have mainly used the water pump in Broad Street (see the illustration on the front cover of this book). At Snow's insistence the pump was sealed off and the cholera epidemic cleared up. It took years to learn that cholera is actually caused by a bacterium, but for Snow this was not an obstacle to taking adequate preventive action. Close study of his writings demonstrates that he was guided by an explicit theoretical notion about the etiology of cholera although the biological mechanism was far from understood.

1.2.2 Epidemiology today: developing methodology and increasing knowledge of diseases

Present-day epidemiology is based on two key elements: the development of epidemiological research methods (scientific methodology) and the substantive medical and biomedical knowledge gained by the application of these methods. Both have gained in strength enormously since World War II. Just compare recent and early editions of medical handbooks and you will soon discover what an enormous contribution epidemiology has made to our knowledge and understanding of the distribution of, determinants of and intervention options for major diseases – including ischaemic heart disease, asthma, various types of cancer and AIDS –, the adverse effects of pharmaceuticals and many other things. The first book on epidemiological methodology was published in 1960; nowadays there are enough such books to fill a large bookcase. All sorts of new study designs and data analysis methods have been developed over the past few decades to meet the needs of epidemiological practice. The rapid developments in medical science, computing and communication technology and their medical applications are also contributing to the development of the field, in particular:

- The incorporation of modern molecular biology techniques in epidemiological research, enabling not only phenotypes but also genotypes to be studied on a large scale.
- The rise of e-epidemiology, the use of digital media such as the internet, wearables and mobile phones in epidemiology: for example, web-based questionnaires, recruiting participants through Facebook, analysis of Google search behaviour and Twitter activity to predict flu epidemics, and wearables to measure air quality or personal physical activity and food intake. These new methods have many uses and benefits and are increasingly being seen as fully fledged alternatives to traditional methods. Nevertheless, the field is still in its infancy, and while developing new devices we need to monitor the internal and external validity of the data they produce, safeguard privacy and deal with ethical issues related to new ways of recruiting participants and collecting data via the internet.
- Progress in clinical medicine, with a host of new diagnostic and treatment options and the rise of evidence-based medicine.
- The rise in the cost of healthcare, making it a high priority to demonstrate the efficacy and efficiency of interventions.
- Increasing awareness among professionals and the public at large that human health depends to a large extent on the quality of the physical and social environment. That quality has deteriorated sharply precisely due to human action, with the result that there is still room for the traditional role of epidemiology in tackling public health problems.
- Thanks to the rapidly growing size of the data sets being created in research and records systems (big data) the epidemiological method is increasingly being used to distil useful information on determinants of diseases from the amorphous mass of data.

Case 1.6 The third-generation contraceptive pill and the risk of venous thrombosis

Soon after the first oral contraceptives were introduced in 1960 there were reports of increased risk of thrombotic events, resulting in the development of new types of contraceptive pills with a lower oestrogen content (but containing more progestogens). The first-generation pills were followed by a second and then a third generation, the aim being to reduce the risk of arterial thrombosis. An increased risk of venous thrombosis remained, however. The World Health Organization concluded in 1998 that users of modern oral contraceptives (containing 30–40 µg of oestradiol) are at a three to six times higher risk of venous thrombosis than similar women who are not on the pill. It should be noted, however, that the absolute risk is still low, from less than one case of venous thrombosis per 10,000 person-years in women not using the pill to 3–6 cases per 10,000 person-years in women using it. A host of epidemiological studies – both cohort studies and case-control studies – have meanwhile been pub-

lished comparing the risk of venous thrombosis from 'third-generation' pills with that from second-generation pills. These show that venous thrombosis is 1.5 to 4.0 times more frequent in users of third-generation pills than in users of second-generation pills. The risk of venous thrombosis is highest in the first few years of using this type of pill, approximately one case per 1,000 first years of use. This is thought to be due to the progestogens in the third-generation pill having an adverse effect on certain blood coagulation factors. The advice, based on these findings, is not to prescribe a third-generation pill to new users of oral contraceptives.

Epidemiology and epidemiologists face two major challenges in the coming decades, which will give rise to both opportunities and potential problems.

Developments in biology

New developments in immunology, molecular biology and genetics will need to be incorporated in epidemiological research: for example, using biomarkers (e.g. protein adducts containing exogenous toxins) instead of questionnaires to measure exposure. The study of individual genetic susceptibility will help to identify the most vulnerable groups. Epidemiology also stands to play a greater role in discovering or confirming the effects of biological factors (e.g. blood hormone levels) on health and disease. Enhanced biomedical technologies are enabling biomedical parameters to be measured more easily, making it easier to examine them in large epidemiological studies. This is bringing the classic 'black box' approach in epidemiology and the mechanistic approach in biology closer together, as can be seen clearly from the rapid developments taking place in genetic epidemiology. New, relatively inexpensive genome sequencing techniques have made a major contribution to this advance (see ▶ chap. 7).

Case 1.7 Prediction of live birth after IVF treatment

In-vitro fertilization (IVF) is a complex and costly assisted reproductive technology that involves a substantial emotional and physical burden for the women at issue. Unfortunately, IVF treatment does not guarantee success. Couples would therefore like to know their prior probability of success before starting IVF treatment. The prediction models that are often used to predict pregnancy followed by live birth after IVF are based on old data, and the model predictions often underestimate the clinical outcomes. These models include among other things the age of the women, previous pregnancies and the use of donor cells. However, other factors that are associated with pregnancy outcome in women undergoing IVF, such as body-mass index (BMI), ovarian reserve and ethnicity are not included in these models. Recently, data from a cohort study among 9,915 women in the UK and Ireland derived and validated a novel prediction model of live birth for women undergoing IVF. The model includes age, duration of infertility, BMI, antral follicle count (AFC), previous miscarriage, previous live birth, cause of infertility and ethnicity. While the average probability of success in the cohort was 31.5%, individual predictions ranged between 18 and 41%, with overall good validity.
This prediction model can be used to create a user-friendly decision-making aid to inform couples and their physicians about the likelihood of a successful pregnancy. This can increase the quality of the decisions and may reduce the number of unsuccessful IVF treatments.

Developments in the discipline itself

Specialization is growing in epidemiology, just as in every maturing scientific discipline. Content and methodological development are growing increasingly far apart, with separate attention being paid to improving methods for exposure monitoring, measurement quality, dose-response models, longitudinal designs and so on. Epidemiologists are also spe-

cializing in particular specialist areas, either focusing on a particular category of diseases (e.g. cancer epidemiology, infectious disease epidemiology or the epidemiology of ageing), a particular category of determinants (e.g. nutritional epidemiology, genetic epidemiology or pharmacoepidemiology), or a particular application (e.g. clinical epidemiology or forensic epidemiology). These types of specialization are an inevitable consequence of the successful development of the discipline, but they also entail the risk of losing sight of the wood for the trees, with nobody being able to tackle a complex public health problem using a general approach, and not enough results of epidemiological research finding their way into healthcare practice.

Recommended reading

Bonita R, Beaglehole R, Kjellstrom T. Basic epidemiology. 2nd ed. Geneva: World Health Organization; 2006.

Carneiro I, Howard N. Introduction to epidemiology. 2nd ed. Maidenhead: Open University Press; 2005.

Fletcher RH, Fletcher SW, Fletcher GS. Clinical epidemiology: the essentials. 5th ed. Baltimore: Lippincott, Williams & Wilkins; 2012.

Grobbee DE, Hoes AW. Clinical epidemiology: principles, methods, and applications for clinical research. 2nd ed. Burlington: Jones and Bartlett Learning; 2015.

Hebel JR, McCarter RJ. Study guide to epidemiology and biostatistics. 7th ed. London: Jones and Bartlett Publishers; 2012.

Holland WW, Olsen O, Florey C du, eds. The development of modern epidemiology: personal reports from those who were there. Oxford: Oxford University Press; 2007.

Morabia A. A history of epidemiologic methods and concepts. Basel: Birkhäuser Verlag; 2004.

Porta M. A dictionary of epidemiology. 5th ed. New York: Oxford University Press; 2008.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2012.

Rothman KJ. Epidemiology: an introduction. 2nd ed. New York: Oxford University Press; 2012.

Szklo M, Nieto FJ. Epidemiology: beyond the basics. 3rd ed. Burlington: Jones and Bartlett Learning; 2014.

Webb P, Bain C. Essential epidemiology: an introduction for students and health professionals. 2nd ed. Cambridge: Cambridge University Press; 2011.

Source reference (cases)

Lind J. A treatise of the scurvy in three parts. Containing an inquiry into the nature, causes and cure of that disease, together with a critical and chronological view of what has been published on the subject. London: A. Millar; 1957 (Case 1.1).

Lev-Ran S, et al. The association between cannabis use and depression: a systematic review and meta-analysis of longitudinal studies. Psychological Medicine 2014;44:797–810 (Case 1.2).

Moitra E, Bradley JA, Stein MD. 2016. Reductions in cannabis use are associated with mood improvement in female emerging adults. Depression and Anxiety 2016;33:332–38 (Case 1.2).

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948;2:169–82 (Case 1.3).

Swainson R, Hodges JR, Galton C, Semple J, Michael A, Dunn BD, et al. Early detection and differential diagnosis of Alzheimer's disease and depression with neuropsychological tasks. Dement Geriatr Cogn Disord. 2001;12:265–80 (Case 1.4).

Frost WH. Snow on cholera. New York: The Commonwealth Fund; 1936 (Case 1.5).

Vandenbroucke JP, Rosing J, Bloemenkamp KWM, Middeldorp S, Helmerhorst FM, Bouma BN, Rosendaal FR. Oral contraceptives and the risk of venous thrombosis. N Engl J Med. 2001;344:1527–35 (Case 1.6).

Dhillon RK. 2016. Predicting the chance of live birth for women undergoing IVF: a novel pretreatment counselling tool. Human Repr. 2016;31:84–92 (Case 1.7).

Frequency

2.1 Definition of disease

2

As the previous chapter made clear, epidemiology is concerned with the study of disease frequency in human populations, usually in relation to one or more determinants. In formal terms this means estimating the epidemiological function.

$$P(O) = f(D_i)$$

The disease outcome (O) is thus examined as the mathematical function of a set of k determinants (D_i, where i=1, …, k), often expressed as a linear relationship, although it could be any other kind of mathematical function.

We shall go into the epidemiological function in more detail in subsequent chapters. This chapter focuses on the left-hand side of the equation (representing the outcome studied, which is typically the occurrence of disease or the preservation of health), as health and disease are core variables in descriptive and analytical epidemiology.

This chapter examines the various measures of frequency used in epidemiology: prevalence, incidence, mortality rates and some derived measures. Obviously the estimated frequency of a disease in a population will depend directly on how the disease – and hence its absence (health) – is defined and measured. We regard health and disease as two sides of the same coin. Many different definitions of 'health' can be found in the literature, the most familiar one being that of the World Health Organization (WHO):

» A state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. «

Interestingly, the tendency in this definition of health – and many others – is to make health synonymous with well-being and to define it in positive terms. Being healthy is actually more than just the absence of disease. Another aspect of many definitions of health is their dynamic, process-based nature: health refers to an individual's successful response to changing challenges from the environment. Seen in this way, ill health or disease is the result of going beyond an individual's limits of adaptability. This can be due to overload or reduced (physical or psychological) capacity. The concept of health is generally regarded as comprising a somatic, a mental and a social component.

Such definitions of health do not provide much of a basis for epidemiological research, where the term 'health' or 'disease' needs to be defined in a measurable form. In other words, the concept needs to be made specific and measurable for which there are three dimensions. The objective dimension looks at the organic level ('disease', I have a disease), insofar as that can be determined from the outside. It is based on a diagnosis made by a competent expert (physician, physiotherapist, psychiatrist, clinical psychologist). The subjective dimension looks at the individual level ('illness', I feel ill). This is about people's self-perception of their health, closely related to quality of life. The third dimension looks at the social level ('sickness', I act sick). This is about behaviour, for example in the form of sickness absenteeism or use of healthcare facilities.

A researcher will select one particular dimension of health, depending on the purpose envisaged. Measurements of health at the individual level can generally be aggregated at a higher level so as to give an impression of the health of groups and subgroups in the population. A specific measure of this dimension of health, or of a particular aspect of that dimension at the aggregate level, is referred to as a health indicator. ◘ Figure 2.1 and 2.2 show examples of health indicators. It goes without saying that there are numerous possible health indicators: infant mortality, deaths from coronary heart disease, occurrence of diabetes, number of appendectomies, admissions to psychiatric hospitals and use of sedatives are just a few examples.

At first sight it would seem that the objective dimension of health is scarcely open to debate: after all, a particular condition or disorder is either present or not. This clarity is useful, both in scientific research and in the clinical setting. In practice, however, disease cannot usually be characterized by a simple dichotomous variable: the borderlines between disease and no disease are not always clearly defined, unfortunately. Disease usually manifests itself as a complex pattern of signs and symptoms, because of which a wide variety of severities are often found in a random patient population. Also, individual patients go through various stages of severity as time passes. We

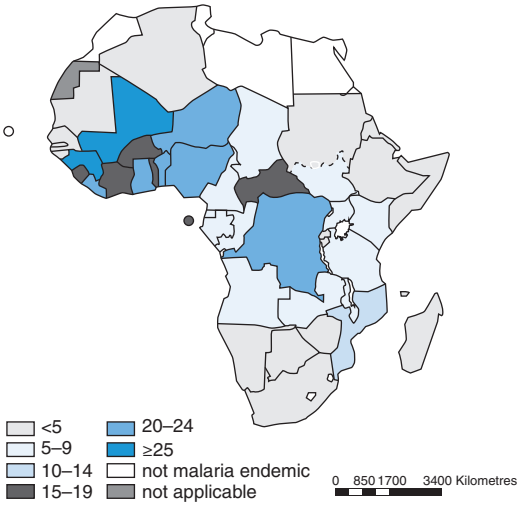
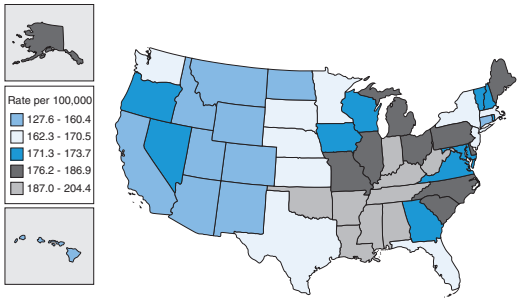


Figure 2.1 Map of the percentage of deaths caused by malaria in children under 5 years of age in sub-Saharan Africa in 2015. The map shows large differences between individual African countries, and a clustering of malaria-related deaths around the equator



Source: National Vital Statistics System. Rates are age-adjusted to the 2000 US standard population using SEER*Stat. More information: https://gis.cancer.gov/geoviewer/data

Figure 2.2 Age-adjusted cancer-related mortality per state in the USA from 2008 to 2013. The map shows a geographical cluster of nine states with the highest mortality rates per 100,000 inhabitants (Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, Tennessee, Kentucky, Indiana, and Ohio)

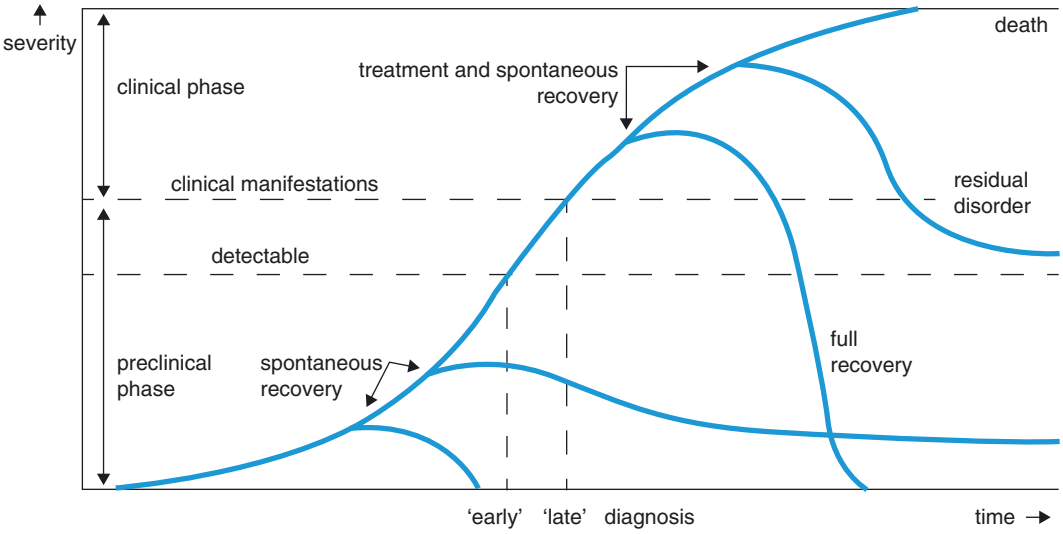


Figure 2.3 Various possibilities for the course of a disease

refer to this disease pattern that changes over time as the course of the disease. Figure 2.3 shows a simplified outline of the course of a condition.

In reality many symptoms and signs will occur to different degrees in different patients. A diagnosis is therefore a crude simplification of the complex phenomena that are seen in reality, and by no means as objective as might be thought. The diagnostic criteria used need to be standardized internationally and made specific and measurable, so that it is clear

what is meant by a particular disease. The International Classification of Diseases (ICD), currently in its tenth revision, performs an important role here. The eleventh revision is set to be published in 2017. What epidemiological researchers need is a detailed description of the diagnostic criteria used, ideally based on international standards.

Indicators of subjective health and sickness are also reported in epidemiological studies: for example, limitations on activities, consumption of medical care, sickness absenteeism or self-rated health. An example of the latter is fig. 2.4, which shows the percentage of local residents that report having good general health in England in 2011. The percentages of the population that reported good general health were by that time higher in central and southern England.

Limitations on activities are a useful measure of the severity of a chronic condition in epidemiological research, but also if we want to assess the effects of occupational therapy or rehabilitation medicine interventions, for instance. A large number of standardized and validated instruments have been developed for this purpose, in the form of interviews, written questionnaires and observation scales, for example those that focus on activities of daily living (ADL).

Also, many instruments exist for measuring health-related quality of life, ranging from disease-specific (applicable to particular conditions and/or patient groups) to generic. The generic questionnaires can be applied across the board to all health problems, regardless of nature and severity. Commonly used are the MHI5 (RAND Mental Health Inventory), designed to measure mental health in the general population, and the SF-36 (Medical Outcomes Study 36-Item Short Form Health Survey), a tool for measuring quality of life.

Indicators of the social dimension of health relate to such things as consumption of medical care, the idea that – at the population level at least – relatively poor health will be associated with increased use of facilities. Utilization of medical care can be expressed in terms of frequency of use or in monetary terms. Another example of an indicator of the social dimension of health is recorded sickness absenteeism, the idea being, roughly speaking, that the higher the absenteeism is the more disease there is expected to be. In principle this argument can apply to an individual or to e.g. the workforce of a particular industry. Sickness absenteeism is difficult to interpret, however, and probably a not particularly valid indicator of health, as there are many reasons for absenteeism that have nothing to do with the occurrence of disease (labour disputes, for instance). Also, many people who are ill want to carry on working as long as they can, so do not take sick leave.

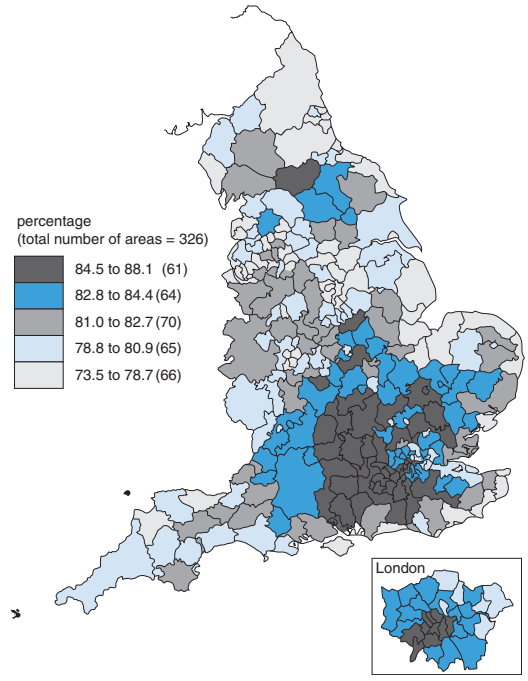


Figure 2.4 Percentage of the United Kingdom population reporting having good general health in 2011. Darker colours are associated with better general health.

2.2 Disease frequency: existing or new cases of disease

Epidemiologists focus on the occurrence of phenomena related to health and disease in groups of people, namely the frequency with which particular states or events occur. They ascertain whether each individual has the disease and then count the number of individuals with the disease in the group as a whole, thus yielding the epidemiological fraction.

$$\frac{\text{Total number of cases}}{\text{Total number of persons in the population}}$$

The epidemiological fraction indicates the frequency of people with a disease (the proportion) in the population.

The epidemiological fraction can take on various forms, depending on how the total number of persons with a disease and the population from which they originate are defined. The main distinction is between prevalence and incidence. Prevalence is a proportion, it relates to the number of persons with a disease at a particular time. This number is then divided by the number of persons in the population containing the cases of the disease. Incidence, on the other hand, relates to the number of persons developing a disease during a particular period. The number of these new cases of the disease is then divided by the number of persons in the population at the start of the period.

$$\frac{\text{Total number of cases at a particular time}}{\text{Total number of persons from which the cases originate}}$$

The epidemiological fraction expressed as a prevalence

$$\frac{\text{Total number of new cases in a particular period}}{\text{Total number of persons from which the cases originate}}$$

The epidemiological fraction expressed as an incidence

Before delving into the various measures of frequency used in epidemiological research in more detail, let us first consider the types of population that can be identified and the importance of the concept of time.

2.3 Types of population: cohort or dynamic population

It is not only the numerator of the epidemiological fraction that is important; so is the denominator, the number of people in the population in which the cases of the disease are being counted. There are two ways of obtaining a population: we either take a cohort or we define a dynamic population.

2.3.1 Cohort

A cohort is a closed population. Membership of a cohort is determined by a particular event and is of unlimited duration. 'Once in a cohort, always in a cohort' is a succinct description of this principle. Examples of cohorts:

1. Children of women who have taken part in a particular study
2. Persons born in a particular area during a particular period
3. All the employees of a particular employer in a particular year
4. All patients who went to a particular GP with flu symptoms during a particular period
5. Persons who started their studies in health sciences in September of a particular year.

For each individual there is a time (t_o) when he or she was added to the cohort: this may be the same calendar time for every member of the cohort (as in Example 5), but this need not be the case (as in the other examples). The point at which events occur in a cohort is usually expressed in relation to t_o, i.e. in terms of follow-up time rather than calendar time. In case of a complete follow-up from t_o to the end of the study period we know exactly what happened to the cohort members (in terms of health outcomes of interest). As time passes, the mean age of the cohort members will go up and the cohort will become smaller as members will die. Also, people sometimes move away from the study area or refuse to continue with the study, making it difficult or impossible to obtain complete follow-up information on all the cohort members. In addition to this unwanted loss to follow-up, the follow-up of cohort members also ends normally when:

- they die
- the health outcome under consideration occurs
- data collection for the study ends.

So although membership of a cohort is of unlimited duration, there are various – unwanted or normal – ways in which the follow-up of a member of a cohort can come to an end.

2.3.2 Dynamic population

Unlike a cohort, a dynamic population is open-ended: membership depends on a particular state and ends once the individual is no longer in that state. The duration of membership of a dynamic population is therefore variable. Examples of dynamic populations:

1. The residents of Florence
2. Residents of a hospital's catchment area
3. Influenza patients aged over 55 in and around Thessaloniki
4. Students at the University of Manchester.

Unlike those of a cohort, the characteristics of a dynamic population do not necessarily change with time: for example the age structure of a dynamic population may remain relatively constant, whereas a cohort always ages over time. The same istrue of the other characteristics of a dynamic population. A dynamic population of this kind is referred to as stable. The fact that the dynamic population is stable is an assumption that always implicitly underlies the measures of frequency discussed below, and also the measures of association that we shall consider in ▶ chap. 3. Different individuals will contribute to the experiences of a dynamic population to different degrees, as different people come and go. In addition to unwanted loss to follow-up due to being unable to obtain the information required on some members of the population, the follow-up in a dynamic population will come to an end normally when:

- the state that defines membership ends
- they die
- the health outcome under consideration occurs
- data collection for the study ends.

Unlike in the case of a cohort, whose membership is of unlimited duration, the first three normal events mentioned above end follow-up also and membership of the dynamic population.

2.4 Time is a difficult concept

The time factor is a difficult phenomenon when studying disease frequencies and interpreting outcomes. Let us illustrate this with a few situations.

Measures of frequency, especially measures of incidence, are also measures of risk: they indicate the likelihood of people in the group in question developing the disease. Although the concept of 'risk' or 'likelihood of disease' is easy for most people to understand, it is often used carelessly. For example, if we read in the press that 60-year-old women have a 2% likelihood of dying of cardiovascular disease, it is impossible to interpret that number, as there is no indication of time. A 2% mortality risk for 60-year-old women would be very high if it relates to the next 24 hours, the next week or even the next year, whereas 2% would be very low if it relates to the mortality risk from cardiovascular disease during the remainder of life. This risk estimate is impossible to interpret without an indication of time.

Obviously, the longer the time period the higher the risk of death. The theoretical values increase from 0% for a very short time interval to 100% for a very long one. This argument is just as true of any other risks, of course. The way a disease risk changes with age can also differ markedly in different populations or for different conditions. The annual risk of chickenpox, for instance, rises sharply during the first few years of life but gets smaller and smaller after childhood, whereas the risk of cardiovascular disease only starts to rise sharply in middle age.

If we follow a cohort over time to ascertain the frequency of new cases of a disease, we have to contend with the problem that some of the cohort members will die of other causes before they get the chance to develop that disease. This 'competing risks' phenomenon will be negligible if the follow-up period is short, but it will cause problems of interpretation if we follow the cohort over a long period. A similar problem is caused by loss to follow-up for other reasons.

A solution to the problem of incomplete follow-up of the cohort is to add up the individual episodes of each subject in the cohort. If a population is followed for 30 years and an individual in that population dies after five years, that person has contributed

five person-years to the follow-up of the cohort. Other persons will have contributed more or fewer years, up to a maximum of 30. This yields the incidence density, a concept that we shall examine in detail later on in this chapter.

Epidemiological studies sometimes focus on a disease that can occur more than once in an individual over the years: contraction of the common cold, for example. The researchers then need to decide whether to include only the first disease episode or also the second and any subsequent episodes. The information in the denominator presents an even more difficult problem: which person-years of the individual cohort members count in such cases and which do not? Generally speaking, it is a good idea to apply the 'at risk' concept in this case and only count the person-time when the subject was actually at risk of contracting the disease. We should not count women's person-time when studying prostate cancer, for example.

If risks are expressed in terms of person-time we can be faced with surprises if the time unit is not stated precisely. Suppose we measure a disease frequency of 47 cases in a population of individuals who have together contributed 1,580 person-months to the follow-up in the study. The disease frequency is then 47 per 1,580 person-months, i.e. 0.03 cases per person-month. We could just as well have expressed the disease frequency in person-years: 47 cases per 132 person-years, i.e. 0.36 cases per person-year. If we then compare numbers of this kind with those given for other populations in the literature and we fail to consider the time units (which might not even be stated), we may well come to confusing conclusions.

Sometimes the researcher is interested not in the disease frequency but in the time to a particular event (e.g. the time to pregnancy in a study of artificial reproductive techniques). In such cases it is useful to know that – given certain assumptions – this 'waiting time' will be equal to the reciprocal of the disease frequency per unit of person-time. In the example above, with a disease frequency of 0.03 cases per person-month the mean waiting time will be 1/0.03 = 33 months.

As these examples show, we need to be careful when using the concept of time in epidemiological research. This has given rise to various measures of disease frequency and derived measures. We need to include the relevant time dimension in every calculation and interpretation. Another take home message is that measures of frequency provide little information without an understanding of how the data were made specific and measurable and the method of calculation adopted.

2.5 Measures of disease frequency[1]

2.5.1 Prevalence: existing cases

The proportion (or percentage) of the population where a particular state of health is present at a particular time is referred to as the prevalence of that state. Thus the prevalence of influenza is the percentage of influenza cases at a particular time in the population in question, regardless of whether we are looking at a cohort or a dynamic population: prevalence implies a cross-section of the population, a snapshot. That snapshot may be taken at the same time for all members of the population, but this is neither necessary nor customary. By way of illustration, ◘ fig. 2.5 shows the occurrence of a condition in a cohort of ten persons in the six years from t_o. This is a non-fatal disease that can occur more than once in the same person. The prevalence of the disease in question is 40, 20 and 20% in years 2, 4 and 6 respectively. Note that a particular follow-up moment can occur at different calendar times for different members of the cohort (for example, in a study of gestational diabetes, where t_o marks the time when women became pregnant). In a similar way we can study the prevalence in a cross-section of the dynamic population, except that in this case the denominator of the epidemiological fraction (see ▶ par. 2.2) – the total number of persons in the population – can differ from one moment to the next. ◘ Figure 2.6 shows a dynamic population: here the prevalence of the disease in question is 43, 25 and 75% in years 2, 4 and 6 respectively.

In practice, however, it is hardly ever possible to ascertain for each member of a population (cohort or dynamic) at exactly the same time whether or not

1 Measure of disease frequency (animation) (Source: Bas Verhage). ▶ http://bit.ly/1tGnFSv.

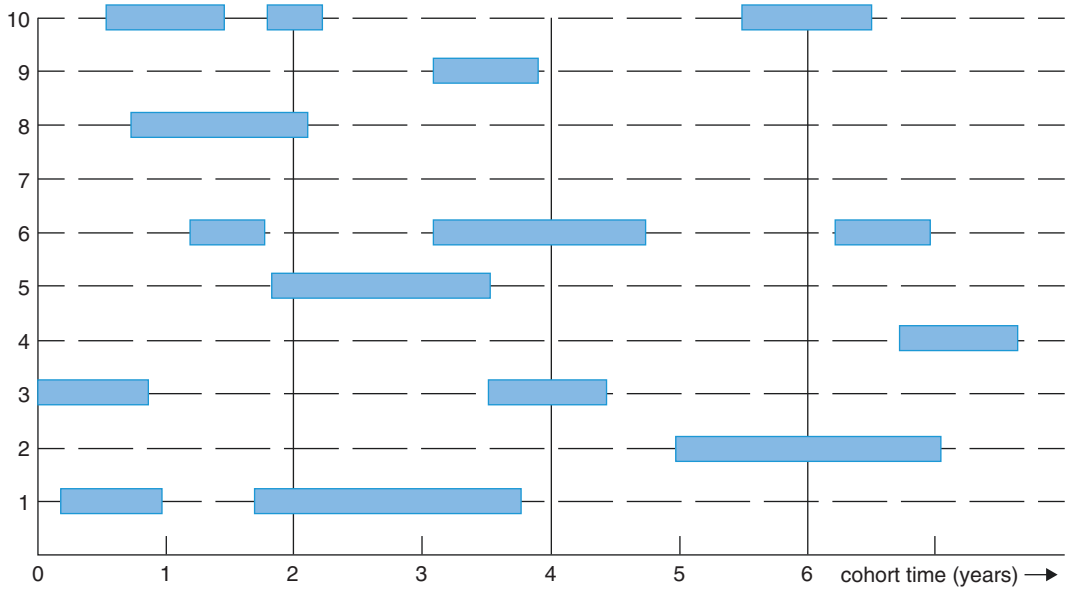


Figure 2.5 Prevalence of a condition in a cohort of 10 persons

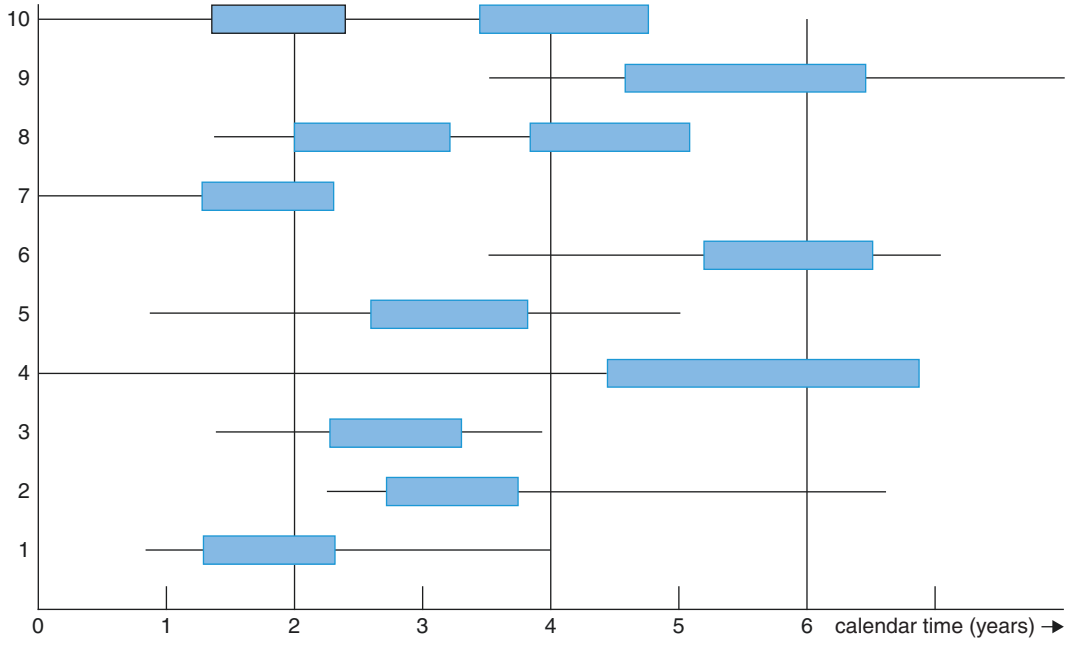


Figure 2.6 Prevalence of a condition in a dynamic population of 10 persons

this member is a prevalent case. The essential point is that we have a snapshot for each individual. In addition to point prevalence some related measures are used in the epidemiological literature. If we examine in a cross-section of a population whether a particular state has ever been present in an individual up to that time, that is referred to as the lifetime prevalence. An example is the percentage of a company's workforce who have ever had an industrial accident by the time they reach pensionable age. Another type of prevalence is period prevalence, the proportion of the population that had the disease in question during a particular period. In fig. 2.5, for instance, 70% of the cohort had the condition in question at least once during the first four years of the follow-up period. If we need to calculate the period prevalence in a dynamic population, we first have to calculate the mean number of persons in the population in that period. In fig. 2.6, for instance, the period prevalence from year 4 to 6 year is 5/6 = 83%. As we can also see from fig. 2.5 and 2.6, the point prevalence of a condition will depend on both the number of new cases that develop as time passes and the mean duration of the condition.

Prevalent cases constitute a heterogenous group: on one hand they have survived the condition so far, but also have not been cured. Because of this selection effect, prevalent cases are less suited to study the causes of diseases (etiological epidemiology). When it comes to determining the demand for healthcare services and the burden on available resources, however, prevalence figures are usually just what we need, as patients who have been cured or who have died are no longer a burden on the healthcare system.

$$\frac{\text{Total number of cases in a particular period}}{\text{Mean number of the total population (average between start and end of the period)}}$$

The epidemiological fraction expressed as a period prevalence in a dynamic population

2.5.2 Incidence: new cases

As we have seen, epidemiologists usually prefer to ascertain the frequency of new cases of a disease. This entails following a population for some time and identifying new cases of the condition that occur during that period. In more general terms the incidence is the proportion of the population under consideration in which a particular disease (or stage of the condition) first occurs during a particular time. We can not only study the incidence of first heart attacks, for example, but also that of second or subsequent heart attacks. Incidence is by definition studied in populations where all the members are in principle candidates for (i.e. are at risk of) the events in question. We can only track down the incidence of first heart attacks, for instance, in a population of individuals who have not yet had a heart attack. Similarly, it makes sense to study the incidence of second heart attacks in a population of persons who have had a heart attack and survived it. Since ascertaining incidence requires events that occur in a population when it is followed over a particular period, it will make a difference whether we study incidence in a cohort or in a dynamic population.

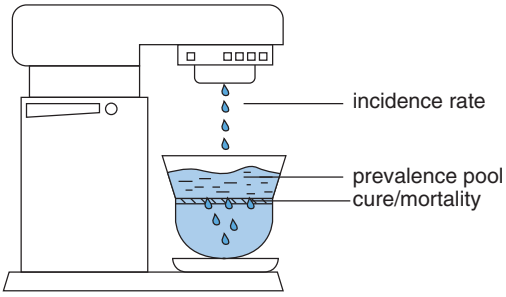


Figure 2.7 Coffee-making: incidence and prevalence

The relationship between incidence and prevalence

The relationship between incidence and prevalence is similar to the way a filter coffee machine works, where the incidence can be compared with the rate at which the water passes through the filter and the prevalence with the amount of water left in the filter. The size of the holes in the filter – i.e the rate of cure or mortality in the population – determines the time it takes for the water to pass through: the faster the time, the lower the prevalence (fig. 2.7).

If we have a low prevalence (P) (less than 10%) and a 'steady state' (the prevalence is constant be-

cause the incidence is in equilibrium with cure and mortality), we can estimate the prevalence from the product of the mean disease duration (T) and the incidence density (ID), discussed below:

$$P = ID \times T$$

In the case of higher prevalences:

$$P = \frac{ID \times T}{1 + (ID \times T)}$$

Note that the denominator in this formula for low prevalences will be approximately 1. At higher values of ID and/or T the corresponding rise in prevalence falls off, as a substantial proportion of the study population then already has the condition, so cannot develop it.

Cumulative incidence: calculating the likelihood of disease

Calculating the cumulative incidence (CI) essentially involves following all the members of a cohort for a particular period. The period could for instance be 1, 5 or 10 years from time t_o – usually a different calendar time for different members of the cohort. At time t_o all members of the cohort are by definition candidates for the event in question, as it is a requirement for membership that the person is at risk at t_o, i.e. susceptible to the condition in question and hence as yet free from it at t_o. The CI comprises the proportion of cohort members (at t_o) who develop the condition during the follow-up. It is therefore a proportion (percentage) with minimum and maximum values of 0 and 1 (0 and 100%) respectively.

$$\frac{\text{Total number of new cases in a particular period}}{\text{Total number of persons in the population at the start of that period}}$$

The epidemiological fraction expressed as a cumulative incidence in a cohort

In order to interpret a cumulative incidence we obviously need to know the length of the period to which it relates. A problem with interpretation is the implicit assumption that there will not be any competing diseases or causes of death, and this assumption is usually not justified, especially in the case of long follow-up periods (see ► par. 2.4).

The proportion of the cohort still at risk of developing the condition under consideration will eventually fall right down to zero, at which point it would be strange to continue looking at the cohort in terms of CI. For this reason the incidence density, discussed below, is often used as the measure of frequency in long-term cohort studies. The length of the period to which the cumulative incidence relates can be either variable or fixed, depending on how the end is defined. Here are some examples of both.

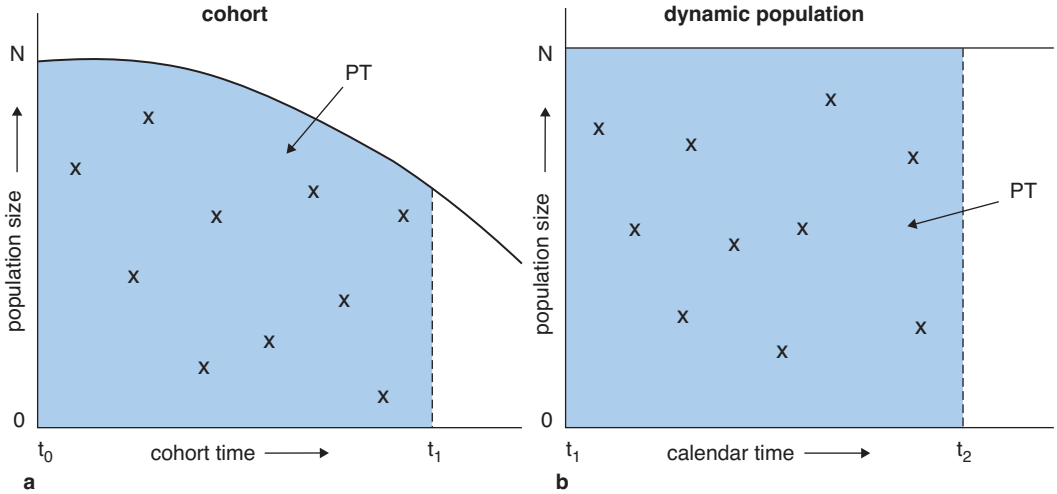
Variable period:
1. Lifetime incidence of prostate cancer
2. Hospital mortality of heart attack patients.

Fixed period:
1. Five-year incidence of second heart attacks in persons with a non-fatal first heart attack
2. Ten-year survival rate following successful surgery of women with breast cancer with metastases in the axillary lymph nodes.

Note that the CI can also be regarded as the individual risk (R) to a member of the cohort at time t_o of undergoing the event in question (developing the disease) during the follow-up period. If the cumulative incidence of second heart attacks within five years is 10%, that is also the risk of having a second heart attack after having (and surviving) a first heart attack. Note that this is a mean risk, which needs to be modified if relevant characteristics of the patient in question are taken into account. The risk will be substantially worse, for instance, if we know that the patient is a persistent heavy smoker aged 65 with high blood pressure.

Incidence density in dynamic populations and with incomplete follow-up of a cohort

If we do not have the opportunity to follow a cohort for some time, we have to fall back on a dynamic population to ascertain the frequency with which new cases occur. The members of a dynamic population are followed over a period of time to ascertain the incidence of the event under consideration. The structure of the candidate population for this event is constantly changing over time, however. Unlike in the case of a cohort, in a dynamic population the occurrence of the event itself is one of the mechan-



CI = n/N
ID = n/PT
CI = ID x ΔT (where ID <0.001 per year)
CI = 1 − $e^{-ID \times \Delta T}$ (where ID remains constant)
CI = 1 − $e^{-\Sigma_i(ID_i \times \Delta T_i)}$ (where ID is variable)

CI: cumulative incidence
ID: incidence density
n: number of new (incident) cases (x)
N: population size
PT: observed person-time (area under the curve)
ΔT: time interval for which CI is estimated from ID
Δt: time interval within which ID remains constant

Figure 2.8 Incidence in a cohort (**a**) and a dynamic population (**b**)

isms by which individuals are removed from the population. For example, persons who have already been admitted to a particular hospital with a heart attack no longer form part of the dynamic population of persons who would be admitted to that hospital in the event of a heart attack. Incidence in a dynamic population is expressed not in relation to the number of members at a particular time (as in the case of a cohort) but in relation to the total person-time observed (usually expressed in person-years). The resulting measure is referred to as the incidence density, incidence rate or hazard rate. The dimension is thus 1 per year (years^{-1}) and the values will be between zero and infinity. The incidence density (ID) is calculated by dividing the number of new cases by the number of person-years observed: for example, observing 15 cases of bladder cancer in a dynamic population where 5,000 person-years are followed yields an ID of 0.003 per year. We cannot tell from this calculation whether an average of 100 persons were followed for an average of 50 years or an average of 500 persons for an average of 10 years. ID is a measure of the speed at which new cases of the disease occur in the population. Sometimes people are only at risk of the conditions under consideration for part of the person-time: the ID of sports injuries is expressed per 100 sports hours and that of accidents at work per 1,000 working days, for instance.

$$\frac{\text{Total number of new cases in a particular period}}{\text{Total number of observed person – years at risk during that period}}$$

The epidemiological fraction expressed as an incidence density

The ID cannot be translated directly into an individual risk during a particular period. Given certain assumptions, however, it is possible to calculate the CI (for a hypothetical cohort) from the ID in a dynamic population. Assuming that the ID is constant over the period in question, in the case of short time intervals the CI (the risk) can be estimated from the product of the ID and the length of the particular

time interval. Figure 2.8 shows how the CI can be calculated from the ID for longer time intervals and/or a variable ID. As we have seen above, it is also sometimes preferable to use ID's in non-hypothetical cohorts with a long follow-up period, since as time passes and an increasing proportion of the cohort members are no longer at risk because they have had the condition or competing diseases or causes of death have occurred, the CI will provide less information. Figure 2.8 shows how the ID in a cohort can be calculated.

A lot of health statistics are incidence figures based on a dynamic population. Recorded frequencies of events are usually expressed in relation to demographic statistics (the number of residents of a particular region by age and gender at a particular time). Examples are: the annual incidence of syphilis in and around Berlin; the monthly incidence of influenza among over-65s in Hamburg; the annual incidence of breast cancer in women aged over 40 in Sweden. With statistics of this kind it is not clear whether all the recorded events actually occurred in the dynamic population for which the incidence is calculated: for example, patients who live elsewhere could end up in a particular hospital by chance (while travelling or on holiday) or intentionally (because the hospital specializes in the condition in question). The size and structure of the population to which the recorded frequencies relate can also fluctuate in time. A dynamic population is usually not completely stable, and we have to rely on the reliability and validity of the demographic data available and when they were collected, since the individuals in the dynamic populations used for health statistics are not followed individually. One has to be aware that estimates of the number of person-years observed could therefore depend on assumptions that are sometimes unverified.

2.5.3 Mortality rates: a special type of incidence

Just as we saw in the previous section with regard to measures of frequency, mortality can be studied in either a cohort or a dynamic population. Mortality is the incidence of death, so everything we have said earlier about calculating the cumulative incidence and incidence density applies to the corresponding mortality rates as well. Analytical epidemiology uses both methods, depending among other factors on the study design adopted (see also ▶ chaps 4 and 6). In descriptive epidemiology questions are often (but not always) formulated in terms of CI. At first sight this seems strange, and on close inspection it is usually incorrect, since the basis is usually the mortality recorded in a dynamic population.

Mortality is probably the indicator of public health most commonly used, the idea being that where the death rate is low the population must be healthy. This argument is of course not always correct. The validity of mortality as a health indicator depends on such things as the type of health problems under consideration. Conditions that are not often fatal (i.e. have a low case fatality rate) but can cause substantial long-term limitations and disabilities cannot be characterized effectively using this indicator. The popularity of mortality as a health indicator at the aggregate level can be explained by the fact that recording deaths is reliable, relatively easy and straightforward. Many countries have indeed been doing this for some considerable time, and comparing death rates is often the only way to compare the health of different countries. Various measures of mortality can be found in publications on descriptive epidemiological research and health statistics, some of which are discussed below.

$$\frac{\text{Total number of deaths in a particular period}}{\text{Total number of persons in the population from which these deaths originate at the start of that period}}$$

The epidemiological fraction expressed as a mortality, in this case the cumulative incidence of mortality

Crude mortality rates

A crude mortality rate describes the number of deaths during a particular period (usually one year) in a particular population. It is usually expressed per 100,000 persons in the population in question. These rates can be calculated for different periods and different populations. Table 2.2 shows an example. The crude mortality rate in index population A is (50 + 4)/(10,000 + 1,000) = 54/11,000 = 491/100,000. The crude mortality rate in index population B is (5 + 40)/(1,000 + 10,000) = 45/11,000 = 409/100,000. Thus the crude mortality rates in the two

Table 2.1 A selection of age-specific measures of mortality, as defined by the World Health Organisation

perinatal mortality	stillbirth or death to live births within the first 7 days of life after a pregnancy duration of at least 22 completed weeks
neonatal mortality	death occurring during the first 4 weeks after birth
early neonatal mortality	death between 0 – 6 days after birth
late neonatal mortality	death between 7 – 27 days after birth
infant mortality	death occurring within the first year of life
under-five mortality	death occurring within the first 5 years of life
post neonatal mortality	death between 1 – 59 months
adolescent mortality	death between 10 and 19 years

Table 2.2 Example of direct standardization of mortality rates (expressed per 100,000 per year)

age		index population A	index population B	standard population
young	deaths	50	5	55
	size	10,000	1,000	11,000
old	deaths	4	40	44
	size	1,000	10,000	11,000
total	deaths	54	45	99
	size	11,000	11,000	22,000
crude mortality rate		491/100,000	409/100,000	450/100,000
standardized mortality rate		450/100,000	450/100,000	450/100,000

index populations are different. In this example the age distribution between the two populations is not the same, so it is not really correct to compare the crude mortality rates directly.

Other types of age-specific mortality rate are shown in tab. 2.1. These too are calculated over a particular period and expressed e.g. per 10,000 or 100,000 persons in the population in question. The WHO regards the age-specific mortality rates in the first year of life shown in tab. 2.1 as a reliable indicator of the quality of healthcare as a whole, especially in developing countries. Maternal mortality is often mentioned in the same breath: this is the rate of deaths among women due to complications during pregnancy or during or soon after delivery. Maternal mortality is in fact not an age-specific but a cause-specific death rate.

Age-specific mortality

As mortality is closely correlated to age, mortality rates are usually specified for particular age groups. Five, 10 and 15-year age groups are generally used, depending on the size of the population for which they are calculated and the purpose.

Direct and indirect standardization

In order to enable us to compare the health of different populations with different age structures, we will want to aggregate the age-specific mortality rates for each of the populations into a single number. This is in effect what happens when we calculate a standar-

dized mortality rate (i.e. standardized for age). The age-specific mortality rates are aggregated in such a way as to enable the mortality rates for the populations to be compared. There are two ways of standardizing rates.

Direct standardization for age involves applying the age-specific mortality rates for the population in question (the index population) to the age distribution of a selected standard population. The result is the expected mortality in the standard population if the age-specific mortality rates for the index population were to apply. The example in tab.2.2 compares two index populations using a standard. Here the standard has simply been calculated as the sum of the two populations. The expected number of deaths in the standard population based on the age-specific mortality in index population A is (50/10,000 × 11,000) + (4/1,000 × 11,000) = 99. The standardized mortality rate in this population is therefore 99/(11,000 + 11,000) = 450/100,000. The standardized mortality rate for population B is calculated similarly: (5/1000 × 11,000) + (40/10,000)]/(11,000 + 11,000) = 450/100,000. After standardizing for age the differences in mortality rates disappear completely in this example. It goes without saying that in other examples standardization can result in the differences in mortality rates going down or up or even being reversed. In the latter case, the population with the highest crude mortality has the lowest mortality after standardization and vice versa.

When comparing different countries or regions within a country the sum population is usually taken as the standard population, as in tab. 2.2. When studying trends over time the demographic structure in a year at the start or in the middle of the set of mortality rates is usually taken as the standard. Standardization can in principle be applied for all sorts of other variables besides age, but this is no longer common practice, as the development and spread of computers has made it easy to perform various kinds of multivariable analysis, which enables adjustments to be made simultaneously for differences in multiple variables.

In order to carry out direct standardization we need to have sufficient information on the numbers of deaths in each age group in the index populations. The numbers must not be too small, as this will make the estimates of the expected numbers of deaths too imprecise. This was not a problem in the example in tab. 2.2, but it often is a problem in practice. An alternative way of comparing mortality rates in two populations is therefore to apply the age-specific mortality rates found in the standard population to the age structure of an index population, as sufficiently precise information is often available on this. The result is the expected mortality in the index population if the age-specific mortality rates for the standard population were to apply. This method is referred to as indirect standardization. We can carry out an example calculation using the numbers in tab. 2.3. Let us take a male population in 2014 as the standard population. Based on the number of deaths in that year in each age group we calculate the mortality per 100,000 men for each group. These mortality risks are shown in column 2 of tab. 2.3. Column 3 of the table shows the numbers of men in each age group in the index population for 2012. Multiplying the mortality risks with the numbers in columns 2 and 3 we arrive at column 4, the number of deaths for each of these age groups that we could have expected in 2012 if the mortality risks for 2014 had been applied. The sum of these numbers – 9,263 – is the total number of deaths expected for 2012 if the mortality risks were the same in that year as in 2014 (the indirect standardized mortality rate). The crude mortality rate in 2012 was 13,319. The crude mortality rate in the index population divided by the indirect standardized mortality rate gives us the standardized mortality ratio (SMR). The SMR in the example is 13,319/9,263 = 1.44. This SMR is indicative of 44% 'excess mortality' in the index population compared with the standard.

In principle indirect standardized mortality rates, and therefore SMRs, for different index populations should not be compared, as they relate to different populations. If the differences in age structure are not too great, however, the resulting error will remain within bounds.

Cause-specific mortality

As an indicator mortality is far too rough and ready for epidemiological research into the causes and effects of disease. Mortality rates are therefore often provided separately for different causes, either standardized for age or not. Such figures are usually based on the records of causes of death that are re-

Table 2.3 Example of indirect standardization of mortality rates (expressed per 100,000 per year)

age (years)	mortality per 100,000 men in 2014	number of men in 2012	number of deaths expected for 2012
0–24	0.1	2,990,488	3
25–29	0.8	552,556	4
30–34	4.7	444,516	21
35–39	15.5	402,241	63
40–44	37.8	387,333	147
45–49	71.5	365,953	262
50–54	132.2	335,549	444
55–59	249.2	303,004	755
60–64	381.1	269,132	1,026
65–69	630.0	223,720	1,409
70–74	940.3	164,614	1,548
75–79	1378.2	111,897	1,542
80–84	1,793.8	63,661	1,142
85+	2,550.3	35,207	898
total	*156.2*	*6,649,871*	*9,263*

quired in many countries. We have already mentioned one example of cause-specific mortality, namely maternal mortality. Other examples are cancer mortality, deaths due to road accidents, and deaths from suicide. How valuable such figures are will depend very much on the validity and reliability of the records on the cause of death in question. In many cases the cause of death is not immediately obvious. First, the correct diagnosis is not always made before death, and autopsies (pathological investigations after death) are not often carried out. Secondly, the primary cause of death recorded is the result of often complex reasoning. The condition that caused the series of events resulting in death needs to be selected, and this selection is sometimes open to question, especially in the elderly, who often suffer from a variety of chronic conditions (comorbidity). The ICD classification, or a simplified version thereof, is generally used when recording causes of death.

Case fatality rate

The case fatality rate is the proportion of the incidence of patients with a particular condition who die of that condition during a particular period. This measure reflects the severity of the condition and the efficacy of the healthcare provided. The case fatality rate is highly dependent on the severity of the condition, however. For this reason the fact that the case fatality rate of home births is lower than that of hospital deliveries is not a valid argument in favour of giving birth at home, for example, as a risk of complications is normally regarded as an indication for hospital delivery. In other words, hospital deliveries are more risky on average.

In effect the case fatality rate is the cause-specific mortality among the incident cases of the condition in question. It can also relate to high-risk interventions such as invasive diagnostic or surgical procedures. The case fatality rate is expressed as a cumulative incidence of cause-specific mortality, but the period to which it relates is not always specified.

Take the case fatality rate of heart and lung transplants, for instance: this can in theory relate either to mortality due to the operation itself or to mortality from complications in the operating theatre, during subsequent hours or days, during the stay in hospital, during the year after the operation, or during the entire remainder of life. As this example shows, case fatality rates are not always easy to interpret.

$$\frac{\text{Total number of deaths from a particular disease}}{\text{Total number of incident patients with the disease}}$$

The epidemiological fraction expressed as a case fatality rate

Proportional mortality rate

The proportional mortality rate is the proportion of total deaths in a population due to a particular condition. It can be expressed not only in relation to total mortality but also in relation to particular groups of causes of death that include the condition under consideration. For instance, we can speak of 'proportional cancer mortality': mortality from e.g. prostate cancer can be expressed either in relation to total deaths or in relation to total deaths from cancer.

Proportional mortality rates are particularly difficult to interpret, as differences in proportional mortality can be due not only to differences in the frequency and fatality of the condition under consideration but also to differences in the frequency and fatality of all other conditions. For example, the proportional mortality due to accidents is approximately 40% for one to four-year-olds and approximately 2% for 70 to 75-year-olds, and yet far more elderly people than children die due to accidents. Because of problems of interpretation of this kind cause-specific mortality rates are generally to be preferred. These, however, require information not only on the deaths but also on the size of the population in which those deaths occur. As that information is not always available we may need to make use of proportional mortality rates, as was done in fig. 2.1.

$$\frac{\text{Total number of deaths from a particular disease}}{\text{Total number of deaths in the population}}$$

The epidemiological fraction expressed as a proportional mortality rate

Life expectancy

An alternative, somewhat more positive, way of representing mortality rates is in the form of life expectancy. This is usually represented as the mean number of years of life expected at birth. It is calculated based on the age-specific mortality rates for the successive age groups as found at the time of birth. The assumption, then, is that these age-specific mortality rates will not change during the lifetime of a group born in the same year. If these figures improve, e.g. as a result of improvements in healthcare, the mean life expectancy calculated at birth will in effect be an underestimate. The remaining life expectancy at any age can be calculated similarly by applying the age-specific mortality rates following that age found at that time. Interestingly, the mean estimated age at death rises with the age at which the remaining life expectancy is estimated: a man who had a life expectancy of 73 years at birth may have a further life expectancy of 14 years at age 65, for instance. The reason for this is that by the age of 65 he has already survived for 65 years, thus increasing the likelihood of getting older. Out of a hundred neonates fewer will reach 80 than out of a hundred 65-year-olds. This is also the case if the age-specific figures do not change during the lifetime of the persons concerned.

Survival rates

The complement of the case fatality rate is the survival rate: the proportion of patients with a particular condition who are still alive after a certain period. Common examples are the five-year and ten-year survival rates of cancer patients after treatment. Here again there is a strong dependency on the respective part of the disease spectrum: lung cancer patients who are diagnosed by chance have a far better five-year survival rate than those who go to a respiratory physician with clear symptoms, for example.

We can take calculating the survival rate a step further by drawing a survival curve, which plots survival rate against time since diagnosis (or some other logical starting point). Figure 2.9 shows an example of a survival curve for Diabetes Mellitus patients with different disease durations. We can use the survival curve to read off the survival rate for any time period (if observations have been made for that per-

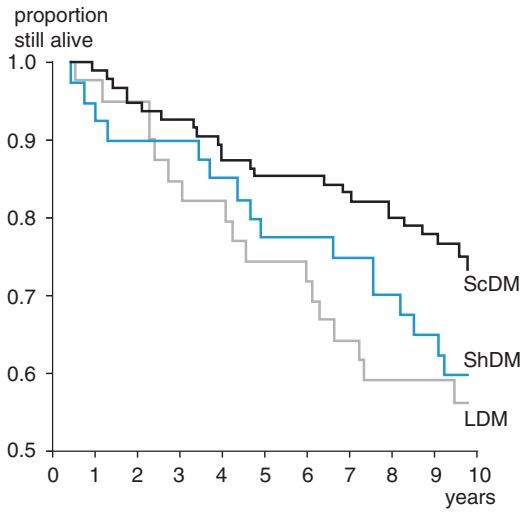


Figure 2.9 Survival curves for patients with Diabetes Mellitus (DM) by disease duration. ScDM: patients detected by means of screening. ShDM: patients with a short disease duration of 6.2 years or less. LDM: patients with a long disease duration of over 6.2 years

iod). Survival rates are the inverse of the cumulative incidence of mortality.

There are various methods for estimating survival curves. A method as described by Kaplan-Meier is often used in epidemiology. This involves recalculating the proportion of survivors each time a member of the cohort dies. The advantage of the survival curve approach is that it shows not only the proportion of the phenomenon being measured at a particular time but also the time that elapses before that phenomenon occurs (in this case mortality, but it can also be used for other outcomes such as disability, cure or pregnancy).

Quality-adjusted and disability-adjusted life years

We can consider not only length of life – quantity – but also quality of life. The aim of healthcare, after all, is not only to prolong life but also to a large extent to make it more pleasant. We therefore need to assign a value to the remaining years of life expected: this weighting factor, or utility, usually ranges from 0 to 1. A utility of 0.75 suggests that we regard four years of life spent in the state in question as interchangeable with three years of life in full health. This enables us to calculate life expectancy adjusted for quality, expressed as quality-adjusted life years (QALYs). Various methods have been developed to estimate utility, but to consider them would be beyond the scope of this book. The QALY concept is increasingly gaining popularity for use at the aggregate level, e.g. for cost-utility analyses. In the area of public health similar but slightly simpler measures are used, for example healthy life expectancy and disability-adjusted life years (DALYs). DALYs is a measure of the number of years lost due to ill-health: it is the sum of the number of years lost due to premature mortality (lost years of life) and the healthy years 'lost' due to living with a disease. We can use DALYs to compare diseases in terms of their effect on public health, as they are calculated based on four important factors: the number of people suffering from the disease, the severity of the disease, mortality from it and the age at which mortality occurs. Weighting factors are used to 'weight' the years spent with the disease for the severity of the disease so as to make them comparable with years of life lost due to mortality. For example, a weighting factor of 0.5 for a particular disease means that one year of life with that disease is regarded as equivalent to 0.5 years lost due to premature mortality.

There are other variations on this theme. Healthy life expectancy is the mean number of years of life that people can expect to spend 'in good health'. This measure of health combines life expectancy and quality of life in a single number, using indicators for particular kinds of healthy life expectancy: life expectancy in good self-rated health (HLE), disability-free life expectancy (DFLE) and life expectancy in good mental health (MHLE).

2.6 Continuous measures of health and disease

The dependent variable in epidemiological research, disease, is often expressed on a dichotomous scale (disease versus no disease). The distribution of a disease in the population can then be described simply based on the numbers (proportions) of people with the disease. Some aspects of health and disease, however, are not measured on a dichotomous scale but on a continuous scale, for example blood pressure, lung function, hearing loss, cognitive function,

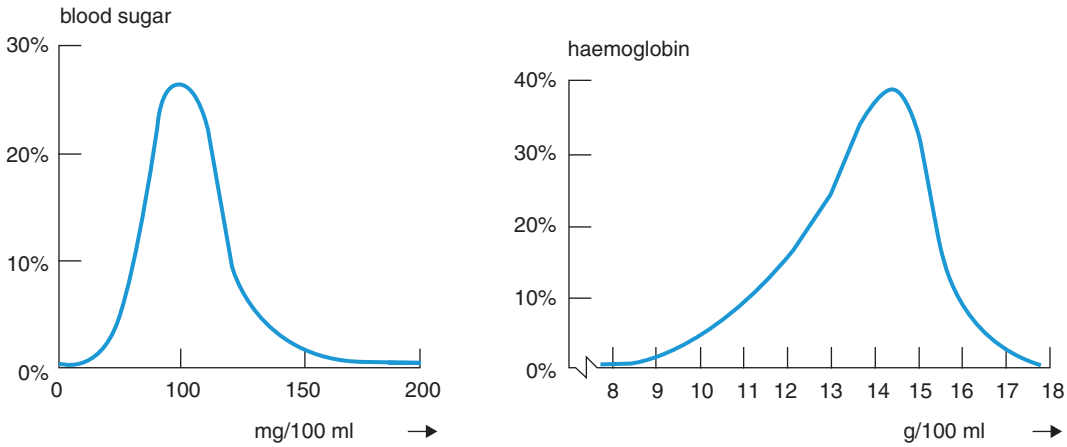


Figure 2.10 Examples of frequency distributions of clinical parameters in the general population: blood glucose and haemoglobin

severity of pain and quality of life. Each of these continuous variables could of course be reduced to a dichotomous variable (e.g. hypertension yes/no, quality of life good/poor), but that would entail loss of information, which is not always desirable. In such cases the continuous variable can itself be incorporated in the epidemiological function as the dependent variable: the left-hand side of the function is then not the risk of disease (calculated as a proportion) but the mean of the variable in question. The form of the function also changes generalized linear model, as the mathematical equation often becomes a model instead of a logistic model (see ▶ par. 3.4). ▶ Case 2.1 gives a simplified example of a linear regression function of this kind and how it can be used to study the association between determinants (in this case various prognostic factors) and outcome (change in functional limitations) in people with acute or chronic low back pain.

The distribution of a variable that is measured on a continuous scale can be summarized succinctly in two measures, one for the central value around which the observed results cluster and one for the spread in the observations.

Examples of central measures:

- The mean, i.e. the sum of the measurement results divided by the number of measurements
- The median, i.e. the value at which 50% of the measurement results are higher and 50% are lower
- The mode i.e. the measured value that occurs most frequently.

Examples of measures of spread:

- The standard deviation (SD) from the mean measured value: this is an approximation of the absolute value of the mean difference between the individual measurement results and the mean value of these results.
- The interpercentile range: an area marked by two measurement results containing a certain percentage of the observations.
- The range: the distance between the highest and lowest measured values.

What central measures and what measures of spread can be calculated will depend on the scale on which the particular variable has been measured and the shape of the distribution curve of all measurements. Figure 2.10 shows an approximation of the distribution curves of two biological variables – blood glucose and blood haemoglobin.

Although the natural distribution of many clinical variables approximates the normal distribution (the Gaussian curve) fairly closely, it rarely coincides precisely. The normal distribution is a theoretical distribution with unique statistical properties, the most important of which is symmetry: i.e. a fixed percentage of the observations are in the area delineated between one, two or more standard devia-

Table 2.4 Analysis of the predictors 'radiating pain' and 'fear of pain'

predictor	acute low back pain (n = 258)		chronic low back pain (n = 668)	
	crude regression coefficient (β_1)	adjusted[a] regression coefficient (β_1)	crude regression coefficient (β_1)	adjusted[a] regression coefficient (β_1)
radiating pain	3.18 (1.85, 4.51)	0.82 (–0.46, 2.10)	3.60 (2.61, 4.59)	0.73 (–0.09, 1.55)
fear of pain	3.46 (1.93, 4.99)	1.73 (0.31, 3.15)	2.97 (1.77, 4.18)	1.03 (0.04, 2.02)

[a]Adjusted for age, gender and severity of back pain at the time of inclusion.

tions from the mean value (X ± 1 SD = approx. 68%; X ± 2 SD = approx. 95%; X ±3 SD = approx. 99%). Many biological variables, however – e.g. blood glucose and haemoglobin –, have a frequency distribution that is unimodal (one peak value) and asymmetrical (skewed due to an excess of high or low values).

A prerequisite for using a continuous variable as the outcome in many epidemiological functions is that it should roughly follow a normal distribution. Skewed distributions can often be 'normalized' by applying a transformation to the original value for each individual, either a logarithmic transformation (ln = natural logarithm), a root transformation or some other type of transformation, depending on how skewed the distribution is.

Case 2.1 Predicting the course of symptoms in people with acute versus chronic low back pain

A study by Grotle et al. investigated whether there were differences in factors (prognostic indicators) predictive of an unfavourable course of symptoms in people with acute low back pain and people with chronic low back pain. The outcome measure was the score after twelve months on the Roland Morris Disability Questionnaire (RMDQ, range 0–24), a 24-item questionnaire that measures the severity of functional limitations due to back pain. The outcome was regarded as continuous and the association with various predictors was examined using linear regression analysis. The study was carried out on participants in two prospective cohort studies into the course of low back pain recruited from general practices in Staffordshire in England, of whom 258 reported transient (acute) low back pain and 668 chronic low back pain. A substantial number of potential predictors were examined, but tab. 2.4 shows the results of analysing two factors: the presence of radiating pain (to below the knee) and a high score for fear of pain.

The regression equation (for each of the predictors and in the two subgroups) will have looked like this:

$$RMDQ_{12m} = \beta_0 + \beta_1(D_x)$$

where D_x represents the determinant, which is often coded as 0 for participants who score negative for the determinant (i.e. who do not report radiating pain or score low for fear of pain) and as 1 for participants who do report radiating pain or fear of pain. In that case the intercept (β_0) indicates the outcome (the RMDQ score after twelve months) for participants who score negative (0) for the determinant. The regression coefficient (β_1) indicates the difference in outcome between people who score positive for the determinant and those who score negative.

From the table we can read off that the difference for radiating pain is 3.18 points (confidence interval 1.85 to 4.51) among participants with acute back pain and 3.60 points (2.61, 4.59) in the case of chronic back pain. The presence of radiating pain at the time of inclusion, then, is correlated to more severe functional limitations after twelve months, and this association is approximately equally strong among participants with acute and participants with chronic

low back pain. The difference is statistically significant: the confidence interval does *not* include zero (no difference).

The investigators state that various factors could be confounders of this association, however. The difference in limitations could be explained partly by the fact that people with radiating pain also have a higher pain score at the time of inclusion, or that they are older, or that more of them are women. The investigators have therefore expanded the regression equation to adjust for potential confounding by age, gender and severity of pain at the time of inclusion:

$$\mathrm{RMDQ}_{12\mathrm{m}} = \beta_0 + \beta_1(\mathrm{Dx}) + \beta_2(\mathrm{age}) + \beta_3(\mathrm{gender}) + \beta_4(\mathrm{pain\ score})$$

The adjusted regression coefficient β_1 indicates the difference in outcome taking the influence of these confounders into account. The table shows that the association between radiating pain and functional limitations after twelve months is much smaller and no longer statistically significant in participants with acute as well as participants with chronic low back pain (the difference is less than 1 point and the confidence interval now includes zero).

The same analysis was carried out on the determinant 'fear of pain'. Even after adjusting for confounding, this determinant was still found to be statistically significantly correlated to the severity of limitations after twelve months, although the differences after adjustment were smaller. Fear of pain was found to be a stronger determinant in the case of participants with acute low back pain (adjusted difference of 1.73 on the RMDQ) than in the case of those with chronic low back pain (adjusted difference of 1.03).

Recommended reading

Huber M, Knottnerus JA, Green L, et al. How should we define health? BMJ, 2011, 343, p. d4163.

International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization, 2001.

International Classification of Diseases. 10th revised ed. Geneva: World Health Organization, 2010.

McDowell I. Measuring health: A guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press, 2006.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins, 2012.

Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. 4th ed. New York: Oxford University Press, 2008.

Source references (fig. 2.1, 2.2 and 2.4, and case 2.1)

Grotle M, Foster F, Dunn K, Croft P. Are prognostic indicators for poor outcome different for acute and chronic low back pain consulters in primary care? Pain, 2010, 151(3), pp. 790–7 (Case 2.1).

- http://bit.ly/1zZ5VWA (fig. 2.1).
- https://gis.cancer.gov/geoviewer/app/ (fig. 2.2).
- http://bit.ly/2hP0r33 (fig. 2.4).

Association

3.1 The epidemiological function describes the association between disease frequency and determinants

As we have seen in previous chapters, the epidemiological function represents the question being addressed by an epidemiological study as a mathematical formula. The epidemiological function links the disease outcome (frequency) to one or more determinants (also referred to as 'exposures'):

$$P(O) = f(D_i)$$

Case 3.1 gives a classical example of an epidemiological function and how it can be used. ▶ Chapter 2 focused on the left-hand side of the equation, the disease outcome. This chapter looks at the relationship between the left and right-hand sides of the equation. Let us now consider the relationship between the specific determinants (D_i) and the disease outcome P(O).

Case 3.1 Risk factors for cardiovascular disease

One of the first large-scale epidemiological studies into the determinants of cardiovascular disease was the Framingham Heart Study, named after the small town of Framingham in the United States, where it began in 1949. The Framingham Heart Study has become a true classic example in epidemiology. The entire adult population was asked to take part in regular health checks. Their blood pressure, serum cholesterol and body height and weight, were measured annually or biennially and with detailed questionnaires various lifestyle aspects (smoking, alcohol consumption, etc.) were assessed. These measurements, plus data from local hospitals and records of deaths, enabled the incidence of and deaths from various types of cardiovascular disease in the population to be determined precisely. The study population was monitored for a number of decades. The following data relate to the first twelve years' follow-up of the cohort of 2,187 men and 2,669 women (aged 30-62) who did not have coronary heart disease at the time of the first health survey. Over the twelve-year period 258 men and 129 women developed coronary heart disease, representing a cumulative incidence of 11.8% for the men and 4.8% for the women. A logistic regression analysis was used to link disease frequency to various risk factors. This function was as follows:

$$P(O) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k)}}$$

To put it in words, the probabilty of getting a myocardial infarction (P(O)) can be estimated based on a mathematical logistic equation with a constant b_0 and a combination of various determinants x_i, weighted by their associated coefficients b_i.

Table 3.1 shows the various determinants with their coding and regression coefficients separately for men and women. Using the formula and the table we can calculate, for example, that a 60-year-old woman with a serum cholesterol level of 250 mg/100 ml, systolic blood pressure of 140 mmHg, relative weight of 140 kg and a haemoglobin level of 120 g/L who smokes one pack of cigarettes a day but whose ECG is normal has an 85% risk of myocardial infarction within twelve years. If the same woman did not smoke but otherwise displayed all the same characteristics, she would have an 83% risk of myocardial infarction within twelve years. Try to calculate this for yourself, using the formula and the data in the table.

Let us first consider the various kinds of determinants and the various types of epidemiological functions.

In epidemiological research, as we saw in ▶ chap. 2, the outcome studied is always an aspect of disease or health, and the presence or absence of disease or a stage of disease is thus regarded as the resultant of one or more factors acting upon the individual, the determinants. In etiological research determinants

Table 3.1 Determinants of cardiovascular disease and associated weights in the Framingham Heart Study

determinant	unit of measurement	coefficient for men	coefficient for women
constant	–	$b_0 = -10.90$	$b_0 = -12.59$
x_1 = age	years	$b_1 = 0.071$	$b_1 = 0.076$
x_2 = serum cholesterol	mg/100 ml	$b_2 = 0.011$	$b_2 = 0.006$
x_3 = systolic blood pressure	mmHg	$b_3 = 0.017$	$b_3 = 0.022$
x_4 = relative weight	100 × weight/standard weight	$b_4 = 0.014$	$b_4 = 0.005$
x_5 = haemoglobin	gram/L	$b_5 = -0.084$	$b_5 = 0.036$
x_6 = cigarette smoking	0 = none, 1 = < 1 pack per day, 2 = 1 pack per day, 3 = > 1 pack per day	$b_6 = 0.361$	$b_6 = 0.077$
x_7 = abnormality on ECG	no = 0, yes = 1	$b_7 = 1.046$	$b_7 = 1.434$

Table 3.2 Examples of determinants of health and disease

behaviour	smoking	drinking	exercise	diet	medication
biology	blood pressure	DNA	cholesterol	age	sex
environment	occupational exposure	social support	socioeconomic class	residential environment	access to care

can be divided into three categories, broadly speaking: behaviour, biology and environment. Table 3.2 gives a few examples from each of these categories.

If we wish to use targeted measures to prevent people developing disease we need to know the etiological determinants of health and disease, and what these are will differ from one population to another. In countries with a low average income the main determinants that can be influenced are such things as malnutrition, unsafe sex and poor hygiene, whereas in countries with a middle to high average income smoking, being overweight and a lack of physical activity are more important determinants of disease (see tab. 3.3). Whether a determinant can be influenced and whether this will actually reduce the risk needs to be determined by research, and research of this kind is vital to effective prevention. This will be discussed in ▶ chap. 6 and 10.

Sometimes determinants are used solely to identify individuals and groups who are at increased risk of particular health problems, leaving aside the question of whether the given factor is a cause of the health problem under consideration. A familiar example is the relationship between differences in socioeconomic status and disease or mortality. Although education, income or status do not as such cause disease or mortality, these factors are strongly correlated with the causal determinants of health problems. A description of socioeconomic health differences can therefore help to identify the areas, neighbourhoods or groups within a population that are most in need of preventive and curative healthcare. Another example of the use of determinants is in life insurances, where the amount of premium payable is increasingly geared to the applicant's profile in terms of health determinants (his or her 'risk profile').

Table 3.3 Ranking of the main modifiable determinants of disease for low, middle and high-income countries

	risk factor	deaths (millions)	percentage of total		risk factor	deaths (millions)	percentage of total
world				**low-income countries**			
1	high blood pressure	7.5	12.8	1	underweight in children	2.0	7.8
2	smoking	5.1	8.7	2	high blood pressure	2.0	7.5
3	high blood sugar	3.4	5.8	3	unsafe sex	1.7	6.6
4	physical inactivity	3.2	5.5	4	contaminated water	1.6	6.1
5	overweight and obesity	2.8	4.8	5	high blood sugar	1.3	4.9
6	high cholesterol	2.6	4.5	6	smoke in the home	1.3	4.8
7	unsafe sex	2.4	4.0	7	smoking	1.0	3.9
8	alcohol consumption	2.3	3.8	8	physical inactivity	1.0	3.8
9	underweight in children	2.2	3.8	9	poor breastfeeding	1.0	3.7
10	smoke in the home	2.0	3.3	10	high cholesterol	0.9	3.4
middle-income countries				**high-income countries**			
1	high blood pressure	4.2	17.2	1	smoking	1.5	17.9
2	smoking	2.6	10.8	2	high blood pressure	1.4	16.8
3	overweight and obesity	1.6	6.7	3	overweight and obesity	0.7	8.4
4	physical inactivity	1.6	6.6	4	physical inactivity	0.6	7.7
5	alcohol consumption	1.6	6.4	5	high blood sugar	0.6	7.0
6	high blood sugar	1.5	6.3	6	high cholesterol	0.5	5.8
7	high cholesterol	1.3	5.2	7	low fruit and vegetable consumption	0.2	2.5
8	low fruit and vegetable consumption	0.9	3.9	8	air pollution	0.2	2.5
9	smoke in the home	0.7	2.8	9	alcohol consumption	0.1	1.6
10	air pollution	0.7	2.8	10	exposure at work	0.1	1.1

Table 3.4 The epidemiological two-by-two table

	cases	controls
exposed	a	b
non-exposed	c	d

3.2 Measures of association for dichotomous health outcomes

Many measures of association in epidemiology are based on comparing disease frequencies (incidence, prevalence: see ▶ chap. 2) between two categories of a dichotomous determinant (exposed or non-exposed). This can be simply visualized in a two-by-two table, also known as a contingency table (see tab. 3.4).

More complex situations in etiological epidemiology will be considered in ▶ chap. 4 and 5, and the way in which the association between diagnostic or prognostic determinants and disease is quantified will be discussed in ▶ chap. 9.

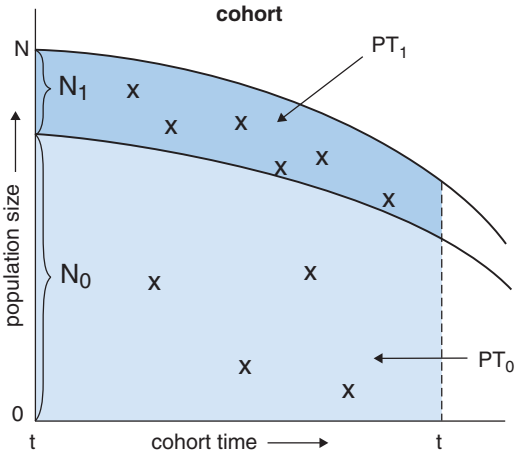


Figure 3.1 Incidence in exposed and non-exposed persons in a cohort study
$CI_1 = n_1 / N_1$ $ID_1 = n_1 / PT_1$
$CI_0 = n_0 / N_0$ $ID_0 = n_0 / PT_0$
CI: cumulative incidence
ID: incidence density
n: number of new incident cases (x)
N: population size
PT: observed person-time (area under the curve)
1: in exposed persons
0: in non-exposed persons

3.2.1 The attributable risk indicates the difference in incidences

Figure 3.1 shows a study population (cohort) divided into two determinant categories, exposed (E_1) and non-exposed (E_0). It also shows how the cumulative incidence (CI) and incidence density (ID) can be calculated for both subgroups.

The obvious way of quantifying the effect of the determinant on the disease outcome is to subtract one incidence from the other, thus yielding the attributable risk (AR). If the exposed persons (E_1) and non-exposed persons (E_0) are comparable in all other respects, this is the additional risk (cumulative incidence) that is attributable to exposure to the determinant.

$$AR = I_1 - I_0$$

For the sake of simplicity, the examples in this chapter are based on the assumption that there is a causal relationship.

Suppose that epidemiological research into serum cholesterol levels and cardiovascular mortality shows that the AR is 26 per 10,000 per ten years, this means that, among persons with a high cholesterol level, 26 per 10,000 will die of a cardiovascular condition within ten years due to their high cholesterol level. In other words, within this group the cardiovascular mortality over ten years would have been 26 per 10,000 lower if everyone had had a low cholesterol level (instead of a high one). The attributable risk is also known as the risk difference (RD). When calculating the attributable risk the incidence (I) can be measured either as the CI or the ID, thus yielding the cumulative incidence difference (CID) and the incidence density difference (IDD).

If the determinant is an intervention the reciprocal (1/AR) can also be used to derive another measure of association: this expresses the average number of persons who need to be treated in order to achieve the desired outcome (e.g. cure) in one person. This measure is referred to as the number needed to treat (NNT = 1/AR). An NNT of 385

(= 10,000/26) in the above example means that 385 persons with a high cholesterol level would need to be treated successfully for ten years in order to prevent one case of cardiovascular mortality.

Table 3.5 shows among other things how the attributable risk can be calculated based on cumulative incidences and incidence densities. The CID is dimensionless and applies to a specified period. The IDD is usually expressed in terms of year^{-1} (= per year). The shorter the follow-up period and the rarer the condition under consideration the closer the CID and IDD will approximate each other numerically. It is standard practice to subtract the lowest risk from the highest so that AR is always a positive number.

3.2.2 The relative risk indicates the ratio between incidences

Another way of representing the determinant's effect on the disease outcome is by dividing the incidence in the exposed subcohort by that in the non-exposed subcohort, thus yielding the relative risk (RR):

$$RR = \frac{I_1}{I_0}$$

A RR of cardiovascular mortality of 2.0 for persons with a high cholesterol level compared with persons with a low cholesterol level means that persons with a high cholesterol level run a risk of cardiovascular mortality that is twice as high. When calculating the RR the incidence (I) can be measured either as the CI or the ID. The terms risk ratio and rate ratio are both abbreviated as RR, although strictly speaking they should be referred to as the cumulative incidence ratio (CIR) and incidence density ratio (IDR) respectively. As a result these terms are often treated – not entirely correctly – as synonyms. Table 3.5 shows how the RR can be calculated from cumulative incidences and incidence densities. Both the CIR and the IDR are dimensionless. Here again, the shorter the follow-up period is and the rarer the condition under consideration, the more the CIR and IDR will approximate each other numerically.

The RR can have values between zero and infinity, where $0 < RR < 1$ suggests a protective effect and $RR > 1$ a risk-increasing effect; $RR = 1$ means that there is no association between the determinant and the occurrence of the disease. There is symmetry between $RR > 1$ and $RR < 1$, as any risk factor can be turned into a protective factor by removing it and vice versa. In the previous example, for instance, the risk of cardiovascular mortality was halved by removing the high cholesterol level.

3.2.3 The hazard rate ratio is the ratio of two incidence densities

Another term for incidence density is hazard rate. The ratio between two hazard rates, the hazard rate ratio, is thus quite simply the IDR already mentioned. The reason for nevertheless devoting a separate section to this is that hazard rate ratios are used when comparing survival curves (see ▶ par. 2.5.3).

If we imagine a survival curve as a linear descending line, the rate at which the line descends is determined by the hazard rate. We assume here that the hazard rate is approximately constant and therefore yields the mean, the incidence density. If we then want to compare the mean hazard rate of an exposed population with that of a non-exposed population we calculate the hazard ratio, which indicates the ratio between the rates at which the two curves fall, i.e. the relative rate at which the control (no disease) population becomes smaller. The Cox proportional hazards model for the statistical analysis of these survival curves is based on the hazard rates and hazard ratio. We would recommend anyone wanting to find out more about this to read a book on survival analysis.

3.2.4 The odds ratio is a convenient measure (albeit indirect and difficult to interpret) when incidences are not available

Sometimes the presence of a determinant among cases with a defined disease is compared with that among a comparable group of healthy controls. ▶ Chapter 4 discusses this type of epidemiological research (case control studies) in detail. As this does not characterize the entire dynamic population

Table 3.5 Calculating measures of association in a cohort study

	O	$\bar{O}$	
D	n_1		N_1
$\bar{D}$	n_0		N_0
	n_T		N_T

$CI_1 = n_1 / N_1$

$CI_0 = n_0 / N_0$

$CI_T = n_T / N_T$

	O	PT
D	n_1	PT_1
$\bar{D}$	n_0	PT_0
	n_T	PT_T

$ID_1 = n_1 / PT_1$

$ID_0 = n_0 / PT_0$

$ID_T = n_T / PT_T$

$$RR = \frac{CI_1}{CI_0} = CI$$

$$RR = \frac{ID_1}{ID_0} = IDR$$

$$AR = CI_1 - CI_0 = CID$$

$$AR = ID_1 - ID_0 = IDD$$

$$EF_e = \frac{CI_1 - CI_0}{CI_1} = 1 - 1/CIR$$

$$EF_e = \frac{ID_1 - ID_0}{ID_1} = 1 - 1/IDR$$

$$PAR = \frac{CI_T - CI_0}{CI_T} = \frac{p(CIR-1)}{p(CIR-1)+1}$$

$$PAR = \frac{ID_T - ID_0}{ID_T} = \frac{q(IDR-1)}{q(IDR-1)+1}$$

$$p = N_1 / N_T$$

$$q = PT_1 / PT_T$$

O: disease (cases)
$\bar{O}$: no disease (controls)
E: exposed
E: non-exposed
N: cohort/subcohort size
n: number of incident cases
PT: observed person-time
1: in exposed persons
0: in non-exposed persons
T: in the total population
p: proportion of exposed persons
q: proportion of exposed person-time
CI: cumulative incidence
ID: incidence density
CIR: cumulative incidence ratio
IDR: incidence density ratio
CID: cumulative incidence difference
IDD: incidence density difference
RR: relative risk
AR: attributable risk
EF_e: etiologic fraction among the exposed
PAR: population attributable risk

from which the cases have been collected but only looks at a sample, it is not possible to determine incidences and thus calculate any of the measures of association mentioned above. Figure 3.2 shows a situation in which ten cases in a dynamic population were identified and twenty controls were selected from the population at random.

Because there is no information on the entire dynamic population it is not possible in a case control study to calculate incidence densities for exposed and non-exposed persons, but the ratio of exposed to non-exposed persons in the patient group can be compared with that in the control group. This ratio is referred to as the odds, a term taken from English horse racing. For example, if the horse Jumping Jack has a 1 in 6 chance of winning (and therefore a 5 in 6 chance of losing), the bookmakers say that the odds are 1:5 (or 1/5 = 0.2). In the same way, in epidemiology we say that if 1/3 of the cases and 1/10 of the controls have been exposed (i.e. 2/3 and 9/10 respectively are non-exposed), the odds of exposure are 1:2 (or 0.5) for the cases and 1:9 (or 0.11) for the controls. The ratio of two odds is the odds ratio (OR). This measure of association is commonly used in epidemiology. As tab. 3.6 shows, the odds ratio is easy to calculate from the results of a case control study expressed in a contingency table.

Interpreting the OR, however, is far less straightforward. Fortunately, it can usually be interpreted as a relative risk or rate ratio (RR).

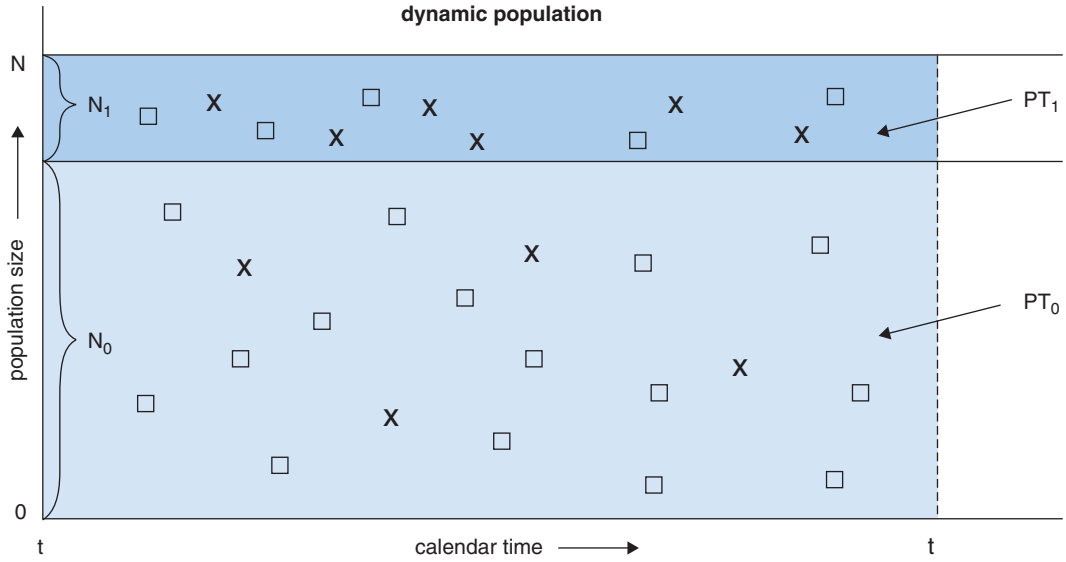


Figure 3.2 Incidence in exposed and non-exposed persons in a dynamic population as the basis for a case control study
$ID_1 = n_1/PT_1$
$ID_0 = n_0/PT_0$
ID: incidence density
n: number of incident cases (x)
N: population size
PT: observed person-time (area under the curve)
1: in exposed persons
0: in non-exposed persons
x: case
□: control

Table 3.6 Odds ratio calculated from a contingency table

	cases	controls
exposed	*a*	*b*
non-exposed	*c*	*d*
	a + *c*	*b* + *d*

$$OR = \frac{odds_{cases}}{odds_{controls}} = \frac{\frac{a}{a+c} \Big/ \frac{c}{a+c}}{\frac{b}{b+d} \Big/ \frac{d}{b+d}} = \frac{ad}{bc}$$

The fact that odds ratios calculated from case control studies can be interpreted as relative risks or rate ratios has made case control studies very popular in epidemiology. The pros and cons of various types of study will be discussed in detail in ▶ chap. 4. For the time being the important point is that relative measures of association (OR, IDR, CIR) can be estimated from case control studies but attributable risks cannot, as there are no cumulative incidences available.

3.2.5 The etiologic fraction among the exposed

Various other measures of association can be calculated from the RR and AR. The attributable risk divided by the incidence in the exposed group yields the proportion of incidence actually attributable to exposure. This measure, which is only relevant if the RR is greater than 1, is referred to as the etiologic fraction among the exposed (EF_e).

$$EF_e = \frac{I_1 - I_0}{I_1} = 1 - \frac{1}{RR} = \frac{RR - 1}{RR}$$

An EF_e of 48% for cardiovascular mortality and high cholesterol level, for instance, means that among persons with a high cholesterol level 48% of the cardiovascular mortality is attributable to that high cholesterol level. In other words, in this group the cardiovascular mortality would have been 48% lower if everyone had had a low cholesterol level (instead of a high one). Translating the EF_e into the case of an individual patient with a high cholesterol level who died of a cardiovascular condition, we could say that the likelihood that death was actually caused by the high cholesterol level is 48%. As the interpretation of the EF_e applies to individuals with a specific exposure, this measure is sometimes used when calculating the likelihood of causality in individual cases, for example when determining medical liability.

Table 3.5 shows how the EF_e can be calculated from both cumulative incidences and incidence densities. The EF_e is dimensionless and is expressed as a proportion (0–1) or percentage (0–100%). The value will be zero if there is no association, and higher the stronger the association between the risk factor and occurrence of the disease is.

3.2.6 The population attributable risk

It is also possible to calculate the population attributable risk (PAR):

$$PAR = \frac{I_T - I_0}{I_T}$$

The total population is a mix of people with and without exposure. The incidence of the disease in the total population is the weighted average of the incidence in the exposed (proportion: p) and non-exposed (proportion: 1 – p) groups respectively:

$$I_T = pI_1 + (1-p)I_0$$

Thus the value of the PAR is dependent not only on the value of the RR but also on the prevalence of exposure in the population (p). The PAR can therefore also be expressed as follows:

$$PAR = \frac{pI_1 + (1-p)I_0 - I_0}{pI_1 + (1-p)I_0}$$

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

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

A PAR of 33% for cardiovascular mortality and high cholesterol level, for instance, means that in the population in question (which contains both persons with a low and persons with a high cholesterol level) 33% of the mortality is attributable to the fact that part of the population has a high cholesterol level. In other words, in this specific population the cardiovascular mortality would have been 33% lower if everyone had had a low cholesterol level. As the interpretation of the PAR applies to an entire population, this measure is sometimes used when setting regional or national policies.

Table 3.5 shows how the PAR can be calculated, again from cumulative incidences and incidence densities. Note, however, that in the alternative method of computation, using the cumulative incidence approach, the proportion of exposed persons (p) is a factor, whereas in the incidence density approach it is the proportion of exposed person-time (q). The PAR is again a dimensionless number that is expressed as a proportion (0–1) or percentage (0–100%). The value zero indicates the absence of any association. The stronger the association, the higher the value.

As already explained, the OR generally gives a good approximation of the RR. Based on this assumption, an EF_e and a PAR can also be calculated in a case control study, by analogy with the formulas used to do this in a cohort study based on the IDR and CIR (see tab. 3.5). To calculate the PAR we

need not only the OR but also an estimate of the prevalence of the risk factor in question. This is estimated based on the prevalence in the control group, assuming that the control group is representative of the total population as regards exposure to the risk factor.

The calculation of a PAR can be extended for a combination of risk factors, in which case the numerical value indicates the proportion of the disease that is attributable to the combination of factors under consideration. The combined PAR is generally lower than the sum of the individual PARs, as a disease is usually caused by a combination of risk factors that together constitute a 'sufficient' cause (see ▶ chap. 6).

$$\text{Combined PAR} = 1-(1-\text{PAR}_1)(1-\text{PAR}_2)(1-\text{PAR}_3)\ldots$$

3.2.7 The potential impact fraction is the attributable proportion that describes the impact of a preventive intervention

An epidemiological measure that is commonly used in prevention policy is the potential impact fraction (PIF). This is a measure of how much of the incidence could be avoided by reducing exposure to a determinant in the population by means of a preventive measure. If this causes the incidence to fall from I_t to I_t', the potential impact fraction is:

$$\text{PIF} = \frac{I_t - I_t'}{I_t}$$

Unlike the PAR, this epidemiological measure allows for the situation that a preventive intervention usually reduces exposure in only part of the population. Therefore, using the PIF avoids overestimating the potential effects of a preventive measure. Like the other epidemiological measures of association, the PIF can also be expressed in terms of the exposed fraction before and after the introduction of the preventive intervention and the RR (CIR or IDR):

$$\text{PIF} = \frac{I_t - I_t'}{I_t}$$

$$= \frac{[p_a \times RR_a \times I_0 + (1-p_a)\times I_0] - [p_a' \times RR_a \times I_0 + (1-p_a')\times I_0]}{p_a \times RR_a \times I_0 + (1-p_a)\times I_0}$$

$$= \frac{(p_a - p_a')\times(RR_a - 1)}{p_a \times (RR_a - 1) + 1}$$

Where:

- I_t: incidence of the disease in the total population before the intervention
- I_t': incidence of the disease in the total population after the intervention
- p_a: proportion of the total population exposed to risk factor 'a' before the intervention
- p_a': proportion of the total population exposed to risk factor 'a' after the intervention
- I_o: incidence of the disease in the total population if risk factor 'a' is absent
- RR_a: relative risk of getting the disease for persons exposed to risk factor 'a' compared with persons not exposed to it

Given an effective preventive measure (vaccination, behaviour change, legislation, infrastructure, screening) a number of healthy persons will never know that they escaped disease or death, as a large number of people take part in a prevention programme whereas only a few of them would have developed disease without it. The big problem is that no-one can say in advance or with hindsight which individuals would have developed disease without the measure and therefore benefit(ed) from it. This is known as the prevention paradox: the potential health benefit for each individual is small on average, but for the population as a whole this small reduction in individual risk nevertheless yields a substantial reduction in incidence. For example, the number of road traffic deaths in the Netherlands has been reduced substantially by requiring all moped riders to wear a helmet, whereas even without this measure the vast majority of them would never have been involved in a serious road accident.

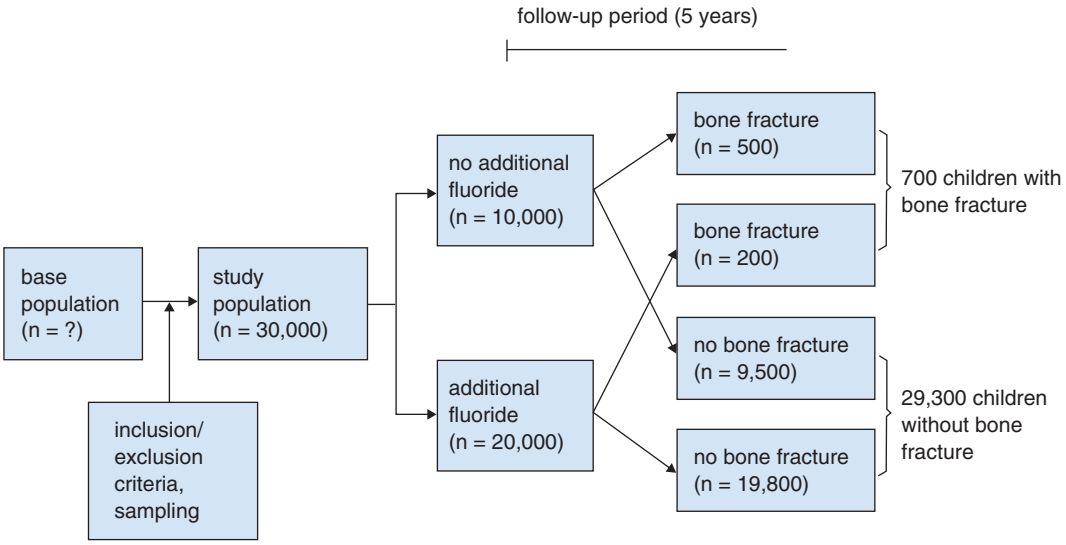


Figure 3.3 Design of a cohort study into fluoride use and bone fractures

Case 3.2 Fluoride and bone fractures (hypothetical example)

Suppose we want to find out whether using additional fluoride (in toothpaste or in tablet form) during childhood reduces the risk of bone fractures by strengthening bone structure. We decide to investigate this question in primary school children. When they see the school doctor, various data are collected routinely, including on the use of additional fluoride. Over a period of a few years all primary school children in a particular region who see the school doctor for the first time are included in the cohort. They are monitored for about five years while at primary school. During the follow-up period the hospitals in the region record all cases of bone fractures in children in the age group concerned. 30,000 children ultimately take part in the study, of whom 20,000 are additional fluoride users and 10,000 are non-users. Figure 3.3 shows the design of the study. An analysis of the research question based on the information available on all the subjects would yield the results shown in tab. 3.7.

Table 3.7 Analysis of all subjects in the cohort study into fluoride use and bone fractures

	fracture	no fracture
F–	500	9,500
F+	200	19,800
	700	29,300

$I_1 = I_{F+}$ = (cumulative) incidence of bone fractures in fluoride users:
= 200/20,000 (in 5 years) = 1% (in 5 years)

$I_0 = I_{F-}$ = (cumulative) incidence of bone fractures in non-fluoride users:
= 500/10,000 (in 5 years) = 5% (in 5 years)

$RR_{F-/F+}$ = relative risk (CIR) of bone fractures in non-fluoride users compared with fluoride users:
$= I_{F-}/I_{F+}$
= (500/10,000):(200/20,000) = 5

$RR_{F+/F-}$ = relative risk (CIR) of bone fractures in fluoride users compared with non-fluoride users::
= I_{F+}/I_{F-}
= (200/20,000):(500/10,000) = 0.2

$AR_{F-/F+}$ = additional risk (CID) of bone fractures in non-fluoride users compared with fluoride users:
= $I_{F-} - I_{F+}$
= (500/10,000) − (200/20,000)
= 800/20,000 (in 5 years)
= 4% (in 5 years)

NNT = number of children who need to use fluoride for five years to avoid 1 bone fracture:
= 1/AR
= 1/(800/20,000)
= 1/0.04
= 25

EF_{F-} = proportion of the risk (CI) of bone fractures among non-fluoride users attributable to non-use:
= $(I_{F-} - I_{F+})/I_{F-}$
= (500/10,000 − 200/20,000)/(500/10,000)
= 0.8 = 80%

EF_T = proportion of the risk (CI) of bone fractures in the total population attributable to non-use of fluoride (by a third of that population):
= $(I_T - I_{F+})/I_T$
= (700/30,000 − 200/20,000)/(700/30,000)
= 0.57
= 57%

Assuming that all the subjects were monitored for precisely five years and that the bone fractures were spread evenly over that period, the relevant measures of association can also be calculated from incidence densities:

$ID_{F-} = 500/[(10{,}000 \times 5) - (500 \times 2.5)]$
$= 0.01026/\text{jaar}$
$ID_{F+} = 200/[(20{,}000 \times 5) - (200 \times 2.5)]$
$= 0.00201/\text{year}$
$RR_{F-/F+} = 5.10$
$RR_{F+/F-} = 0.196$

$AR_{F-/F+} = 0.00825/\text{year}$
NNT = 121 (= number of children who need to use fluoride for 1 year to avoid 1 bone fracture)
$EF_{F-} = 0.804 = 80.4\%$
$EF_T = 0.574 = 57.4\%$

Based on the information available on all bone fracture cases and a sample of the remainder of the study population, a nested case control study carried out on this cohort would yield the following results:

- Patients (bone fractures): n = 700, which breaks down into 500 fluoride non-users and 200 users based on the exposure data
- Controls selected from the disease-free population (no bone fractures; two controls per patient).

Assuming that the bone fractures are spread evenly over the five-year period and that whenever a fracture occurred two children from the remaining disease-free population were recruited to the control group, the ratio of fluoride non-users to users in the control group will be (10,000+9,500)/2 to (20,000+19,800)/2, yielding 460 fluoride non-users and 940 users. The results are shown in tab. 3.8.
The OR (IDR) can be calculated from this table, after which the EF_e and PAR can be estimated using the formulas in tab. 3.5:

$OR_{F-/F+} = (a \times d)/(b \times c) = 5.10$
$OR_{F+/F-} = 0.196$
$EF_{F-} = 1 - 1/OR = 0.804 = 80.4\%$
$EF_T = q(OR - 1)/[q(OR - 1) + 1]$
$= 0.574 = 57.4\%$
where $q = b/(b + d) = 460/1{,}400$

Table 3.8 Analysis of a nested case control study into fluoride use and bone fractures

	patients	controls
F–	500	460
F+	200	940
	700	1,400

An analysis of a subgroup of 2,100 subjects thus yields the same estimates for a number of measures of association as an analysis based on the total study population of 30,000 persons. Note that in this example the way in which the controls are recruited determines the value of the OR. If the control group had been recruited at t_0 (thus making the ratio of fluoride non-users to users 1:2), the OR (=CIR) would have been 5. However, if the control group had been recruited from the children without bone fractures after the five-year follow-up period (thus making the ratio of fluoride non-users to users 95:198) the OR would have been 5.2. A similar overestimation of the IDR based on the OR will also occur in a case control study based on a dynamic population in which cases are compared with a sample from the remaining disease-free population at the end of the follow-up period. The overestimation will be worse, the higher the incidence density of the condition under consideration is.

3.3 Measures of association for continuous health outcomes

Epidemiology examines the association between one or more determinants and the frequency of a disease. Some characteristics of health and disease, however, are expressed not as frequencies of a dichotomous disease variable but on a continuous scale, for example blood pressure, birth weight and quality of life. This section discusses the measures of association for continuous health outcomes.

3.3.1 The difference of means

We introduced the centrum measures (mean, median, mode) and measures of spread (standard deviation, interpercentile range, range) in ▶ par. 2.6. Of these measures, a combination of mean and standard deviation is most commonly used, in particular because of its attractive computational properties.

The obvious way of comparing means and standard deviations between two groups is to subtract one mean from the other and interpret the difference in the light of the two standard deviations. A large difference with a small spread is more impressive than a large difference with a large spread, and certainly more impressive than a small difference with a large spread. In the literature on randomized clinical trials and in systematic reviews we therefore often come across the standardized mean difference (SMD) as a measure of the size of the effect of an intervention,

$$\text{SMD} = \frac{M_1 - M_0}{SD_{1-0}}$$

where M_1 and M_o are the means for the two intervention groups and SD_{1-o} is the standard deviation of the difference between the two groups.

This principle also forms the basis for the statistical procedures for comparing means: calculating a confidence interval for the difference of means, Student's t-test, one-way analysis of variance, simple linear regression etc., for which we refer to a basic statistics book for further explanations.

3.3.2 Correlation and regression

There are times when we wish to link two continuous variables in a population to each other, e.g. length of pregnancy and birth weight, salt consumption and blood pressure, or lung function and quality of life. In such cases we start by drawing a chart, a scatter plot, plotting the continuous determinant against the continuous health variable. Each point in the scatter plot represents an individual in the population with its particular combination of values for the determinant and the health variable. ◘ Figure 3.4 is an example of a scatter plot.

A scatter plot will often take the form of a flattened cloud with a lot of observations in the middle and not many at the ends. The density of the cloud is indicative of the correlation between the two variables: if the points form a (rising or falling) more or less straight line there is a strong correlation; on the other hand, if the points are completely spread out between the two axes of the chart and the cloud has no discernible structure, the two variables are apparently unrelated. Every conceivable variation can oc-

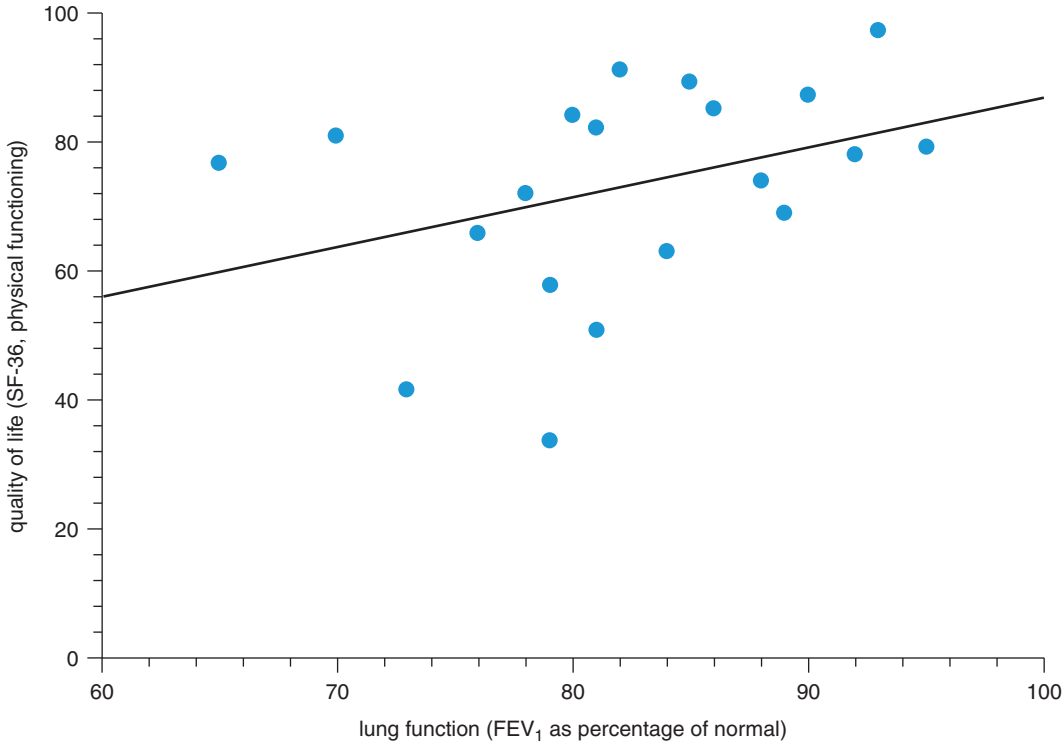


Figure 3.4 Scatter plot with regression line and calculated measures of association
Quality-of-life score = 10.86 + 0.75 × lung function (FEV_1)
Regression coefficient = 0.75
Correlation coefficient (Pearson's r) = 0.36

cur between these two extremes (a non-horizontal straight line and an amorphous cloud). The closer the points are to an imaginary line the stronger the correlation, and we can actually draw this line in such a way that it runs through the middle of the cloud in the direction indicated by the cloud. Some statistical measures are based on these principles.

- The correlation coefficient provides a measure of the density of the cloud: the closer the points are to the best fitted line, the higher the correlation coefficient is.
- A completely different measure is the regression coefficient, which quantifies the slope between the best fitted line (known as the regression line) and the x-axis. The steeper the slope, the stronger the association between a continuous determinant and a continuous health outcome is (see also ▶ par. 3.4).

Thus the correlation coefficient is indicative of the extent to which the points can be represented by a line, whereas the regression coefficient is indicative of the slope of this line. Both these measures are important. Calculations of this kind are subject to several preconditions (e.g. a normal distribution of the two continuous variables and a straight line as the best fitted function).

3.4 Regression analysis

Epidemiological functions of the type $P(O) = f(D_i)$ describe the relationship between the dependent disease outcome O and one or more specific determinants D_i. This type of function can also be expressed as a mathematical regression equation. In its simplest form a regression equation describes a

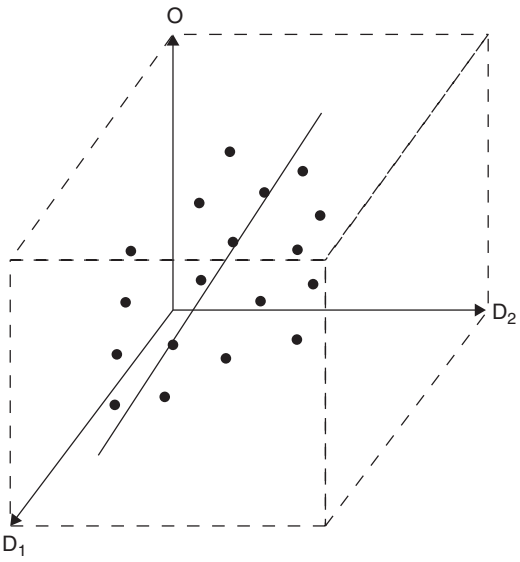


Figure 3.5 A 3D linear regression function

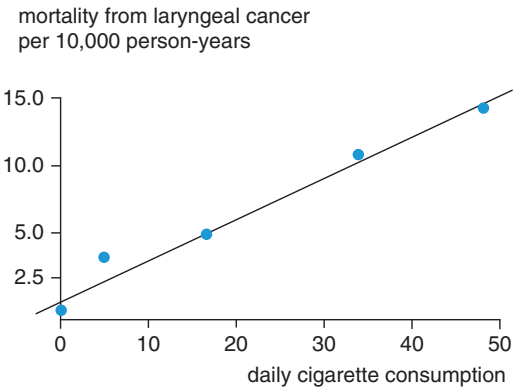


Figure 3.6 A 2D linear regression function

linear relationship between a disease parameter and a single exposure:

$$P(O) = b_0 + b_1 D_1 \quad (3.1)$$

Case 3.3 provides an example of a simple linear regression of this kind. The mortality rate from laryngeal cancer (O) is calculated and plotted for five levels of daily cigarette consumption (D_1). The regression line $P(O) = 1.15 + 0.282 \times D_1$ is an almost perfect fit with the data from the study on which fig. 3.6 is based.

It is not difficult to add a second exposure to this simple linear regression function:

$$P(O) = b_0 + b_1 D_1 + b_2 D_2 \quad (3.2)$$

This too is a straight line, but in a three-dimensional space with the disease parameter O on the vertical y-axis and the two determinants D_1 and D_2 on the horizontal x and z-axes (see fig. 3.5).

If we want to examine the joint and separate contributions that two (or more) determinants make to the disease we need multidimensional regression models of this kind. For example, daily alcohol consumption – also a determinant of laryngeal cancer, and moreover associated with cigarette consumption – would be a worthwhile second determinant in the example in ▶ case 3.3. By including both determinants in the model we can gain a good idea of the separate contributions made by each of the determinants to the occurrence of the disease.

Straight lines as described with formulas (▶ 3.1) and (▶ 3.2) will not be encountered very often in epidemiological research, as a straight line is likely to conflict with the range of possible values for the disease parameter D. The incidence can never be negative, for instance, whereas a linear regression function like that described in ▶ case 3.3 does not rule this out. Also, the disease frequency, expressed as a risk, is by definition limited by the values zero and 1. For this kind of common situations we use simple mathematical transformations of the outcome variable (O) so that the regression function fits better with the actual data from epidemiological research.

Common transformations are the logarithmic regression function:

$$\ln(O) = b_0 + b_1 D_1 + b_2 D_2 + \ldots \quad (3.3)$$

and the logistic regression function:

$$\ln\left(\frac{O}{1-O}\right) = b_0 + b_1 D_1 + b_2 D_2 + \ldots \quad (3.4)$$

Note that the transformed equations (▶ 3.3) and (▶ 3.4) are to some extent examples of linear functions, as the right-hand side of the equation describes a straight line when the y-axis is transformed. The regression functions described in this section are therefore referred to as generalized linear models.

A simple linear regression function is useful pri-

marily in the case of a continuous normally distributed outcome.

If the primary determinant is dichotomous (exposed: $D_1 = 1$; non-exposed: $D_1 = 0$), interpretation is fairly simple. The disease outcome in the exposed persons is thus:

$$P(O_1) = b_0 + (b_1 \times 1) = b_0 + b_1$$

For the non-exposed persons the disease outcome is:

$$P(O_0) = b_0 + (b_1 \times 0) = b_0$$

The effect of the determinant on the disease outcome is then expressed simply by looking at the difference between the exposed and non-exposed groups:

$$P(O_1) - P(O_0) = (b_0 + b_1) - b_0 = b_1$$

The regression coefficient b_1 thus directly indicates the difference in outcome between the exposed and non-exposed groups when the data fits a straight line.

Case 3.3 Laryngeal cancer and smoking

Figure 3.6 shows the data from a study into the number of cigarettes smoked per day (cig) and the age-standardized mortality rate per 10,000 person-years for laryngeal cancer (mort). The data fits a simple straight line, which is described mathematically as:

$$\text{mort} = 1.15 + 0.282 \times \text{cig}$$

The intercept (1.15) in fig. 3.6 represents deaths from laryngeal cancer when nobody smokes. The regression coefficient (0.282) indicates that in this model the mortality rate per 10,000 person-years increases by 0.282 deaths when the average daily number of cigarettes smoked in the population increases by one. Assuming that no other variables (including confounding variables) are involved, the regression coefficient of 0.282 indicates the effect of cigarette-smoking on laryngeal cancer mortality. According to the model a person who smokes two packs a day (50 cigarettes) has a mortality risk of 0.00152 per year. Compared with 10,000 person-years of non-smoking, 10,000 person-years of smoking (two packs a day) result in approximately 14 additional cases of death from laryngeal cancer. We could also say that people who smoke two packs a day run 15.2/1.15 = 13.3 times as much risk of dying from laryngeal cancer compared to non-smokers.

Recommended reading

Carneiro I, Howard N. Introduction to epidemiology. 2nd ed. Maidenhead: Open University Press; 2005.

Gordis L. Epidemiology. 5th ed. Philadelphia: Elsevier Saunders; 2014.

Grobbee DE, Hoes AW. Clinical epidemiology: principles, methods, and applications for clinical research. 2nd ed. Burlington: Jones and Bartlett Learning; 2015.

Kleinbaum D, Klein M. Survival analysis: a self-learning text (Statistics for biology and health). 3rd ed. New York: Springer; 2012.

Webb P, Bain C. Essential epidemiology: an introduction for students and health professionals. 2nd ed. Cambridge: Cambridge University Press; 2011.

Source references (case and table)

Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chron Dis. 1967;20:511–24 (Case 3.1).

Rothman KJ, Lash TL, Greenland S. Modern epidemiology. 3rd edition. Philadelphia: Lippincott, Williams & Wilkins; 2012 (Case 3.3).

Global Health Risks: Mortality and burden of disease attributable to selected major risks. Geneva: World Health Organization; 2009 (tab. 3.3).

Study design

4.1 Introduction: the research question determines the design

This chapter explains how analytical epidemiological studies should be designed in order to find the link between a determinant and a health outcome. A variety of approaches can be used: for example, investigating the influence of factors on the development of a particular condition (an etiological study); investigating the influence of factors on the course of a disease (a prognostic study); investigating the extent to which medical tests give correct predictions (a diagnostic study) or the efficacy of preventive or therapeutic interventions (an intervention study).

This chapter is concerned with the 'architecture' of epidemiological studies and the 'styles' that the researcher can choose from. Although each stage of construction and each style is underpinned by specific mathematical models, all styles are susceptible to errors, albeit to different degrees. Choosing a particular study design therefore has direct consequences for the validity and precision of the results, and hence for the credibility of the conclusions that can be drawn from the study. The question of sources of error and systematic confounding will be discussed in more detail in the next chapter; in this chapter we simply describe the standard study designs.

We cannot adequately design and carry out an epidemiological study unless the research question has been formulated in advance and made specific and measurable. Getting off to a good start is half the battle.

4.1.1 The research question sets out the exposures and outcomes under consideration, the target population and the relationships between these factors

The research question needs to be formulated in specific, measurable terms. PICO provides a useful guide here:

- **Population:** What people are we talking about? What is the domain for which we want to draw a valid conclusion?
- **Intervention:** What is the intervention, diagnostic test or exposure that we want to link with the health outcome?
- **Comparator:** With what intervention, test or exposure do we want to make a comparison?
- **Outcome:** What health outcome are we studying?

Examples of research questions for different types of study:

- Etiological study: In Caucasian adults (P) how large is the additional risk of developing bladder cancer (O) associated with smoking cigarettes (I) compared with never smoking (C)?
- Prognostic study: What is the 10-year risk of death (O) of bladder cancer patients (P) who stop smoking (I) compared with continuing to smoke (C)?
- Diagnostic study: To what extent does an abdominal CT scan (I) reduce the percentage of false negative diagnoses (O) in patients with appendicitis symptoms (P) compared with a doctor's clinical opinion (C)?
- Intervention study: In long-term heroin addicts (P) what effect does prescribing methadone (I) compared with standard care (C) have on the success of an attempt to stop (O)?

4.1.2 Situational research questions are place and time-related; abstract research questions are time and place-independent

When choosing the study design it is important to distinguish between situational and abstract problems. Situational research questions are place and time-related, usually based on a practical question: for example, a descriptive study of the population's state of health in a particular region for the purpose of determining care needs or assessing the quality of the care provided. Research into the possible cause of a Salmonella epidemic at a nursing home is also situational, as the results are relevant to the place and time period stated in the research question.

Abstract research questions, on the other hand, are independent of place and time. The aim of abstract research, insofar as it is into cause and effect, is to gain an understanding of the causal relationship

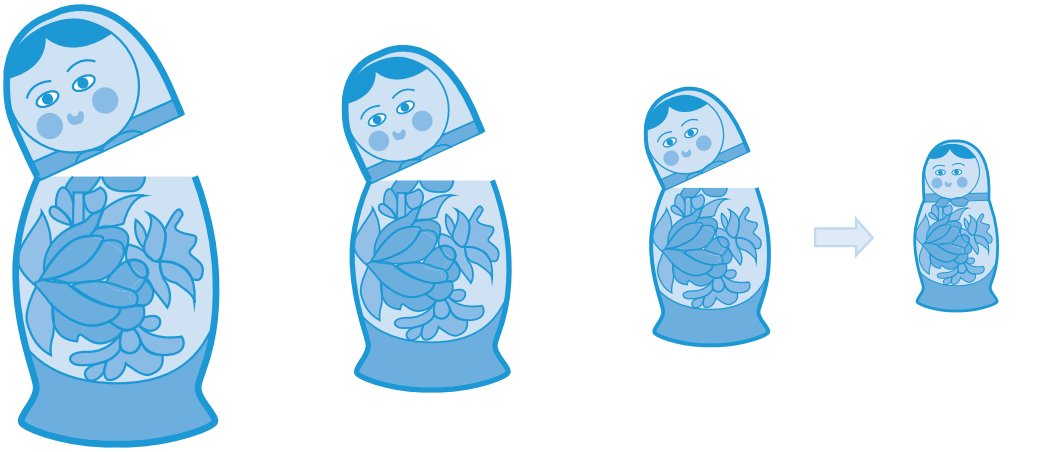


Figure 4.1 Relationship between total population, target population, source population and sample (with acknowledgments to Martijn Boers)

between an exposure or exposures and a disease in abstract reality: for example, an etiological study into whether alcohol consumption causes cirrhosis of the liver, or a prognostic study into whether smoking leads to a poorer prognosis for people with existing occlusion of the leg arteries. Although not place and time-related, an abstract research question is still very specific: we want to know to what extent the effect of the exposure on the disease outcome (e.g. alcohol consumption on cirrhosis of the liver) depends on the dose (the quantity and type of alcohol consumption) or on personal factors (e.g. constitution, behaviour and environment). The researcher decides whether to incorporate this variation in the research question or to focus it on specific doses and/or specific population factors. These choices can have far-reaching effects on the study design.

4.1.3 The epidemiological function is the formal notation of the research question

Once we have formulated a specific research question it is relatively easy to translate it into an epidemiological function. The epidemiological function is in effect the formal notation (in mathematical terms) of the research question (in words). The epidemiological function for the etiological research question in ▶ par. 4.1.1, for instance, could be:

$$P\,(\text{bladder cancer}_{\text{in Caucasian adults}}) = f(b_0 + b_1(\text{smoking vs. not smoking}))$$

Opting for this formal notation has various advantages: it makes the research question quantifiable; it draws attention to elements in the research question that are not entirely clear yet, and it provides an immediate guide to making the study specific and measurable, including the design of the data analysis.

4.1.4 Select the study population that will efficiently give you a valid answer to the research question

Representativeness is an absolute prerequisite for a study with a situational research question, but for what group should the results be generalized? The target population is the (selective) part of the total population for which we wish to extrapolate the results (see fig. 4.1). It is usually not worthwhile or feasible to include the entire target population on which we wish to draw a conclusion in the study; a sample of a well-defined source population will gen-

erally do. The source population is a suitable selection enabling an answer to the research question applicable to the target population; the source population is thus not necessarily (or desirably) statistically representative of the target population. The sample is our actual study population and is statistically representative of the source population, so that the results in the random sample are also valid for the source population.

In a study based on an abstract research question representativeness needs to be defined differently, as we are using results observed in a situational study population to say something about the abstract reality (this generalization process is referred to as 'inference'). In other words, we are not taking a sample from a specific population for which the results can be generalized; rather, our aim is to show that there is a causal link between a particular determinant and a disease, regardless of any specific population. To analyse the abstract problem, however, we need to find a study population and setting that gives us the maximum chance of uncovering the link that we are looking for – assuming that it actually exists, of course. The study design must also sufficiently guarantee that only the determinant in question can be responsible for the effect measured, not other factors (i.e. the internal validity of the design).

A suitable study population in epidemiological research is characterized by:

- a clear contrast in the determinant under consideration (the presumed cause)
- maximum comparability of the study groups in terms of other determinants for the outcome being studied (potential confounders)
- maximum comparability of the study groups in terms of how the various variables are measured
- maximum information content per participant (per unit of time or money).

A clear contrast in the determinant means that the study population is selected in such a way that all the relevant values of the determinant are adequately represented. In many cases we might focus on the effect of extreme values for the determinant (maximum contrast). Sometimes, however, we might prefer an even distribution among the various determinant levels. For example, a researcher wishing to find out whether eating plenty of dietary fibre protects against appendicitis will ideally have a study population consisting mainly of people with low or high consumption of fibre-rich foods. Although most people's level of consumption will be fairly close to the population average, that category provides the least information for etiological research; extreme categories are far more interesting and should ideally be over-represented in the study population. Some of the in-between categories can also be included in the study to give an idea of the entire range.

For maximum comparability on all other determinants for the disease other than the determinant being studied, the study groups being compared should have a similar distribution of all other factors that affect the occurrence of the disease (potential confounders) in each of the categories of the primary determinant. This idea is developed in ▶ chap. 5.

The third criterion is to select the study population in such a way as to obtain measurements of comparable quality: a difference in disease frequency between two subpopulations with different levels of the primary determinant must not be due to a difference in measuring procedures. Suppose, for example, that an occupational physician has made regular lung X-rays to study the additional lung cancer risk among butchers. The incidence of lung cancer found in this way cannot be compared with the incidence figures for lung cancer in the general population produced by the regional cancer centres based on hospital admissions; instead we should compare the butchers with another category of employees that had regular lung X-rays done by their occupational physician as well.

Lastly, the choice of study population can improve the efficiency of the study as well. We argued above already for sufficient contrast in the exposure status. In addition it is recommended to focus on subpopulations where the association being studied will appear as clearly as possible; this is not necessarily the population with the highest disease frequency.

Figure 4.2 illustrates this principle based on two subpopulations with different background risks I_o for the disease being studied, i.e. a different incidence among unexposed persons (D_o). The likelihood of remaining healthy is $(1 - I_o)$. If these people, regardless of the background risk, have a 20% risk of

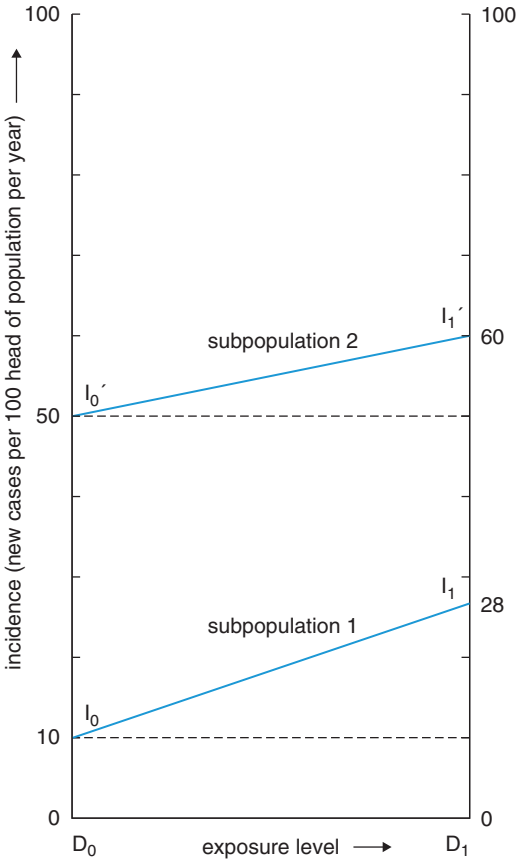


Figure 4.2 Disease incidences among exposed and non-exposed persons in two subpopulations with different background risks of the disease being studied

developing the disease due to exposure to the primary determinant (D_1), we can calculate the AR and RR for both subpopulations:

$$AR = I_1 - I_0 = p(1 - I_0)$$
$$RR = I_1/I_0 = [p(1 - I_0) + I_0]/I_0 = (p/I_0) - p + 1.....$$

subpopulation 1	subpopulation 2
($I_0 = 10/100$ pers.yr)	($I_0 = 50/100$ pers.yr)
AR = 18/100 pers.yr)	AR = 10/100 pers.yr)
RR = 2.8	RR = 1.2

It shows that, with p remaining the same, AR and RR will increase when the background risk (I_o) decreases. This also explains why substantial differences in the strength of an association between determinant and outcome are sometimes found when it is studied in populations with different background risks.

In cause-and-effect research we are looking for a clean, undistorted picture of the effect of exposure to a determinant, starting with the disease frequency in people who have been exposed. Ideally we would want to compare this with the disease frequency in the theoretical situation where these same people (in precisely the same circumstances) would not have been exposed to the determinant – in effect, a perfect mirror image (counterfactual). A perfect counterfactual can never be achieved in empirical research; we always have to work with approximations, either by studying the same persons again at a point in time at which they were not exposed (and assuming that there have been no substantial other changes in the meantime) or by including other comparable unexposed persons in the study (and assuming they otherwise do not differ substantially from the exposed persons).

4.1.5 Classifications of study designs

Study designs can be classified on the basis of their design characteristics (see fig. 4.3). The first distinction is between individual studies and ecological studies. An ecological study – also referred to as a correlation study – examines the relationship between a disease and other phenomena based on aggregate data (e.g. average alcohol consumption per capita and annual mortality from laryngeal cancer per 100,000 head of population in the country in question). Two subtypes of ecological studies are time trend studies (analysis of time series) and geographical correlation studies. We shall discuss these types of study in more detail in ▶ par. 4.3.4.

Individual studies can in turn be divided into cross-sectional and longitudinal studies. As noted in ▶ par. 4.3.3, cross-sectional studies are usually inferior to longitudinal studies when studying cause-and-effect relationships, as exposure to the determinant needs to take place before the occurrence of the disease outcome: this can be done in a longitudinal study but not in a cross-sectional study. Longitudinal studies can be divided into experimental and non-experimental studies. The crucial aspect of an

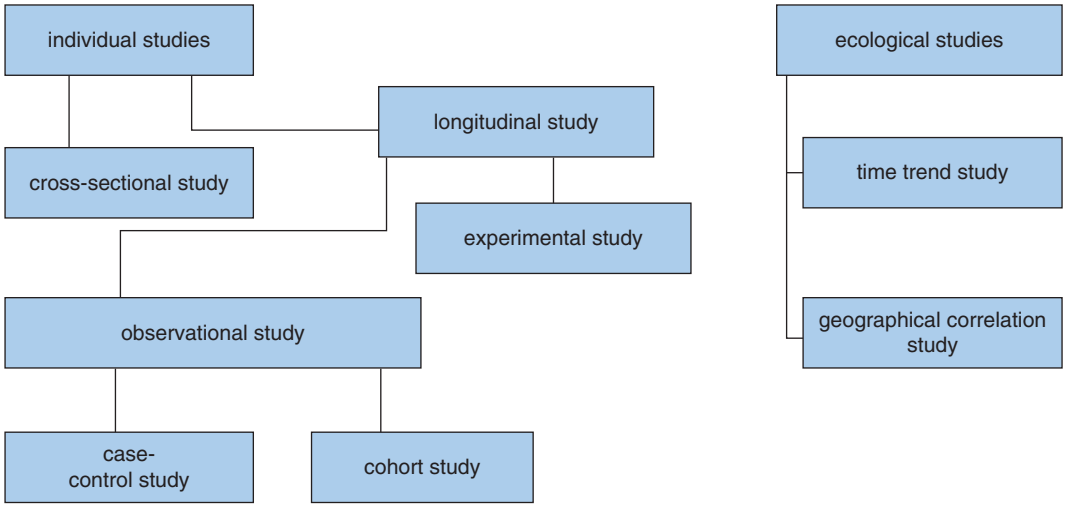


Figure 4.3 Classification of epidemiological study designs based on design characteristics

experiment is that there is an intervention on the part of the researcher, who manipulates one or more determinants (e.g. the therapy given) in order to answer a scientific question. An experiment which also involves random allocation of the intervention to the participants is referred to as a randomized experiment or randomized controlled trial (RCT). RCTs are discussed in ▶ par. 4.2 and ▶ chap. 10. In the case of non-experimental studies (observational studies) the researcher does not allocate different levels of the determinant; instead, he takes the distribution of individuals among the various categories of the determinant that has come about 'spontaneously'. The researcher merely observes and tries to identify as neatly and efficiently as possible what is going on or has already happened, without actually intervening. Longitudinal observational studies can in turn be divided into cohort studies and case-control studies, which are discussed in ▶ par. 4.3.1 and 4.3.2. A cohort study is based on a particular determinant, and the study population is assembled on the basis of that determinant. Subgroups with different determinant statuses at the start of the observation period are compared to see whether these differences are in due course accompanied by differences in the occurrence of the outcome being studied. A case-control study is based on a particular disease, and the study population is assembled on the basis of that disease. Subgroups with different disease statuses (usually cases versus healthy controls) are compared in terms of the prevalence of the various categories of the determinant at a relevant time in the past.

Although the above classification suggests that there are fundamental differences between these study designs, the dividing lines between them are less clear-cut than it might seem at first sight. For example, if we are interested in determinants that remain constant over time (e.g. specific genes, gender, blood group, congenital enzyme deficiencies) – which can therefore be measured at the same time as the hypothesized effect without any problem – the dividing line between a cross-sectional study and a case-control study becomes less clear. A randomized experiment can also be regarded as a special type of cohort study, with random allocation being the main additional element (see fig. 4.3).

Epidemiological study designs can also be classified in terms of their potential evidential value in establishing a causal link between a determinant or intervention and a health outcome (see fig. 4.4). Case reports are often based on a single case, or only a small number of cases, and are therefore regarded as the study design with the lowest evidential value. Case series are better in that respect, but they are often small-scale and do not include a comparison

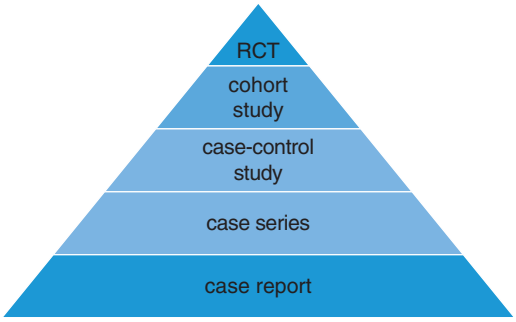


Figure 4.4 Hierarchy of epidemiological study designs in terms of evidential value

group. Case-control studies and cohort studies do include a comparison group and are usually on a larger scale. Cohort studies are regarded as having higher evidential value because of their prospective nature. A randomized experiment is generally regarded as the study design with the highest evidential value, as it enables the researcher to exert maximum control over confounding.

4.2 The experiment as the paradigm for all causal study designs

As in other empirical research disciplines, in epidemiology the randomized experiment is the paradigm, the standard design for cause-and-effect research. Meticulous application of the rules for experimental research based on Good Clinical Practice[1] provides the maximum guarantee of internally valid results. However, practical or ethical considerations will often prevent an experiment being carried out (see fig. 4.5). Yet even then the experimental design provides a conceptual model for cause-and-effect research. It is for this reason that we discuss randomized experiments in this chapter. Randomized experiments or randomized controlled trials (RCTs) carried out on patients to assess the efficacy of therapeutic measures are referred to as clinical trials. Experiments carried out on healthy people to assess the efficacy of preventive measures are referred to as preventive trials. ▶ Chapter 10 looks at the practical aspects of experimental studies.

4.2.1 The randomized experiment

The basic structure of an experiment is shown in fig. 4.6. The participants in an RCT[2] are recruited from the source population, restricted to a more or less homogeneous group (e.g. people in the same age group, of the same gender, with the same disease or at the same stage of the disease). The candidates are checked to see whether they meet the inclusion and exclusion criteria and are told what the proposed experiment entails. They are then asked to consent to take part on a voluntary basis (informed consent). The researcher determines in advance how many participants will be needed to estimate the intervention effect with sufficient precision.

The intervention being studied is allocated to the participants using a random allocation procedure (randomization) which gives each participant an equal chance of being assigned to the intervention group. As a result of this random allocation based on the drawing of lots, the intervention group and the control group(s) will be equal on average in all respects at the start of the experiment, including other determinants (and potential confounders) that, like the intervention factor, could affect the results. Randomization deals with all confounders, both known, unknown and difficult-to-measure ones. The randomized experiment therefore meets one of the essential requirements for cause-and-effect research, namely comparability of the trial groups. Randomization does not guarantee that the various treatment groups will be completely comparable, however, as chance can have unfortunate effects. The likelihood of remaining differences between subgroups in an experiment becomes larger when the study population becomes smaller. Pre-stratification can be a preventive measure here: this involves dividing the study population into 'strata' for major potentially confounding factors (e.g. men and women; 35 to 44-year-olds, 45 to 54-year-olds and 55 to 64-year-olds) and carrying out randomization on each stratum.

1 Good-clinical-practice compliance, European Medicines Agency (website). ▶ http://bit.ly/1FjyW4E

2 Randomized Control Trial (animation). ▶ http://bit.ly/1Fjz30f

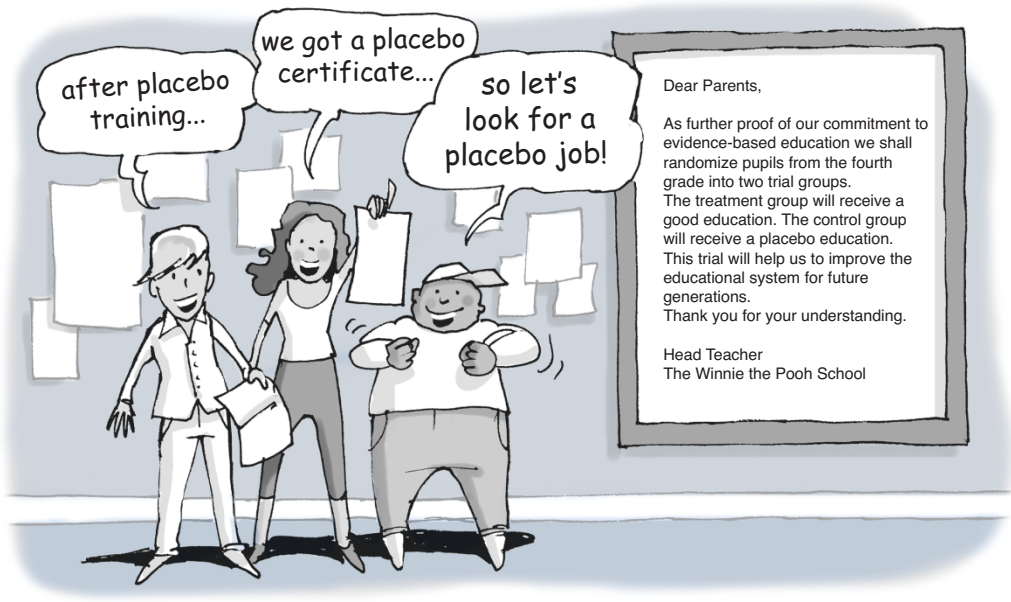


Figure 4.5 An RCT is not always ethically feasible (with acknowledgments to freshspectrum.com)

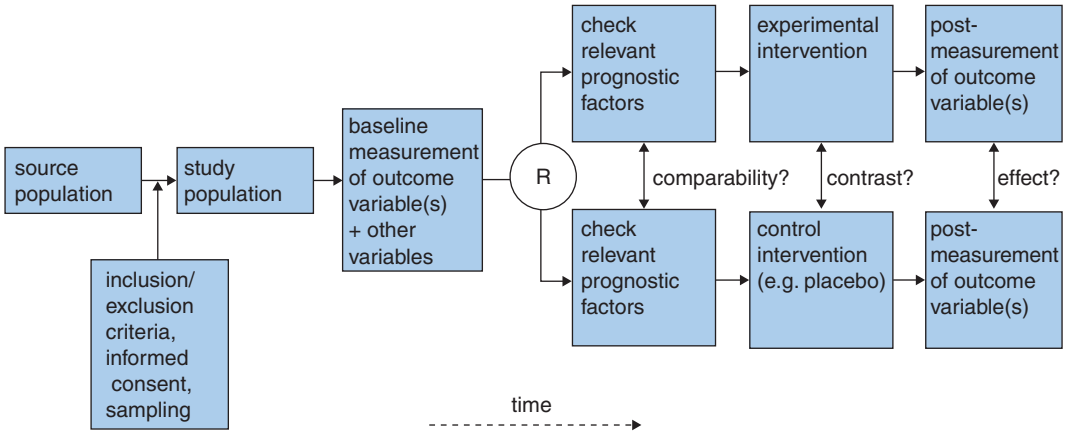


Figure 4.6 Basic structure of an experiment

This substantially increases the likelihood that the treatment groups will be equal for these important factors at the start of the experiment.

Once all the participants have been allocated to one of the categories of the intervention being studied, the researcher may want to check whether the resulting groups actually are comparable using baseline data, although some might argue that is it better to postpone this until after evaluating the overall effect of the trial. This usually involves measuring the initial value of the outcome variable (the baseline), sometimes also the background level of the intervention factor, and always the level of potential confounders, and other indicators for comparability of subgroups. Baseline measurement before the randomization procedure ensures that the baseline is

comparable for all participants (as nobody knows which group they will be assigned to).

Once randomization and baseline measurement have been performed, the intervention is applied, well prepared (dosage, method of administration), clearly described and standardized in a treatment protocol. Steps must be taken to ensure that participants adhere to the prescribed treatment (compliance) and any non-compliance is recorded. The control group is given a control intervention that is consistent with the research question (the standard treatment, no treatment, a placebo, or a different dosage of the same treatment). A placebo intervention (i.e. a control intervention that is indistinguishable from the real intervention in terms of appearance, taste, etc.) provides the best guarantee that the trial groups are comparable for co-interventions and outcome measurements. There are situations, however, where a placebo intervention is not feasible or ethically unacceptable, for example in surgery or in the case of a life-threatening condition such as cancer.

Placebo treatments enable the trial participants and researchers to be blinded, in which case the participants and/or researchers do not know which participants are given the intervention and which are given the control treatment. This prevents them from possibly acting in a way that could defeat the experimental design. If the effect assessor or the participant himself knows the intervention category to which the participant has been assigned, that can result in conscious or unconscious manipulation of the observations (observer bias). If it is not possible to blind the patients and the treating professional, observer bias can be avoided by having the outcome measurements carried out by someone else and blind this person carefully for the treatment intervention. Observation bias is also reduced by using measurement instruments that leave little opportunity for subjective interpretation.

Although randomization produces trial groups that are comparable at the start of a trial, differences can develop along the way as a result of participants dropping out during the follow-up period. Dropout that is selective (attrition) and is associated with factors that affect the outcome poses a threat to the validity of the study. The problem is particularly serious if there are different reasons for dropout, or different dropout numbers, in the trial groups. A placebo intervention prohibits selective dropout to some extent, as participants will not act based on dissatisfaction with their allocation.

There are a number of variations on the basic RCT structure, which will be discussed in ▶ Chapter 10.

Case 4.1 A randomized dietary experiment in premature infants

Children born prematurely often experience problems after birth because of immature organ systems. One of these problems is dietary intolerance in the gastrointestinal tract; another is a greater risk of infection. These problems occur particularly in the first month of life. Before birth children receive amino acids that are important to the function of those organs, including glutamine, through the placenta.

To investigate whether adding glutamine to the diet from Day 3 to Day 30 after birth has a beneficial effect on dietary tolerance (the primary outcome) and the occurrence of serious infections (one of the secondary outcomes), a randomized controlled trial (RCT) was conducted on premature children at increased risk of these problems (the inclusion criteria were a length of pregnancy of less than 32 weeks and/or a birth weight less than 1,500 grams). Informed consent had been obtained from the parents. Children who were at increased risk of these problems for other reasons were not included (the exclusion criteria included congenital abnormalities and genetic defects).

The study (sponsored by a large baby food company) took place at a neonatal intensive care unit of a teaching hospital. It involved a total of over 100 children. Although there was no difference in dietary tolerance between the two groups, far fewer serious infections occurred in the group of children who were given glutamine (the intervention group) than the group given maltodextrin (the placebo group). The follow-up showed among other things, by means of brain imaging (MRI), that the children in the intervention group had greater brain volume at the age of eight than the children in the control group. Statistical analysis was able to show that this effect was

attributable to the reduction in the number of serious infections brought about by adding glutamine.

4.2.2 Guidelines for experimental studies

Most editors of scientific journals now require articles on clinical trials to comply with the CONSORT guidelines.[3] The 2010 CONSORT guidelines recommend:

- A flow diagram showing the changes in the number of participants in the various phases of the experiment (see fig. 4.7)
- Organizing the article as follows:
  - Title and abstract
  - Introduction (background and research question)
  - Method (design, participants, interventions, health outcomes, sample size, randomization, blinding and statistical methods used)
  - Results (flow of participants, baseline data, results for primary and secondary outcome measures and side effects)
  - Discussion (limitations, generalizability and interpretation).

There are also checklists for assessing the quality of experimental research, which can be used in systematic literature searches and when designing new studies: the Cochrane Risk of Bias Tool,[4] for example.

4.3 Observational study designs when a randomized experiment is not feasible

Although the experiment is the paradigm for epidemiological research, experimental designs are often not feasible in epidemiology. Non-experimental study designs, where the researcher has no influence over the distribution of the determinant and instead uses a distribution that has come about in some other way, are more the rule than the exception in epidemiology. The most common types of these observational study designs are discussed below.

4.3.1 Cohort studies

Of all the types of observational epidemiological research, cohort studies come closest to the randomized experiment. Essentially the randomized experiment is a special type of cohort study (see fig. 4.8). The basic structure of a cohort study is shown in fig. 4.9. Each of the main characteristics is discussed below.

Defining the study population

A cohort study has two or more subcohorts as its study population: exposed and non-exposed persons in a sample from the population register, patient records, personnel records, etc. The outcome being studied has not occurred in any of the participants at the start of the study. In some cases a subcohort of non-exposed subjects is not easily identifiable (and/or would be too expensive); in these cases the exposed cohort is compared with the general population.

Measuring exposure

A cohort study focuses on the exposure of the participants to the particular determinant. The participants are divided into a minimum of two subcohorts based on the measured values of that determinant. The other determinants still need to be measured in a cohort study, as the primary determinant is often (due to lack of randomization) associated with other risk factors for the disease. If we are studying the effects of alcohol consumption on the condition of the heart, for instance, we need to realize that drinkers and teetotallers may well have different dietary habits, smoking habits, mental characteristics and exercise patterns. By measuring these other determinants in all the participants we can adjust for any confounding effects of these factors in the analysis. We can also anticipate potential confounding when selecting the participants, e.g. by selecting only smo-

3 CONSORT guidelines for the publication of experimental studies (website). ▶ http://bit.ly/1DXuYMS

4 Cochrane Risk of Bias Tool (PDF). ▶ http://bit.ly/2ykMP6Y Cochrane Risk of Bias Tool

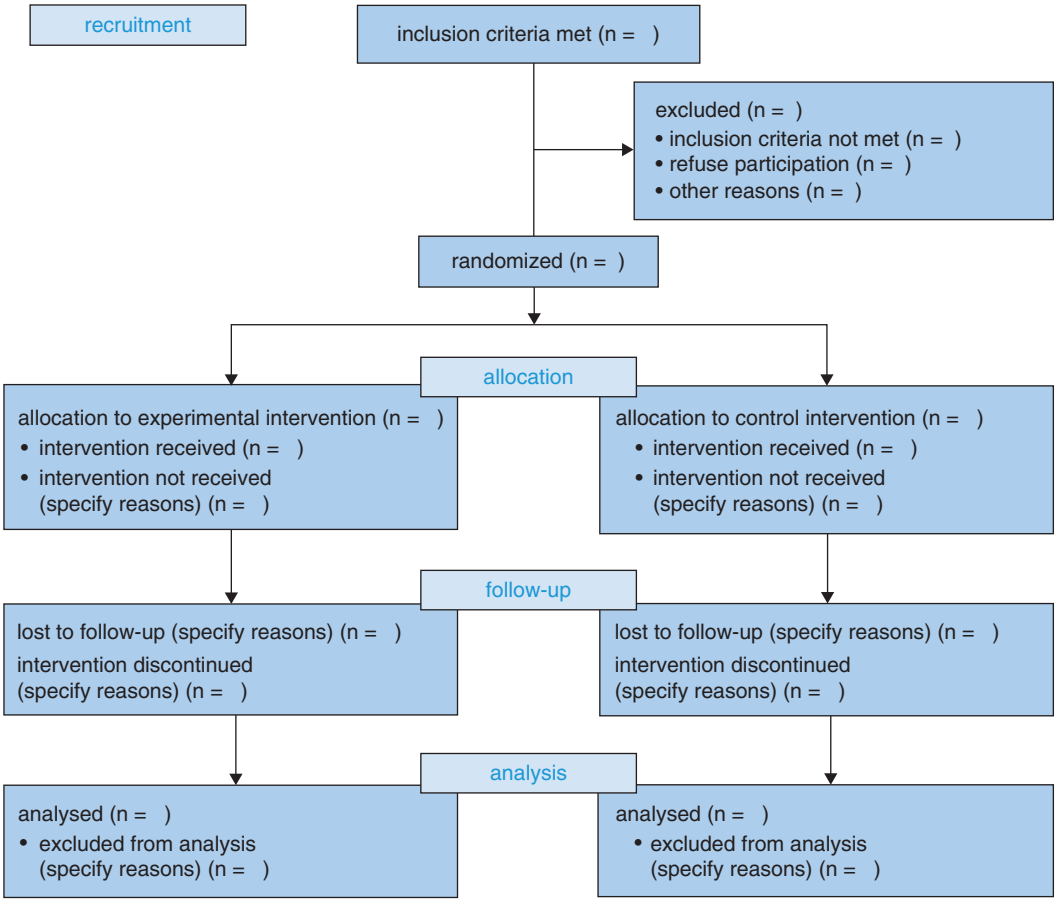


Figure 4.7 CONSORT 2010 flow diagram for the progress of trial participants

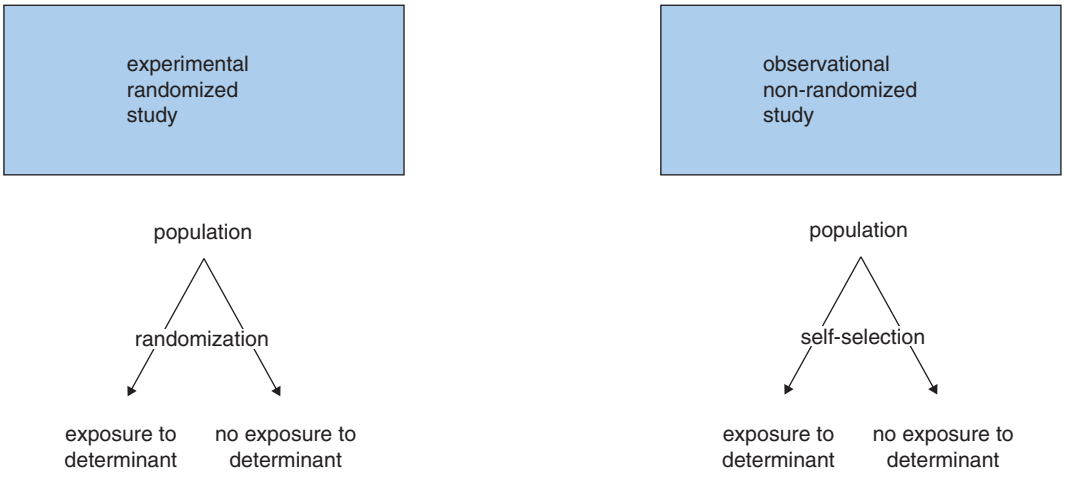


Figure 4.8 Similarities between an RCT and a cohort study

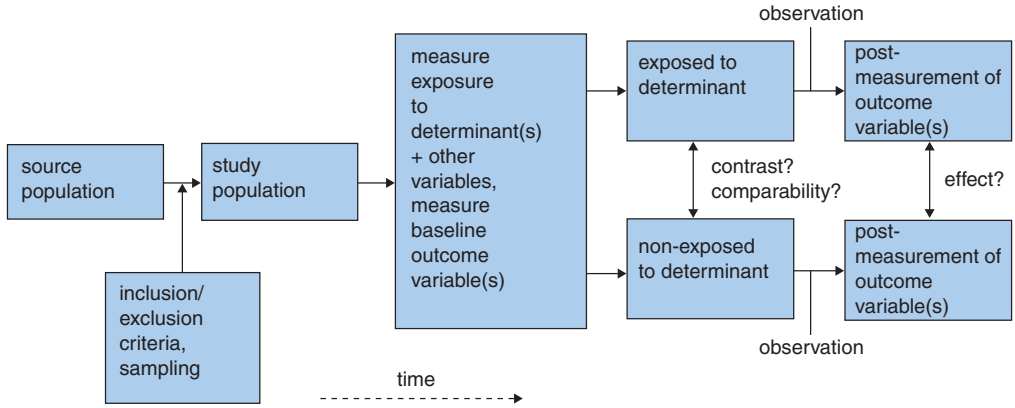


Figure 4.9 Basic structure of a cohort study

kers among the teetotallers (restriction) or ensuring that the percentage of smokers is the same among the drinkers and the teetotallers (matching).

Follow-up and outcome measurement

The follow-up period for which the outcome in the participants is recorded must be long enough for any effect of the primary determinant to be able to manifest itself. By recording the number of incident cases of the disease accurately we can calculate the cumulative incidence (CI) or incidence density (ID) in each subcohort (see ▶ par. 2.5.2). In the case of continuous outcome variables, for example, we calculate the mean value of the outcome at the end of the follow-up period, or the mean change in the value of the outcome during the follow-up period (see ▶ par. 2.6).

In a cohort study the outcome should ideally be measured 'blind', i.e. the assessor must not have prior knowledge of the participants' exposure status. The need for blind measurements will depend on the nature of the outcome being studied and the type of measuring procedure to capture this outcome. Blinding is more important in the case of subjective measurements.

The evidential value of a cohort study is determined mainly by the following factors:

1. The relevant determinants and outcomes are measured at the individual level.
2. The order in which the observations take place coincides with the natural course of the events.
3. Study participants are free from the primary disease at baseline. This primary outcome does not therefore play any role in the formation of the cohort (i.e. no selection bias: see ▶ par. 5.4.1).
4. The determinants (the primary determinant as well as the confounders) are measured at a time when the relevant health outcome is as yet absent. Any measurement errors in determining the determinant status are therefore independent of future disease status (there is no differential misclassification of the exposure measurement: see ▶ par. 5.4.1).
5. In many cases disease status can be measured blind (for the primary determinant), so that any measurement errors are not related to the determinant status (i.e. no differential misclassification of the outcome measurement).
6. The effect of the determinant on various disease frequencies can be assessed at the same time, provided that important potential confounders have been recorded for all these diseases.

Cohort studies have some limitations:

1. Because of the absence of randomization, placebo exposure and blinding of participants, there are fewer opportunities for creating comparable subcohorts and comparable circumstances as in an experiment. Confounding thus poses a serious threat to a cohort study. Any differences in these factors will have to be adjusted for in the statistical analysis – assuming that these factors have been properly measured.

2. A cohort study is not suitable for exploration; it can only be designed once well-founded hypotheses have been developed.
3. The classification of study participants according to determinant status as the basis for assembling cohorts is a snapshot. For example, people with high exposure to a harmful factor conceivably constitute a selective group, as many potential participants have already dropped out due to disease or death, and the remainder may be relatively immune to the disease. This is less of a problem with new exposures to a harmful factor (e.g. new employees of a company), where cohorts have a comparable past exposure status. But after formation of the cohorts the exposure pattern may change substantially, diluting the original contrast in determinant status. Exposure levels to the primary determinant need to be monitored during the follow-up period.
4. It is difficult to conduct a cohort study for diseases with low incidence and a long preclinical stage. Cohort studies therefore usually have large numbers of participants and long follow-up periods, although there are exceptions. For common conditions – e.g. the common cold, haemorrhoids or hypertension –, small cohorts may suffice, especially if the effect is substantial (i.e. a strong association). Cohort studies do not always need to take a long time either. For a study of congenital abnormalities due to exposures during pregnancy the follow-up period is generally confined to a maximum of nine months. If the incidence of the disease is low and/or it develops slowly a larger study population and/or a longer follow-up period is needed to accrue enough observation time.

There are various ways of collecting as much information as possible using as few participants as possible and at minimum cost:

1. By ensuring that there is sufficient contrast in the distribution of the primary determinant. As explained in ▶ par. 4.1.4, the size of the study population can be reduced substantially by selecting exposed and non-exposed persons.
2. By using determinant data that have been recorded in the past it is sometimes possible to project a cohort study into the past: this is referred to as a historical cohort study. In a historical cohort study the study population is defined and the cohorts created based on information that was collected and documented in the past. In practical terms this means having access to records or databases that bring together relevant data on large groups of people: for example, company records with information on working conditions and exposures, specified in periods for all employees and ex-employees. Based on these records we can then check which participants subsequently developed the disease. This type of study follows the basic pattern of a cohort study precisely, except that the time scale is unusual (see ◘ fig. 4.10). A historical cohort study is particularly useful to study the etiology of diseases with a long preclinical stage. In those situations a prospective study would take far too long to obtain enough incident cases of the disease. As historical records containing enough individual information on the primary determinant and potential confounders are relatively rare there are not so many opportunities for historical cohort studies. The opportunities for such studies have been substantially improved – at least potentially – with the introduction of large-scale storage of body materials in biobanks (see ▶ case 4.2). In addition, large quantities of digital information (Big Data) are collected and stored e.g. via the internet and social media and in electronic files. These data files often contain longitudinal records of demographic data, diagnoses, medical treatments and medication on millions of people – anonymously or in encoded form. Using data from these big datasets could make historical cohort studies very efficient, but lack of complete and valid data on exposure and confounders and/or the the fact that the data cannot be linked to information on disease status will often force researchers towards regular (prospective) cohort studies.
3. By carrying out case-control analysis within a cohort study (nested case-control study). As in a cohort study of a rare disease only a few participants will develop the disease during the follow-up period, it is inefficient to include all the participants who are still healthy at the end of the follow-up period in the analysis. A more effi-

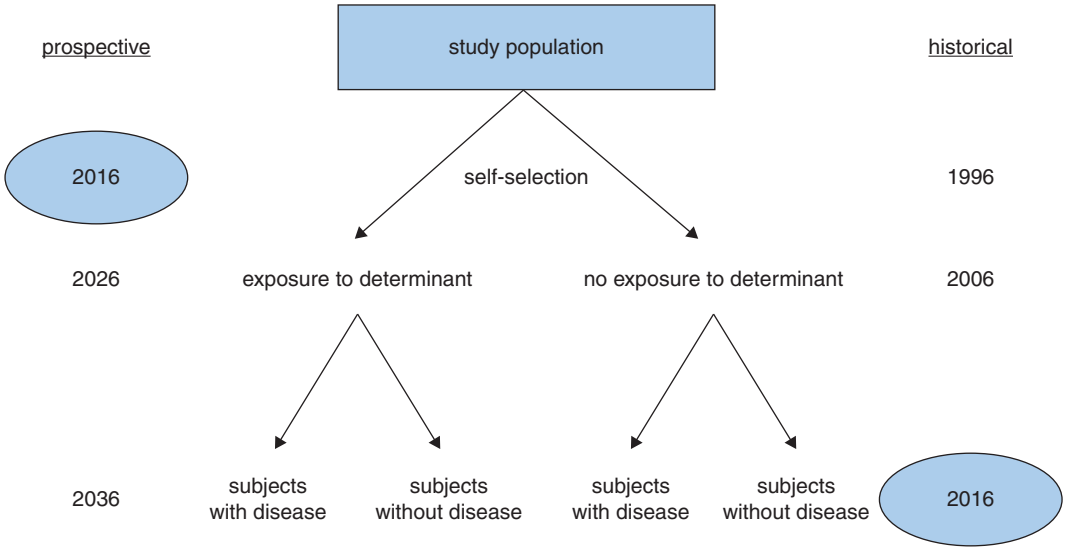


Figure 4.10 Similarities between a prospective and a historical cohort study

cient approach can sometimes be adopted: collect and store the relevant exposure data on all the members of the cohort at the start of the study if collection of data is easily feasible, but do not analyse and process them all. Record all the new cases of the disease during the follow-up period. Analyse all the available data and materials (determinant, confounders) for these cases, along with data and material from a random sample of the original cohort (which could theoretically include people with the disease). Compare the data on the primary determinant and the confounders for the cases and those for the reference group (see fig. 4.11). Performing case-control analysis within a cohort study improves efficiency especially when:

- analysis of all raw exposure data would be expensive or labour-intensive (e.g. processing dietary questionnaires, or analysing DNA in high densities or concentrations of substances in biological samples such as blood, urine, faeces, hair or nails)
- the primary media can be stored for a long time (e.g. frozen serum samples).

▶ Cases 4.2 and 4.3 provide some examples of cohort studies.

Case 4.2 UK Biobank, a large cohort study in the United Kingdom

UK Biobank is a major national population cohort enabling longitudinal studies in a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. UK Biobank recruited 500,000 people aged between 40-69 years in 2006-2010 from across the United Kingdom to take part in this project. They have undergone measurements, provided blood, urine and saliva samples for future analysis, detailed information about themselves and agreed to have their health followed. Over many years this biobank builds into a powerful resource to help scientists discover why some people develop particular diseases and others do not.

UK Biobank is hosted by the University of Manchester and open to bona fide researchers anywhere in the world. In addition to information collected during the baseline assessment, participants wear wrist activity monitors and complete detailed web-based questionnaires on their diet, cognitive function and work history.

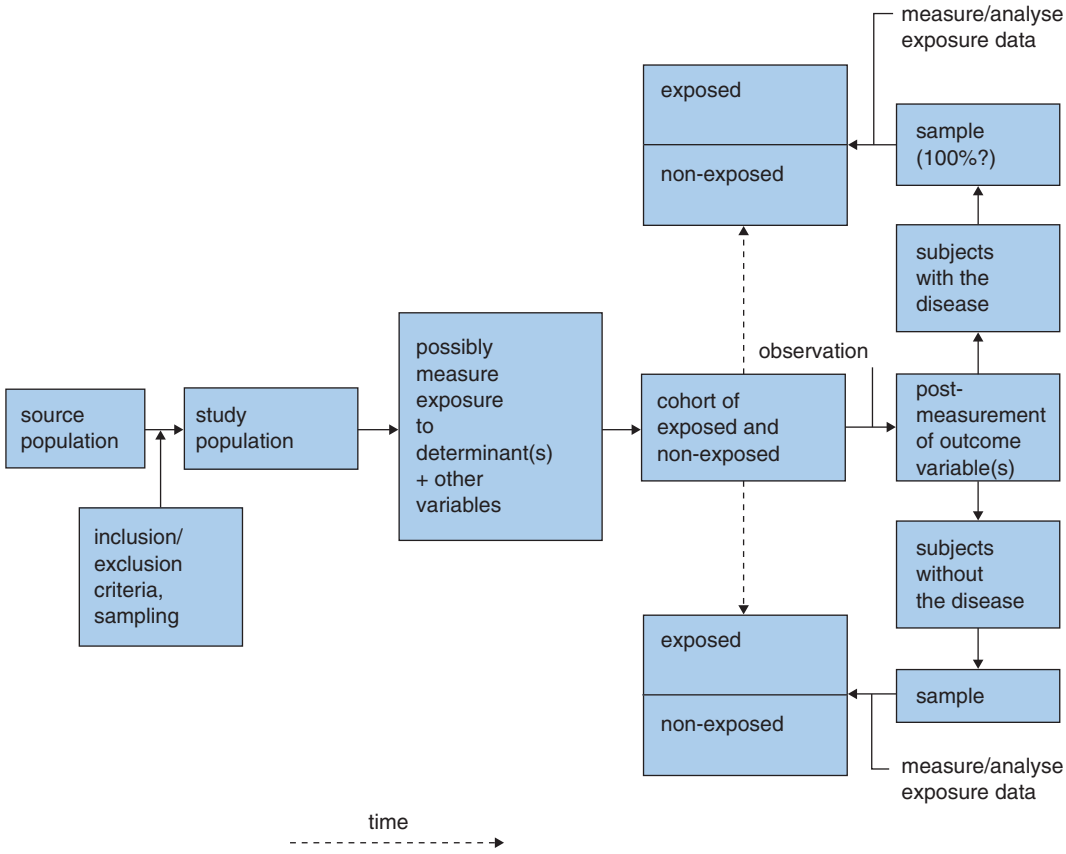


Figure 4.11 Structure of a case-control analysis within a cohort study

A substantial number of the participants have undergone MRI body and DEXA bone scanning. All data, including genetic, biochemistry and imaging data, are being made available for research as they become ready.

Case 4.3 Ionising radiation and risk of death from leukaemia and lymphoma in radiation-monitored workers (INWORKS): an international historical cohort study

To quantify the risks of leukaemia and lymphoma after repeated or protracted low-dose radiation exposure in occupational, environmental, and diagnostic medical settings an international consortium set up a large scale historical cohort study on leukaemia, lymphoma, and multiple myeloma mortality among radiation-monitored adults employed in France, the UK, and the USA. The investigators assembled a cohort of 308,297 radiation-monitored workers employed for at least 1 year by atomic energy companies in France, by the ministries of energy and defence in the USA, and by nuclear industries in the UK. The cohort was followed up for a total of 8.22 million person-years. All registered deaths caused by leukaemia, lymphoma, and multiple myeloma were recorded and Poisson regression was applied to quantify associations between estimated red bone marrow absorbed dose and leukaemia and lymphoma mortality.

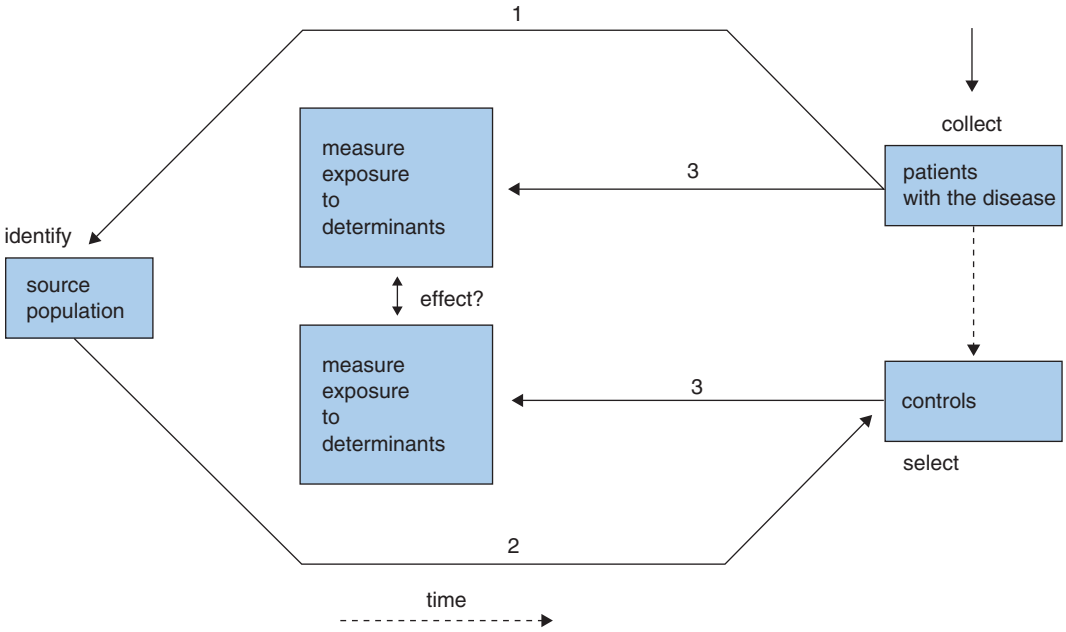


Figure 4.12 Basic structure of a case-control study

Radiation doses were very low (mean 1.1 mGy per year, SD 2.6). The excess relative risk of leukaemia mortality (excluding chronic lymphocytic leukaemia) was 2.96 per Gy (90% CI 1.17–5.21; lagged 2 years), most notably because of an association between radiation dose and mortality from chronic myeloid leukaemia (excess relative risk per Gy 10.45, 90% CI 4.48–19.65).

This study provided strong evidence of a causal association between protracted low-dose radiation exposure and leukaemia.

4.3.2 Case-control studies

As we saw in the previous paragraph, a case-control analysis can be embedded in a cohort study. Case-control studies[5] are usually carried out without a predefined population from which cases and controls are taken, however. The cases are collected e.g. from a hospital department, a general practice, a pathologist's records or an insurance company's records. The determinant status of these cases at a relevant time in the past then has to be ascertained retrospectively. The next challenge is to define the underlying source population from which the cases are taken (see fig. 4.12, Step 1). In essence this source population is the collection of all individuals (the dynamic population) who would have been assigned to the patient group if they had developed the disease under study. Once the source population is identified, we have the ideal sampling framework (study base) for a suitable control group (see fig. 4.12, Step 2). In practice this turns out to be far from simple, and we always need to satisfactorily demonstrate that the comparison group selected properly reflects the study base (the imaginary source population). Once we have selected a patient group and a suitable control group, we need to collect information on the past relevant determinants for all individuals in both groups (see fig. 4.12, Step 3). So a case-control study does not follow the course of life events over time but looks back at the individual history of the participants. This feature of case-con-

5 What is a case-control study? (animation). ▶ http://bit.ly/1EybOAL

trol studies makes them less transparent, but quick to carry out, as all the relevant events have already taken place. Another important feature of case-control studies is that, by approximating an efficient sample of the source population, only a limited number of participants is needed. This advantage is lost, however, if we are dealing with a rare exposure.

Designing and carrying out a case-control study presents a substantial challenge in practice. Let us briefly discuss the main complications.

Recruiting patients with the condition in question

The patients that we find e.g. in a hospital department will usually constitute a selection of the total number of people with the disease. Minor cases do not generally end up in hospital and the worst cases may already have died, for instance. Some hospitals attract patients from outside their region because they specialize in the particular disease. A proportion of the patients will be missed as a result of incorrect diagnosis, and patients with the risk factor may be over-represented because the presence of that factor has been used to diagnose them.

These various mechanisms can select in such a way as to potentially jeopardize the validity of the study. To avoid selection bias and information bias, incident (new) cases of the disease should preferably be used. Confounding can be prevented to some extent by applying strict inclusion and exclusion criteria when selecting patients.

Selecting the control groups

The purpose of having a control group is to show the frequency of the primary determinant in the population from which the cases are taken. We then compare this 'normal' distribution of the various determinants with that found among the cases. The cases and controls will usually be taken from the population of people who could develop the disease: in a study on the use of oestrogen and endometrial cancer we would look for female controls with a uterus, for instance. Strictly speaking, however, a control group can be selected from a population other than that from which the cases are taken, provided that the distribution of the exposure being studied is the same in both populations. For example, if we are examining the relationship between blood group and endometrial cancer, male infants could provide a suitable control group.

While there are often several options for the selection of a control group, we need to choose which is the best from the point of view of validity and efficiency. Some common options:

- A *population-based control group*. This is a sample from the general population (e.g. taken from the population register). We then assume that the population from which the controls have been selected is the same as the population from which all the cases have been taken (the same source population). The assumption is valid when incident cases are recruited, for example, from a population cancer registry. Often, however, it cannot usually be tested properly, as hospital patients have been accrued through unknown referral patterns, crossing of geographical borders and personal preferences. A disadvantage of this kind of control group from the general population, then, is selection bias if cases are recruited at hospital level. Selective non-response (smokers tend not to participate in a control group, for instance) and recall bias (when patients remember certain events associated with the disease better than healthy controls) can cause bias. Collecting controls from the general population is also often time-consuming and expensive.
- A *hospital-based control group*. This is made up of patients from the same hospital but suffering from a different condition. Such patients are easier to find and contact, as they are usually more motivated to take part in a study than healthy controls. Another advantage is that the risk of recall bias is probably smaller (as the controls are themselves patients). It is important, however, to select for the control group only patients suffering from conditions not related to the determinant being studied, otherwise we may be comparing two diseases related to the same exposure, which will cause bias. This phenomenon is known as Berkson's fallacy. For example, if we were examining what risk of having a heart attack is associated with smoking and were to use lung cancer patients as the control group, obviously the controls would not accurately reflect the source population.

Table 4.1 Major differences between a case-control study and a cohort study

case-control study	cohort study
relatively inexpensive	often expensive
quick results	often a long wait
relatively small study population	relatively large study population
suitable for rare diseases	suitable for common diseases
unsuitable for rare exposures	suitable for rare exposures
complete information	risk of loss to follow-up, selective dropout
susceptible to bias, especially information bias (exposure measurement), selection bias, confounding	less susceptible to bias, but possible information bias (outcome measurement) Confounding, selective dropout, changes in measuring procedures
only OR can be calculated (RR estimated), not incidences (risks)	incidences (risks), AR and RR can be calculated directly

- A *peer-based control group*. A quick, simple and inexpensive way of collecting controls is to ask patients to suggest friends or contacts from their social network. This provides a control group containing people who are comparable with the patients. The danger, however, is that the controls will also be comparable in terms of the determinant being studied.
- A *family-based control group*. If we are prepared to ignore the effects of genetic factors, it may be useful to select family (blood relatives) of the patient as controls: in studies of twins, for example. There is family clustering of many environmental factors, however, so this method is not suitable for studying exposures that correlate highly within families (tab. 4.1).

Measuring exposure

In a case-control study we usually have to rely on questioning to measure the determinant under consideration in the time period relevant to the development of the disease, which means relying on the respondents' memories. Some determinants cannot be ascertained at all using this method (e.g. concentrations of certain substances in the body), while others cannot be ascertained accurately (complex behaviours, e.g. diet during the teenage years). There is a real danger of exposure measurement being affected by the fact that the researcher and the respondent with the disease know what relationship the study is looking for. Blinding is by no means always feasible. When measuring the determinant it is also vital to ask about the correct time period in the past, which means gauging the etiological moment correctly. What we have said about measuring the primary determinant also applies mutatis mutandis to potential confounders and effect modifiers (see ▶ chap. 5). Similar errors can be made when measuring these in case-control studies, with all that this entails.

Matching

Matching is sometimes used in case-control studies. This involves selecting controls who are similar as possible to the patients as regards important confounding variables. When it comes to genuine confounders this strategy is more efficient than subsequently adjusting for confounding in a multivariable data analysis within a non-matched case-control study. Matching can be done in two ways. Individual matching involves finding a suitable control with the same characteristics for each patient. Frequency matching involves selecting controls in such a way that the distribution of the matching variables among them is comparable to that among the patients. Matching also has its disadvantages. Once

Table 4.2 Calculating an odds ratio in a matched case-control study

controls	cases	
	exposure present	exposure absent
exposure present	a	b
exposure absent	c	d
OR = b/c		

matching has been done for a particular factor, the effect of that factor cannot be studied, for instance. Matching factors strongly associated with the primary exposure cause unintended masking of the relationship being studied (overmatching). When designing the study we need to check whether the advantages of matching outweigh the additional time, money and energy involved in finding suitable controls. Note that when calculating odds ratios from case-control data with individual matching there is a special statistical method of analysis based on matched case-control pairs (see tab. 4.2). If we were to mistakenly calculate odds ratios using the standard method (see ▶ chap. 3) we would underestimate the strength of the relationship. Matching is not only used in case-control studies. It can also take place in cohort studies (where exposed and non-exposed groups are matched to similar characteristics) or in RCTs, as block randomization is also some sort of matching (see ▶ par. 10.2.3).

Case 4.4 Traveller's diarrhoea, a case-control study

Suppose we want to examine whether consumption of tequila, a local hard liquor, is the cause of traveller's diarrhoea in German tourists on holiday in the city of Acapulco in Mexico. To answer this question we collect all the German patients admitted to hospital in Acapulco with traveller's diarrhoea. We then investigate their tequila consumption prior to the diarrhoea episode. The next question is what is most suitable as a control group. A representative sample of the inhabitants of Mexico is not suitable, as their tequila consumption is different from that of the average tourist. A representative sample of Germans will not do either, as only Germans who travel to Mexico could drink tequila there or be admitted to hospital in Acapulco. The best control group is other German travellers who arrived in Acapulco at the same time as the patients and did not have traveller's diarrhoea during their stay there.

Case 4.5 Smoking and lung cancer, a case-control study

Richard Doll and Bradford Hill, two of the best-known epidemiologists of the twentieth century, were the first to track down the link between smoking and lung cancer using a case-control study – one of the most important epidemiological findings to date, which had a major impact on public health. In order to explain the high mortality from lung cancer in London, in 1950 Doll and Hill started questioning all patients with lung cancer in twenty London hospitals. When visiting a lung cancer patient they looked for a control at the same hospital of the same gender and approximately the same age but without cancer. If they were unable to find suitable hospital controls at that hospital they selected controls from other hospitals in the area. Each patient's diagnosis was checked using histopathology. In this way Doll and Hill were able to include 709 lung cancer patients and an equal number of matched controls in their study. Although they initially hypothesized that lung cancer was related to car exhaust fumes they found that the association with smoking was much stronger. Virtually all the lung cancer patients had smoked (99.7% of the men and 68.3% of the women), whereas the figures were far lower among the controls (95.8% of the men and 46.6% of the women). When they looked at the number of cigarettes

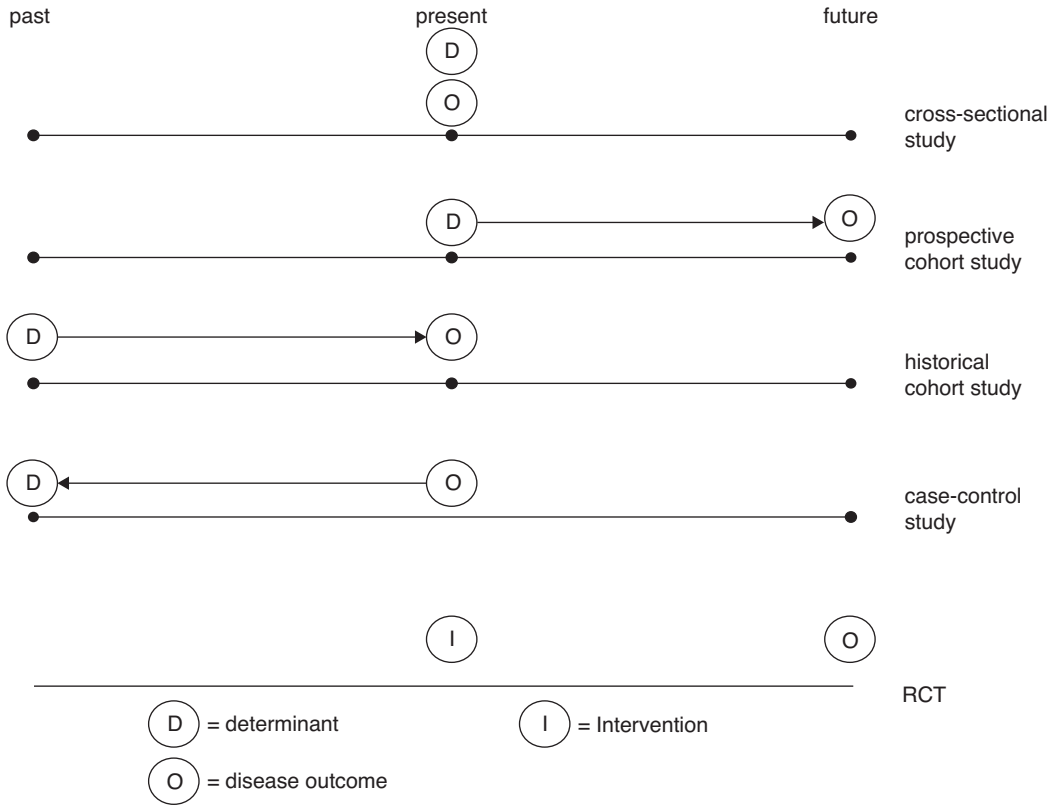


Figure 4.13 The temporal relationship between observation times in the various epidemiological study designs

smoked they found the effect was even stronger: the risk of developing lung cancer was 50 times greater in those who smoked 25 cigarettes a day than in non-smokers. Based on his findings Bradford Hill stopped smoking himself.[6]

4.3.3 Cross-sectional studies

The study designs discussed earlier are all longitudinal. There are at least two different observation times for each individual, one for the determinant under consideration and the confounders, and one for the outcome. In a cross-sectional study (transversal study, survey) the primary determinant, the outcome and all confounders are measured at the same point in each individual's history (see fig. 4.13), which will not necessarily be the same chronological time for all the individuals at issue. In a cross-sectional study all variables are measured and their associations studied in a random sample from the general population. Or we can select the cases and controls and measure the current value of the determinant: this would be a case-control study, with a cross-sectional rather than a longitudinal time dimension. Alternatively, determinant status can be used as an initial criterion for recruiting the participants: this would be a cohort study with a cross-sectional time dimension. The weak point of a cross-sectional study is that we are usually not certain whether the exposure to the determinant took place before the occurrence of the disease. This is

6 Doll and Hill's original publication on the relationship between smoking and lung cancer (PDF). ▶ http://1.usa.gov/1HvJwat

usually not a problem in the case of descriptive epidemiological research questions, e.g. in diagnostic research (see ▶ chap. 9), but it is a problem in cause-and-effect research (etiology, intervention research). If we want to know whether stress is a cause of heart attacks, there is no point in measuring the degree of tension in a group of patients recently admitted to hospital with an acute heart attack and a group of controls. Nor will a researcher find an answer to whether there is a causal relationship between regularly wearing a hat and baldness by measuring the degree of baldness in a sample of men with and without a hat. Only if we can satisfactorily demonstrate that the measured value of the determinant under consideration remains stable over a long period (e.g. blood group, gender, personality and other genetically determined characteristics) can we use a cross-sectional study for etiological research.

4.3.4 Ecological studies are useful to explore a topic, but fallacies lie in wait if we want to delve deeper

In the types of study described so far the observations are always carried out on individuals and the comparison is between individuals. An ecological study compares groups of individuals (populations). Examples of groups include countries, smaller regions, general practices and schools. Disease outcome and determinant are regarded as group characteristics. Population averages are presented that are usually obtained from disease and determinant data collected as a matter of routine, such as consumption data for foods and stimulants and statistics on mortality or hospital admissions. To gain an understanding of the relationship between disease and particular determinants we can compare two or more populations – usually geographical units – at a particular point in time (geographical correlation studies), or we can examine the same population at two or more different times (time trend studies). A hybrid approach can also be adopted. We could look, for example, at the differences in disease frequency between a region where a new medical facility has been introduced and a comparable region where this has not been done (simultaneous comparison). Or we could look for differences in disease frequency in a region between the periods before and after the introduction of a new facility (pre-post comparison). The fact that aggregate data are used to examine the relationship between the exposure under consideration and the disease is the Achilles heel of ecological studies. If we incorrectly translate associations at the population level to the individual level we fall into the ecological fallacy, for example because there is a systematic difference in background risk (prevalence) between the groups we are comparing. Suppose we want to investigate whether people who drink excessively are at greater risk of dying from an accident or violence, and we want to answer this question using an ecological study design. We have both the mortality rates for accidents and violence and the percentage of excessive drinkers in three cities. Mortality from accidents and violence is 35% in City A, 45% in City B and 55% in City C. 20% of the population are alcoholics in City A, 40% in City B and 60% in City C. By comparing these numbers we can calculate that the risk of death from accidents or violence is four times higher for excessive drinkers than for non-excessive drinkers. In reality the relative risk in this hypothetical example will be much lower. In any epidemiological study we need to check whether the research question is about individual risk or group risk. If it is about individual risk, the most that an ecological study can do is generate ideas about possible causes of the disease. For the actual study we will need to select a study design that has the individual as the unit of observation and analysis, so as not to fall into the ecological fallacy. If we have no information on the drinking habits of individuals we cannot be sure that it is drinkers who have high mortality in City C. Conversely, using individual data to answer a research question at group level (e.g. does a population including a large number of people who drink alcohol excessively have an increased mortality rate from accidents and violence?) can result in falling into a similar trap, where an ecological study would produce a better answer.

Guidelines for observational studies

There are now STROBE guidelines on the publication of observational studies.[7] Like the CONSORT guidelines on experimental studies, they provide specific advice on the elements that should be included in an article on a cohort study, case-control study or cross-sectional study. There are also checklists to rate the quality of observational studies, the best-known being the Newcastle-Ottawa Scale. It is difficult to assess the quality of observational studies using a single checklist, however, as study designs can differ substantially.

Recommended reading

Armstrong BK, White E, Saracci R. Principles of exposure measurement in epidemiology. Oxford: Oxford University Press, 2008.

Ahrens W, Pigeot I. Handbook of epidemiology. 2nd ed. New York: Springer, 2014.

Bonita R, Beaglehole R, Kjellstrom T. Basic Epidemiology. 2nd ed. Geneva: World Health Organization, 2006.

Carneiro I, Howard N. Introduction to epidemiology. 2nd ed. Maidenhead: Open University Press, 2005.

Elwood JM. Critical appraisal of epidemiological studies and clinical trials. 3rd ed. New York: Oxford University Press, 2007.

Fletcher RH, Fletcher SW, Fletcher GS. Clinical epidemiology: the essentials. 5th ed. Baltimore: Lippincott, Williams & Wilkins, 2012.

Gordis L. Epidemiology. 5th ed. Philadelphia: Elsevier Saunders, 2014.

Gregg MB. Field epidemiology. 3rd ed. New York: Oxford University Press, 2008.

Grobbee DE, Hoes AW. Clinical epidemiology: principles, methods, and applications for clinical research. 2nd ed. Burlington: Jones and Bartlett Learning, 2015.

Guyatt G, Rennie D, Meade MO, Cook DJ. Users' guides to the medical literature: a manual for evidence-based clinical practice. 2nd ed. Chicago: AMA Press, 2008.

Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: how to do clinical practice research. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins, 2006.

Keating C. Smoking kills: the revolutionary life of Richard Doll. Oxford: Signal Books, 2009.

Piantadosi S. Clinical trials: a methodologic perspective. 2nd ed. New York: John Wiley & Sons, 2005.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins, 2012.

Webb P, Bain C. Essential epidemiology: an introduction for students and health professionals. 2nd ed. Cambridge: Cambridge University Press, 2011.

Source references (cases)

Kieviet JF de, Vuijk PJ, Berg A van den, Lafeber HN, Oosterlaan J, Elburg RM van. Glutamine effects on brain growth in very preterm children in the first year of life. Clin Nutr. 2014, 33 (1), pp. 69–74 (Case 4.1).

► www.ukbiobank.ac.uk (Case 4.2).

Leurand K, Richardson DB, Cardis E, Daniels RD, Gillies M, O'Hagan JA, Hamra GB, Haylock R, Laurier D, Moissonnier M, Schubauer-Berigan MK, Thierry-Chef I, Kesminiene A. Ionising radiation and risk of death from leukaemia and lymphoma in radiation-monitored workers (INWORKS): an international cohort study. *Lancet Haematol*. 2015, 2, pp. e276–e281 (Case 4.3).

Miettinen OS. The "case-control" study: valid selection of subjects. J Chron Dis. 1985, 38, pp. 543–8 (Case 4.4).

Doll R, Hill AB. Smoking and carcinoma of the lung: preliminary report. Br Med J. 1950, 2 (4682), pp. 739–48 (Case 4.5).

7 STROBE guidelines for the publication of observational studies (website). ► http://bit.ly/1DGDWe0

Validity and reliability

5.1 Introduction: does the parameter estimate give a valid and reliable picture of what is going on?

You will undoubtedly have noticed that characterizing the epidemiological function is pivotal in epidemiological research. Here is that function once again:

$$P(O) = f(b_0 + b_1D_1 + b_2D_2 + \ldots + b_kD_k)$$

To put it in words: the probability of a disease outcome (P(O)) is described as a mathematical function of a series of determinants D_i (where i = 1, …, k). The regression coefficients in the function indicate the strength of the link between the determinant in question and the disease outcome. We want the estimated regression coefficient, and the epidemiological measure of association derived from it, to accurately reflect the true link between the determinant and the outcome. For example, if a logistic regression equation yields a regression coefficient of 0.6 (corresponding to an OR of Exp(0.6)=1.8), that number must be a valid and reliable representation of the link between that determinant and the disease outcome being studied. It must not be the case that systematic or random distortion causes the observed odds ratio to be much higher or lower than the true value. A small discrepancy between the estimated parameter and the true strength of the link is not disastrous (e.g. an odds ratio of 1.7 or 1.9 in this case), but the discrepancy must not be so large as to lead to completely different conclusions about the nature of the link. An odds ratio of 1.8 is generally regarded as a fairly weak association. An odds ratio of 1.0 would be interpreted as no link, whereas one of 6.0 would be interpreted as a strong link.

5.2 Validity and reliability in metaphor

When estimating measures of frequency and measures of association in epidemiological research, just as when measuring variables, two types of errors can occur: random errors and systematic errors. These errors lie at the root of reliability and validity problems in epidemiological research.

The difference between validity and reliability can be easily illustrated with the results of a series of gunshots at a target: someone with a trembling (unreliable) hand shooting with a straight (valid) gun may hit the bull's-eye by chance but will be more likely to hit around it. A steady (reliable) hand with a warped (invalid) gun will always produce hits at the same – wrong – spot unless the person shooting receives feedback on previous shots. Only someone with a steady hand and a straight gun will always hit the bull's-eye. The bull's-eye of a shooting target corresponds to the true value of the measure of frequency or measure of association in epidemiological research (see fig. 5.1).[1] Systematic errors are due to wrong decisions made during the design or execution of the study or the data analysis. Random errors result in unreliable estimates of effect, and systematic errors result in invalid estimates of effect. This chapter looks first at validity and reliability in study designs, moving on to the validity and reliability of instruments in ▶ par. 5.7.

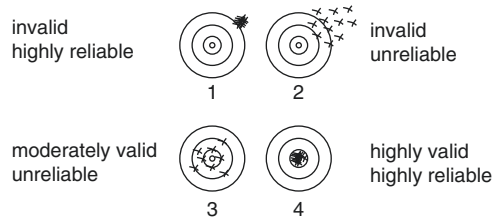


Figure 5.1 Validity and reliability expressed as a series of gun shots at a target

5.3 Reliability: the same results on repetition

A study design that repeatedly produces the same result (an incidence, relative risk, difference of means, etc.) has high reliability. In other words, a reliable study must be reproducible, i.e. the result should always be more or less the same if the study is repeated (with the same design and method).

1 Validity and reliability in an American rural area (video). http://bit.ly/1Eydt9r

5.3.1 Sample size and distribution determine the degree of reliability

Random errors are due to the fact that the phenomena that epidemiologists study (e.g. the link between the occurrence of a disease and exposure to a risk factor) vary around the true value. Random errors can be due not only to measuring errors in individual observations (see ▶ par. 5.7) but also to the fact that every epidemiological study is based on observations of samples. The individuals on whom the observations are carried out (the study population) are a more or less random representation of a larger source population on which the researchers want to collect information. The study population could have contained all the individuals who met the study criteria, but only a certain number of them have actually been selected. The result of the sampling is based purely on chance, so sampling error (due to that random effect) is a major factor in the reliability of epidemiological research.

Estimates of measures of frequency and association must be sufficiently reliable, i.e. the magnitude of the random errors must be limited in relation to the true value of the parameter. In statistics, sampling error is expressed as a standard error (SE), which can be interpreted as the standard deviation of the measured parameter if the study was to be repeated many times using other samples from the same source population. The formulas for calculating the standard error of a proportion, the mean, the difference of means, the odds ratio and the relative risk are shown in ◘ fig. 5.2. We can see from each formula that – as we would expect – the larger the sample, the smaller the standard error will be.

Not only the measures of frequency and association already mentioned but any population parameter estimated from a sample will have a standard error. Generally speaking, the standard error is determined by three factors:

- The measuring error of the separate observations of the individuals in the sample
- The distribution of the parameter in the population from which the sample has been taken
- The size of the sample.

SE of a proportion $\sqrt{\frac{p\,(1-p)}{n}}$

SE of the mean $\frac{s}{\sqrt{n}}$

SE of a difference of means $\sqrt{\frac{s_1^2 + s_2^2}{n}}$

SE ln(OR) $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

SE ln(RR) $\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}$

	cases	controls
exposed	a	b
non-exposed	c	d

◘ **Figure 5.2** Calculation of the standard error (SE) for a proportion, a mean, the difference of means, the logarithm of the odds ratio (OR) and the logarithm of the relative risk (RR). p: proportion; s: standard deviation of the measurement in the sample; n: sample size; a–d the cells in the above contingency table

The standard deviation, then, should not be confused with the standard error. The standard error indicates the variation in a population parameter (mean, proportion, etc.) estimated from a sample. The standard deviation indicates the variation in the individual measurements (see ▶ par. 2.6).

5.3.2 Reliability is expressed as a confidence interval

To indicate how precise the estimate of an epidemiological parameter is we use a confidence interval (CI). A 95% confidence interval is formally defined as the interval in which the true value would lie in 95% of cases if the study were to be repeated many times. Describing it more loosely, we could say that the true value of the measure of association is 95% likely to lie within the 95% confidence interval. In other words, the confidence interval gives a good idea of how reliable the results are. The smaller the interval, the more precise the estimate. More (e.g. 99%) or less strict (e.g. 90%) confidence intervals can also be calculated. The higher the reliability required, the stricter the interval should be. A confidence interval only provides information on the re-

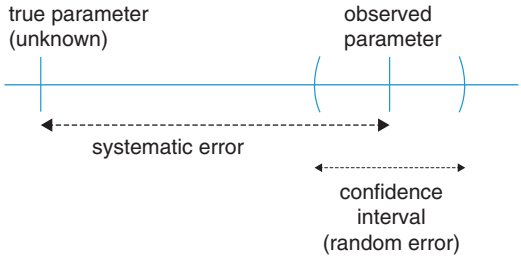


Figure 5.3 A confidence interval only provides information on the reliability (random error) of an epidemiological parameter, not on its validity

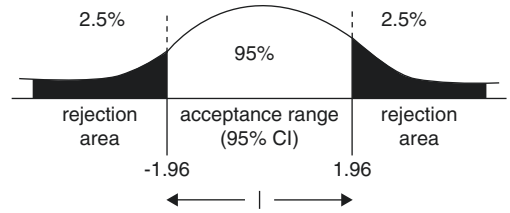


Figure 5.4 The standard normal distribution

liability (random error) of an epidemiological parameter, not on its validity (systematic error), as it is quite possible that the true value of the estimated parameter lies outside the confidence interval (see fig. 5.3).

In order to calculate confidence intervals we need an estimate of the standard error (SE) of the estimated parameter (P) (see fig. 5.2) and a translation of the confidence selected into a z-score on the standard normal distribution (e.g. 95% confidence corresponds to a z-score of 1.96) (see fig. 5.4). The 95% confidence interval of an epidemiological parameter P which has an approximately normal distribution is thus:

$$95\%\text{–CI} = \text{P} \pm 1.96\,\text{SE}_\text{P}$$

Parameters such as RR and OR do not have a normal distribution but a log-normal distribution. The 95% confidence interval for ln(P) is thus:

$$95\%\text{–CI} = \ln(\text{P}) \pm 1.96\,\text{SE}_{\ln(\text{P})}$$

The result of these calculations can be interpreted as follows. Say an estimate of a cumulative incidence (CI) in a population yields the following result: CI = 5 per 10,000 = 0.0005. At a particular size of study population the 95% confidence interval would then be e.g.:

$$95\%\text{–CI} = 0.0005 \pm 0.0001 = (0.0004 - 0.0006).$$

This means that we can assume a 95% likelihood that the true cumulative incidence in the population from which the random sample has been selected lies between 4 and 6 per 10,000.

An estimate of the RR in a particular study, for example, yields the following result: RR = 1.52 (95% CI 1.27–1.81). This means that we can be fairly certain that there is a real, slightly increased risk. That would not have been the case if we had obtained a 95% CI of 0.82–2.33 for this RR of 1.52, as such a confidence interval includes cases where there is no increased risk (RR = 1), where there is a slightly increased risk (RR between 1 and 2) and where there is a moderately increased risk (RR between 2 and 3).

Sometimes reliability is expressed not by a confidence interval for a point estimate of an epidemiological parameter but as the result of a statistical test, for example in the form of a p-value. A p-value of 0.03, for instance, expresses that, if the null hypothesis (that a determinant has no effect on the disease) is correct, there is only a 3% likelihood (p = probability) of encountering the value found – or an even more extreme value of the effect measure – by chance. Given the criterion (α) that indicates what probability of error is regarded as acceptable (generally 5%), this probability is regarded as low enough for us to say that the determinant has a significant effect on the disease (see fig. 5.4).

The use of statistical tests is based on the idea that we ultimately want to distinguish solely between effect parameters (difference of means, RR, etc.) that can or cannot be explained by random variation with a degree of plausibility. A p-value of 0.03, then, indicates that the likelihood that this result is purely due to chance (sampling error) is so small that the null hypothesis should be rejected. Similarly, it is standard practice to interpret a p-value higher than 0.05 as indicating that the result obtained could have been caused by chance. Clearly, the p-value is not a gauge of the magnitude of an effect. Statistical tests answer the question of

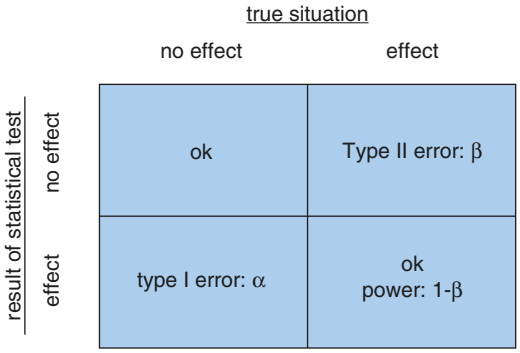


Figure 5.5 The relationship between type I errors, type II errors and power

whether there is an effect, not how large it is or could be. To estimate the magnitude of an effect, then, the p-value will not suffice; we are better off using a confidence interval.

5.3.3 Reliability is improved by a larger sample and more efficient designs

To calculate the number of participants needed to obtain results with a particular reliability we have a large arsenal of sample size formulas at our disposal. These are included in most statistical software packages as well, for instance in the iPhone app WhatStat.[2]

In order to use the sample size formulas we need information on the following parameters in the case of dichotomous outcomes and determinants:

1. The disease frequency in the reference group, if necessary estimated based on the frequency in the source population.
2. The minimum magnitude of the effect that we would consider to be worthwhile to demonstrate. Obviously, we need to include fewer people in the study when the difference (in disease frequency or exposure frequency) that we expect to find between groups is larger.
3. The probability that we regard as acceptable of finding an effect that is actually absent (type I error or α-error). For example, a type I error of 0.05 means a 5% risk of a statistically significant effect that does not actually exist. The lower we want to make this probability, the more people we need to include in the study (see also fig. 5.5).
4. The probability that we regard as acceptable of missing a true effect (type II error or β-error). A type II error of 0.20, for instance, means a 20% risk of missing an effect that actually exists. The lower we want to make this probability, the more people we need to include in the study (see also fig. 5.5). The complement of the type II error (1-β) is the power, i.e. the probability of detecting a specified true effect in the study.
5. The desired ratio between the numbers of participants in each of the groups being compared, e.g. three times more controls than patients in a case-control study.

Reliability in a study depends not only on the size of the study population but also on how that population is assembled and divided into the various groups being compared. The efficiency of an epidemiological study is essentially related to the total amount of factual information contained in the study data. The more participants – but also the larger the average amount of information per individual –, the higher the statistical efficiency. We have already seen in ▶ chap. 4 how the efficiency of a study can be increased by the careful selection of informative study groups, an example being matching participants as part of a cohort study or case-control study. One aim of matching is to enable us to say as much as possible about the presence and magnitude of an effect based on the smallest possible number of participants.

5.4 Validity: absence of bias

Although a result may be very reliable (i.e. yield the same estimate of effect on repetition), it may nevertheless be completely wrong because of a systematic error in the study design – bias. Research of this kind is not valid, it is distorted. This section

2 The iPhone app WhatStat, for calculating e.g. sample size (University of Utrecht) (app). http://apple.co/1aXGu09

discusses bias in measures of association, not in frequencies (means) of determinants or outcomes.

5.4.1 The three threats to validity: selection bias, information bias and confounding

Bias is when associations systematically differ from reality (see fig. 5.3). A host of mistakes can be made that lead to invalid conclusions on the relationship between determinants and outcome. Broadly speaking, these can be classified into three types of bias: selection bias, information bias and confounding. Bias can cause the effect to be overestimated, underestimated or even observed as going in the opposite direction.

Suppose we want to examine the link between a determinant and an outcome using a case-control study. The true association is OR = 2.5. Errors could result in the study producing an OR of 5.0: this is an overestimate of the effect, also referred to as positive bias or bias away from the null value (= no effect). The study could alternatively produce a result of OR = 1.5: this is an underestimate, also referred to as negative bias or bias towards the null value (= no effect) (see also fig. 5.6). The study could even produce an OR of 0.7: here the effect changes into the opposite direction (switch-over bias). In this example the determinant was actually a harmful factor, but as a result of errors in the design the study suggests that it is a protective factor (see also fig. 5.7).

We need to anticipate potential bias during the design and execution of the study, since once it has been conducted there is not much that can be done about any remaining bias. Although the direction of the resulting systematic error might be indicated in most instances, it is usually not possible to ascertain the magnitude of the bias. Nevertheless, all information on the direction and potential size of bias, however limited and incomplete it may be, is vital to the interpretation of the results. For example, if a case-control study finds a fairly clear increased risk (OR = 3.0) and it can be argued that any bias must be towards the null hypothesis, the true OR will be greater than 3.0 and there is almost certainly an association. If only a very slightly increased risk is found (OR = 1.3) and it can be argued that any bias is away from the null, the true OR is evidently less than 1.3 and what remains has little if any value. The true OR could even be less than 1.0, in which case the true effect is in the opposite direction. And if a substantially increased risk is found (e.g. OR = 8.0), there must be a very strong bias away from the null hypothesis if there is actually no effect.

The ensuing subsections look at the trio of selection bias, information bias and confounding in more detail.

Selection bias is a distortion of the estimate of effect due to errors in the selection or follow-up of the participants. Information bias is a distortion of the estimate of effect due to errors in measuring the variables in the study (determinant, outcome, confounders, effect modifiers) and hence in the classification of the study population into subcategories based on the measured variables. Confounding results in a distorted estimate of effect because the groups being compared in the study have been exposed in various ways to other risk factors for the disease. This causes the effect of the primary determinant to become mixed up with that of the other determinants ('confounders' or 'confounding variables') that we are not currently interested in.

These three categories of distortion can overlap to some extent. Confounding can be introduced, for instance, by the selection of the participants. Nor is it always possible to make a clear distinction between confounding and information bias. The term 'confounding' is usually reserved for known measurable risk factors for the disease in question that are unequally distributed among the groups being compared. If these variables have actually been measured, it is usually possible to adjust for them at the analysis stage of the study. Selection bias and information bias cannot be eradicated once the study data have been collected, unless the exact magnitude and direction of the selection and measuring errors are known, which is seldom the case in practice.

Selection bias

Selection bias occurs when the probabilities of people being included in the study population depend on the outcome being studied (in a cohort study) or on the determinant being studied (in a case-control study). We shall discuss this phenomenon separately for cohort studies and case-control studies.

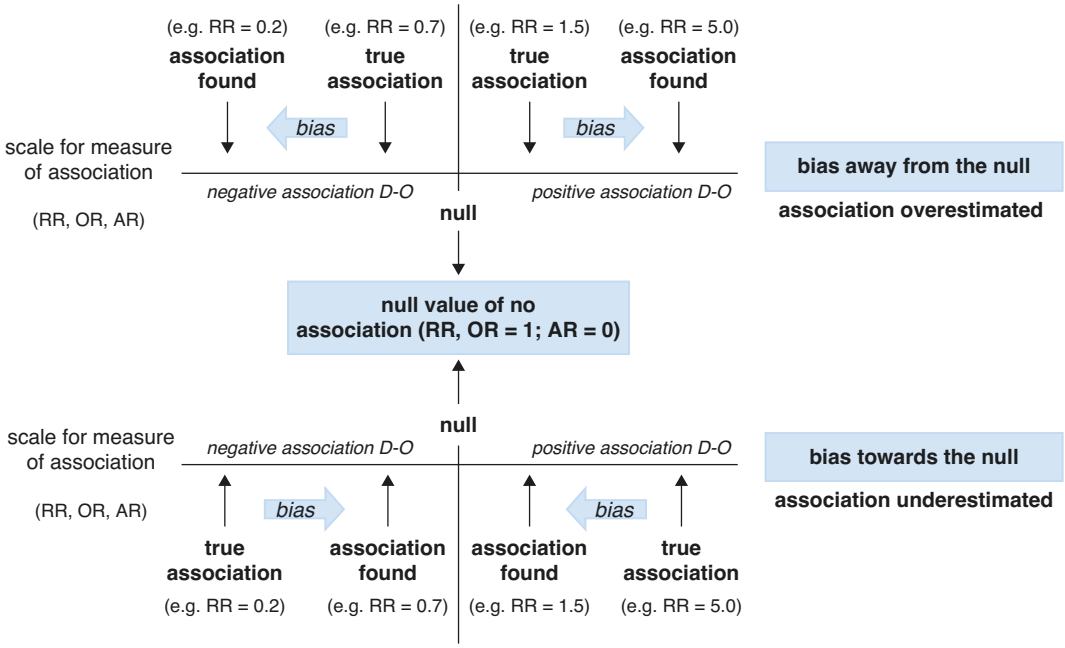


Figure 5.6 Quantitative bias in the association between determinant (D) and disease outcome (O)

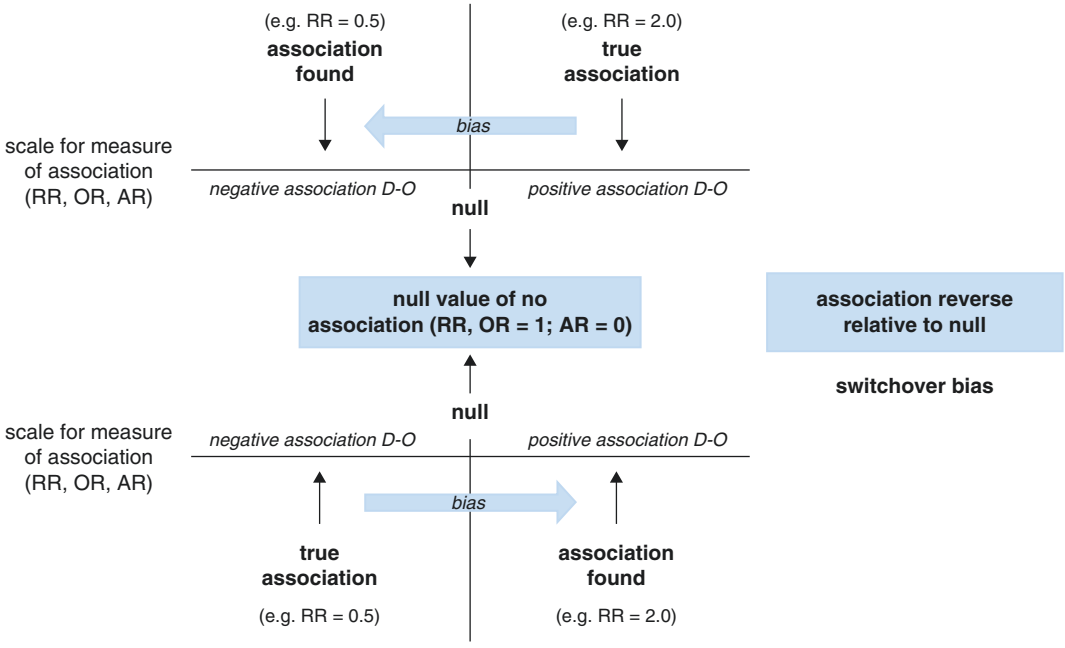


Figure 5.7 Qualitative bias in the association between determinant (D) and disease outcome (O)

In a cohort study the participants are recruited and classified on the basis of their determinant status (in the simplest case either exposed or not exposed to the primary determinant in the study). In this type of study selection bias would occur if the probability of being included in the study population as a representative of a particular determinant category (e.g. exposed/non-exposed) depended on the disease status at the end of the follow-up period. To put it another way, selection bias in a cohort study occurs if a person with/without exposure to the determinant is more or less likely to be included in the study if he or she is unhealthy than an exposed/non-exposed person who is healthy.

This is a case of differential selection, i.e. unhealthy and healthy exposed persons have different selection probabilities (as do unhealthy and healthy non-exposed persons). Non-differential selection can also occur in a study when the selection probability, although different for exposed and non-exposed potential participants in the study, does not differ systematically for unhealthy and healthy persons. Only differential selection causes selection bias.

Selection bias is not really a factor in the context of a prospective cohort study, as it is characteristic of a cohort study that the determinant status of the potential participants in the study is established at a time when the disease is not yet present. Cohort studies are therefore highly resistant to selection bias, although it cannot be ruled out. Take, for example, a cohort study into the effect of physical activity on life expectancy in middle-aged people. Among the potential participants there are people with a congenital or heart defect or some other heart problem. These people will generally be at increased risk of mortality (from cardiovascular disease). Many of them could also conceivably have been advised to take it easy with regard to physical exercise because of the increased risk of complications. This causes a problem: in the subcohort of people with little physical activity the category with short life expectancy due to a congenital heart defect will be over-represented compared with the general population. To avoid this type of selection bias everyone with a known heart problem could be kept out of the study population by restriction. We need to be particularly wary of the possibility of selection bias in a cohort study if preliminary stages or particular preclinical characteristics of the disease being studied are related to exposure to the determinant in question.

Differential selection can also occur in a cohort study due to selective dropout of participants after the cohort has been assembled (attrition bias). Selective dropout occurs when some of the cohort members stop taking part in the follow-up measurements after a while, for reasons related to the disease that we are interested in. This type of selection bias can also distort the results of a randomized intervention study. Take, for example, an experiment looking at the effects of a slimming course: it is quite conceivable that particularly those people who do not detect any results in themselves will abandon the course after a while.

The participants in a case-control study are selected on the basis of their disease status (with or without the disease). There is a danger of selection bias in this type of study if the probability of being included in the study population as a patient (with the disease) or control (without the disease) depends on exposure to the primary determinant in the study. To put it another way, selection bias in a case-control study occurs if a person (with/without the disease) is more or less likely to be included in the study as a patient/control if he or she has been exposed to the determinant than a person (with/without the disease) who has not been exposed to the determinant.

This is another case of differential selection, as the selection probabilities differ for exposed and non-exposed patients or controls. Unlike a cohort study, a case-control study is particularly susceptible to selection bias, as the participants are recruited after the relevant exposure to the determinant being studied has taken place. The examples below therefore relate to selection bias in case-control studies.

Case 5.1 Venous thromboembolism and use of the pill (hypothetical example)

Suppose there is a study into the association between the use of oral contraception (OCP) in adult women and the risk of venous thromboembolism (VTE), a vascular disorder. The study is carried out in a particular region where

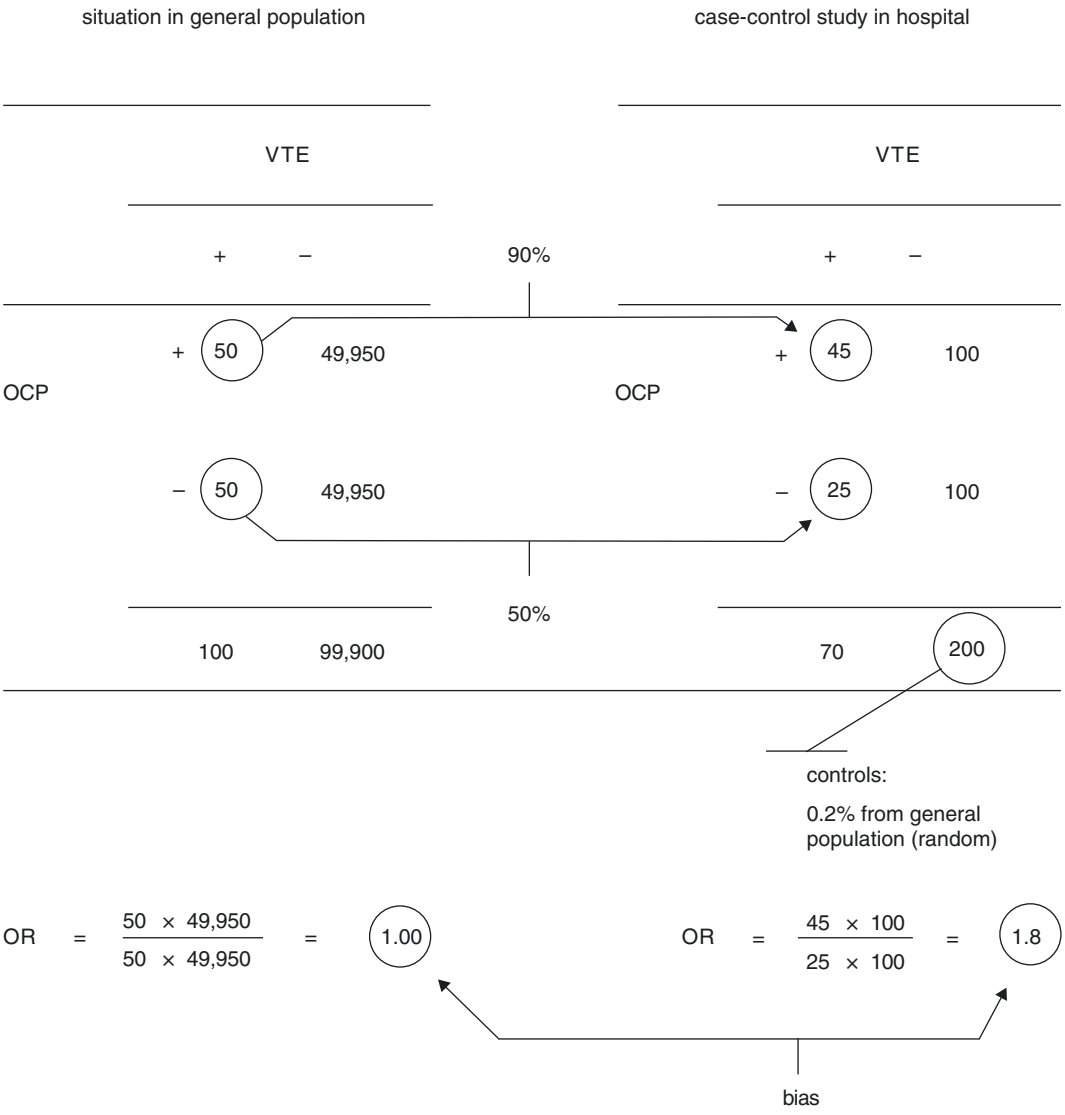


Figure 5.8 Selection bias in a case-control study into venous thromboembolism and use of the pill

100,000 women in the 25-45 age group (the inclusion criterion) live. Over a period of two years 100 cases of venous thromboembolism occur in this population. Half of all women take the 'pill' (OCP).

In reality use of the pill is not related to the development of thromboembolism (see the left half of fig. 5.8). If all women were to be included in the study it would find an OR or RR of 1.0. For the sake of efficiency, however, the researchers opt for a case-control study, recruiting the patients from hospital. At the time the study was initiated the idea that use of the pill could be a risk factor for venous thromboembolism had already been put forward in the medical press. GPs faced with women with

symptoms suggestive of thromboembolism therefore tend to refer those who are also on the pill for specialist investigation in hospital more often (90%) than those who are not on the pill (50%). As a result of this selective referral policy, of the aforementioned 100 women with symptoms of venous thromboembolism 70 are ultimately admitted to hospital, the majority of them pill users. These women are included in the case-control study as patients. The control group is a random 0.2% sample from the general population (excluding the VTE patients). The response rate for both patients and controls is 100%. The analysis of the case-control study produces an estimated OR of 1.8, i.e. a substantial overestimate of the true association (positive selection bias). The patient group in the study does not accurately reflect the total patient population about which the researchers wish to draw conclusions, as it contains an excessive proportion of exposed patients (pill users).

Case 5.2 Cervical cancer and pap smear frequency (hypothetical example)

Suppose there is a case-control study into whether having regular pap smears done can prevent cervical cancer. The study is carried out on women in the 35-65 age group in a particular region. Pap smear frequency is divided into two categories: pap smears/no pap smears during the past five years. A total of over 10,000 women are potentially eligible to take part in the study. In this population 200 women die of cervical cancer within a few years. The distribution of pap smear frequency (PAP) in the total population is shown in the top part of ◘ fig. 5.9. Suppose research is done based on deaths from cervical cancer during this period. The control group is a 10% sample from the general population. These women are contacted for a telephone survey, in which they are asked among other things about the frequency of pap smears in the past. The response rate is poor, only 20%. A non-selective response would result in the results shown at the bottom left of the diagram (no bias). In practice, however, it is mainly women in the lower socioeconomic group, who often do not have pap smears done, who are contactable by telephone. As a result the control group contains 18% of the women who did have pap smears done and 40% of those who did not. The result, as the right-hand side of the diagram shows, is a substantial underestimate of the protective effect of pap smears (negative selection bias).

Selection bias occurs when the researchers recruit the participants for the study, but the seed may already have been sown at an earlier stage. An example of this is the healthy worker effect: it has been found that morbidity and mortality rates in industrial cohorts are usually lower (more favourable) than in the general population. This is related to the fact that people with serious health limitations are less likely to participate in regular employment, or do not have the required physical capabilities for particular occupations. For the same reasons the general state of health of other groups too – e.g. vegetarians, people who take part in sporting activities, migrants – may be more favourable than that of the population as a whole (membership bias). The consequence of this differential selection is that certain health risks can escape attention. The incidence of non-specific respiratory tract disorders among upholsterers, for instance, will be lower than that in the population as a whole, although their working conditions are presumably associated with an increased risk of respiratory tract disorders.

Selective survival or selective mortality of part of the population being studied before the participants are recruited can also result in selection bias. In case-control studies in particular there is a danger that patients whose disease lasts only a short time (they die soon or get better quickly) and patients with a mild variety of the disease will be underrepresented in the study population. Take, for example, a case-control study into risk factors for acute myocardial infarction based on patients admitted to hospital after an infarction. The risk profile of those who survive an infarction will presumably be very different from that of those who die of it immediately.

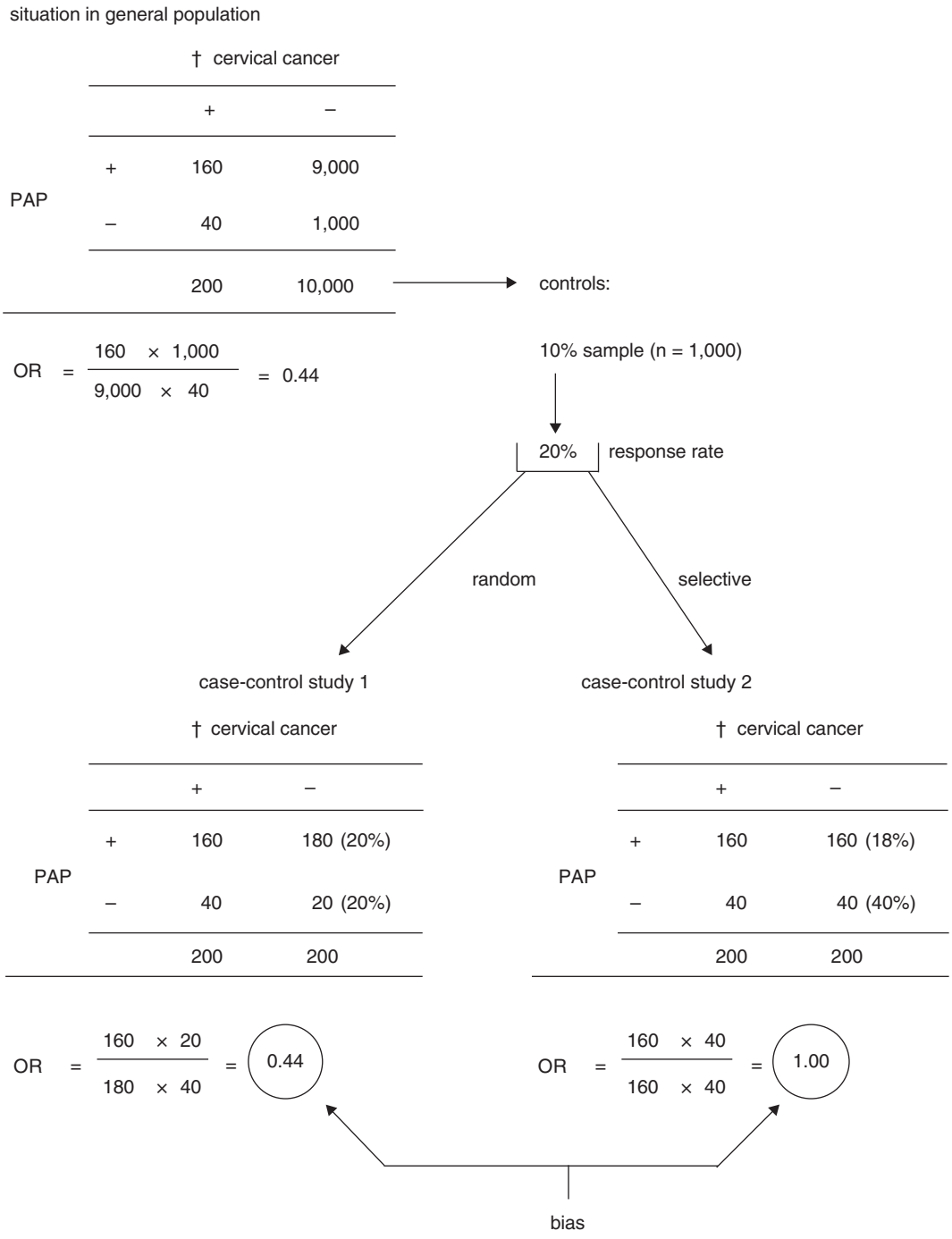


Figure 5.9 Selection bias in a case-control study into cervical cancer and pap smear frequency

► Case 5.1 also illustrates the fact that hospitalized patients are generally different from those who are not hospitalized. If this selection is related to the exposure that we are interested in, the effect is biased (admission rate bias). This is also known as Berkson's fallacy. This bias is sometimes due to a particular referral policy (referral bias), for example because complicated patients (with a non-standard risk profile) are referred to a university medical centre and uncomplicated patients to a general hospital. Selection bias can also occur after the study population has been selected, because the people who have been selected and agree to take part in the study differ systematically from those who decline to take part (non-respondent bias), or the other way round (volunteer bias). We have already mentioned the danger of selective dropout during the follow-up period (withdrawal bias). Lastly, incomparability between groups can occur as a result of exposure to the determinant in one group affecting exposure in the other group. In experimental studies this contamination bias is a serious danger, as are systematic differences in compliance (compliance bias).

Information bias

Information bias is caused by measuring errors, which can relate to the determinant, the disease outcome or both. In the case of variables that are not measured and analysed on a continuous scale these result in the participants being classified into the wrong determinant and/or disease category (misclassification). The misclassification can differ from one group to another. This is referred to as differential misclassification (systematic measuring error): the error frequency in recording the disease depends on the determinant level measured, or conversely the error frequency in measuring the determinant depends on the disease state recorded. In a cohort study, for example, the disease that we are interested in may be diagnosed more accurately in participants who have been exposed to the determinant than participants not exposed to it. And in a case-control study more effort may be put into finding out about prior exposure to a determinant with regard to patients than with regard to controls. Differential misclassification of the determinant and/or outcome can result in an underestimate, overestimate or even reversal of the true effect of the exposure on the outcome.

The term non-differential misclassification (random measuring error) is used when the errors in measuring one variable are independent of the measured values of the other variable.

Non-differential misclassification usually distorts the effect too, but the direction is generally clear: if there is non-differential misclassification of a dichotomous variable (assuming there is no misclassification of confounders) the true effect will always be underestimated. This is referred to as bias towards the null (= no effect). There are exceptional cases where non-differential misclassification does not distort the result, for example when estimating the RR in a cohort study in which some of the persons that develop the disease are missed. In the case of random measuring errors on a continuous variable there can also be exceptional cases where there is no distortion. In most cases, however, the rule is that non-differential misclassification results in an underestimate of the true effects. The consequences of measuring errors and misclassification can be considerable, even if the measuring error is non-differential and of limited magnitude.

Misclassification not only affects exposures and outcomes but also confounders (see below). When such errors in confounder measurements occur adjustment for confounding during analysis can only be partially implemented. Particularly for strong confounders misclassification will lead to substantial residual confounding.

In principle, information bias can affect all types of study. The danger is particularly large, however, when collecting information on the past based on self-reporting (in case-control studies). Knowing the disease status can for instance lead the researchers to search fanatically for the suspected determinant in patients, especially if there is already a strong suspicion that that determinant is harmful. For instance, in a case-control study looking for teratogenic risks, women with a baby with a congenital abnormality (cases) will search their memories more thoroughly for potentially hazardous substances than women with a normal pregnancy outcome, resulting in systematic differences in the amount and quality (accuracy) of the information (recall bias). Information bias occurs in a cohort study when prior knowledge about the harmful nature of a particular determinant affects the intensity

removal of kidney stones

	successful	unsuccessful	
open surgery	185 (75%)	60	245
laparoscopy	408 (82%)	87	495

small stones

	successful	unsuccessful	
open surgery	28 (93%)	2	30
laparoscopy	394 (83%)	81	475

large stones

	successful	unsuccessful	
open surgery	157 (73%)	58	215
laparoscopy	14 (70%)	6	20

Figure 5.10 Simpson's paradox

and outcome of the diagnostic process. The source of information bias, then, can be the researcher, the participant with his limited or biased memory, or the instrument used.

Confounding

Confounding is caused by mixing up the effect of the primary determinant with that of one or more other external determinants.

Take, for example, an intervention study into the removal of kidney stones which finds that open surgery has a success probability of 75% and percutaneous nephrolithotripsy (an endoscopic procedure) a success probability of 82%. Based on the study we would conclude that an endoscopic procedure is to be preferred. If the size of the kidney stone is included in the analysis, however, the effect is very different. In the case of small stones (under 2 cm) the success probability is 93% for open surgery and 83% for an endoscopic procedure; in the case of larger stones too the success probability is higher for open surgery (73% as against 69%). As endoscopic procedures are performed mainly on small kidney stones they appear to have a higher success probability than open surgery (fig. 5.10). If the confounder (size of the stones) had not been taken into account in the analysis, preference would have been given to the wrong treatment. We call this reversal of effect due to confounding Simpson's paradox. Reversal of effect is an extreme example; confounding can also cause overestimation or underestimation of the true effect.

Another example is a case-control study into whether alcohol consumption is a cause of laryngeal cancer. As cigarette smoking is also associated with an increased risk of laryngeal cancer and people who drink a lot of alcohol are also more likely to be smokers, the effect of alcohol consumption will be overestimated if the smoking factor is not included in the analysis, since part of the effect is due to the drinkers' smoking behaviour.

In order to act as a confounder of the relationship between the primary determinant and the outcome being studied a factor must meet the following criteria:

1. The factor in question is itself an independent determinant of the disease being studied.
2. People without the factor have a different disease frequency than people with the factor, including those not exposed to the primary determinant. This association need not be causal; a marker of a causal determinant (age, socioeconomic status, etc.) can also act as a confounder. Prior knowledge (from the literature)is needed to determine what the suspected risk factors for the particular disease are.
3. The factor is associated with the primary determinant, regardless of the disease being studied. Such an association exists if the values of the factor in question are unevenly distributed among the values of the primary determinant. In a cohort study an association between the potential confounder and the determinant can be assessed based on the study data, as these will have been collected on the basis of a predefined source population. This is not usually the case in a case-control study. If the control group is sufficiently large and representative of the source population it is best to ascertain in that group whether there is any association between the potential confounding factor and the determinant.

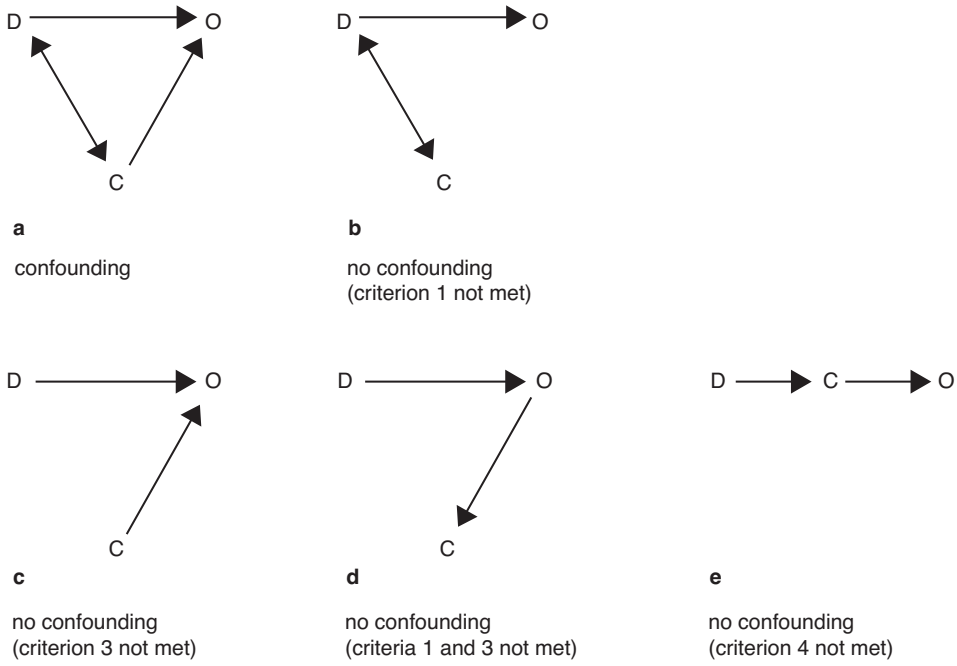


Figure 5.11 Possible relationships between primary determinant, potential confounder and disease.
D: determinant
O: disease outcome;
C: potential confounding factor;
↔: association;
→: causal relationship

4. The factor in question is not a connecting link in the causal chain between the primary determinant and the outcome. If the potential confounder is part of the mechanism linking the determinant to the disease (an intermediary factor) this is not confounding but a chain of successive effects referred to as mediation. Although an association may then be found between the potential confounder and the primary determinant, this is due to the relationship between the primary determinant and the disease itself. In such cases we must not adjust for confounding; we can carry out a mediation analysis instead. The situation is similar if the potential confounder is due to the disease (e.g. a symptom): here again there is a secondary association between the primary determinant and the disease that is not due to confounding.

Figure 5.11 outlines a number of situations where confounding is or is not present.

Examples of potential confounders (situation a):

- Smoking in a study into the relationship between alcohol consumption and laryngeal cancer (see above)
- Physical exercise in a study into the relationship between fat consumption and overweight
- Driver's age in a study into the relationship between wearing seat belts and the risk of a fatal road accident.

Examples of situations where an external variable could wrongly be regarded as a confounder:

- Alcohol consumption in a study into the relationship between cigarette smoking and pulmonary emphysema (situation b in fig. 5.11)

- Chronic cough in a study into the relationship between smoking and lung cancer (situation b in ◘ fig. 5.11)
- Dietary habits in a study into the relationship between blood group and the risk of an infectious disease (situation c in ◘ fig. 5.11)
- Consumption of tissues in a study into the relationship between dietary vitamin C intake and the common cold (situation d in ◘ fig. 5.11)
- Birth weight of the baby in a study into the relationship between malnutrition in pregnant women and infant mortality (situation e in ◘ fig. 5.11)

If a variable is wrongly treated as a confounder – i.e. if the criteria above are not met – this generally has adverse effects only on the efficiency of the study. Adjusting for suspected confounding is not necessary in the above case, nor can it do any harm, unless Criterion 4 is not met. Intermediary factors (situation e) must never be adjusted for, as that would 'adjust out' the true relationship between the exposure and the disease outcome.

Confounding, then, can cause positive bias (overestimation), negative bias (underestimation) or even reversal of the direction of the effect (Simpson's paradox). What effect confounding has in a specific study situation will depend on the direction and strength of the associations between the variables concerned: the stronger the associations, the greater the bias in the estimation of effect. Confounding presents a real danger in all types of causation-oriented observational research and non-randomized experimental research.

To assess whether there actually is confounding in a study a stratified analysis can be carried out. The following steps are involved:

1. Calculate the association between the primary determinant (D) and the disease outcome (O) for the study population as a whole (crude association).
2. Calculate the association between the determinant and the disease for each category (stratum) of the suspected confounder (C) (stratum-specific association).
3. Compare the crude association with the stratum-specific associations: there is confounding only if the crude association significantly differs from the stratum-specific associations.

Stratified analysis can also be used to adjust for confounding (▶ par. 5.4.2). ▶ Cases 5.3 and ▶ 5.4 further illustrate the evaluation of confounding. These are based on the simplest possible situation: the determinant, the disease outcome and the potential confounder all are dichotomous (two categories) and only one confounder is considered at a time.

Substantial differences between stratum-specific associations are called effect modification, a concept discussed in ▶ par. 5.5.

▶ Chapter 6 returns to the subject of confounding in detail in relation to the question of causality.

Case 5.3 Physiotherapy or GP treatment for back problems (hypothetical example)

Suppose a cohort study is done among patients with back problems to discover if patients who opt for treatment by a physiotherapist (D) are more likely to be symptom-free ($\overline{O}$) one year after therapy than people who opt for treatment by their GP ($\overline{D}$). At the start of the treatment the researchers ascertain among other things whether the participants engage in regular sporting activities (C_1) and for how long they have had the back problems (C_2). ◘ Figure 5.12 shows the results of the analysis.

Case 5.4 Smoking and risk of a second heart attack (hypothetical example)

Assume a case-control study into the association between second heart attacks (O) within five years of the first attack and continued smoking of cigarettes (D). The severity of the first attack (C_1) and the age of the patients (C_2) are regarded as potential confounders. The results of the analysis are summarized in ◘ fig. 5.13.

Assessing confounding in a cohort study into the effectiveness of treatment for back problems

D = treatment by physiotherapist
$\bar{D}$ = treatment by GP
O = treatment unsuccessful: still symptoms after 1 year
$\bar{O}$ = treatment successful: no more symptoms after 1 year
C_1 = regular sports
$\bar{C}_1$ = no regular sports
C_2 = duration of symptoms > 1 month
$\bar{C}_2$ = duration of symptoms < 1 month

5

1

	O	$\bar{O}$	
D	160	240	400
$\bar{D}$	480	720	1,200
			1,600

$$RR_{DO} = CRR = \text{crude } RR = \frac{160/400}{480/1,200} = 1.0$$

a. Is regular sports a confounder? (details: 2a–5a)

b. Is duration of symptoms a confounder? (details: 2b–5b)

2a

	C_1		
	O	$\bar{O}$	
D	20	100	120
$\bar{D}$	60	300	360
			480

	$\bar{C}_1$		
	O	$\bar{O}$	
D	140	140	280
$\bar{D}$	420	420	840
			1,120

$$RR_{DO|C_1} = RR_1 = \frac{20/120}{60/360} = 1.0 \qquad RR_{DO|\bar{C}_1} = RR_2 = \frac{140/280}{420/840} = 1.0$$

3a $RR_{DO} = 1.0$: physiotherapy is no more successful than treatment by a GP.
$RR_{DO|C_1} = RR_{DO|\bar{C}_1} = 1.0$: GPs and physiotherapists achieve the same results in both people who play sports regularly and people who do not play sport regularly; there is no confounding and no effect modification.

4a

	C_1	$\bar{C}_1$	
D	120	280	400
$\bar{D}$	360	840	1,200
	480	1,120	1,600

$$\frac{P(C_1|D)}{P(C_1|\bar{D})} = \frac{120/400}{360/1,200} = 1.0$$

no link between intensity of sports and nature of treatment.

Figure 5.12 Assessing confounding in a cohort study into the effectiveness of treatment for back problems

5a

	$\bar{D}$		
	O_1	$\bar{O}$	
C_1	60	300	360
$\bar{C}_1$	420	420	840
			1,200

$$\mathbf{RR_{C_1|\bar{D}} = \frac{60 / 360}{420 / 840} = 0.33 \; (\neq 1.0)}$$

back pain patients treated by a GP who play sports regularly are more likely to get better than patients who do not play sports regularly.

in summary:
no confounding (see Fig. 5.12a).

2b

	C_2		
	O	$\bar{O}$	
D	147	173	320
$\bar{D}$	253	147	400
			720

$$\mathbf{RR_{DO|C_2} = \frac{147 / 320}{253 / 400} = 0.73}$$

	$\bar{C}_2$		
	O	$\bar{O}$	
D	13	67	80
$\bar{D}$	227	573	800
			880

$$\mathbf{RR_{DO|\bar{C}_2} = \frac{13 / 80}{227 / 800} = 0.57}$$

3b $RR_{DO} = 1.0$: physiotherapy is no more successful than treatment by a GP if the duration of the symptoms is ignored.

$RR_{DO|C_2} = 0.73$; $RR_{DO|\bar{C}_2} = 0.57$: physiotherapists achieve better results than GPs in both patients with long-standing symptoms and patients with recent symptoms; the efficacy of physiotherapy was initially underestimated (bias towards the null); there is also some effect modification.

4b

	C_2	$\bar{C}_2$	
D	320	80	400
$\bar{D}$	400	800	1,200
	720	880	1,600

$$\mathbf{\frac{P(C_2|D)}{P(C_2|\bar{D})} = \frac{320 / 400}{400 / 1,200} = 2.4 \; (\neq 1.0)}$$

patients treated by a physiotherapist are more likely to have long-standing symptoms than patients treated by a GP.

Figure 5.12 Continued

5b

	$\bar{D}$		
	O_1	$\bar{O}$	
C_2	253	147	400
$\bar{C}_2$	227	573	800
			1,200

$$RR_{C_2|\bar{D}} = \frac{253/400}{227/800} = 2.23 \; (\neq 1.0)$$

back pain patients with long-standing symptoms treated by a GP are less likely to get better than patients without long-standing symptoms.

in summary:
confounding, resulting in a biased estimate of the effect towards the null (effect of physiotherapy underestimated) (see Fig. 5.12b)

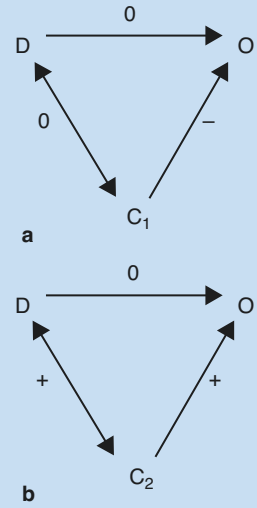


Fig. 5.12 summary of the conclusions from case 5.3

Figure 5.12 Continued

5.4.2 Bias needs to be tackled with a smart study design

A number of measures to deal with bias were discussed in ▶ chap. 4. Randomization, restriction and matching are examples of strategies to avoid selection bias and confounding during the design and execution stages. Information bias can be avoided in various ways. The blinding of researchers and participants has already been discussed. To avoid differential misclassification, objective instruments that leave little room for interpretation by the researcher or the participant should be used wherever possible. The following measures to avoid distortion due to confounding and selection bias are discussed below: randomization, restriction and matching as measures taken at the design stage to improve validity. Other measures that can be used at the analysis stage to eliminate confounding in the collected data are stratification, standardization and multivariable regression analysis.

Randomization

An experimental study in which the researchers can influence the allocation of the determinant being studied offers the possibility of randomization (random allocation). In principle, random allocation based on the drawing of lots results in the intervention group and the comparison group (or groups) being equal in all respects at the start of the experiment, including the status of potential confounders. Randomization does away with both known and unknown as well as difficult-to-measure confounders. It does not guarantee that the various treatment groups will be completely comparable, however: in small study populations in particular chance can have unfortunate effects. Pre-stratification can improve comparability in such instances. This involves classifying members of the study population into subcohorts (strata), based on their values for the main confounders, and subsequently carrying out randomization on each stratum. This will substantially increase the likelihood that each treatment group has the same distribution for these important confounders.

Assessing confounding in a case-control study into smoking and risk of heart attack

D = smoking after first heart attack
$\bar{D}$ = no smoking after first attack
O = second attack
$\bar{O}$ = no second attack
C_1 = severe first attack
$\bar{C}_1$ = minor first attack
C_2 = age ≥ 55
$\bar{C}_2$ = age < 55

1

	O	$\bar{O}$	
D	100	200	
$\bar{D}$	100	600	
	200	800	1.000

$$\mathbf{OR_{DO} = cOR = crude\ OR = \frac{100 \times 600}{200 \times 100} = 3.00}$$

a. is the severity of the first attack a confounder? (details: 2a–5a)
b. is age a confounder? (details: 2b–5b)

2a

	C_1		
	O	$\bar{O}$	
D	50	61	
$\bar{D}$	20	60	
	70	121	191

$$\mathbf{OR_{DO|C_1} = OR_1 = \frac{50 \times 60}{61 \times 20} = 2.46}$$

	$\bar{C}_1$		
	O	$\bar{O}$	
D	50	139	
$\bar{D}$	80	540	
	130	679	809

$$\mathbf{OR_{DO|\bar{C}_1} = OR_2 = \frac{50 \times 540}{139 \times 80} = 2.43}$$

3a $OR_{DO} = 3.00$: patients with a second heart attack are more likely to have continued smoking after the first attack.

$OR_{DO|C_1} = OR_{DO|\bar{C}_1} = 2.44$: smoking is a prognostic factor for a second attack in both patients with a severe first attack and patients with a minor first attack; the effect is not as strong as in the group as a whole.

Figure 5.13 Assessing confounding in a case-control study into smoking and risk of heart attack

4a

	C_1	$\bar{C}_1$	
D	111	189	300
$\bar{D}$	80	620	700
	191	809	1.000

$$\frac{P(C_1|D)}{P(C_1|\bar{D})} = \frac{111/300}{80/700} = 3.24\ (\neq 1.0)$$

O			
	C_1	$\bar{C}_1$	
D	50	50	100
$\bar{D}$	20	80	100
	70	130	200

$$\frac{P(C_1|DO)}{P(C_1|\bar{D}O)} = \frac{50/100}{20/100} = 2.50\ (\neq 1.0)$$

$\bar{O}$			
	C_1	$\bar{C}_1$	
D	61	139	200
$\bar{D}$	60	540	600
	121	679	800

$$\frac{P(C_1|D\bar{O})}{P(C_1|\bar{D}\bar{O})} = \frac{61/200}{60/600} = 3.05\ (\neq 1.0)$$

patients with a severe first heart attack are more likely to smoke, both in the group with a second attack and in the control group.

5a

	$\bar{D}$		
	O	$\bar{O}$	
C_1	20	60	
$\bar{C}_1$	80	540	
	100	600	700

$$OR_{C_1O|\bar{D}} = 2.25\ (\neq 1.0)$$

non-smoking patients with a second heart attack are more likely to have had a severe first attack than patients without a second attack.

in summary:
confounding, resulting in a biased estimate of the effect away from the null (effect of smoking overestimated) (see Fig. 5.13a).

Figure 5.13 Continued

2b

	C_2		
	O	$\bar{O}$	
D	85	115	
$\bar{D}$	65	235	
	150	350	500

$OR_{DO|C_2} = 2.67$

	$\bar{C}_2$		
	O	$\bar{O}$	
D	15	85	
$\bar{D}$	35	365	
	50	450	500

$OR_{DO|\bar{C}_2} = 1.84$

3b $OR_{DO} = 3.00$

$OR_{DO|C_2} = 2.67$; $OR_{DO|\bar{C}_2} = 1.84$: smoking after the first heart attack is a prognostic factor for a second attack in both older and younger patients; the effect is not as strong as in the group as a whole, however: age is a confounder; the magnitude of the effect also differs between the two age groups: age is also an effect modifier.

4b

	C_2	$\bar{C}_2$	
D	200	100	300
$\bar{D}$	300	400	700
	500	500	1.000

$$\frac{P(C_2|D)}{P(C_2|\bar{D})} = \frac{200/300}{300/700} = 1.56\ (\neq 1.0)$$

	O		
	C_2	$\bar{C}_2$	
D	85	15	100
$\bar{D}$	65	35	100
	150	50	200

$$\frac{P(C_2|DO)}{P(C_2|\bar{D}O)} = \frac{85/100}{65/100} = 1.3\ (\neq 1.0)$$

	$\bar{O}$		
	C_2	$\bar{C}_2$	
D	115	85	200
$\bar{D}$	235	365	600
	350	450	800

$$\frac{P(C_2|D\bar{O})}{P(C_2|\bar{D}\bar{O})} = \frac{115/200}{235/600} = 1.47\ (\neq 1.0)$$

older patients who have had a first heart attack are more likely to smoke than younger patients, both in the group with a second attack and in the control group.

Figure 5.13 Continued

5b

	$\bar{D}$		
	O	$\bar{O}$	
C_2	65	235	
$\bar{C}_2$	35	365	
	100	600	700

$OR_{C_2O|\bar{D}} = 2.88 \ (\neq 1.0)$

non-smoking patients with a second heart attack are older than non-smoking patients without a second attack.

in summary:
confounding, resulting in biased estimate of effect away from the null (effect of smoking overestimated); also effect modification.

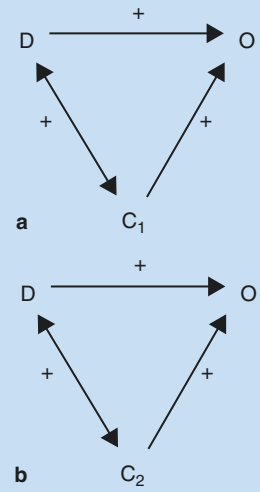


Fig. 5.13 summary of the conclusions from case 5.4

Figure 5.13 Continued

Randomization is critical to the validity of an intervention study into the intended effect of a treatment or prevention measure, as without random allocation the person giving the treatment will be likely to allocate the experimental intervention to those patients expected to benefit most from it (e.g. those with a favourable – or unfavourable – prognosis). This causes serious bias in the results, confounding by indication, since differences in effect are no longer related solely to the intervention chosen but also to differences in prognosis between the patients to whom the various interventions are allocated. The best solution to this problem is to leave the allocation of the various interventions to chance. Randomization is the crucial element in a treatment experiment, but as we have seen it does not guarantee that the various exposures will actually be distributed equally among the various interventions. Should an unequal distribution of exposures occur, a stratified analysis will enable adjustment for the resulting distortion. A stratified analysis can only be used, however, if the confounder in question is known and has been measured, which is not always the case. Unknown and unmeasurable confounders are in fact the main reason for randomizing.

Restriction

The most rigorous method to avoid confounding in a study design is restriction, ensuring that these variables no longer vary. As the entire study population will then have the same value for the confounder in question, it cannot cause bias. Examples of restrictions are allowing only men to take part in the study, or only non-smokers. We can then be certain that gender or smoking cannot act as confounders, thus increasing the internal validity of the study. On the other hand, restriction of a variable to a single category will often affect the scope for generalization (see ▶ par. 5.6). If a study involving only men does find an effect, it will not be possible to generalize this to women without qualification. In addition to avoiding confounding, restriction of the study population can also be used to obtain an optimal (efficient) distribution among the determinant categories, so as to guarantee sufficient contrast in exposure status (see ▶ chap. 4). In an experiment restriction is used to reduce the source population to a more or less homogeneous group (e.g. people in the same age group, of the same gender, with the same disease, at the same stage of the disease). This involves setting inclusion and exclusion criteria that

the people who will be included in the study population must meet.

Matching

Matching refers to the selection of a comparison group – persons not exposed to the determinant in a cohort study, or controls in a case-control study – who are identical (as far as possible) to the index group – persons with exposure to the determinant in a cohort study, or the patients in a case-control study – in terms of the distribution of one or more potential confounders. Intuitively, matching would seem to be a very attractive and obvious measure to improve the quality of a study design. As we shall see below, however, intuition is deceptive in this respect, notably in case-control studies where matching is quite popular.

Matching in a cohort study involves selecting the members of the subcohorts to be compared (exposed/not exposed to the primary determinant) in such a way that there is an equal distribution of one or more potential confounders between the subcohorts. This avoids confounding due to the matching factors. On the other hand, matching in a cohort study is often time-consuming and therefore expensive, as many potential participants in the study have to be assessed, measured – and often rejected for participation – in order to arrive at the final selection. The validity of a study can be achieved just as well by measuring confounder status in all participants followed by adjustment procedures at the data analysis stage. In most cohort studies this is more productive than matching in advance. ▶ Case 5.5 gives a detailed example of the use of matching in a cohort study.

Matching in a case-control study involves selecting healthy controls from the population from which the cases originate in such a way that the patients and controls have an equal distribution of one or more potential confounders. As the second part of ▶ case 5.5 shows, matching in a case-control study does not increase validity and sometimes even reduces it. Matching can improve the efficiency of a case-control study, but this benefit is soon lost if a lot of energy has to be invested in the matching process itself. It is important to take into account whether or not matching has been carried out when analysing the data in a case-control study, as that eliminates the problem of any confounding introduced by the matching.

The same arguments apply in an experimental study as in a cohort study, except that an experiment also offers the possibility of controlling confounding by means of randomization. Nevertheless, in the case of very strong confounders it might be worthwhile to match the experimental groups (as groups or individuals) for these important confounders in advance. Chance then decides which people in the strata or matched pairs are given the experimental intervention and which are not. This is referred to as pre-stratification with block randomization. Block randomization is discussed in ▶ chap. 10. Experiments in which a single organ (eye, ear, hand) is treated and the parallel organ in the same person is used as a control can be regarded as the ultimate form of individually matched experiments.

Case 5.5 Alcohol and road traffic accidents (hypothetical example)

Suppose we have a source population comprising 500,000 men and 500,000 women – the adult population of a European region. A study is initiated in this region to look for the influence of alcohol consumption on the risk of becoming a road accident victim. Alcohol consumption is estimated from a personal interview and expressed as the average number of alcoholic drinks per day (for the sake of simplicity we assume that there is no information bias). Based on the results the participants are classified into 'high alcohol consumption' (two or more glasses of alcohol per day, A in ◘ fig. 5.14) and 'low alcohol consumption' (less than two glasses of alcohol per day, $\overline{A}$). Records are then kept over a period of one year of who has fallen victim to a road traffic accident ('accident injury' = I) and who has not ('no accident injury' = $\overline{I}$). Now let us look at what consequences matching in the study design would have, given the events outlined for the source population. ◘ Figure 5.14 shows what the results of the study would be if the entire source population were to be followed over a period of one year. Clearly, there is confounding in the population in question. The RR of accident injury for 'high

		I	$\bar{I}$	Total	risk of I in 1 year
males (500,000)	A	2,000	398,000	400,000	0.0050
	$\bar{A}$	250	99,750	100,000	0.0025
females (500,000)	A	200	99,800	100,000	0.0020
	$\bar{A}$	400	399,600	400,000	0.0010
total (1,000,000)	A	2,200	497,800	500,000	0.0044
	$\bar{A}$	650	499,350	500,000	0.0013

$$RR_{A/\bar{A}} = RR_{crude} = 0.0044 / 0.0013 = 3.38$$
$$RR_{A/\bar{A}\,|\,male} = 0.0050 / 0.0025 = 2.00$$
$$RR_{A/\bar{A}\,|\,female} = 0.0020 / 0.0010 = 2.00$$
$$RR_{m/f} = \frac{2{,}250 / 500{,}000}{600 / 500{,}000} = 3.75$$

Figure 5.14 Alcohol consumption and accident injuries in a hypothetical source population

alcohol consumption' is 3.38, against 2.00 for men and women separately. The gender factor is a confounder because (1) there is a link in the population between gender and the amount of alcohol consumption (80% of the people with high alcohol consumption and only 20% of those with low alcohol consumption are male) and (2) the risk of accident injury is higher for men than women regardless of the amount of alcohol consumption. If a cohort study were to be carried out to examine the association between alcohol consumption and the risk of accident injury, and the cohort – e.g. comprising 100,000 people (10%) – were to be selected at random from the source population, the confounding present in the source population would manifest itself in the same way in the study population (ignoring the influence of sample variability).

Now suppose that we decide to carry out a cohort study on 100,000 people in which we match individually for the gender variable. This means recruiting a subcohort of 50,000 people with 'high alcohol consumption' and an equally large subcohort of people with 'low alcohol consumption' from the source population in such a way that each 'exposed' male is matched with a 'non-exposed' male and each 'exposed' female is matched with a 'non-exposed' female. This strategy would produce a study population comprising 80% men and 20% women, i.e. a completely different gender distribution than in a cohort study based on a random sample. The expected results of a matched cohort study of this kind are shown in fig. 5.15. These calculations show that matching has eliminated confounding due to gender. The RR is no longer being overestimated. This non-confounded estimate of the effect of alcohol consumption on the accident risk could also have been achieved, however, by analysing the data from a sufficiently large random sample using stratified analysis. Another striking point is that the matched cohort study finds a larger number of accident injuries (330) than the non-matched version (285). This shows the higher statistical efficiency of a matched approach.

Now suppose that the population described above is used as the basis for a case-control study into the relationship between alcohol consumption and accident risk. The patients in the study are the 2,850 people from the source population identified as having an accident injury during a period of one year. They com-

		I	$\bar{I}$	total	risk of I in 1 year
A (50,000)	M	200	39,800	40,000	0.0050
	F	20	9,980	10,000	0.0020
$\bar{A}$ (50,000)	M	100	39,900	40,000	0.0025
	F	10	9,990	10,000	0.0010
total (100,000)	M	300	79,700	80,000	0.0037
	F	30	19,070	20,000	0.0015

$$RR_{A/\bar{A}} = RR_{crude} = \frac{220 / 50{,}000}{110 / 50{,}000} = 2.00$$

$$RR_{A/\bar{A}\,|\,male} = \frac{200 / 40{,}000}{100 / 40{,}000} = 2.00$$

$$RR_{A/\bar{A}\,|\,female} = \frac{20 / 10{,}000}{10 / 10{,}000} = 2.00$$

Figure 5.15 Cohort study into alcohol consumption and accident injuries in 100,000 people (matched for gender)

prise 2,250 (79%) males and 600 females, of which 2,200 (77%) with high alcohol consumption and 650 with low alcohol consumption. In a non-matched case-control study with an equal number of controls from the general population the controls are randomly selected from the source population from which the patients are taken, resulting in a control group of 2,850 people containing 50% men and 50% people with high alcohol consumption. The crude OR is (2,200 × 1,425)/(1,425 × 650) = 3.38. The confounding present in the source population is thus 'copied' to the study population, resulting in a substantial overestimate of the odds ratio. In a matched case-control study, for the 2,250 male patients 2,250 male controls are recruited from the source population. The expectation is that 1,800 (80%) of them will have a high level of alcohol consumption. Similarly, for the 600 female patients 600 female controls are recruited, of whom 120 (20%) are expected to have a high level of alcohol consumption. As fig. 5.16 shows, the matching results in an OR of 1.64, which is an underestimate of the true RR of 2.00.

Evidently, matching introduces confounding that does not reflect the confounding that was initially present in the source population, causing bias in the opposite direction. The explanation for this phenomenon is as follows: the purpose of a control group in a case-control study is to estimate the distribution of exposure levels in the population from which the patients are taken. However, if the controls are matched with the patients for a factor related to the primary determinant, this will make the distribution of the primary determinant in the control group similar to that in the patient group. The result is an underestimate of the true association between the determinant and the disease. If the correlation between the matching factor and the exposure factor is very strong, the effect of the determinant on the disease could even be matched out entirely. In this example, if all the men in the population have high and all the women have low alcohol consumption (a perfect correlation), matching for gender in a case-control study into the association between alcohol consumption and accident risk would result in an OR of 1. In a case-control study, then,

		I		$\bar{I}$	
males	A	2,000		1,800	
	$\bar{A}$	250		450	
	total		2,250		2,250
females	A	200		120	
	$\bar{A}$	400		480	
	total		600		600
total		2,850		2,850	

$$OR_{A/\bar{A}} = RR_{crude} = \frac{(2{,}000 + 200) \times (450 + 480)}{(1{,}800 + 120) \times (250 + 400)} = \frac{2{,}200 \times 930}{1{,}920 \times 650} = 1.64$$

$$OR_{A/\bar{A}\,|\,male} = RR_{A/\bar{A}\,|\,male} = \frac{2{,}000 \times 450}{1{,}800 \times 250} = 2.00$$

$$OR_{A/\bar{A}\,|\,female} = RR_{A/\bar{A}\,|\,female} = \frac{200 \times 480}{120 \times 400} = 2.00$$

Figure 5.16 Case-control study into alcohol consumption and accident injuries in 2,850 people with accident injuries and 2,850 controls from the general population (matched for gender)

matching can introduce new confounding. This can even happen if we match for a factor that is not a confounder but is merely associated with the disease, as shown in fig. 5.17 and 5.18. Confounding introduced by matching in a case-control study can be eliminated at the analysis stage by carrying out a stratified analysis to control for the same matching variables (see later on in this chapter). As fig. 5.14-5.18 show, the stratum-specific odds ratios give a correct indication of the effect being studied. For that matter, we would have carried out a stratified analysis even if we had not matched for the confounders in question. It should be noted in passing that a stratified analysis to control for the matching variables is not needed in a matched cohort study.

In case-control studies, then, the importance of matching lies not in the ability to avoid confounding but in the conditions it creates to deal efficiently with confounding at the analysis stage. Matching can be used, for example, to avoid the creation of strata that contain patients but hardly any controls. Strata of this kind will not make any contribution to a stratified analysis. It is often difficult to find matches. In the worst case it will be necessary to screen a large number of potential candidates for the comparison group to find a single person who is suitable. The energy put into tracing matches is generally better employed in collecting information on a larger number of non-matched participants.

As ▶ case 5.5 clearly shows, matching is usually not warranted. The advantages of matching will only outweigh the drawbacks when confounders are very strong and very unequally distributed among the primary determinant categories.

		I	$\bar{I}$	total	risk of I in 1 year
males (500,000)	A	2,000	398,000	400,000	0.0050
	$\bar{A}$	250	99,750	100,000	0.0025
females (500,000)	A	500	99,500	100,000	0.0050
	$\bar{A}$	1,000	399,000	400,000	0.0025
total (1,000,000)	A	2,500	497,500	500,000	0.0050
	$\bar{A}$	1,250	498,750	500,000	0.0025

$$RR_{A/\bar{A}} = RR_{crude} = 0.0050 / 0.0025 = 2.00$$

$$RR_{A/\bar{A} \mid male} = 0.0050 / 0.0025 = 2.00$$

$$RR_{A/\bar{A} \mid female} = 0.0050 / 0.0025 = 2.00$$

$$RR_{M/F} = \frac{2{,}250 / 500{,}000}{1{,}500 / 500{,}000} = 1.50$$

$$RR_{M/F \mid A} = \frac{2{,}000 / 400{,}000}{500 / 100{,}000} = 1.00$$

$$RR_{M/F \mid \bar{A}} = \frac{250 / 100{,}000}{1{,}000 / 400{,}000} = 1.00$$

Figure 5.17 Alcohol consumption and risk of accident injury in a source population without confounding

		I		$\bar{I}$	
males	A	2,000		1,800	
	$\bar{A}$	250		450	
	total		2,250		2,250
females	A	500		300	
	$\bar{A}$	1,000		1200	
	total		1,500		1,500
total		3,750		3,750	

$$OR_{A/\bar{A}} \approx RR_{crude} = \frac{(2{,}000 + 500)\,(450 + 1{,}200)}{(1{,}800 + 300)\,(250 + 1{,}000)} = \frac{2{,}500 \times 1{,}650}{2{,}100 \times 1{,}250} = 1.57$$

$$OR_{A/\bar{A} \mid male} \approx RR_{A/\bar{A} \mid male} = \frac{2{,}000 \times 450}{1{,}800 \times 250} = 2.00$$

$$OR_{A/\bar{A} \mid female} \approx RR_{A/\bar{A} \mid female} = \frac{500 \times 1{,}200}{300 \times 1{,}000} = 2.00$$

Figure 5.18 Case-control study into alcohol consumption and risk of accident injury without confounding (matched for gender)

Stratification and standardization

In epidemiological research stratified analysis is one way of dealing with confounding once it is detected. The steps involved are as follows:

1. Calculate the association between the determinant under consideration and the effect being studied in the total population (the crude effect).
2. Create strata based on the potential confounder categories.
3. Calculate the effect for each stratum and assess whether there is any substantial difference in the stratum-specific effects (to rule out effect modification: see ▶ par. 5.5).
4. Calculate the weighted average of the stratum-specific effects. This is the overall effect adjusted for confounding.

There are various ways of averaging out stratum-specific effects to obtain an effect adjusted for confounding, with differences in the weighting factor used. A method commonly used in epidemiology is the Mantel-Haenszel estimate, which can be used to pool stratum-specific odds ratios, relative risks or risk differences. The Epidemiology Calculator iPhone app, for example, provides a convenient way of doing this.

By way of illustration we show how a Mantel-Haenszel odds ratio can be calculated manually:

$$OR_{MH} = \frac{\sum(a_i d_i / N_i)}{\sum(b_i c_i / N_i)}$$

where a_i, b_i, c_i, d_i and N_i represent the figures in cells *a, b, c*, and d and the total number of participants in the ith stratum respectively. Applying this to the data from ▶ case 5.4 we obtain the following result:

$$OR_{MH} = \frac{[(50 \times 60)/191] + [(50 \times 540)/809]}{[(61 \times 20)/191] + [(139 \times 80)/809]}$$

The crude OR is 3.00 and the OR_{MH} adjusted for confounding is 2.44.

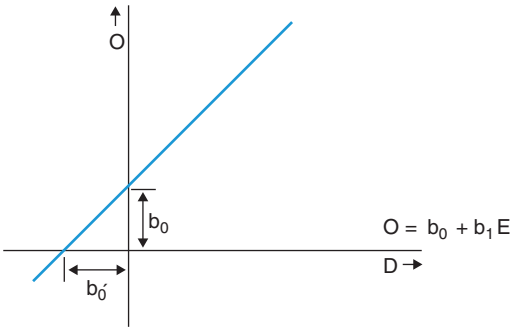
Another way of obtaining a weighted average of stratum-specific effects is to use the standardization technique for RRs or RDs already discussed in ▶ par. 2.5.3. Standardization is not traditionally used for odds ratios. Although the procedure is very different, the effect is the same: it eliminates confounding by the factor for which standardization is carried out. From the statistical point of view the Mantel-Haenszel method is preferable to standardization, as it lends the greatest weight to the strata containing the most information, and therefore the smallest sampling error. This is not the case with standardization, where the weighting is based on the distribution in an arbitrarily selected standard population. Therefore standardization is nowadays used as a technique to adjust for confounding only when analysing and interpreting health statistics.

Mathematical modelling

If there are a lot of confounders or a lot of categories for each confounder, the information available on each stratum will be too limited for stratified analysis. We are then forced to use multivariable analysis techniques based on mathematical models. This is in fact a very logical choice, which links up directly with the epidemiological function that lies at the heart of this book: a mathematical description of a model that describes the interrelationships between the disease outcome (O), the primary determinant (D) and the other determinants (confounders, C) as completely and accurately as possible. Of all the statistical analysis models available, the various types of multivariable regression analysis have therefore become the standard for confounder adjustment in epidemiology. The starting point is the linear model, which describes e.g. the relationship between the dependent variable (disease) and the independent variables (determinants) as a straight line (see ◘ fig. 5.19).

Here the distribution of the disease outcome variable is regressed to the values of the determinant. The straight line that shows the relationship between D and O is therefore referred to as the 'regression line'. This line can also be used to estimate the disease frequency for a given determinant. Linking the disease variable to a single determinant is done in univariate regression analysis.

Analysing more than one determinant at the same time is referred to as 'multiple' or 'multivariable' regression. Thus multivariable linear regression entails describing the occurrence of the disease using a linear combination of determinants (primary determinant, confounders):



O = frequency of disease outcome
D = level of exposure to determinant
b_0 = value of O if D = 0; the intercept of the regression line
b_1 = regression coefficient = tangent of the slope of the regression line with the line O = 0; $b_1 = b_0 / b_0'$

Figure 5.19 Linear regression model describing the relationship between determinant and disease

$$P(O) = b_0 + b_1 D + b_2 C_1 + b_3 C_2 + \ldots$$

To describe a relationship of this kind we need a multidimensional instead of a two-dimensional space (see e.g. fig. 3.5).

In epidemiology we generally study disease frequencies, i.e. probabilities with a value between zero and 1. Because probabilities do not produce straight lines a transformation is necessary to obtain a linear model. A model that is commonly used in epidemiology, especially in case-control studies, is multivariable logistic regression, represented by the following general formula:

$$P[O] = 1/[1 + e^{-(b_0 + b_1 D + b_2 C_1 + b_3 C_2 + \ldots)}]$$

Log-linear transformation of this function describes the disease variable based on a linear combination of determinants:

$$\ln[O/(1 - O)] = b_0 + b_1 D + b_2 C_1 + b_3 C_2 + \ldots$$

This model is particularly suitable for analyses where the outcome variable is dichotomous (e.g. with/without the disease). Other examples of non-linear multivariable models are the proportional hazards model and the log-linear models. All these models are derived from this basic linear model and are closely related to the multivariable logistic regression model.

Mathematical modelling has certain limitations: in particular the difficulty of ascertaining the biological plausibility of the underlying assumptions, and the problem of interpreting the results correctly. Strictly speaking, mathematical models should be based on prior knowledge of the underlying biological or pathophysiological mechanisms. There are various ways of ascertaining which model is best suited to a particular data set. The logistic model in particular has a wide range of applications. In practice popular models are often used indiscriminately, without questioning the whether the underlying assumptions are appropriate. Without careful checks these analyses may produce misleading results.

Causal diagrams

The definitions of confounding can be easily be applied in the case of a single disease variable O, a single determinant D and a single potential confounder C. Things might become complex, however, when several interdependent variables are involved and the scenarios in fig. 5.11 do not apply. Take the case of four variables in fig. 5.20a. Do we need to adjust for both C_1 and C_2 in order to study the relationship between D and O? And in the case shown in fig. 5.20b, with five variables, do we need to adjust for C if there are two – latent, unmeasured – variables L_1 and L_2?

If we adjust for too many confounders, we run the risk of overcorrection (resulting in unnecessarily wide confidence intervals). If we adjust for too few confounders, there is a risk of undercorrection (resulting in residual bias). It is important, therefore, to look at all the covariates relevant to the study – including those that are not being measured – and determine what set of variables needs to be used to avoid bias (the adjustment set). These variables are best set out in a causal diagram as shown in fig. 5.20. These diagrams are referred to as Directed Acyclic Graphs (DAGs) in the literature. A DAG contains all the variables relevant to the research question and the causal relationships between them.

If the DAG rules are applied strictly, the DAGs will help to identify the adjustment sets. These rules are as follows:

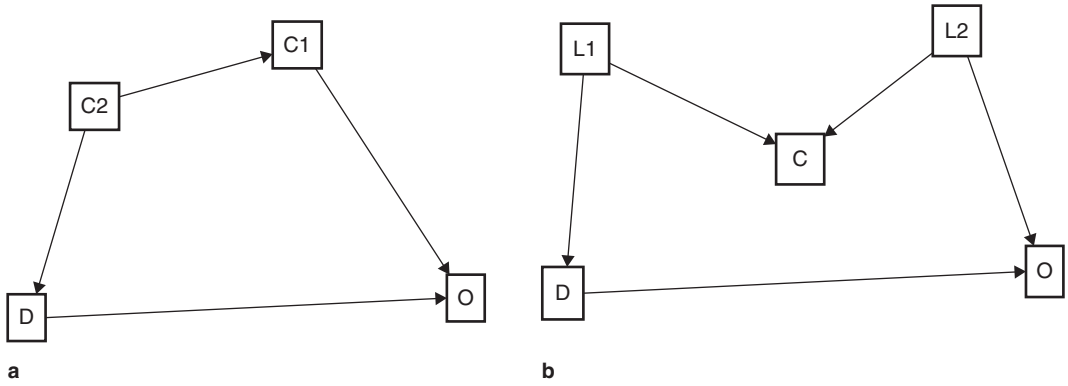


Figure 5.20 Causal diagrams

1. Include all the variables, including the unmeasured ones.
2. Show only direct causal relationships between variables using an arrow (an arrow A → B in a DAG means that A has a direct causal influence on B).
3. Successive arrows are referred to as a 'path'; more than one variable may lie on a path.
4. Only paths that connect D and O and that start with an arrow in the direction of D can yield an adjustment set.

Adjusting for one of the variables in an adjustment set blocks the path and adjusts for the confounding effect of all the variables on that path. In fig. 5.20a, for example, variables C1 and C2 both meet the criteria for confounding (determinant of the disease and associated with D), but there is only one adjustment set because both variables lie on the same path D ← C2 → C1 → O, so it is enough to adjust for C1 or C2. Once we have controlled for one of the two confounders we no longer need to adjust for the other one.

A variable on a path with two arrows pointing to that variable is referred to as a collider. In fig. 5.20b, for example, the variable C is a collider on the path D ← L1 → C ← L2 → O. C is not a confounder, even though it is independently associated with D (via L1) and O (via L2), as the diagram does not show any correlation between L1 and L2. This path does not therefore meet the criterion for an adjustment set. If 'just to be on the safe side' we were to control for C, we would in fact create a link between L1 and L2, thus introducing bias that was not there before. This is referred to as collider bias. As this example also shows, it is not a good idea to control for all sorts of 'pre-treatment' variables, as is often recommended, unless a properly described DAG indicates that there is an adjustment set.

For a full description of the rules for determining adjustment sets we refer to the literature listed at the end of this chapter. In practice it is often better – especially in the case of large DAGs – to carry out software analyses. A useful free software program is DAGitty.[3] Causal diagrams can also be used to examine effect modification, which is discussed in the next section.

5.5 Effect modification: different effects in subgroups

Although effect modification is not really a topic that should be discussed in a chapter about the validity and reliability of study designs, we nevertheless present it here, in particular to show how it differs from confounding. The term effect modification is used when the effect of a determinant on the disease frequency differs for the various categories of another variable (often another determinant of the disease).

3 The free software package DAGitty, for drawing and analysing causal diagrams (Textor et al.) (application). http://bit.ly/1FjBVdu

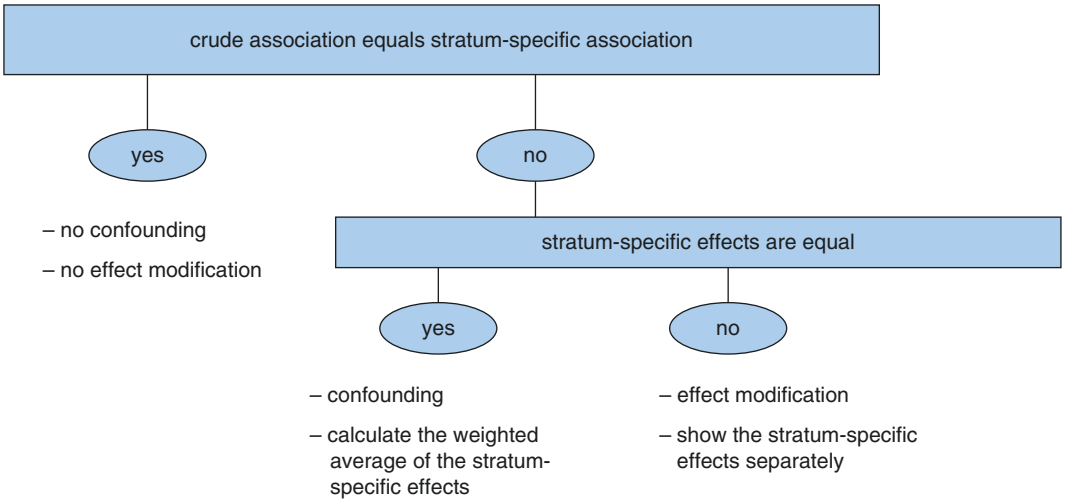


Figure 5.21 Rules for decision-making on effect modification and confounding in stratified analysis

Effect modification occurs in the language of stratified analysis, as described in ▶ par. 5.4.2, when the stratum-specific effects are substantially different (see fig. 5.21). If we use a multivariable regression model to study the data from an epidemiological study we can examine effect modification by including interaction terms in the model.

One of the clearest examples of effect modification is the combined effect of alcohol and car-driving on the risk of having a road traffic accident. Both factors are also separate risk factors for having an accident – a pedestrian or cyclist can have a road accident due to alcohol too – but the combination of the two is a much stronger determinant of accidents than each one separately. We find similar phenomena if we include skin colour when investigating whether exposure to sunlight causes skin cancer: in this example the effect of sunlight on skin cancer is modified by skin colour. Figure 5.22 describes an arrow diagram representing effect modification. Compare this with the arrow diagram for confounding (fig. 5.11).

Effect modification is an aspect of reality that can be examined, just like the effect of the determinant on the development of a disease. This is based on two biological concepts: synergism, i.e. interaction between two or more factors in bringing about a biological effect, and antagonism, i.e. opposition between two or more factors in bringing about a biological effect. In other words, synergism and antagonism act at the biological level. The extent to which synergism and antagonism are expressed as effect modifiers in an epidemiological study will depend on whether the study is capable of uncovering these more complex types of cause-and-effect relationships. It will need a sufficient sample size to distinguish effect modification from random variation, for instance. ▶ Case 5.6 describes a hypothetical example of effect modification by smoking on the relationship between exposure to asbestos and lung cancer.

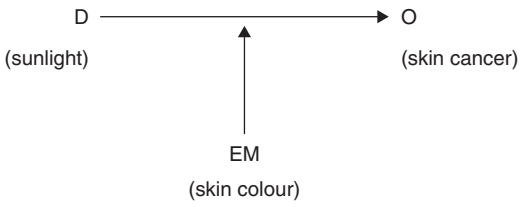


Figure 5.22 Effect modification represented as an arrow diagram

Case 5.6 Asbestos, lung cancer and smoking (hypothetical example)

Suppose we are carrying out a cohort study into the relationship between exposure to asbestos and lung cancer in an industrial population that has been exposed to asbestos (A_1) and a control group of employees who have not been exposed to asbestos (A_0). Both groups include cigarette smokers (S_1) and non-smokers (S_0). Based on exposure to these two determinants of lung cancer we can identify four different determinant groups, A_0S_0, A_1S_0, A_0S_1 and A_1S_1. Let us assume that the study finds the following risks of lung cancer for these subpopulations:

- $R_{A0S0} = R_{00}$ = 23/100,000 per year
- $R_{A1S0} = R_{10}$ = 117/100,000 per year
- $R_{A0S1} = R_{01}$ = 244/100,000 per year
- $R_{A1S1} = R_{11}$ = 1,244/100,000 per year

For these subpopulations, comprising people with different combinations of smoking behaviour and exposure to asbestos, we can calculate ARs and RRs of lung cancer compared with the subpopulation with the lowest lung cancer incidence, whose members do not smoke cigarettes and have not been exposed to asbestos (background risk):

- $RR_{00} = R_{00}/R_{00}$ =1.0
- $RR_{10} = R_{10}/R_{00}$ = 5.1
- $RR_{01} = R_{01}/R_{00}$ = 10.6
- $RR_{11} = R_{11}/R_{00}$ = 54.1
- $AR_{00} = R_{00} - R_{00}$ = 0
- $AR_{10} = R_{10} - R_{00}$ = 94/100,000 per year
- $AR_{01} = R_{01} - R_{00}$ = 221/100,000 per year
- $AR_{11} = R_{11} - R_{00}$ = 1,221/100,000 per year

Clearly, the combination of asbestos and cigarettes present a far higher risk than each of these determinants separately. So does that mean there is effect modification? If we look at the RRs and multiply the effect of asbestos (RR = 5.1) with that of smoking (RR = 10.6), the result is close to the increased risk that we find in the group with combined exposure (RR = 54.1). In other words, there does not appear to be any interaction on the scale of RRs. If we look at the attributable risks, on the other hand, there clearly is interaction, since the additional risk in the group with combined exposure (AR = 1,221/100,000 per year) is far higher than the sum of the separate effects (AR = 94+221 = 315/100,000 per year). So it depends on the model being used: the relative risk model (also referred to as the multiplicative model) or the attributable risk model (also referred to as the additive model).

Looking at this from the point of view of the participants in the population we find that there are a whole lot of people (1,244 – 224 – 117 + 23 = 926/100,000 per year) who would not have developed the disease without the combined exposure. For interpretation at the individual level, then, the additive model is generally more informative. The absence of effect modification on the multiplicative scale provides information on the causal effects of smoking: the risk is increased more than tenfold, regardless of the presence of asbestos as an additional important risk factor. Note, however, that this is a special case, and there will usually be some effect modification on both the additive and the multiplicative scale. Also, a case-control study can only provide information on the multiplicative scale, as it only enables a RR or OR to be calculated, not an AR.

Various combinations of confounding and effect modification due to an external variable can occur. Sometimes there will be no confounding and no effect modification. In other cases there will only be confounding, no effect modification. In yet other cases there will be both confounding and effect modification. Lastly, the effect modification can be so strong that confounding is a 'non-issue': if the effect under consideration varies widely for the various strata of a variable, there may be confounding due to that variable but there is no point in calculating an overall effect measure adjusted for confounding. Evidently there is not a single effect, but different effects for different subgroups in the study population, which need to be reported separately. See the decision tree in fig. 5.21.

Effect modification could be said to be of a

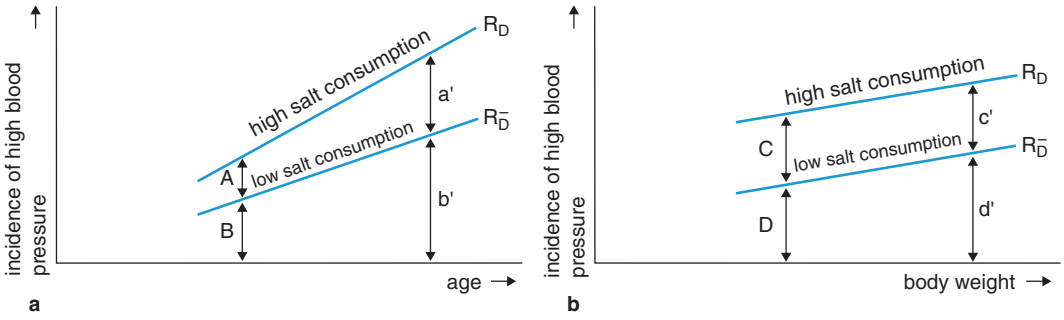


Figure 5.23 Illustration of effect modification on an additive scale (**a**) and a multiplicative scale (**b**)

higher order than confounding. Confounding is a nuisance, a smokescreen within the data set that obscures the view of the true relationships between the variables. We can gain an impression of the magnitude of the distortion by making an estimate of the effect parameter adjusted for the influence of a potential confounder. Effect modification reflects the true situation in the population, showing the natural variability of the extent to which people react to specific determinants.

Investigators may decide to study effect modification, if they are actually interested in that. If so, the next step is to prove the existence of effect modification by inspecting the data and examining the differences between the stratum-specific effects, taking reliability into account. The number of participants in each stratum must be large enough to enable a sufficiently reliable estimate of the stratum-specific effect to be made. But even if researchers are not particularly interested in effect modification, they will still need to ascertain whether there are large differences between the strata, since there is no point in calculating the weighted average of two totally different effects.

Figure 5.23 shows the principle of effect modification in diagrammatic form. Panel a shows that the effect of salt consumption on the incidence of high blood pressure affects the various age groups in line with a multiplicative model, as the effect at each age expressed in terms of RR ((a+b)/b) is the same. In other words, there is no effect modification on the multiplicative scale in panel a. Looked at in terms of an additive model considering the difference between risks, the effect of salt consumption on the incidence of hypertension does depend on age (the difference increases with age). To put it another way, on the additive scale in panel a age is an effect modifier of the relationship between salt consumption and the incidence of hypertension. The reverse is true in panel b, where the difference in the incidence of hypertension between high and low salt consumption is the same for each body weight category. On the additive scale, then, body weight is not an effect modifier, but on the multiplicative scale the effect of salt consumption on the incidence of high blood pressure does depend on body weight. This is why there is effect modification on the multiplicative scale in panel b.

5.6 External validity: degree of generalizability

The internal validity of a study indicates the extent to which the results obtained are correct for everyone who ought to have been included in the study. This is referred to as the study domain or target population. The external validity of a study, on the other hand, refers to the generalizability of the results. This is highly dependent on the extent to which people are kept outside the study domain (in the external population) by applying inclusion and exclusion criteria. Strict criteria improve the homogeneity of the study population and may help to obtain interpretable results for the subdomain. Subsequently the question is whether the results can still be generalized to other subdomains. If we apply very broad criteria, we will obtain a motley mix of parti-

cipants, which may well make the results generalizable but will also make them difficult to interpret because of the heterogeneity.

5.6.1 From abstract domains and target populations to a study population

Both the internal validity and generalizability of a study are strongly influenced by the choice of study population. As we saw earlier, the aim of epidemiological research is to answer an epidemiological question, on e.g. the occurrence of a disease or the relationship between a disease and one or more determinants. This relationship can be represented in the form of an epidemiological function (see ▶ par. 1.1.4) showing how the health parameter in question (the dependent variable) relates to a series of determinants (the independent variables).

When selecting the study population for a scientific study the first step is to define the domain (see fig. 4.1). The domain is an abstract concept referring to the type of person to be studied. It can be narrowed down by individual characteristics such as gender, race and age. The domain needs to be narrowed down particularly if we expect the effect of the exposure under consideration on the outcome studied to be substantially different in one subdomain than in the other subdomain: in other words, the subdomains differ in terms of an effect modifier (see ▶ par. 5.5). For example, if it is clear that the effect will be different between men and women, light-skinned and dark-skinned people, people with and without an accompanying chronic condition, etc., we will want to express this in the definition of the domain, either by limiting it to a single category (only women, only dark-skinned people, only people without a chronic condition), or by explicitly having two subdomains on which to gather separate – and sufficiently precise – epidemiological parameter estimates.

Once the domain has been defined, the next step is to find a suitable source for identifying people of the kind required for the study. Sources (sampling frameworks) could be a civil population register, the patient population of a hospital, people registered with one or more general practices, or members of a student association, for example. The source needs to contain a sufficient number of representatives of the domain. The people in the source who fall into the domain constitute the source population for the study. Within that population we will want to examine events relevant to the study (exposure to determinants, development of the outcome), as these are the people who can provide information for the epidemiological function. The ultimate study population is that part of the source population from which we actually collect information for the study.

When identifying the study population for a situational study (see ▶ par. 4.1.2) the first step is not to define an abstract domain but to identify the specific target population. This definition can be based on strict geographical criteria (e.g. place of residence) and other relevant characteristics. A target population for a study into the use of XTC among young people in Belgian Limburg could be 12 to 18-year-old Limburg residents, for example. Here again we need a specific source population to enable us to carry out the study in practice, in the form of a list, register or database containing all or part of the target population. The important thing is that the people in the target population represented in the source should together be representative of everyone in the target population. From this point of view it is doubtful whether the population register of the regional capital, Hasselt, would suffice as the source for a study into the use of XTC among young people in Limburg.

Errors can easily be made when translating the domain (for an abstract scientific study) or target population (for a situational study) into a study population. If these errors are systematic they can affect generalizability. Systematic differences between the study population and the domain or target population can creep in at various levels and at various stages of the recruitment process:

- Some municipal authorities may refuse to grant the researchers access to their population register.
- Some people in the source population may have to be excluded from the study because they are unable to take part (due to comorbidity, comedication, temporary absence, visual, hearing, speech or motor disorders, etc.).

- Other people, who are qualified to take part, may decide not to do so for personal reasons.
- The people finally included in the sample may ultimately not all be included in the analysis and counted in the final results (due to lack of cooperation, dropout, poor patient compliance).

5.6.2 Representativeness in frequency research is different from representativeness in research into cause-and-effect relationships

High internal validity can be guaranteed by ensuring that the study meets certain methodological standards. The level of external validity, on the other hand, is far more difficult to assess in terms of objective scientific rules and criteria. Generalizing and abstracting on the basis of research information must involve a carefully considered judgment as to whether the populations concerned and the conditions regarding relevant aspects are actually comparable.

In order to generalize on the basis of a descriptive study of disease frequency we will need to have a study population that is representative with respect to all the determinants of the disease in question. It will not be possible to generalize from a study into the frequency of HIV among drug users in Barcelona to all drug users in Europe, and a study of nursing home residents in Vienna will not suffice in order to study the frequency of falls among the elderly. The situation is different when it comes to generalizing about cause-and-effect relationships. Here again the study population needs to be representative of the domain, but it does not need to be representative of all the determinants of the disease in question, only of the relationship under consideration. It could be argued, for instance, that an effect of smoking on the development of lung cancer found in a study of men will also apply to women (although more men are smokers), as women's lungs will display the same reaction to the chemical stimulus of cigarette smoke as men's. We would be less inclined, on the other hand, to extrapolate a relationship between physical activity and osteoporosis (bone decalcification) found in a study of women to the male population, as we can assume that the specific hormonal status of women may affect the effect of physical activity on osteoporosis. In other words, in the case of the smoking-and-lung-cancer example gender is probably not an effect modifier, but it is likely to modify the effect of physical exertion on osteoporosis (see ▶ par. 5.5).

5.6.3 Generalizing from qualitative conclusions works better than generalizing from quantitative conclusions

The generalizability of study results is an elastic concept. The example in the previous section, where we suggested that the effect of smoking on lung cancer among men could also be applied to women, can be contested in the sense that the strength of the effect could well differ somewhat between men and women. Quantitatively speaking ('how strong is the effect?'), the result obtained among men can perhaps not be generalized to women exactly, but this does not affect the conclusion in practice: smoking is such a strong determinant of the risk of lung cancer in both men and women that everyone should be discouraged from it. So, qualitatively speaking ('is or isn't there an effect?') it is perfectly acceptable to generalize the results in this case.

A study that sets out to answer a clear quantitative question ('how large, how strong, how many?') requires a study population that is representative of all the factors that can influence the strength of the effect. Many studies, however, are designed to find qualitative links between determinants and disease outcomes, in which case a study population that is representative of the parameters that very strongly influence the relationship will suffice. Thanks to this phenomenon it is possible to use research carried out in the United States in Europe, and the results of research conducted fifty years ago are still useful. And also thanks to this phenomenon it can even be justified to use only animal experiments to answer certain medical questions: in those cases the laboratory animals can serve as a model for the relationship that we would really like to examine in humans.

5.6.4 Selection is often a useful tool, not just a threat

One of the threats to the internal validity of an epidemiological study is selection bias (see ▶ par. 5.4.1). Having read the previous section, however, readers might conclude that selection is unavoidable. In fact, we saw in ▶ Chapter 4 that selection can actually be used as a way of improving the validity and efficiency of a study. This apparent contradiction is easy to resolve: selection is a tool for researchers that enables them to design a good study and achieve maximum internal validity with minimum effort. They must however ensure that the selection is carried out equally in the groups being compared: in other words, it must be non-differential (see ▶ par. 5.4.1). Differential selection, i.e. different degrees of selection between the groups being compared, causes bias and adversely affects the internal validity of the study.

For example, in a cohort study into the effect of using a computer mouse on the development of arm, neck and shoulder problems it is a good idea to carry out the study on 30 to 50-year-old full-time office workers at a large insurance company who use a mouse for more than five hours a day, provided that the comparison group (with low mouse usage) also comprises 30 to 50-year-old full-time office workers at the same company. This selection will increase the probability of obtaining a valid answer to the research question with relatively low effort compared with taking a random sample of all adult residents of a particular country and classifying them into mouse users and non-mouse users. The results will however probably be generalizable to the total working population of that country.

A case-control study into the effect of folic acid on the occurrence of congenital heart defects could well be carried out among children with such defects operated on at St George's Hospital in the UK, provided that the control group accurately reflects the population from which the patients are taken – for example a control group of children with another severe congenital abnormality assumed not to be related to folic acid intake. The results of such a study could be published and used internationally, in spite of the selection.

5.7 Validity, reliability and responsiveness of instruments

Sometimes epidemiological researchers aiming to carry out an epidemiological study will first need to demonstrate that the instruments envisaged are suitable for that purpose. In particular, they will want to test whether newly developed instruments or modifications to existing instruments produce reproducible and valid data. Reproducibility means being able to achieve more or less the same results on repeated measurement. Validity of an instrument means that it measures what it is intended to measure. Sometimes researchers want answers to other questions: for example, whether the instrument is sufficiently responsive (i.e. sensitive enough to detect relevant changes over time).

5.7.1 Making information specific and measurable: from concept to instrument

Variables that need to be measured in an epidemiological study correspond to theoretical constructs (concepts). Everyone will have some idea of what is meant by concepts such as 'heart attack', 'hearing loss', 'alcohol consumption', 'age' and so on, but this will not suffice if we want to examine these parameters. The theoretical constructs need to be defined precisely if they are to be measured. In an epidemiological study it is important to assign a value to each unit of observation – usually an individual person – for each parameter (outcome, determinants, confounders, effect modifiers). A good theoretical definition of the parameter in question (the concept) and a satisfactory operational definition (that can be used in practice) should result in a suitable instrument.

The following steps need to be taken to make a parameter specific and measurable:

1. Decide what you want to measure: a conceptual definition of the variable (e.g. blood pressure, headache, heroin use).
2. Decide what possible values you wish to identify with this concept. This is referred to as the conceptual scale (e.g. high, normal or low blood

pressure; presence or absence of headache; presence or absence of heroin use).

3. Translate the conceptual definition into an operational definition of the variable. This essentially boils down to choosing or developing an instrument that can provide valid, precise information efficiently on the parameter in question (e.g. systolic and diastolic blood pressure measured using a sphygmomanometer after five minutes in a recumbent position; use of a headache powder; visible signs of use of an intravenous needle on the forearm).
4. Choose the empirical scale on which you will actually measure the participants, i.e. the possible values that you wish to identify when measuring in practice (e.g. diastolic and systolic blood pressure in mmHg measured using a sphygmomanometer after five minutes in a recumbent position; the number of times that headache medication has been used during the past year; presence or absence of intravenous needle scars on the forearm). The empirical scale, then, refers to the instrument ultimately used, including the response options, assessment criteria and scoring rules.

Ensure that the theoretical concept, the conceptual scale and the empirical (or operational) scale are linked as closely as possible. For example, in a study into alcohol and accident risk we would like to have information on the amount of ethanol in the brain tissue during the last few minutes before the accident (the theoretical concept). This is translated into the number of glasses of alcohol drunk during the hours preceding the accident (the operational definition). We could measure this by asking 'How many glasses of alcohol did you drink during the past six hours?' If we were interested in alcohol consumption as part of a study into the development of cirrhosis of the liver, the theoretical concept would be completely different (the lifetime cumulative amount of alcohol that the liver has had to process). In operational terms this could be translated into total alcohol consumption over the past twenty years. The instrument ultimately used, however, is likely to ask about the number of years that the person has been drinking alcohol and the average alcohol consumption per week. The product of these two numbers (multiplied by 52) yields the value of the variable in question.

A whole range of instruments are used in epidemiological research, for example:

- Observation of behaviour or signs of disease
- Questionnaires or interviews (self-reported behaviour or self-reported symptoms, as entered by the respondents themselves or told to an interviewer)
- Physical examination based on direct sensory observation (looking, listening, feeling, smelling, tasting)
- Physical measurements of the body (body weight, blood pressure, muscle strength, etc.)
- Biochemical and other laboratory tests on biological material (serum, urine, hair, tissue biopsies, etc.)
- Imaging (X-rays, CT scans, angiograms, MRIs, PET scans, etc.)

5.7.2 Quality of measurement: precision

Testing for the reproducibility of the results produced by an instrument involves looking at the extent to which the results obtained when carrying out a test more than once on the same participants correspond or differ. It goes without saying that such testing will be carried out on people who are more or less representative of the people we ultimately want to include in the epidemiological study and on whom we shall carry out the measurement in question. Provided that the measuring conditions remain constant, and assuming that the parameter being measured does not change, we can expect repeated tests to produce the same results.

Various terms are used to indicate the degree of correspondence between repeated measurements and the absence of random measuring errors, e.g. reproducibility, repeatability, reliability, agreement, consistency and precision. We prefer the term precision when referring to instruments (and 'reliability' when referring to study design).

There are various ways of assessing the precision of an instrument. First, we can check whether a particular observer repeating the test on the same people always comes up with the same findings (low intra-

observer variability). This entails at least carrying out the test at two different times. Successive observations should be independent of each other and not too close together, to avoid knowledge of the first observation influencing the second one. They should not be too far apart either, as an actual change in the biological parameter could play tricks with the measurement. Secondly, we can check whether two or more observers carrying out the test on the same people – simultaneously as far as possible – reach the same conclusions (low interobserver variability).

Various measures can be used to quantify reproducibility. A commonly used measure for categorical test results is the agreement rate (between the first and second rating or between the first and second rater). Cohen's kappa is a measure of inter-rater and intra-rater agreement that indicates actual agreement as a proportion of potential agreement after adjusting for random agreement. This topic is discussed in more detail in ▶ chap. 9.

5.7.3 Quality of measurement: validity

To test the validity of an instrument we need to compare the test results obtained in a series of individuals with the independent results in these same individuals of another objective instrument providing a more or less accurate picture of the real value of that parameter. An instrument with 100% validity is termed the gold standard. In contrast to testing for reproducibility, testing for validity has a clear dependent variable (the result from the gold standard, Y) and a clear independent variable (the result from the instrument being tested, X). The question of validity can therefore be simply translated into an epidemiological function ($Y = b_o + b_1X$), where the distribution (SD) in the regression coefficient b_1 is a measure of the instrument's validity. This distribution should ideally be zero, as we can then perfectly predict what the true value will be once we have carried out the measurement. This is also true if the regression line does not pass through the zero point at an angle of precisely 45°. In other words, as long as the points over the interval required for the study are on a single straight line, the instrument is valid and can therefore be used for epidemiological research. The true value could be deduced from each measurement using a calibration line obtained from the above test. Calibration is not even needed for an etiological study comparing the average measurements between groups. Systematic measuring errors will be eliminated when the groups are compared. The validity of instruments is discussed in more detail in ▶ chap. 9.

Recommended reading

Altman DG. Practical statistics for medical research. 2nd edition. London: Chapman & Hall; 2010.

Armstrong BK, White E, Saracci R. Principles of exposure measurement in epidemiology. Oxford: Oxford University Press; 2008.

Carneiro I, Howard N. Introduction to epidemiology. 2nd edition. Maidenhead: Open University Press; 2005.

Elwood JM. Critical appraisal of epidemiological studies and clinical trials. 3rd edition. New York: Oxford University Press; 2007.

Gordis L. Epidemiology. 5th edition. Philadelphia: Elsevier Saunders; 2014.

Grobbee DE, Hoes AW. Clinical epidemiology: principles, methods, and applications for clinical research. 2nd edition. Burlington: Jones and Bartlett Learning; 2015.

Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: howto do clinical practice research. 3rd edition. Philadelphia: Lippincott, Williams & Wilkins; 2006.

Kleinbaum D, Klein M. Survival analysis: a self-learning text (statistics for biology and health). 3rd edition. New York: Springer; 2012.

Morgan SL. Handbook of causal analysis for social research. Dordrecht: Springer; 2013.

Pearl J. Causality: models, reasoning and inference. New York: Cambridge University Press; 2009.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd edition. Philadelphia: Lippincott, Williams & Wilkins; 2012.

Simon SD. Statistical evidence in medical trials: what do the data really tell us? New York: Oxford University Press; 2006.

Szklo M, Nieto FJ. Epidemiology: beyond the basics. 3rd edition. Burlington: Jones and Bartlett Learning; 2014.

Vet HCW de, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge: Cambridge University Press; 2011.

Webb P, Bain C. Essential Epidemiology: an introduction for students and health professionals. 2nd edition. Cambridge: Cambridge University Press; 2011.

Weiss NS. Clinical epidemiology: the study of the outcome of illness. 3rd edition. New York: Oxford University Press; 2006.

Etiology and causality

6.1 Introduction

What does it mean if we say that A is the cause of B? The simplest idea of causality is that we mean that B always follows A, but there are two reasons why this cannot be correct. First, a causal relationship need not be perfect: smoking is a cause of lung cancer, but not all smokers get lung cancer. Secondly, even a perfect association is not enough to be regarded as causality: every time the cock crows the sun rises, but this does not of course mean that the cock's crowing causes the sun to rise.

If we consider the matter in more depth we can see that causality has something to do with what would have happened if the cause had been different. If the cock had not crowed the sun would still have risen, but many lung cancer patients would not have developed lung cancer if they had not smoked. That is why we speak of causality in the case of smoking but not in the case of the crow.

Epidemiologists are interested in differentiating between causal and non-causal relationships because they are looking for ways of intervening. We cannot make the sun rise earlier by getting the cock to crow earlier, but we can improve people's health by discouraging smoking. Interventions are only possible where there is causality, so causal relationships are quintessential in epidemiology.

The purpose of observational epidemiological research is to establish whether there is a statistical relationship (an association) between the presumed cause and the outcome under consideration. Identifying an association is not enough to prove causality, however: people who carry a cigarette lighter tend more strongly to develop lung cancer, but we cannot say that possessing a lighter leads to the development of lung cancer. In this example smoking is of course the 'real' cause, and carrying a lighter is merely associated with this. In many cases the situation is not so obvious, and it remains uncertain whether we have tracked down the 'real' cause. We epidemiologists need to be particularly careful here, as associations in epidemiological research are fairly likely to be reported as news, even when no causal relationship has been established (see fig. 6.1). Of course it is an interesting fact that higher socioeconomic status is associated with longer life expectancy, for example, but in etiological epidemiology we need to know what the explanation for this association is: genetic selection, lifestyle factors (smoking, physical exercise, particular dietary factors) and/or particular living conditions (air pollution, hygiene, occupational exposure).

Once we have identified the main causal determinants, preventive measures can be taken: a vaccine against the bacterium that causes the disease, education to change lifestyle, statutory measures to ensure that environmental pollution is reduced, and so on.

It goes without saying that there will not be an appropriate preventive intervention for all etiological factors (genetic variation, age and gender, for instance). It is nevertheless useful to know about these non-modifiable causal determinants, for example to help us understand why under apparently similar circumstances some people develop a disease and others do not.

6.1.1 Risk factors for the development of a disease do not necessarily explain its course

Sometimes the prognostic factors for the course of a disease will be the same as the risk factors for developing it. We know, for example, that smoking plays a role not only in the development of coronary heart disease but also in the course of the disease after the first heart attack, but this is by no means always the case. As ▶ case 6.1 shows, some risk factors for breast cancer also play a role in the prognosis for the disease, whereas others do not.

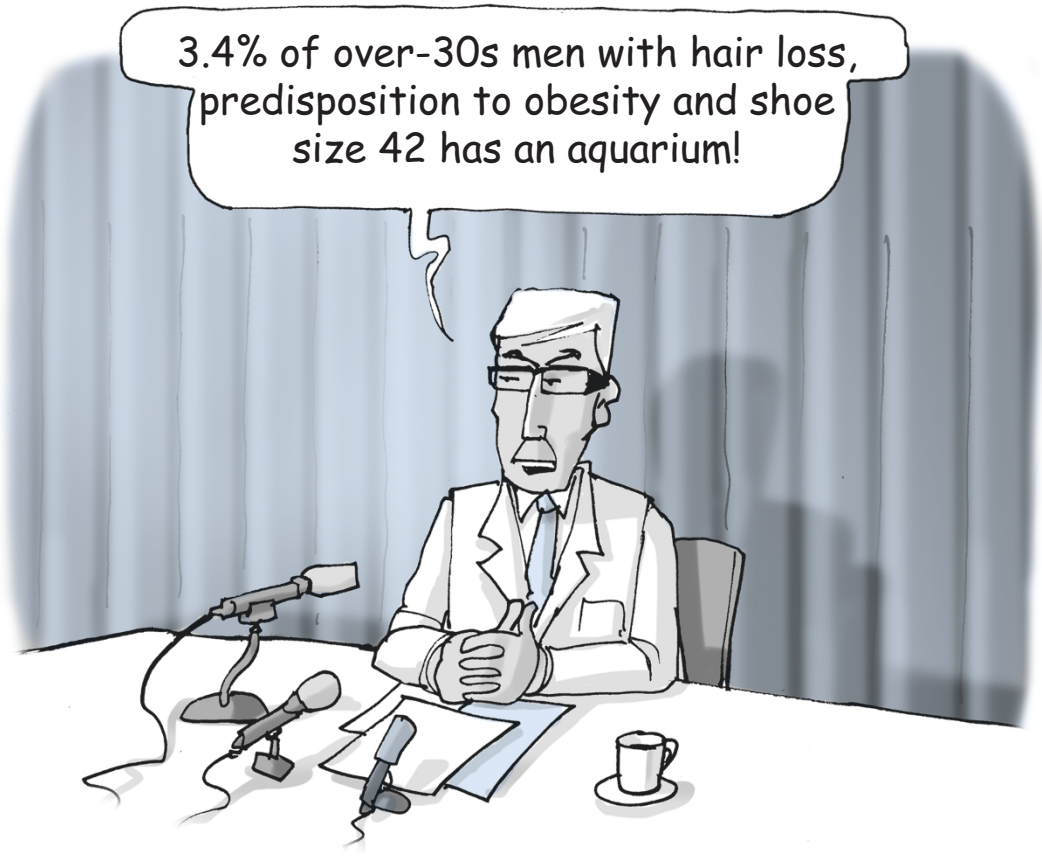


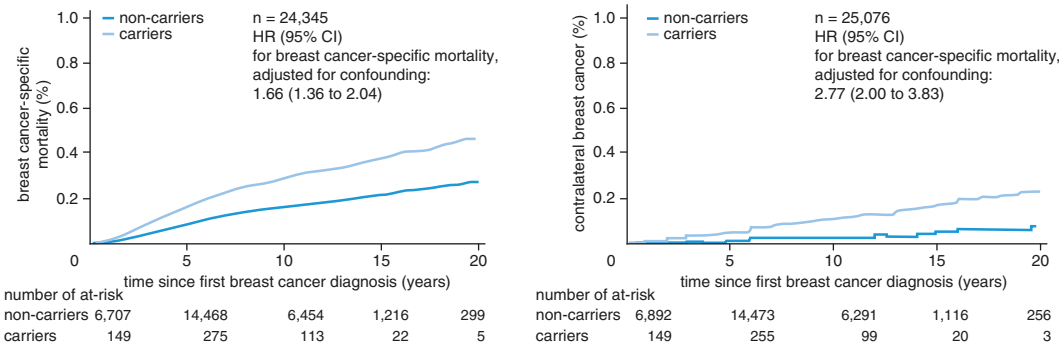
Figure 6.1 Epidemiological associations in the news

Case 6.1 Risk factors and prognostic factors for breast cancer in women

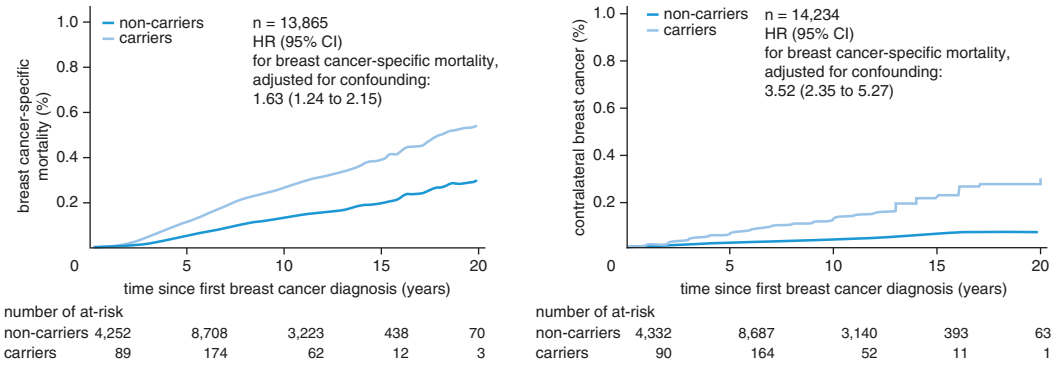
Overweight is a risk factor for the development of breast cancer in postmenopausal women. It is also associated with poorer breast cancer survival rates, partly because overweight women have worse survival rates due to other diseases associated with it. Even if we adjust for these comorbidities, however, overweight remains a prognostic factor for the survival of breast cancer patients.

Age, on the other hand, is a factor that has different effects on the development and course of breast cancer respectively. Higher age is associated with a higher risk of breast cancer, whereas young age is associated with higher mortality from it, probably because women who develop it at a young age have a more aggressive type of tumour.

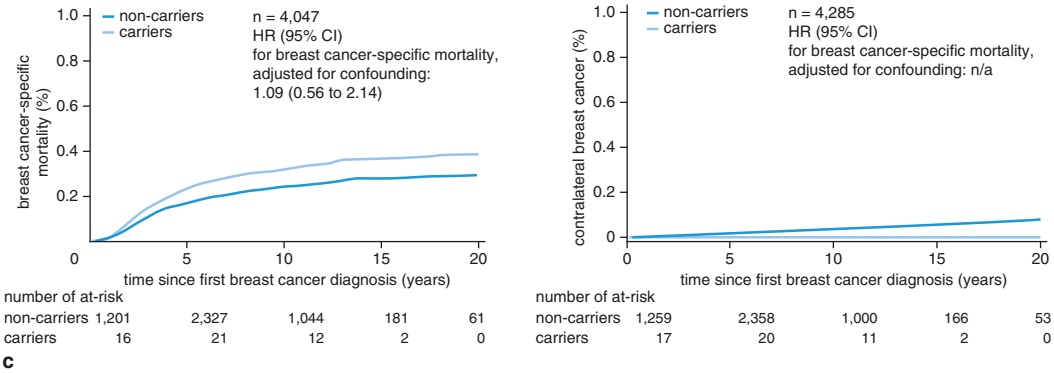
We know that various hereditary variants play a role in the development and/or survivability of breast cancer. Variant 1100delC in the *CHEK2* gene is a fairly rare variant, occurring in up to 1.5% of individuals in North-West Europe. Women with this variant in their DNA have more than double the risk of developing breast cancer, both the primary tumour and the second primary tumour in the other breast. Figure 6.2 shows the results of a study of the effects of breast cancer in women with a first tumour. Some of them were carriers of the *CHEK2**1100-delC variant. The results show that



a



b



c

Figure 6.2 Breast cancer-specific mortality (*left*-hand side of figure) and cumulative incidence of contralateral breast cancer (*right*-hand side of figure) in carriers and non-carriers of the CHEK2*1100delC genetic variant, stratified for oestrogen receptor status (**a** all participants, **b** receptor-positive patients, **c** receptor-negative patients)

CHEK2*1100delC is also associated with a poorer prognosis: women with a hormone receptor-positive breast tumour and the *CHEK2*1100delC variant in particular were at increased risk of dying from breast cancer, compared with breast cancer patients who did not have the *CHEK2*1100delC variant; the multivariate corrected hazard ratio was 1.63 (95%-CI 1.24-2.15).

6.1.2 Causal diagrams as a theoretical explanation for the development of disease

There are always multiple factors involved in the causal chain leading to the disease (multicausality). The various determinants affect one another, and each determinant in turn has its own determinants. If we try to show these various factors schematically we end up with a causal diagram, a simplified theoretical representation of how the disease in question develops due to the interplay of various factors. Directed Acyclic Graphs (DAGs), which we introduced in the previous chapter as a means of detecting confounding, are also useful when teasing out complex questions of causality.

Case 6.2 uses a simplified model of the development of tuberculosis. We can deduce from the corresponding causal diagram that although exposure to the tuberculosis bacterium (*Mycobacterium tuberculosis*) is an essential factor, there are other factors that determine whether an individual will actually develop a tuberculosis infection. As the causal diagram clearly shows, socioeconomic factors are also important when combating tuberculosis in the population as the tuberculosis bacterium itself. This contention is supported by the observation that deaths from tuberculosis fell sharply in the course of the 19th and 20th centuries, whereas the tuberculosis bacterium was not discovered until 1885, and antibiotic therapy did not become available until 1948.

In principle, a causal diagram can be drawn for any disease. The results of epidemiological and pathophysiological research enable the models to be constantly improved and refined. With sufficient empirical support they can be used as a basis for preventive or therapeutic interventions, even if the precise mechanism of action has not yet been identified. Conversely, the availability of an etiological model, as shown in a causal diagram, will help epidemiological researchers to formulate the precise research question (usually a single relationship in this model) and design the study required to answer that question.

Case 6.2 The etiology of tuberculosis

Infection experiments have been used to prove that *Mycobacterium tuberculosis* infection of susceptible hosts results in the clinical picture long known as 'tuberculosis'. The bacterium invades and grows in the host tissue (lungs, bones and brain), eventually causing the disease. From the pathogenic point of view, therefore, combating the disease is a question of eliminating the bacterium with an effective antibiotic or vaccine.

These drugs have been developed successfully through biomedical research, but the development of tuberculosis is far more complex (see ◘ fig. 6.3). Whether it actually develops in an individual depends on how susceptible the host is and the degree of exposure to the bacterium. Whether individual susceptibility is a 'cause' of tuberculosis is a matter for debate, but it is certainly an important factor when it comes to prevention and treatment. Indeed, improved diet and living conditions – and preventing HIV infection – have proved more important worldwide in combating tuberculosis epidemics than antibiotics and vaccines.

6.2 Causality

6.2.1 Causality: a vital part of the causal complex

Our concept of causality is based first and foremost on simple, everyday observations: when we flip the light switch on the wall we see straight away that the light goes on or off. If the bulb is faulty or there is a power failure, however, we realize that it takes more than just flipping the switch to get the light to come on. There would be no electric light if the home did not have wiring. Clearly, the position of the light switch is not the sole cause; it forms part of a complex of factors (electricity in the grid, good wiring, a working bulb). The tendency to regard the switch as the unique cause of the light coming on is understandable in that it is the last factor in the causal

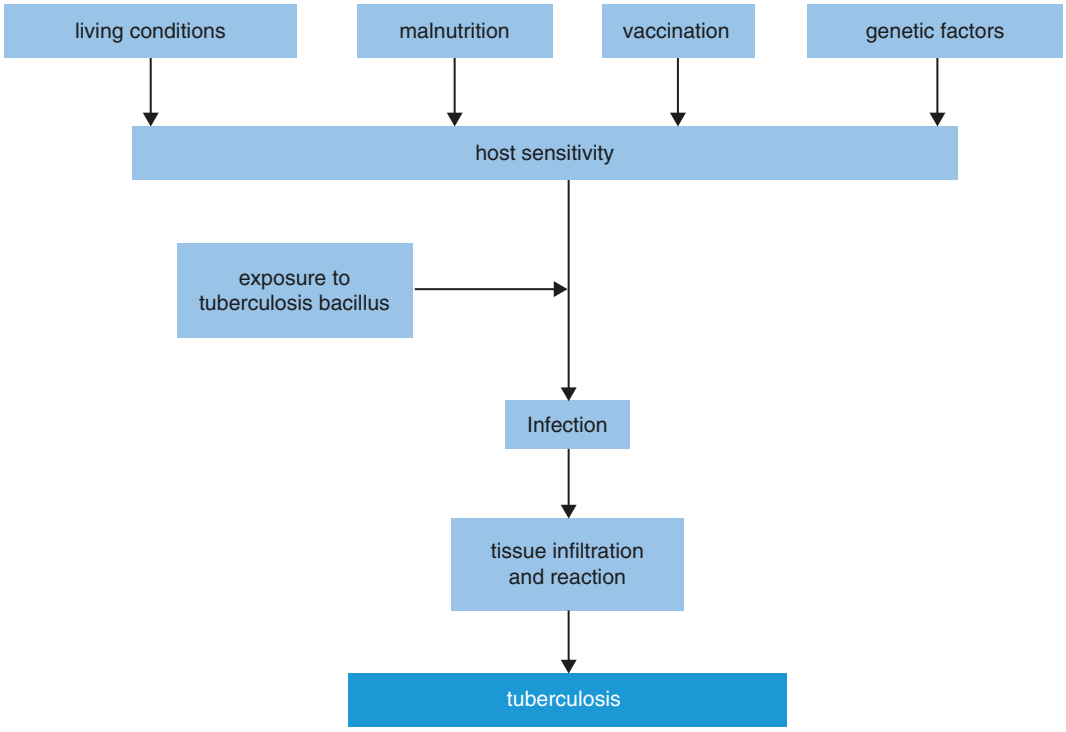


Figure 6.3 Causal diagram of tuberculosis

complex of factors that eventually produce the result, light.

Etiological epidemiology involves searching for the causes of a health problem.The research focuses on the factors that we can use to 'disable' the process that results in disease, but we must not forget that all sorts of other factors contribute to that process. In many cases these factors are not modifiable (e.g. genetic predisposition), but this does not make them any less important in the search for causality.

Ideally we would want to investigate whether a particular factor makes a causal contribution to the development or course of a particular disease by manipulating the presumed cause and seeing whether the effect under consideration (the health problem) actually occurs or in fact disappears. Obtaining empirical evidence of causality is often not a practical proposition in epidemiology, so we need to fall back on interpreting results from observation.

6.2.2 A model of causality: a combination of factors constitutes a sufficient cause

We refer to a set of determinants that together will inevitably result in the medical condition under consideration as a sufficient cause. If one of the risk factors is missing from the set – e.g. the light switch in our example above – the outcome will not occur.

Figure 6.4 shows a model of three such sufficient causes, all three of which result in the same hypothetical condition. In other words, there are three ways in which the disease can develop. Each of the three sufficient causes is made up of eight determinants that together – provided they are all present – result in the condition. Some of these determinants (C, E and H) occur in all three of the sufficient causes, and these are referred to as the necessary causes. Without these determinants the condition cannot develop.

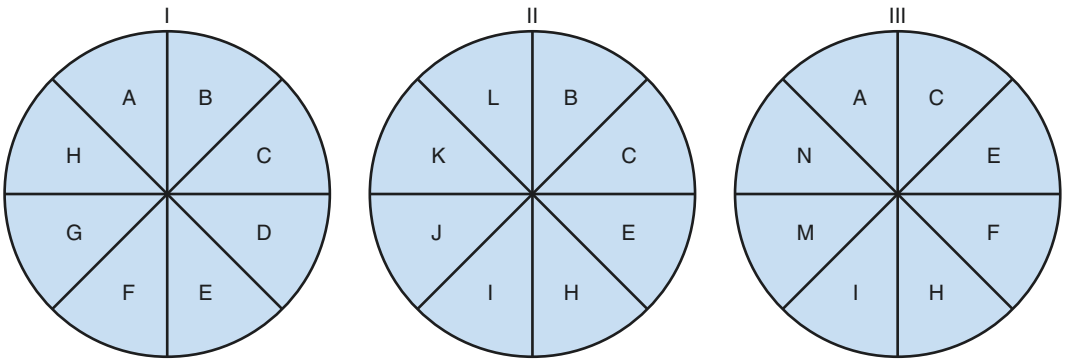


Figure 6.4 Causal model of the development of a disease with three sufficient causes

Sufficient causes, then, are in fact always combinations of component causes, necessary or otherwise. An example of a necessary cause is deep vein thrombosis in the case of pulmonary embolism: an embolism will not develop without thrombosis in the leg or pelvis. Clearly, once we have discovered a necessary cause, in principle we hold the key to the effective prevention of the condition in question. We have succeeded, for example, in eliminating malaria in large parts of the world by killing the mosquito that transfers the pathogen (a microorganism) from diseased to healthy people, since the malaria mosquito is a necessary cause for the development of malaria (as is the microorganism itself, of course). Similarly, vaccination has been successful because inadequate immune resistance is a necessary cause for the development of infectious diseases.

Smoking will serve to illustrate the difference between a necessary cause and a sufficient cause. Smoking is a major risk factor for lung cancer, but it is certainly not a sufficient cause, as over 90% of smokers will not develop it. Nor is smoking a necessary cause of lung cancer, as it also occurs in non-smokers. This is relatively rare, however, and lung cancer is usually the result of a sufficient cause including smoking, so abstaining from smoking could nevertheless prevent over 90% of lung cancer cases.

The causal model outlined in fig. 6.4 is mainly educational, as in reality it is extremely difficult to identify all the components of the various sufficient causes. As a rule, all that we have is the associations (usually expressed as AR or RR) between particular risk factors and the development of the condition under consideration. The known risk factors will generally be only part of the totality of factors that influence the development of the condition.

6.2.3 Causal interpretation: ruling out the possibility that an association is explained by chance or bias

Etiological research is sometimes compared to a court case, with the epidemiologist playing the role of the prosecution. Just as it is the job of the prosecution to prove (beyond reasonable doubt) that the defendant, not someone else, is guilty (alone or with others) of the criminal offence, an etiological researcher needs to convince the users of the study that the determinant is causally responsible (or co-responsible) for the development of the disease in question. Nowadays forensic epidemiologists are indeed consulted in court cases involving damage to health (side effects of medical drugs, medical errors, food contamination, occupational diseases and so on). For more information see ▶ par. 6.3.1.

The causal model discussed in ▶ par. 6.2.2 also makes it clear when there is confounding and/or effect modification, concepts explained in the previous chapter. A confounder is not part of the set of determinants that, together with the primary determinant, constitute the sufficient cause; on the contrary, confounders are found in the other mechanisms (sufficient causes) that can result in the disease. In fig. 6.4, for example, the risk factors I, J, K, L, M and N are potential confounders of the relationship

between risk factor G and the development of the condition. Whether these risk factors actually act as confounders will depend on whether they are associated with the presence of G in the study population. If there is an association of this kind between the primary determinant (G) and the confounders (I, J, K, L, M and N), the difference between the incidences of the condition when G is present and when it is absent will be due not only to factor G (and sufficient cause I) but also to the other sufficient causes (II and III in ◘ fig. 6.4) that do not include G but do include the confounders.

Determinants that together constitute a sufficient cause are dependent on one another to achieve their effect. If there is a relationship of this kind between risk factors this is referred to as effect modification (see ▶ par. 5.5). This does not mean that the two determinants cannot separately form part of other sufficient causes, but there is evidently at least one mechanism (sufficient cause) where it is necessary for both factors to be present. In ◘ fig. 6.4, for example, it is possible for determinant A to result in the disease without determinant B (via mechanism III); conversely, determinant B could contribute to the disease via mechanism II without A being present. But there is another mechanism (I) which only results in the disease if both A and B are present: this is referred to as modification of the effect of determinant A by determinant B (and vice versa). Thus A has a greater effect on the development of the disease if determinant B is also present (and vice versa). A familiar example of effect modification is the way smoking and asbestos influence the development of lung cancer (see ▶ case 5.6). Both smoking and exposure to asbestos each increase that risk substantially, but smokers who are also exposed to asbestos run a far greater risk of lung cancer than we would expect from simply adding the separate risks due to smoking and asbestos.

In the example in ◘ fig. 6.4 the determinants K and J, and also M and N, display effect modification, but here the effect of one determinant is entirely dependent on the presence of the other determinant: J has no effect without K (and vice versa), and N has no effect without M (and vice versa).

The situation where a determinant acts as both a confounder and an effect modifier corresponds to risk factors B and E, I and J, and H and I in ◘ fig. 6.4, which have some sufficient causes in common but not all. Age is an example of a risk factor that is both a confounder and an effect modifier in the case of the association between total serum cholesterol and coronary heart disease in women. Age is an effect modifier because the association between total serum cholesterol and coronary heart disease in women is much stronger after the menopause than before it. But age is also a confounder, because it is a determinant of coronary heart disease in both premenopausal and postmenopausal women and is also associated with total serum cholesterol.

To establish whether an association found in an observational study can be interpreted causally, so as to draw conclusions on the strength of the effect and any effect modification, we try to rule out other possible explanations step by step. Specifically, the alternative explanations are as follows:

1. The association found can be explained (at least partly) by selection bias or information bias (see ▶ par. 5.4.1). To find out whether this alternative is likely we need to discuss the design and procedures of the study in detail. Even if no association is found it is important to look at bias, as this can make a relationship that exists in reality invisible in the study.
2. The association found can be attributed wholly or partly to confounding (see ▶ par. 5.4.1). We can use stratified analysis of the data to see whether this is a likely explanation of the association found, provided that for all the study participants information on these confounders is available in the data set. If the crude association does not differ substantially from the association within the strata of the confounders, confounding is probably not the explanation.
3. The association found may be due to chance. We can use statistical techniques (confidence intervals) to determine whether this alternative is likely. A narrow confidence interval that does not include the value of 1 (for RR and OR) or zero (for AR) does not provide much support for this alternative explanation.

If all three of these explanations are unlikely, a causal relationship is more likely, but this does not give us absolute certainty regarding the presence of a causal relationship. It could be said, following the philoso-

pher of science Karl Popper, that researchers are constantly trying to reject (i.e. falsify) the hypothesis of causality based on the results of the research (see ▶ par. 1.1.5). The credibility of the hypothesis that there is a causal relationship increases the more rigorous and varied the researchers' attempts to test and rule out such alternative explanations are.

To support the discussion of causality, sensitivity analyses are often used to explore the quantitative influence of bias and random measurement errors on the estimation of effect. To identify selection bias, misclassification and confounding we examine what values could be realistic and use computer simulation to calculate the resulting distribution of possible outcomes. If this range of alternative values does not lead to substantially different conclusions, causality is more likely. It will not be possible to interpret a weak link between alcohol consumption and lung cancer correctly if we have not carefully controlled for the influence of cigarette smoking. Conversely, a strong link between smoking and lung cancer will not vary greatly if we add realistic assumptions about distortion due to the absence of good data on alcohol consumption. Sensitivity analysis is commonly used in situations where there are uncertainties in the design and execution of the study, for example in comparative research into the cost-effectiveness of two or more treatments. A conclusion as to the superiority of one of the treatments based on research of this kind will gain in strength if we can show by a sensitivity analysis that this conclusion does not change if we use other realistic values for uncertain factors that have a substantial influence on the results.

▶ Par. 3.2.6 introduced the concept of population attributable risk (PAR). This is a measure of the contribution made by a particular determinant to the occurrence of the condition under consideration in a particular population. The model of causality described in ◘ fig. 6.4 shows that the PAR depends on the frequency with which the other determinants in the same sufficient cause occur in the population. Suppose the three sufficient causes in ◘ fig. 6.4 are responsible for 80, 15 and 5% respectively of the cases of the condition under consideration. The population attributable proportion for risk factor K, for instance, is thus 15%, since eliminating cause K in the population eliminates sufficient cause II (which accounts for 15% of the disease cases). For a necessary cause (e.g. C, E and H in ◘ fig. 6.4) the population attributable proportion is always 100%. As these examples show, there is no point in adding up population attributable proportions, as the sum will soon exceed 100%. This is because – despite the fact that a conjunction of various factors is needed for a disease to develop – eliminating just one of those factors can prevent cases of the disease.

◘ **Table 6.1** Hypothetical incidence density of head and neck cancer per 100,000 person-years

	non-smokers	smokers
non-drinkers	1	4
drinkers	3	12

The same principle is illustrated by the following example. Suppose the incidence density (ID) of head and neck cancer per 100,000 person-years among smokers and drinkers is distributed as shown in ◘ tab. 6.1.

We can calculate from the table that in the group of smoking drinkers 12−3 = 9 of the 12 incident cases of head and neck cancer per 100,000 persons can be attributed to smoking and 12−4 = 8 of the 12 to drinking. The etiologic fraction among the exposed (EFe) is thus 75% for smoking and 67% for drinking. The sum is significantly more than 100%, because effect modification (see ▶ par. 5.4) is taking place: without smoking and without drinking there would have been 1 case of cancer per 100,000 person-years. With smoking alone there would have been 3 additional cases, and with drinking alone 2 additional cases. This means that the combination of smoking and drinking has caused 12−1−2−3=6 cancer cases per 100,000 person-years. Without this effect modification the attributable proportions for the group of smoking drinkers would have been (6−3)/6 = 50% for drinking and (6−4)/6 = 33% for smoking respectively.

We can also deduce from the model in ◘ fig. 6.4 that the attributable risk (AR) for a necessary cause is equal to the incidence of the disease in question. In the absence of a necessary cause the incidence is zero by definition, so the relative risk (RR) will then be infinitely large. In the case of non-necessary risk

Figure 6.5 Biostatistician Austin Bradford Hill formulated criteria for arguing the likelihood of causation in the 1960s

factors the magnitudes of AR and RR depend on the relative contributions made by each of the sufficient causes, with and without the risk factor, to the total incidence. This means, then, that the strength of association (expressed in terms of AR or RR) depends on the population under consideration.

6.2.4 Arguments that make causality more likely

It will be clear from the foregoing that epidemiological research cannot produce conclusive proof of a causal relationship (as is true for any scientific discipline). To enable conclusions on causality to be drawn despite this disclaimer, over the years guidelines have been developed on how to argue the likelihood of causality. Well-known for example are the criteria for causation of Austin Bradford Hill (fig. 6.5). These criteria, as set out below, provide an aid to the systematic analysis of matters of causality.

Study type

As some study designs provide more ways of controlling bias and confounding than others (see ▶ chap. 4 and ▶ chap. 5), the various types of study have different levels of evidence when it comes to causality. The hierarchy, from low to high, is as follows: case report, patient series, case control study, cohort study and RCT (see fig. 4.4). There is still some debate as to the relative merits of case control studies and cohort studies, as the causal explanatory power of these two designs is potentially the same. In practice it is often easier to avoid bias (including confounding) in cohort studies than in case control studies.

Coherence

When discussing the coherence of findings of which we attempt to gauge their causality we look for arguments – preferably based on the presumed biological mechanism of action – that a causal relationship found for one category in the population is likely to apply to other categories as well. The more coherent an association is, the more likely it is to reflect a causal effect.

Biological plausibility

The question of theoretical or biological plausibility is whether a causal relationship is likely, given our current knowledge of human biology. For example, the idea that there is a link between smoking and lung cancer makes sense to everyone, but a suggested association between smoking and cervical cancer is less self-evident. No-one will be surprised to be told that there is a relationship between aircraft noise and sleep disorders, but it makes less sense to attribute the occurrence of hare lip to aircraft noise. Biological plausibility is not a hard and fast criterion: what seems like a reasonable biological explanation to one researcher may be rejected as nonsense by another. The history of medicine is riddled with examples of inventive descriptions of biological mechanisms that eventually turned out to be incorrect. Relevant findings from basic biomedical research, especially research on animals, are very important when assessing this criterion. A proven causal relationship in laboratory animals does not of course automatically mean that the relationship also exists

in humans, but it often provides a hint in that direction.

Temporal relationship

A cause must precede its effect: this is a necessary precondition, and in fact the only absolute criterion of causality. If it can be shown that the presumed effect preceded the cause, there cannot be causality. Unfortunately, the converse of this rule is not so strong: the correct time sequence does not guarantee a causal relationship.

Strength of the association

A strong association tends to suggest a causal relationship more than a weak one, as a strong effect is not very likely to be explained by bias and confounding. As we saw in ▶ par. 6.2.3, the strength of an association is determined to a large extent by the frequency of the other determinants in the complex of sufficient causes. Also, a weak association does not rule out causality.

Dose-response relationship

Dose-response relationships are an argument in favour of causality, so researchers investigate whether a higher dose or longer exposure coincides with a higher frequency of the response. A dose-response relationship can also be due to a confounder, however. Moreover, the absence of a dose-response relationship cannot generally be interpreted as an argument against causality in and of itself. There could be a threshold below which there is no response, for instance. Or there could be a threshold for the risk factor above which the response is always maximized. Another possibility is that the causal relationship between dose and response is U-shaped.

Consistency

If the same association has been shown by different researchers, at different times and places, in different populations and using different study designs, that is an argument in favour of causality. However, this criterion – consistency – is not absolute either. In addition, if different subgroups in the study population (both men and women, both older and younger people, both rural and urban areas etc.) demonstrate a similar association, it is an argument in favour of causality. However, a lack of consistency can be due to effect modification, as particular causal relationships could conceivably manifest themselves only in special circumstances. If different studies or subgroups show opposite effects (increased risk in one subgroup and reduced risk in another), causality is less likely.

6.2.5 Levels of evidence

A hierarchy of study designs with ascending levels of evidence was presented in ▶ chap. 4 (see ◘ fig. 4.4). The Bradford Hill criteria also use a hierarchy of this kind, with 'experiment' providing the top level of causal credibility. Over the years many kinds of levels of evidence have been published, often in the form of a pyramid, with systematic reviews or randomized trials at the top and case reports or expert opinion at the bottom. Levels of evidence are nowadays used routinely when devising healthcare guidelines: recommendations on treatment are stronger, the higher the levels of evidence available in the research data are. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) system provides a further refinement.[1]

6.2.6 Causality at the individual versus the population level

When interpreting associations it is useful to argue separately for each element in the causal chain what would happen in the population if the factor in question was not present. Would the effect – the disease, for example – occur with the same frequency if the situation were completely identical except for the absence of that one factor? In a thought experiment of this kind, then, the argument is based on the converse situation (counterfactual). If the disease only occurs in the thought experiment if the factor is present, and not if it is absent (as in the light switch example in ▶ par. 6.2.1), the factor is a necessary cause. If the disease still does occur, but to a lesser extent, the factor is evidently one of the set of deter-

1 The Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (website). ▶ http://bit.ly/1IQ69Ub

minants that together constitute a sufficient cause for the condition in question, but it is not a necessary factor. If the disease occurs with the same frequency with or without the presence of the factor (all other factors remaining equal), that factor evidently does not influence the development of the disease in question. In this case we need to find some other way of explaining the association found – by chance, for example, or by showing that the factor in question often coincides with another factor that is causally related to the disease in question (take for instance carrying a cigarette lighter in relation to the development of lung cancer as described in ► par. 6.1).

At the individual level it is very difficult to show what the causal factor is. For example, did granddad get lung cancer because he smoked a lot of cigarettes during his lifetime? Smoking is not a necessary cause of lung cancer, so we do not know what would have happened if he had not smoked. Only if we are dealing with a necessary cause can we say with certainty that the individual would not have developed the disease if he had not been exposed to the factor in question.

The question of causality is easier to answer at the population level. 'Was the lung cancer epidemic caused by large numbers of people starting to smoke cigarettes?' Yes, because we know for sure that there would be far less lung cancer in the population if – in the converse situation, but in otherwise similar circumstances – everyone were to refrain from smoking cigarettes. It is precisely these population-based conclusions, founded on sound epidemiological research which has shown that the incidence of lung cancer increases sharply under the influence of the number of smokers in the population, that can be applied to individuals: everyone who smokes is at greater risk of developing lung cancer and dying from it than if he or she did not smoke. But just as the population would not be immune from lung cancer if nobody smoked, there is no guarantee that a non-smoking individual will be immune from the disease. The converse is also true: smokers do not always develop lung cancer, and the incidence of lung cancer in a population of hardened smokers is far below 100%. Smoking is undoubtedly a cause of the disease, but it is neither a necessary cause nor a sufficient one. Population data can be translated into disease risks, but not into absolute conclusions on individual causality.

Case study 6.3 Caffeine intake and risk of melanoma

Ultraviolet (UV) radiation is the major environmental cause of cutaneous malignant melanoma, a common type of skin cancer. Caffeine, a central nervous stimulant that is present in coffee and tea, had previously been shown to inhibit UV-induced thymine dimers and sunburn lesions in the epidermis of mice, acting as a sunscreen. Also, earlier studies had concluded that caffeine enhances UV-induced cell apoptosis and had shown that caffeine inhibits growth of melanoma cells both in vivo and in vitro. This led to the hypothesis that caffeine intake could reduce the risk of melanoma on locations that are most exposed to sunlight.

To test this hypothesis, data from three large cohorts were combined. The Health Professionals Follow-Up Study (HPFS) was established in 1986 and provided data on 39,424 male health professionals aged 40 to 75 years. The purpose of the HPFS study was to evaluate a series of hypotheses about men's health relating nutritional factors to the incidence of serious diseases such as cancer, heart disease and other vascular diseases. This all-male study was designed to complement the all-female Nurses' Health Study that began in 1976. The Nurses' Health Study and the Nurses' Health Study 2 (1989) contributed data on 74,666 and 89,220 female nurses aged 30 to 55 and 25 to 42 respectively. Participants were selected based on the availability of a completed dietary questionnaire, and had to have no history of any cancer at baseline. Participants in all three cohorts received biannual questionnaires on disease outcomes.

Caffeine intake was categorized as quintiles. The lowest quintile (<60 mg) amounts to less than half a cup of coffee per day; the highest quintile (≥393 mg) amounts to drinking at least 3 cups of coffee each day. In total, 2,254 new cases of melanomas were observed in over four million person-years of follow-up. Cox proportional

Table 6.2 Risk of cutaneous malignant melanoma for quintiles of caffeine intake

	Q1 (<60 mg/day)	Q2 (60–140 mg/day)	Q3 (141–246 mg/day)	Q4 (247–392 mg/day)	Q5 (≥393 mg/day)
pooled for women and men	1.0	0.83 (0.70, 0.99)	0.79 (0.67, 0.95)	0.87 (0.66, 0.93)	0.71 (0.59, 0.86)
women	1.0	0.79 (0.63, 0.99)	0.82 (0.66, 1.0)	0.73 (0.59, 0.91)	0.66 (0.53, 0.83)
men	1.0	0.91 (0.68, 1.2)	0.74 (0.54, 1.0)	0.89 (0.66, 1.2)	0.82 (0.59, 1.1)

[a]Adjusted for family history of melanoma, personal history of non-skin cancer, natural hair colour, number of moles on legs or arms, sunburn reaction as a child/adolescent, number of blistering sunburns, time spent in direct sunlight since high school, cumulative ultraviolet flux since high school, body mass index, smoking status, physical activity, total energy intake and alcohol intake. HRs are computed including 95% confidence intervals. Q1 served as the reference quintile.

hazards regression was used to compute hazard ratios (HRs), which were adjusted for known melanoma risk factors and potential confounders.

Table 6.2 shows that people consuming ≥393 mg of caffeine per day as compared to <60 mg/day were less likely to report incident melanomas (HR = 0.71). However, this inverse association was more pronounced in women (HR = 0.66) than in men (HR = 0.82). The difference between men and women could be due to differing sun exposure patterns because of different dress styles, but it might also be due to effect modification by gender. This was also suggested by the finding that women reported more melanomas on extremities (53.5% of all melanomas) compared to men (29% of all melanomas). The generalizability of the results may be limited, since both cohorts comprise mostly white, well-educated health professionals. Skin colour is a major determinant of melanoma risk, making these results difficult to apply to people of different ethnicities. In terms of the Hill criteria for causality this example is based on a strong study design, argues that a protective effect of caffeine is plausible and provides some evidence of a dose-response relationship. Furthermore, the temporal sequence is probably correct, unless pre-clinical melanomas were already present at the baseline measurement.

6.3 Applications of epidemiological causality research

6.3.1 Forensic epidemiology

Questions of causality in epidemiological research are increasingly being used in court cases, recently resulting in a new spin-off from the discipline: forensic epidemiology. This subdiscipline uses epidemiological and statistical methods to assess and quantify causal relationships in court cases with a medical connection. It is applied in both criminal and civil law, although the majority of cases involve civil law (in particular liability law), for example with respect to medical negligence, occupational safety negligence, side effects of medical drugs, faults in medical devices and serious road accidents. In liability cases the claimant has to prove that the medical condition was caused by the determinant in question (e.g. a medical drug, medical device, food product or working condition) and not by something else. Conversely, the defence will do everything in its power to prove that the research leaves a lot of doubt (potential bias) as to the alleged causality of the relationship. The result of forensic epidemiological research is a well-founded quantitative estimate of the likelihood that the relationship in question is based on a causal effect.

As a rule, forensic epidemiologists act in a sequence of four steps:

1. Assessing all the relevant elements in the case file, including conflicting theories concerning causality, liability, guilt and innocence.
2. Carrying out a quantitative epidemiological analysis of the theories put forward by both parties. Are the medical estimations of likelihood correct and properly substantiated? Can they be quantified and compared?
3. Carrying out an additional systematic examination of the medical literature and/or additional data.
4. Making a quantitative, substantiated estimate of the likelihood that the association is a causal relationship, based on all the available evidence.

6.3.2 Prevention

Etiological research primarily produces scientific knowledge of the causes of diseases and determinants that influence their course. This knowledge is used to devise preventive and therapeutic measures. After all, if one of the determinants can be eliminated from the complex of sufficient causes, the disease will no longer be able to develop through that mechanism. It is not necessarily true, however, that removing a cause of a disease will eliminate that disease in patients who have it: sometimes this will be the case (e.g. antibiotics kill the bacterium responsible for the persistence of the infection), but often it will not (e.g. noise-induced hearing loss will not be cured by starting to wear hearing protectors).

Knowledge of the causes of a disease does not automatically lead to effective preventive measures; conversely, full knowledge of the causes is not always needed for prevention to be effective. Take safe sex and AIDS prevention, for example: this measure was advocated before the cause of the disease (the HIV virus) was known. Generally speaking, prevention and knowledge of the etiological factors go hand in hand, however. Examples of preventive interventions are changing the physical environment (sewers, crash barriers), laying down statutory rules (food safety, speed restrictions), prescribing prophylactic medication (folic acid, fluoride), carrying out preventive surgery (circumcision, angioplasty), eradicating pathogenic microorganisms (vaccination) and health education.

There are three levels of prevention: primary, secondary and tertiary. The purpose of primary prevention is to prevent the condition developing. The focus here is on etiological factors. Quitting smoking – or better still, not starting – so as to prevent such things as asthma, coronary heart disease and lung cancer is an example. The aim of secondary prevention is to detect a condition early once it has developed, so as to increase the likelihood of cure. An example is advocating breast self-examination so as to detect breast cancer at an early stage. Screening programmes carried out by professionals and case-finding by GPs and specialists are also part and parcel of secondary prevention. Tertiary prevention is carried out on patients who have the condition with the aim of curing it, minimizing its effects and reducing the likelihood of relapse. Health education provided to diabetes patients is an example. In essence, all forms of treatment could be regarded as tertiary prevention.

Recommended reading

Bonita R., Beaglehole R., Kjellstrom T. Basic epidemiology. 2nd ed. Geneva: World Health Organization, 2006.
Freeman M., Zeegers M. Forensic epidemiology: principles and practice. San Diego: Elsevier, 2016.
Grobbee D.E., Hoes A.W. Clinical epidemiology: principles, methods, and applications for clinical research. 2nd ed. Burlington: Jones and Bartlett Learning, 2015.
Hill A.B. The environment and disease: association or causation? Proc R Soc Med. 1965, 58, pp. 295–300.
Morabia A. A history of epidemiologic methods and concepts. Basel: Birkhäuser Verlag, 2004.
Rothman K.J., Greenland S., Lash T.L. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins, 2012.
Szklo M., Nieto F.J. Epidemiology: beyond the basics. 3rd ed. Burlington: Jones and Bartlett Learning, 2014.
Webb P., Bain C. Essential epidemiology: an introduction for students and health professionals. 2nd ed. Cambridge: Cambridge University Press, 2011.

Source references (case and figure)

Weischer M., Nordestgaard B.G., Pharoah P., Bolla M.K., Nevanlinna H., Veer L.J. van 't, et al. CHEK2*1100delC heterozygosity in women with breast cancer associated with early death, breast cancer-specific death, and increased risk of a second breast cancer. J Clin Oncol. 2012, 30(35), pp. 4308–16 (Case 6.1).

Recommended reading

Fletcher R.H., Fletcher S.W., Fletcher G.S. Clinical epidemiology: the essentials. 5th ed. Baltimore: Lippincott, Williams & Wilkins, 2012 (Case 6.2).

Wu S., Han J., Song F., Cho E., Gao X., Hunter D., Qureshi A.A. Caffeine intake, coffee consumption, and risk of cutaneous malignant melanoma. Epidemiology 2015 Nov;26(6):898–908. (Case 6.3).

► http://jco.ascopubs.org/content/30/35/4308.long (fig. 6.2).

Genetic epidemiology

The co-authors of this chapter are dr. ir. Sita Vemeulen en prof. dr. Bart Kiemeney (Department for Health Evidence, Radboudumc, Nijmegen).

7.1 Introduction: looking for variation in the human genome as a determinant of disease requires a different approach

When looking for determinants of a disease we divide them, broadly speaking, into three categories: biology, behaviour and environment (see ▶ chap. 3). Genetic epidemiology is specifically concerned with the first category, and with the complex interactions between genetic factors and environmental or behavioural factors. It is thus closely related to molecular epidemiology, which is concerned with the role of biological or molecular markers of exposure and susceptibility, including DNA. Once the human genome had been mapped, interest in genetic determinants of disease – and hence in genetic epidemiological research – increased enormously. This is not the reason for devoting a separate chapter to this subspecialty in a general textbook of epidemiology, however: after all, this book does not include chapters devoted to the epidemiology of cardiovascular disease, psychiatric epidemiology, nutritional epidemiology or pharmacoepidemiology, to name but a few of many subspecialities. The reason for having a separate chapter on genetic epidemiology lies in the existence of specific biological mechanisms that apply to the transfer of genetic traits from one generation to the next. These make it possible to tackle questions regarding the genetic determinants of disease differently and more efficiently than other types of determinants. Consequently, genetic epidemiology uses some substantially different study designs and analytical methods in addition to the general epidemiological methods.

7.1.1 Genetic variants influence the development of disease in a variety of ways

Genetic epidemiology is the study of how hereditary variations in the human genome influence the frequency of diseases in the population, either in interplay with behavioural and environmental factors or on their own.[1] Hereditary information is stored in chromosomes, which are composed of DNA. Except for men having only one copy of the X and Y chromosomes, everyone has two copies (alleles) of each chromosome – and therefore of each gene: one from the father and one from the mother. Alleles can occur in several variants, which is why people differ in eye colour, hair colour and many other physical characteristics. The combination of paternal and maternal alleles is known as the genotype. Any two individuals' genomes will be 99.9% identical, but there are millions of places in the remainder of the DNA where people differ from one another. If a variant allele has a frequency of less than 1% in the population it is referred to as a mutation. If the frequency of an allele is greater than 1% it is referred to as a polymorphism (see ◘ fig. 7.1).

Mutations or polymorphisms can be associated with the likelihood of contracting certain disorders. The classic type is the monogenic disorder, where a structural defect in a chromosome or a mutation in a gene is a necessary cause for developing the disease. Familiar examples of this are Down's syndrome (trisomy 21) and Huntington's disease (caused by an abnormal gene on chromosome 4). Without the genetic defect the disease cannot develop. These monogenic defects may be caused by a chance mutation with no prior history in the family, as in the case of Down's syndrome (a 'de novo' mutation). In other cases a disease develops due to heredity, as is almost always the case with Huntington's disease. Where there are strong associations between a genotype and a phenotype (a measurable trait or disease) there is said to be high penetrance. Penetrance is a term used by geneticists that is identical to the term etiologic fraction among the exposed (see ▶ par. 3.2.5). If highly penetrant variants are passed on from one generation to the next, we will see an increased incidence of the disorder in families. These variants, which increase the likelihood of the disease and are associated with short survival and/or low fertility (referred to in the field of population genetics as low fitness), will obviously be far less common than variants that have only a small influence on the likelihood of the disease (low-penetrance variants).

Far more common than monogenic disorders are those disorders where multiple genetic traits of

0 The structure and function of the human genome (animation). ▶ http://bit.ly/1Jh2nGO

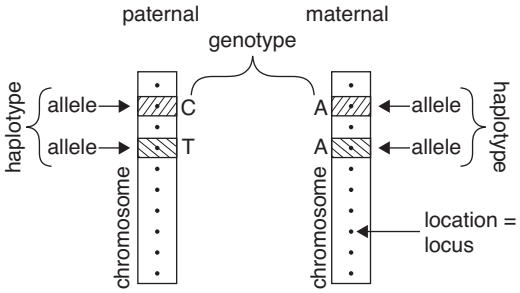


Figure 7.1 Genetic terminology

an individual combined with environmental and behavioural factors determine the likelihood of developing a particular disease. These are referred to as multifactorial disorders. The number of genetic determinants and their penetrance and frequency can vary substantially from one disease to another. Until recently it was thought that only a few underlying genetic determinants were involved in these multifactorial disorders, namely fairly common determinants (with an allele frequency of over 5%) with only mild effects (the 'common disease-common variant' hypothesis). In recent years, however, it has become clear that it is more likely that rare, relatively low-penetrance genetic determinants are be important in the development of multifactorial diseases (the common disease-rare variant hypothesis). In some cases different genetic profiles can cause a phenotypically identical disease (genetic heterogeneity), or the disease can be due to effect modification between genes and environmental factors. As these diseases, because of their complex etiology, are far less common in patients' families, they are referred to as complex disorders, and it is mainly this type of disorders that genetic epidemiologists study nowadays. Hence the greater emphasis in this chapter on the genetic epidemiology of multifactorial disorders than on that of monogenic diseases.

7.1.2 The approach depends on the research question

This chapter looks at various methods and techniques used in genetic epidemiological research. Genetic epidemiology is more than just a collection of tools, however; the methods and techniques are used to examine specific research questions. The study designs that are suitable can be divided, broadly speaking, into family-based studies and population-based studies. The study design chosen will depend on the research question. The type of disorder that the researcher is interested in will also affect the choice of design.

The questions that can be posed in genetic epidemiology follow a certain natural order, gradually zooming in on the genetic determinant.

The first question is whether genetic variation is of relevance to the particular disorder. If there is familial aggregation, i.e. the disease is more common among family members than would be expected based on the frequency of the disease in the general population, this points to a role for factors shared by family members – environmental, behavioural or genetic. The use of case-control studies, family-based cohort studies and twin studies to identify this aggregation is discussed in ▶ par. 7.2. If we are convinced that genetic variation is of relevance to the disease, the next question is *which* genetic determinants are relevant to the disease. If we have no idea where the genetic determinant or determinants is/are located, or if we do not wish to make assumptions about the location, we carry out a genome-wide linkage analysis, in which the segregation of large numbers of markers (small pieces of variable DNA at known locations) in the genome in families is observed and compared with the segregation of the disease in those families. This is an efficient way of tracking down monogenic disorders that are strongly aggregated in families. These linkage analyses are often supplemented with or replaced by techniques that identify *all* the variants in the genome or exome (the coding part of the genome) in a family. In practice these analyses tend to fall in the domain of genetics rather than that of epidemiology, and they are therefore described only briefly (in ▶ par. 7.3). Population-based studies are generally used to localize low-penetrance variations for multifactorial diseases. The classic case-control study is often used in the design of genome-wide population or association studies: this is described in detail in ▶ par. 7.4. If the approximate location is known, we can zoom in using more markers that are increasingly close together. Specific bioinformatics meth-

ods can be used to track down the possible genes involved, but these last methods fall outside the domain of epidemiology. Once these questions have been answered, further bioinformatics analysis is carried out into the possible biological pathways by which the variant influences the development of the disease, the underlying inheritance model, the prevalence of the variant in highly predisposed families, and in sporadic cases (where the patient has no affected family members). That information is used to decide whether it would be worthwhile to screen families and/or populations for the particular variant.

It should be noted that the scope for genetic epidemiological research has been and still is highly dependent on advances in technology for measuring genetic variation in the human genome.[2] A number of large-scale international projects have been important in increasing our knowledge of variation in the human genome (see ▶ case 7.1).

Case 7.1 Technological advances have increased the scope for genetic epidemiological research in recent decades

Genetic epidemiology is a relatively young subdiscipline that started in the 1960s. In the early years the focus was on estimating the relative contributions of genetic components to familial aggregation, which did not require measurement of the genetic variants themselves. Once sets of hundreds of markers became available in the early 1980s this paved the way for genome-wide linkage analyses for monogenic disorders. Not until some ten years later, when it became easier to characterize these variants on a large scale, did the focus shift to genetic association studies of multifactorial diseases. Until the early years of the twenty-first century this research focused mainly on the association between a particular disease and a small number of genetic variants in 'candidate genes' that were already suspected to be involved in the development of that disease. Since 2005 hundreds of thousands or even millions of genetic variants in large numbers of samples can be measured quickly and accurately using DNA microarrays (also known as 'chips'). This enables exploration of the entire genome for locations associated with multifactorial diseases. Thanks to the recent major advances in sequencing technology, enabling the complete sequence of DNA building blocks in a genome to be efficiently mapped, the scope for genetic epidemiological research is likely to change again. The knowledge of existing variation in the human genome that has become available from the Human Genome Project (HGP),[3] the International HapMap Project and the 1000 Genomes Project[4] has also been important. The HGP, which was completed in 2003, mapped the entire human genome for the first time down to the level of the smallest building block, the base pair. The HapMap Project measured millions of single nucleotide variants (SNVs) in populations with different geographical origins so as to describe population-level variation, focusing particularly on similarities and differences in the occurrence of combinations of multiple alleles on a chromosome, known as haplotypes. The 1000 Genomes Project can be regarded as building upon the HapMap Project to provide an even more comprehensive picture of the existing variation – specifically rare variation – in a wide variety of populations by sequencing the genomes of 2,500 individuals.

7.2 Family-based studies: estimating the contribution of genetic variation

Familial aggregation of a disorder can be indicative of genetic variation playing a major role as a determinant of that disorder. The existence of such clustering can be ascertained by assessing the risk asso-

2 Genetic technologies (animations). ▶ http://bit.ly/1GpUZrZ

3 The Human Genome Project (animation). ▶ http://bit.ly/1lc7D2f

4 The 1000 Genomes Project (website). ▶ http://bit.ly/1JB8GSv

ciated with familial predisposition. Classic case-control studies (see ▶ chap. 4) identify the familial predisposition for the patients and controls, usually summarized as a dichotomous variable (negative/positive family history). This variable is included in the epidemiological function as a determinant. The odds ratio (see ▶ chap. 3) then gives an approximation of the RR associated with a positive family history. This approach, however, can give a distorted picture of the degree of familial aggregation of a disease, as the likelihood of it occurring in the family will depend on the size of the family, the ages and gender distribution of the members, their genetic relationship with the participants and each family member's risk pattern due to other determinants. Familial aggregation of diseases is therefore ideally studied in family-based cohorts of patients and controls. The incidence of the disease in the two cohorts is estimated. The ratio between these incidences – the RR – indicates the additional risk run by patients' family members compared with the controls' family members (the basic risk). The RR is also referred to as the familial relative risk. As the occurrence of the disease among family members may of course also depend on age (there is more competition from other causes of death in later life), suitable measures of frequency and analytical strategies need to be used (incidence densities, survival rates: see ▶ chap. 2), just as in classic cohort studies. The central determinant in these family-based cohort studies is therefore the familial relationship with the patient or the control. The analysis can be stratified by degree of kinship if necessary: this enables the risk of the disease among a patient's first-degree relatives to be compared with that of a control's first-degree relative. Information on potential confounders or effect modifiers can also be collected for each family member and included in the analysis: regression models can then be used to estimate incidence densities or hazard ratios, adjusted for relevant confounders, enabling an initial impression of the relative contribution of the genetic component to the causation of the disorder. These analyses need to take the statistical dependence of family members into account, however. ▶ Case 7.2 gives an example of a family-based cohort study.

Case 7.2 Clustering of breast cancer in families

Various case-control studies and family-based cohort studies in the closing decades of the twentieth century showed that a positive family history is a risk factor for breast cancer and that the risk is greater the earlier the age at which the family member was diagnosed with it. One of these was a family-based cohort study from 1990, which included 4,730 breast cancer patients and 4,688 controls in the 20–54 age range. The patients had been recorded in the cancer registries of the National Cancer Institute in the United States between 1 December 1980 and 31 December 1982. Information on the incidence of breast cancer in female family members (first-degree relatives and paternal and maternal grandmothers and aunts) was obtained from interviews with the patients and controls within six months of the first primary breast cancer diagnosis in the patient. To study the family clustering the incidence of breast cancer among the patients' family members was compared with that among the controls' family members. The study showed that in all the age groups the risk of breast cancer for patients' mothers was approximately twice as high as that for controls' mothers. It also showed that the risk was highest among people whose family members had been diagnosed at an early age. The familial relative risk (in this case, the cumulative incidence ratio (CIR)) of breast cancer in mothers of patients who had been diagnosed before the age of 30, 40 or 50 was 4.3, 2.7 and 1.7 respectively.

To study the contribution of genetic variation to continuous traits such as blood pressure or cholesterol level, the correlations between family members with various types of kinship are often compared. If there is a strong correlation among relatives who share a lot of their genetic material (e.g. monozygotic twin pairs) and a weak correlation in pairs who share little genetic variation (e.g. second cousins), that is indicative of a strong genetic component in the trait concerned. Most traits are affected by a

combination of genetic and environmental and/or behavioural factors. To study the relative contributions of these categories of determinants based on correlations between family members with different degrees of kinship we ideally need families that vary in shared genes and shared behavioural and environmental factors. The classic Mendelian laws of inheritance are taken as a reference. A parent-child pair share half of their genetic material, a grandparent-child pair a quarter. Brothers and sisters in a family display far more similarity in behavioural and environmental factors in childhood than second cousins. Research linking this data on kinship and shared behavioural and environmental factors to measurements of continuous traits provides information on the relative contributions of heredity (nature) and behaviour/environment (nurture). Good data on kinship and shared environment will increase the validity of this research. Studies of twins provide an efficient way of studying the genetic component in continuous traits, for one thing because they enable monozygotic and dizygotic twin pairs to be compared. ▶ Case 7.3 gives an example.

Case 7.3 All human traits are heritable

Scientists have always been interested in the causes of individual differences in human traits. This has led to a lively 'nature versus nurture' debate going back to the 18th century, but only in the past few decades have adequate methods to study the etiology of human trait variation become available. The 'workhorses' of this type of research have been twins, as they enable high-quality natural experiments to disentangle the effects of nature (heritability), and nurture (environmental influences). Monozygotic (MZ) twins are genetically identical, while dizygotic (DZ) twins share 50% of their genes on average. Usually, the environment that twins have in common (i.e., 'the shared environment', such as the home environment) is 100% for MZ as well as DZ twins, at least during childhood. Comparison of the within-pair resemblance for a given trait of MZ twins versus DZ twins – usually presented as MZ and DZ twin correlations – indicates to what extent genes or the environment contribute to trait variation. Higher within-pair MZ correlations in comparison to DZ correlations suggest the presence of genetic influences, as the only difference between MZ and DZ twin pairs is the within-pair genetic similarity.

A large meta-analysis of all twin correlations and heritability estimates of the past 50 years was published in 2015. This was based on 2,748 twin studies that reported on 17,804 traits of a total of 14.5 million twin pairs. One of the main findings was that all the traits investigated were heritable: not one trait had a heritability estimate of zero. The results of the meta-analysis, across all traits and regardless of age and gender, showed that the contribution of genetic and environmental influences was equal. Thus genetic and environmental influences each contribute about 50% to overall human variation. When zooming in on different traits it became clear that heritability estimates differ across traits. High heritability estimates were reported for 'diseases of the skin' (69%), while one of the lowest heritability estimates was noted for voice and speech functions (15%). Heritability estimates may also be dynamic over time. For instance, intelligence is highly heritable in adulthood, while in childhood environmental influences are equally important. Heritability estimates for same sex pares (SS) and males (M) and females (F) were very similar across traits, indicating no gender differences in the contribution of genetic and environmental effects to trait variation. All the results [by trait, gender, or age group] of this study can be visualized using the MATCH web tool. ◘ Figure 7.2 shows an example of results obtained from the website.[5]

5 ▶ http://match.ctglab.nl

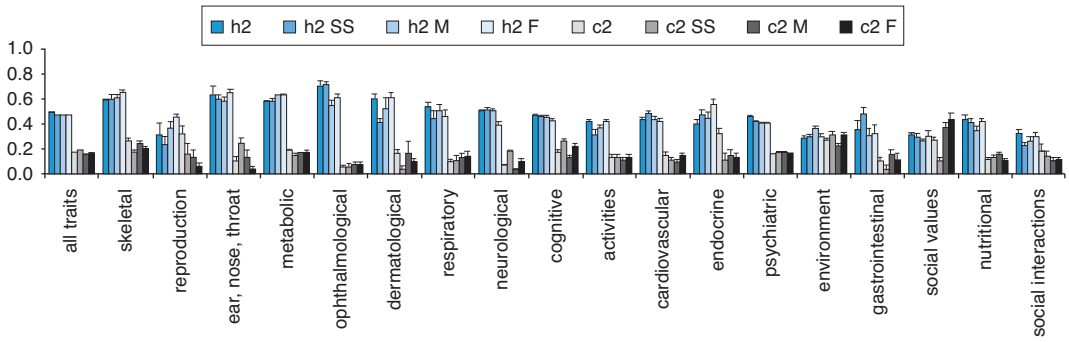


Figure 7.2 shows the heritability estimates (h2) for 'all traits' and other investigated trait domains. The h2 estimates are shown in shades of blue and shared environmental influences in shades of grey. The remainder of the variation is due to the contribution of unique environmental influences.

7.3 Linkage analysis: finding highly penetrant mutations in highly selected families

As already noted, finding highly penetrant mutations for monogenic disorders is primarily the province of genetics. Genome-wide linkage analyses of selected families with several affected members have in the past resulted in various highly penetrant disease genes being localized. The principle of linkage analysis is simple. Many markers with known locations in the genome are measured (the location of the causal mutation is not known, of course). Next, the co-segregation of the alleles of the markers and the disease is assessed. If the marker is very close to the disease gene, recombinants (rearrangements of the alleles on a chromosome) will hardly ever be found for the marker alleles (whose location is known) and the disease (the location of whose gene is not known). This is referred to as linkage between the marker and the disease. If the marker is not inherited together with the disease, it is evidently far enough apart from the disease variant for recombination to have broken the link between the alleles of the two variants. The concept of recombination is illustrated by an example in ▶ par. 7.4.3. The strength of the linkage between the marker and the disease variant can be quantified using the log-odds score (LOD score). As family members share long segments of their genome, only a few thousand markers need to be measured for a genome-wide linkage analysis.

Figure 7.3 shows a family for which the genotype of a single marker with two alleles (1 and 2) has been determined for a grandparent pair, the parents and four children. The squares represent males, the circles females. A blackened symbol for a family member means that the person has the disease. Where this is the case it is assumed that the person carries the D disease allele (in other words, has the genotype Dd), so there is full penetrance and dominant inheritance.

The sick mother is found to have received a 1-allele from the healthy grandmother and therefore both the disease and a 2-allele from the grandfather. This means that the mutated D allele of the disease gene has to be inherited along with the 2-allele of the marker in this family. The father has both two 1 alleles and two d-alleles and does therefore not provide any information on recombination during meiosis; the children have all received a chromosome with a 1-allele and a d-allele from the father. The leftmost child left has the disease and has received a 2-allele from the mother, which means that there was no recombination between D and 2 during this child's maternal meiosis. The second child from the left is healthy and has received a 1-allele from the mother, so this is also a non-recombinant. Only the rightmost child deviates from this pattern: she has received the 1-allele from the mother but neverthe-

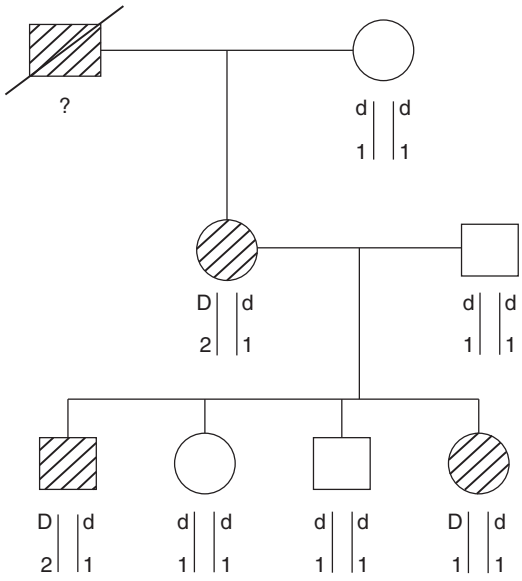


Figure 7.3 Family tree of a family with the disease gene and the marker on the same chromosome

less has the disease, which means that there must have occured a recombination during meiosis. The likelihood of linkage in this family (with r = 1 recombinant from n = 4 informative meioses), assuming for example that the recombination fraction (θ) between the disease gene and the marker is 0.05, can be calculated as follows:

$$L(\theta) = (\theta/2)^{r} \times [(1-\theta)/2]^{n-r}$$
$$= (0.05/2)^{1} \times [(1-0.05)/2]^{4-1} = 0.0027$$

The likelihood of the alternative hypothesis of no linkage is:

$$L(\theta = 0.5) = (0.50/2)^{4} = 0.0039$$

The LOD score is the ^{10}log of the likelihood ratio, in this case:

$$LOD = \log[L(\theta = 0.05)/L(\theta = 0.50)] = 0.69$$

Of course we do not know whether the recombination fraction is 0.05, so we calculate the LOD score for several recombination fractions between 0.005 and 0.25 (higher fractions would provide little evidence that there is a marker near the gene). The recombination fraction for the highest LOD score indicates the distance between the marker and the gene. If more families have been described, each with a particular LOD score for a specific recombination fraction, the LOD scores for different families can be added up. Fewer recombinations will lead to higher LOD scores for low recombination fractions, and hence to more evidence that the marker is linked to the disease. An LOD score higher than zero suggests that the marker is linked to the unknown disease gene, but because of random variation we can only assume linkage if the LOD score is high (say 3 or higher).

It is important to realize that it is only the relationship between the location of the marker and that of the unknown disease gene that is relevant in linkage analyses; an allele variant of the marker does not relate to any pathophysiological function (they do not cause the disease). So it can be the case that in one family allele 2 of a particular marker is co-inherited with the disease and in another family allele 1. The marker is merely a coordinate on the genomic map, as it were; whether that coordinate has been marked on the map in red, blue or black is not important in linkage analysis. This is very different from genetic association studies, which are discussed later on: they examine the potential relationship between a disease and a specific allele. It is also important to realize that this technique provides only a rough indication of the location of the gene concerned: the region that displays linkage is likely to contain dozens of genes. Further research will be needed to zoom in on the region (fine mapping) and ultimately lead to the identification of the disease gene. ▶ Case 7.4 describes a classic study of the localization of the *BRCA1* gene using linkage analyses before it was possible to clone this gene. Nowadays, when searching for highly penetrant mutations for monogenetic disorders, modern sequencing techniques that map all the variants in the human genome are often used. Here again the search is for the one causal, highly penetrant mutation that explains the presence of the disease in the family. One of the advantages of this technique is that it enables the causal variant to be identified directly.

Case 7.4 Localization of the highly penetrant breast cancer gene

Family-based studies have shown that a genetic component is important in the development of breast cancer. Segregation analyses carried out by the group of the American researcher Marie-Claire King pointed to a rare, highly penetrant variant in combination with an autosomal dominant inheritance model as the best explanation for the spread of breast cancer in a large selection of high-risk families and families of young breast cancer patients. The data were used for a linkage analysis. Blood was collected from 329 people in 23 families with a large number of breast cancers (a total of 146 patients) from North America, Puerto Rico, the United Kingdom and Colombia. The genotypes of 183 markers in each person were identified. It was found that a marker called D17S74, on chromosome 17q21, was inherited along with the disease in the families, especially in those where the patients were diagnosed at a very early age. The linkage with the marker was not found in the families of old patients. The LOD score was just under 6, which is a very strong indication of a gene with the causal mutation (or mutations) near the marker. In the race to find the gene, King's group was beaten by Mark Skolnick's, which first identified it in 1994. This gene, now known as *BRCA1*, turned out to be a tumour suppressor gene that is important in repairing DNA damage. A host of different mutations were found in it, and women with such mutations were at high risk not only of breast cancer but also of cancer of the ovaries, where the gene is also expressed. Nowadays women with a positive family history can be tested for mutations in the gene at clinical genetics centres. Women found to have a relevant *BRCA1* mutation are monitored very closely for breast cancer and ovarian cancer. Some women who test positive have their breasts amputated as a preventive measure. They can also have their ovaries removed.

7.4 Association studies: finding genetic variants for multifactorial disorders

The linkage studies described above are a very powerful method for localizing unknown genetic variants on the genome for rare disorders that are highly clustered in families. Because of the complexity of the underlying genetic etiology, however, with particular variants generally having small effects and a lack of high-risk families, that method is inefficient for multifactorial disorders. Genetic determinants of disorders of this type are generally identified using population or association studies, usually 'classic' case-control studies in which DNA is collected from patients and controls and the prevalence of variant alleles in the two groups is compared. This design of genetic association studies is discussed in ▶ par. 7.4.1. In essence, the starting point for an association study is that patients with the disease are descendants of one and the same distant ancestor, who was the first to receive the allele in question through a mutation (see ▶ par. 7.4.3). Among the controls who do not have the disease we would expect fewer people to be descendants of that ancestor.

These association studies require genetic variants to be measured. Unlike those in a linkage study, however, these variants can be functional (i.e. affect the functionality of a protein) and be located in genes already suspected to influence the risk of the disease. This is referred to as a direct association study and a candidate gene approach. In an indirect association study the genetic variant merely serves as a coordinate in the genome, and if an association is found it is indicative of linkage disequilibrium between the measured variant and the non-measured causal variant. This is explained in more detail in ▶ par. 7.4.2. An association between a genetic variant and a disease found in a case-control study can hence be indicative of causality or linkage disequilibrium with the causal variant, but also of two other phenomena, namely population stratification or chance (in small-scale studies). The phenomenon of population stratification is explained in ▶ par. 7.4.4.

7.4.1 Case-control designs are well-suited for genetic association studies

Although any type of observational study can be used for genetic epidemiological research, case-control studies are particularly suitable, as they enable a whole series of genetic variants to be studied at the same time, either on their own or in interaction with environmental and behavioural factors. Moreover, there is no problem with differential misclassification (information bias), since the genetic information stored in the DNA is stable. Also, the risk of selection bias in genetic research is very low, as the likelihood of selection as a control or a patient in the study is unlikely to depend on the presence of an underlying – as yet unobserved – genetic variant. Lastly, disorders whose genetic determinants we wish to study will generally have low prevalence, which makes case-control studies far more efficient than e.g. cohort studies (see ▶ chap. 4). ▶ Case 7.5 gives an example of a case-control study of a variant in a biologically plausible gene (candidate gene). Although case-control studies are very efficient when searching for genetic determinants of a disease, cohort studies are also recommended, especially when studying gene-environment interactions, as a cohort study is in principle better able to characterize the exposure to the environmental factor, independent of the presence or absence of the disease.

Case 7.5 APOE and the risk of diabetic nephropathy (1)

Apolipoprotein E (APOE) plays a major role in the metabolism of blood fats. There are three different functional alleles of the APOE gene that commonly occur: ε2, ε3 and ε4. As various studies have suggested associations between the APOE ε2 allele and the occurrence of hyperlipoproteinemia and lipoprotein glomerulopathy, new research has been initiated into the relationship between this APOE variant and the risk of renal failure in diabetic patients, known as 'diabetic nephropathy' (DN). One of these studies was a case-control study of 223 DN patients and 192 controls. The DNA analysis of the APOE alleles showed that carriers of the ε2 allele had a highly increased risk of developing DN, compared with individuals who did not have the ε2 allele (OR = 3.1, 95% CI 1.6–5.9).

7.4.2 Linkage disequilibrium can help to identify genetic determinants

The rationale behind an association study is that the patients are descendants of one and the same distant ancestor, who developed the disease allele at one time or another, and that the controls are less closely related to that ancestor. Variants found in the current population that are located near the disease variant will display linkage disequilibrium (LD) with the disease locus as a result of lack of recombinations. The specific alleles of these nearby variants, like the disease alleles themselves, will then be more common among patients. Linkage disequilibrium can hence be used to track down a disease variant without actually measuring it. In order to use this LD approach we need to know what LD patterns are present in the human genome in populations with different genetic backgrounds. Generating this knowledge was one of the aims of the 1000 Genomes Project. Based on these data (which are freely available), we can select a set of SNVs that represent all variation for a gene or region that we are interested in. By looking for indirect associations with these 'tagging markers' in a case-control study we can track down that part of the genome where the actual causal variant is most likely to be found. ▶ Case 7.6 gives an example of an association study using tag SNPs. Single Nucleotide Polymorphisms (SNPs) are SNVs whose rare allele occurs in at least 1% of the population.

7.4.3 Crossing-over causes recombination of genetic material

For a proper understanding of the term 'Linkage Disequilibrium' (LD) we need to be familiar with the concept of recombination. During meiosis – the

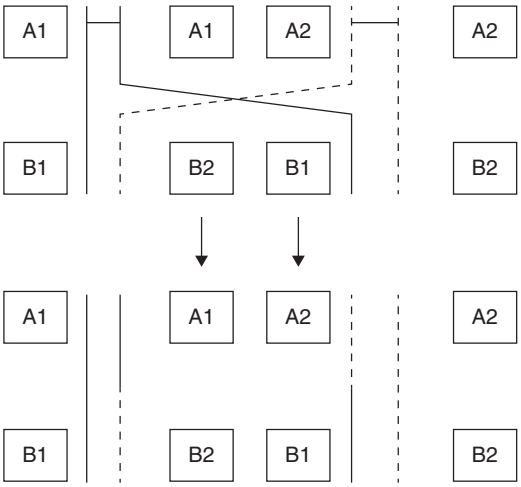


Figure 7.4 Schematic representation of the development of recombinants (middle two chromosomes) and non-recombinants (outer two chromosomes) during meiosis

cell division process in which reproductive cells are generated – the two homologous chromosomes come together and recombine as a result of 'crossing-over' between two chromatids. If two markers are far apart there is a good chance that crossing-over will take place between them, resulting in a recombination. If the number of crossovers between two markers is zero (or two; on average slightly more than one crossover per chromosome takes place during meiosis) there will be a non-recombinant. The likelihood of recombination of two markers that are far apart is therefore 50%. It follows that a child has a 25% likelihood of receiving each of the four possible combinations of the two alleles of the two markers (see fig. 7.4: A1 B1, A2 B2, A1 B2, A2 B1, where A1 and A2 and B1 and B2 refer to the two alleles of markers A and B). If two markers are close together on the chromosome (linked), however, crossing-over between them is very unlikely. In other words, the marker alleles are more likely to be passed on together from the parent to the child (non-recombinant).

In genetics we assume that each allele of a gene developed thousands of generations ago in one of the ancestors. We can then assume that there has been sufficient opportunity for crossing-overs between the particular place on the genome and nearby places, and recombinations have therefore occurred (this is unlikely in the case of a family that goes back only a few generations). The prevalence of the combination of alleles at two locations on a chromosome (haplotype) in a population is therefore in principle equal to the product of the separate allele frequencies. The population is then in equilibrium for these loci and the alleles of the two loci are not associated: this is referred to as linkage equilibrium (LE). If an allele developed more recently, however (hundreds rather than thousands of generations ago), or if the places on the genome are very close together, it may be that little or no recombination has taken place for this marker and nearby locations in the population (see fig. 7.5). Combinations of particular alleles of these markers and/or genes will then be more common, or less common, than we would expect based on the separate allele frequencies. This dependency between nearby alleles is referred to as linkage disequilibrium (LD), allele association or correlation between alleles. The strength of LD is often expressed as the measure r^2, which can range from 0 to 1, where 0 means complete LE and 1 complete LD (see fig. 7.6).

Knowledge of the existence of LD between variants in the human genome is needed, for example, for the efficient design of genome-wide DNA microarrays: precisely those markers can be selected that are not strongly correlated themselves but for which there is strong LD with multiple unmeasured genetic variants in the genome. Thus by measuring the genotype for a limited number of markers we can indirectly determine the genotype for a far larger number of existing variants in the human genome.

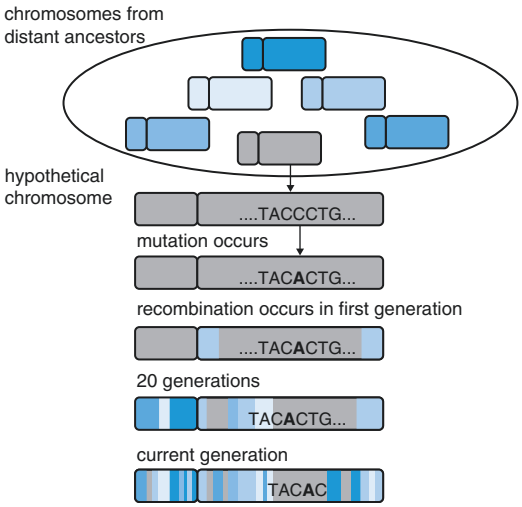


Figure 7.5 Situation where six haplotypes for a genomic region occur in a population. A new allele is introduced into the grey haplotype. After a few generations (and therefore meioses) crossing-over causes haplotypes comprising a mix of the six original haplotypes (recombinants) to develop. Because of lack of recombination between nearby variants, however, after 50 generations the new variant allele A still occurs only on a grey haplotype background (haplotype TACAC)

Case 7.6 Variants in TCF7L2 and risk of type 2 diabetes

In 2003 Icelandic researchers showed in a linkage analysis that the chromosomal region 10q is linked to type 2 diabetes (DT2). This region was studied more closely in 2006 in a genetic association study among the Icelandic population (and in two other case-control studies). Associations were found between DT2 and markers in the gene 'transcription factor 7-like 2' (*TCF7L2*), which is located in the 10q region. These markers were DG10S478 and five SNVs that displayed moderate to strong LD with DG10S478. *TCF7L2* codes for a transcription factor that plays a role in the 'Wnt signalling pathway', an important regulatory mechanism in cell development and growth. Other researchers subsequently carried out a replication study on *TCF7L2* in a Finnish population. Of the five SNVs that emerged from the Icelandic study, rs12255372 and rs7903146 displayed the strongest association. Twelve additional SNVs were selected in *TCF7L2*, as these were able to indirectly measure (or 'tag') all of the 63 SNVs that showed LD with rs12255372 in the HapMap data, the SNV that displayed the strongest association with DT2 in the Icelandic study. The twelve additional SNVs displayed a weaker association with DT2 than rs12255372, however. The researchers therefore concluded that rs12255372 and rs7903146 in *TCF7L2* display the strongest – indirect – association with DT2. They could not, however, rule out the possibility that other variants in *TCF7L2* influence the risk of DT2, as the twelve selected SNVs were only able to tag some of the variants located in *TCF7L2*. In a subsequent study of the association signal of SNVs in TCF7L2 among other populations the Icelandic researchers found the strongest signal for rs7903146. To ascertain whether this signal was actually a direct or an indirect one, Palmer et al. sequenced the gene region in DT2 patients and controls of African ethnicity in 2011. In this control population with a relatively long population genetic history the LD between variants is relatively weak, promoting the discovery of the causal variant. The results of this study again pointed to variation in rs7903146 as the explanation for the association between gene *TCG7L2* and DT2.

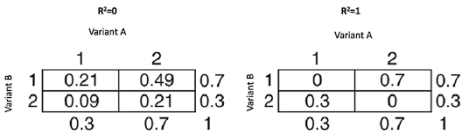
The introduction in the early twenty-first century of microarrays that enable large numbers of SNVs in large numbers of samples to be measured at relatively low cost paved the way for genome-wide association studies (GWASs), which search the entire genome for disease variants without a prior hypothesis. Based among other things on information from the HapMap Project, various companies have developed arrays containing a selection of SNVs that can provide information on existing genetic variation in the human genome via LD. It is possible, based on the measured variants and knowledge of patterns of correlation between variants in the human genome, to make a very good estimate of the genotypes for unmeasured variants, or to impute unmeasured var-

iants. Use of these arrays in large-scale case-control studies has led to a rapid increase in the discovery of new associated regions and genes for multifactorial disorders.[6]

Case 7.7 gives an example of one of the first GWASs. The principles underlying GWASs or LD studies and linkage studies (see ▶ par. 7.3) are related. Whereas in a linkage study we follow the segregation of the disease and the marker within a family, in an LD study we only look at the association between the marker and the disease in the current generation (of a virtual enormous family). For these GWASs we therefore need far more markers, as the region over which LD extends in populations is far smaller than the region that displays linkage in families, because of the many meioses that have taken place before the current population. Whereas a few thousand markers will suffice for a linkage study within families, GWASs require a minimum of hundreds of thousands of SNVs to map the majority of existing variation in the human genome. These results are often shown in a scatter plot with the loci of the SNV on the x-axis and the negative logarithm of the p-values of the marker on the y-axis: a Manhattan plot (see ◘ fig. 7.7). The final result of a GWAS is one or more candidate regions each containing a genetic variant that influences the likelihood of the disease. We can zoom in on the candidate region using even more markers to track down the actual causal variant.

Sequencing techniques enable particular genomic regions to be looked at in great detail. This method can be used to track down the actual causal susceptibility variants, guided by the results of GWASs. Molecular biology and bioinformatics methods can also be used for this purpose. Past GWASs have shown, however, that it is not always easy to go from an associated variant to an underlying causal mechanism.

The search for genetic determinants of multifactorial disorders is gradually being supplemented with or replaced by next-generation sequencing techniques (which are still expensive at present), providing researchers with a complete picture of all the variants, including very rare ones. While not substantially changing the study design, this technique has brought with it some new challenges, such as dealing with potential genotyping errors and processing the large quantities of data that are generated.



◘ **Figure 7.6** Numerical example of pairwise linkage disequilibrium (LD) where LD is expressed as r^2. This measure can range from 0 to 1, where 0 means linkage equilibrium and 1 linkage disequilibrium. Where $r^2 = 1$ there is perfect correlation: by measuring the genotype of variant A we can perfectly predict the genotype for variant B. In this example, for instance, we know that if we find genotype 11 for variant A the genotype for variant B will be 22

Case 7.7 GWAS for type 2 diabetes

One of the first large-scale case-control GWASs was conducted by the Wellcome Trust Case Control Consortium (WTCCC), a partnership of various research groups in Great Britain. It looked at seven different disorders, including type 2 diabetes (DT2). The DNA of 2,000 DT2 patients and 3,000 controls was studied using an array for over 500,000 SNVs. Twelve SNVs showed an association with DT2 based on a probability threshold (p-value) of below 0.00001. Three of these association signals were SNVs in genes which had already been shown in previous linkage and association studies to be linked to the risk of DT2. The GWAS also revealed an association with *TCF7L2* (see Case 7.6), as the strongest association signal was found for rs4506565. This SNV displayed strong LD with rs7903146 (which had not been measured directly), the variant located in *TCF7L2* that had been detected in previous linkage and association studies. The nine additional signals were SNVs of which some were located in or near candidate genes for DT2. Some of these SNVs have now been confirmed in additional case-control populations.

6 Genome-wide association studies (website). ▶ http://1.usa.gov/1EyquzA

7.4.4 Population stratification can cause confounding

The association between a genetic variant and a disease can be confounded if the patients and the controls have different genetic backgrounds and the researcher is not aware of this. This phenomenon is known as population stratification. Suppose we want to investigate the causes of prostate cancer. We take a group of one hundred prostate cancer patients and compare them with a group of one hundred controls. As people of African ancestry run a higher risk of prostate cancer than those of European ancestry, the likelihood is that there will be more individuals of African ancestry in the patient group. After analysing the DNA of all the participants it was found that 70 of the 100 patients had specific allele 1, as against only ten of the controls. The OR of (70/30):(10/90) = 21 suggests that people with allele 1 are at a much higher risk than those without it. It is quite possible, however, that allele 1 in itself does not have anything to do with the disease but is simply more common in individuals of African ancestry. Other random markers could also be associated with the disease, merely because the allele frequency differs for different ancestries. This, of course, is a classic example of confounding (see ▶ Chapter 5). While an epidemiologist will never make the mistake of not taking such an obvious factor as ancestry into account, genetic ancestry goes back many generations and is not easy for researchers to trace. It is often impossible to make a clear classification in the mix of underlying genetic backgrounds of participants.

If we are unsure about the presence of population stratification, we can characterize a large number of markers known not to be associated (or potentially associated) with the disease in the groups that we are comparing (patients and controls). Since any genetic selection will relate to the entire genome, this enables us to track down that selection.

As an alternative we can opt for a study design using immediate family members as controls, for example a trio design. This involves collecting DNA from patients and their healthy parents. As alleles are inherited through a known – Mendelian – mechanism, we can determine which alleles patients have and have not received from their parents.

Table 7.1 Display of data in a trio design for the transmission disequilibrium test (TDT)

	untransmitted alleles		
		1	2
transmitted alleles	1	a	b
	2	c	d
total = 2n (with n children)			

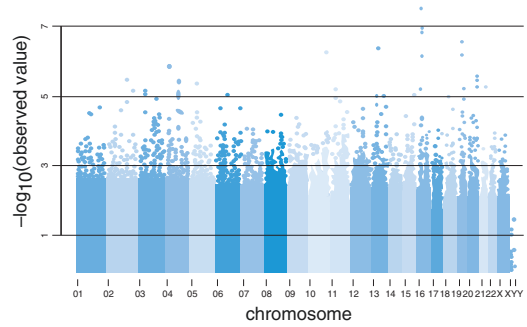


Figure 7.7 The Manhattan plot

This boils down to a case-control study in which alleles that are not passed on from the patients' parents are used as controls. The trio design, using family members as controls, is not susceptible to confounding due to population stratification, as the family members come from the same population. In other words, this is a case-control design with matching for genetic background. The trio design is also known by the name of the statistical test – the transmission disequilibrium test (TDT) – that is used to check whether there is an association between a particular genotype of the child and the disease. If an allele is passed on to the affected offspring more frequently than would be expected, that is a strong indication that that allele increases the risk of the disease. Because of the matched study population the odds ratios for a matched design are calculated (see ▶ Chapter 4) and only heterozygous parents can contribute meaningful information to this type of study. The design is shown in tab. 7.1 and fig. 7.8. The trio design, of which ▶ case 7.8 gives

Table 7.2 Distribution of alleles in children with diabetic nephropathy and their heterozygous parents

	untransmitted alleles				
		ε2	ε3/ε4		
transmitted alleles	ε2	–	38		
	ε3/ε4	21	–		
			total: 59		
	number of transmitted alleles				
	ε2	ε3/ε4	total	TDT χ^2	p-value
observed	38	21	59	7.47	0.006
expected	29.5	29.5	59		

an example, is simple and easy to use. The big disadvantage, of course, is that the parents must still be available to provide DNA, which is usually not the case with diseases with old age of onset. There are versions of the trio design, however, where several brothers and sisters can be used as controls instead of the parents. Another disadvantage is that the effects of environmental factors cannot be identified, as there are no true controls.

Figure 7.8 illustrates this with an example where both parents of an affected child are heterozygous (1||2) for a particular marker. Both of them have passed on allele 1 but not allele 2. The contribution to tab. 7.1 is b = 2.

By characterizing a large number of patient-parent trios we can find out whether allele 1 is in fact being passed on to affected offspring more frequently. We can calculate a matched OR (= b/c) with a corresponding confidence interval using the summary table (tab. 7.1). If necessary a McNemar Chi-square test can be added (referred to here as the TDT Chi-square):

$$T_{TDT} = (b-c)^2/(b+c)$$

Case 7.8 APOE and the risk of diabetic nephropathy (2)

The article that is the source for ▶ case 7.5 also describes a sub-study with a trio design to rule out confounding due to population stratification. It looked at the transmission of APOE alleles from 59 heterozygous parents (ε2/ε3 or ε2/ε4 genotype) to DN patients.

The results in tab. 7.2 show that heterozygous parents passed on allele ε2 of the APOE gene to their affected children more frequently than allele ε3 or ε4. This confirmed the results of the case-control study (see Case 7.5) and led to the conclusion that susceptibility to diabetic nephropathy is evidently linked to the APOE gene and that carriers of the ε2 allele run a higher risk.

7.4.5 The importance of collaboration

An untargeted search for disease genes in the entire genome, i.e. without any specific prior hypotheses, as is the case with GWASs, entails some particular problems. For example, conducting hundreds of thousands or millions of association tests of minor genetic effects (involving low-penetrance variants, possibly measured with some errors) requires large study populations of thousands of people in order to achieve sufficient power and not miss findings. When performing 1,000,000 standard statistical tests with an α (type 1 error) of 0.05 you would expect 50,000 false positive associations. To reduce that amount much smaller confidence intervals

must be used. Even then, the results of the study need to be replicated in other populations in order to distinguish false positive from true positive associations.

The need for large-scale studies and replication of results in independent study populations has vastly encouraged collaboration between research groups in consortia and the creation of biobanks. An example of a consortium is the Wellcome Trust Case Control Consortium[7] (see ▶ case 7.7). This is a partnership that was set up in 2005 and now includes fifty British research groups. The original study population comprised 14,000 patients (2,000 each for seven different disorders) and two control populations of 1,500 controls each, but it has now grown much larger.

The term biobank refers to a collection of biological material (e.g. blood, DNA, tumour material) linked to detailed descriptions of the characteristics of a large number of people. Biobanks provide a rich source for genetic epidemiological research. The importance of integrating knowledge and data from various 'omics' (genomics, transcriptomics, proteomics, metabolomics) in order to clarify the genetic etiology and pathophysiology of multifactorial disorders is becoming increasingly clear, and here too biobanks can play an important role. Most biobanks are based on a cohort design. The study population can be a sample of the population (e.g. of a geographical region) or a particular group of patients.

The data collected and managed by consortia and biobanks are often made available (subject to certain conditions) to other researchers so as to make the best use of those data and promote scientific progress. This is also true of some important large-scale cross-sectional studies designed to describe genetic variation in human populations, for example the 1000 Genomes Project mentioned earlier. A major challenge for the coming decades is to create order and extract the relevant information from the data that are becoming available thanks to these large-scale initiatives.

7 Wellcome Trust Case Control Consortium (website). ▶ http://bit.ly/1HvYYDy

7.5 Using genetic epidemiological research to clarify biological pathways and track down susceptible groups

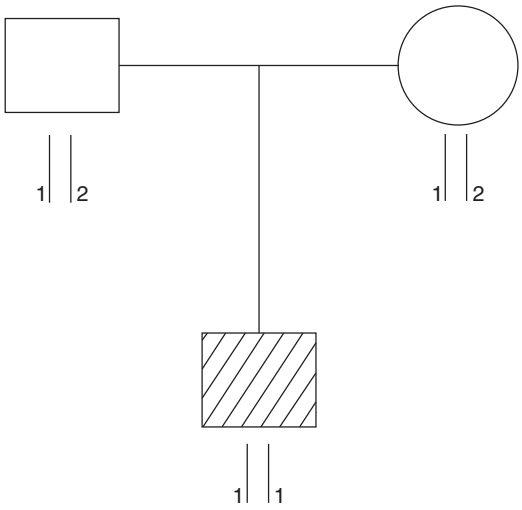
GWASs have rapidly increased our knowledge of the genetic determinants of multifactorial disorders. Some of these determinants are genes whose involvement in pathophysiological mechanisms was previously not known, and this knowledge provides an important basis for the development of new preventive measures and treatments and biomarkers for diagnosis and prognosis.

Particularly important is research into effect modification between genetic factors and modifiable behavioural and environmental factors, as knowledge from these studies enables groups of people to be selected who are likely, given their genetic profile, to benefit more from preventive and treatment measures designed to influence those behavioural and environmental factors. This is referred to as personalized prevention and treatment. Pharmacogenetic epidemiological research, for instance, focuses specifically on identifying genetic determinants of a patient's drug response. The hope is that there will ultimately be enough affordable genetic tests available to determine before the start of treatment which patients will benefit from a particular drug and which will not, or will need a different dosage. As yet there are few examples of this in clinical practice. One is the use of a genetic test for the thiopurine methyltransferase (TPMT) gene in patients eligible for treatment with thiopurines. These are inactive prodrugs with an immunosuppressive effect used for various disorders (e.g. rheumatoid arthritis and Crohn's disease). Certain genetic variants in the gene that codes for thiopurine methyltransferase (TPMT) – an enzyme important in converting thiopurines into active metabolites – are known to reduce TPMT activity. Adjusting the thiopurine dose before the start of treatment based on the results of this genetic test can avoid many of the undesirable side effects of thiopurines.

Knowledge of genetic determinants of an environmental factor can be used as evidence for or against a causal relationship between exposure to the environmental factor (e.g. vitamin D or LDL cholesterol) and the occurrence of a disease. This

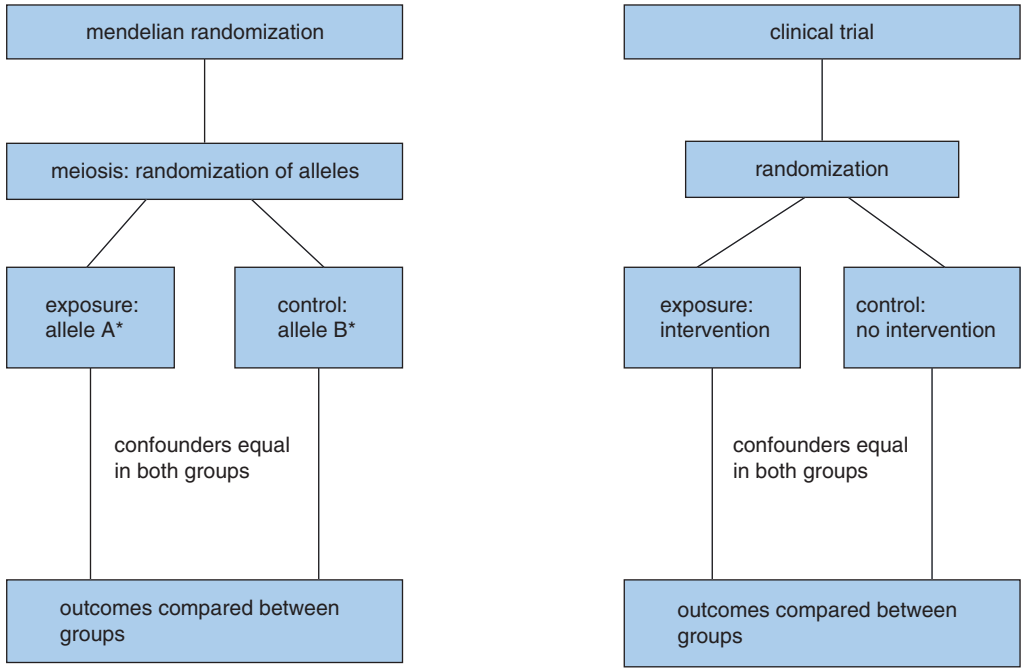
principle, known as Mendelian randomization, is based on the idea that various types of bias that occur in epidemiological research into determinants of disease (see ▶ Chapter 5) do not occur when the association between the genotype and the disease is examined. There will be no information or selection bias, for instance, and no doubt as to what is cause and what is effect. Also, confounding is unlikely when examining the association between the genotype and the disease, because of the randomization of alleles that takes place during meiosis. It makes no sense, for instance, for the genotype of a marker being studied in the context of the occurrence of cardiovascular disease to be associated with the amount of dietary cholesterol intake. The only exception to this rule is confounding due to population stratification, discussed in ▶ par. 7.4.4. The principle of Mendelian randomization is based on the fact that a causal relationship between an environmental factor and a disease means that a genetic determinant of that environmental factor must display an association with the disease. Using the genetic determinant of the environmental factor as a substitute for exposure to that factor thus produces the same effect in an observational setting as randomization in a clinical trial (see ◘ fig. 7.9). This is very important, of course, when trying to prove a causal relationship.

Knowledge of highly penetrant genetic determinants of monogenic disorders has been used for some considerable time now in clinical practice for genetic counselling. For instance, specific genetic tests can be carried out, under the supervision of clinical geneticists, on individuals with a positive family history of a particular hereditary disorder. This targeted genetic screening and counselling can help in the early diagnosis of a particular disorder or to estimate the risk of its occurrence. An example is presymptomatic screening for *BRCA1* and *BRCA2* mutations in families in which these mutations for hereditary breast and/or ovarian cancer have been detected (see ▶ case 7.4). The increase in our knowledge of genetic determinants of multifactorial disorders using GWASs has given rise to a debate on population screening for risk estimation of particular multifactorial disorders based on individual genetic profiles. This type of prognostic genetic testing is also being offered by commercial companies. Unlike monogenic disorders, however, the risk of multifactorial disorders is influenced only partially by genetic determinants, and with regard to virtually all these disorders only a small part of the genetic etiology has been clarified so far. Age-related macular degeneration, a major cause of blindness, is an exception: GWASs have identified a number of genetic variants with a relatively strong effect that predict the genetic component of this disorder fairly well. In other cases, multiple genetic variants and environmental risk factors need to be taken into account simultaneously.



◘ **Figure 7.8** Example of a parental couple with an affected child

For most disorders, the clinical value (validity, precision and efficiency: see ▶ chap. 9) of this type of genetic tests is still too low to warrant their use in practice, however. Moreover, in many cases it is still unclear whether knowledge of genetic risk profiles will actually lead to effective treatment measures. The advice to stop smoking, for example, applies to any smoker, irrespective of his or her genetic risk profile, and knowledge of that profile will not necessarily lead to an improvement in lifestyle.



* is a proxy for exposure to an environmental factor

Figure 7.9 Simulating a randomized trial using the principle of Mendelian randomization

7.6 Guidelines for the publication of genetic epidemiological research

There are now STREGA guidelines on the publication of genetic epidemiological studies. STREGA (STrengthening the REporting of Genetic Association studies) builds upon the STROBE guidelines on the publication of observational studies (see ▶ chap. 4). The STREGA guidelines provide specific advice on the elements that should be included in an article on genetic association research.[8]

8 STREGA guidelines on the publication of observational studies (website). ▶ http://bit.ly/1HzyaRp

Recommended reading

Khoury MJ, Bedrosian SR., Gwinn M, Higgins JPT, Ioannidis JPA, Little J. Human genome epidemiology: building the evidence for using genetic information to improve health and prevent disease. 2nd ed. Oxford: Oxford University Press; 2010.

Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2012.

Thomas DC. Statistical methods in genetic epidemiology. New York: Oxford University Press; 2004.

Source references (cases)

Claus EB, Risch NJ, Thompson WD. Age at onset as an indicator of familial risk in breast cancer. Am J Epidemiol. 1990;131:961–72 (Case 7.2).

Polderman TJC, Benyamin B, De Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM & Posthuma D. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics, 47, 702–709 (Case 7.3).

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990;250:1684–9(Case 7.4).

Recommended reading

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266:66–71 (Case 7.4).

Newman B, Austin MA, Lee M, King MC. Inheritance of human breast cancer: evidence for autosomal transmission in high risk families. Proc Natl Acad Sci U S A. 1988;85:3044–8 (Case 7.4).

Araki S, Dariusz KM, Hanna L, Scott LJ, Warram JH, Krolewski AS. APOE polymorphisms and the development of diabetic nephropathy in type 1 diabetes. Diabetes. 2000;49:2190–5 (Cases 7.5 and 7.8).

Grant SFA, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 2006;38:320–3 (Case 7.6).

Helgason A, Pálsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, et al. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet. 2007;39:218–25 (Case 7.6).

Palmer ND, Hester JM, An SS, Adeyemo A, Rotimi C, Langefeld C.D et al. Resequencing and analysis of variation in the TCF7L2 gene in African Americans suggests that SNP rs7903146 is the causal diabetes susceptibility variant. Diabetes. 2011;60:662–8 (Case 7.6).

Scott LJ, Bonnycastle LL, Willer CJ, Sprau AG, Jackson AU, Narisu N, et al. Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. Diabetes. 2006;55:2649–53(Case 7.6).

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78 (Case 7.7).

Outbreak epidemiology

The co-authors of this chapter are Dr. Teun Bousema (Department of Medical Microbiology, Radboud University Medical Centre, Nijmegen, the Netherlands), Dr. Susan Hahné (Centre for Infectious Disease Control (CIb), National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands) and Dr. Jeannine Hautvast (Department of Primary and Community Care, Radboud University Medical Centre, Nijmegen, the Netherlands)

8.1 Introduction: investigating disease outbreaks is complicated but very exciting

An outbreak occurs when the number of new cases of a disease observed in a particular situation and a particular geographical area over a relatively short period is larger than would be expected based on the descriptive epidemiology of that disease (▶ chap. 1). An increase in the incidence of a disease in a large population and/or area is referred to as an epidemic. We therefore need to have a good idea of the number of cases that can be expected in a particular period, place and population, and what variability is normal. Proper surveillance is very important, then, if we want to be able to recognize and investigate true outbreaks.

The aim of investigating outbreaks is to identify their characteristics and determinants. Ideally we will want to track down both the agent (the main determinant) and the source as soon as possible, as both of these can provide a basis for interventions and disease control, for example:

1. Detecting and/or treating the disease faster or better,
2. Taking steps to halt the spread of the disease, and/or
3. Making recommendations to reduce the risk of similar outbreaks occurring here or elsewhere in future.

▶ Chapter 4 distinguished between situational research questions, related to a particular time and/or population, and abstract scientific research questions, designed to generate theoretical knowledge. Investigations of outbreaks will almost always be situational. Usually the aim is not to discover new determinants of the disease in general but to find out what has caused a particular time, person and place-related outbreak (and what species and type of microorganism in the case of infectious diseases). Based on this knowledge it may be possible to take steps locally to halt the outbreak and prevent similar ones. If the investigation also brings novel insights, e.g. of a transmission route not previously identified, new (acquired) properties of a microorganism (e.g. resistance or escape mutants) or new risk groups, it will also have produced theoretical knowledge that can improve public health elsewhere. Sometimes the primary aim is to gather this theoretical knowledge, for instance because epidemiological research can only be carried out during an outbreak, when people are exposed: for example, a study on measles during an outbreak in a religious community with a low vaccination rate. In this case, gaining knowledge of how to fight the disease is not the primary aim of the study, as we already have a good deal of knowledge of measles transmission and prevention; an investigation into the outbreak can however contribute to new knowledge. A measles outbreak in a group of this kind, for instance, provides an opportunity to carry out behavioural studies into the effect of interventions to promote the acceptance of vaccination against e.g. mumps, measles and rubella (MMR) in unvaccinated children. These are exceptions, however; outbreak epidemiology is usually concerned with situational research questions.

When we talk about the determinants of outbreaks we are usually referring to infectious disease outbreaks (foodborne infections such as *Salmonella*, open tuberculosis, Q fever, methicillin-resistant *Staphylococcus aureus* (MRSA) on livestock farms, measles in populations with religious objections to vaccination, mumps among students, severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), Ebola or influenza pandemics). An outbreak can also be caused by an environmental incident, however (children playing on contaminated land, unintentional emissions of toxic substances or contaminated drinking water), or errors in the treatment and care of patients (iatrogenic outbreaks, e.g. congenital abnormalities due to the use of thalidomide during pregnancy, or eye abnormalities due to excessive oxygen levels in premature babies' incubators). Not all outbreaks are infectious, and not all infectious diseases are prone to causing outbreaks.

As these two types of epidemiology – infectious disease epidemiology and the epidemiology of outbreaks – often coincide, this chapter will focus mainly on infectious disease outbreaks. A specific feature of infectious disease epidemiology is that a microorganism (virus, bacterium, parasite, prion, fungus, etc.) is involved. Many of the microorganisms that are pathological in humans can only survive if they are transmitted from one host to another.

Figure 8.1 The 'excitement' in outbreak investigation derives partly from the social unrest and resulting political pressure that can sometimes develop, which makes the investigation more urgent. This photograph is an illustration of the social unrest that developed during a Listeria outbreak caused by raw milk from an organic farm in Pennsylvania in 2011

Figure 8.2 Disease transmission from animal to human (e.g. in Lyme disease)

This transmission can take place directly from one human to another (e.g. in measles), through an 'intermediate host' (e.g. an insect, as in malaria and in Lyme disease: see fig. 8.2), or directly from animal to human (zoonosis, e.g. Q fever and psittacosis). The term 'contagious disease' is only used if there is direct human-to-human transmission. Other microorganisms infect humans through the environment, e.g. water (*Legionella*, for instance). In such cases the human is usually a 'dead-end' host, humans cannot infect other humans.

The infectious agent's tendency to reproduce is countered by the host's tendency to defend itself. For this purpose it has a general innate immune system, supplemented by acquired specific resistance (immunity) to particular agents. This acquired immunity can result from previous exposure to the agent in question (or a related microorganism in the case of cross-immunity) or vaccination. An impaired immune system or lack of resistance makes the host particularly susceptible to infections. The behaviour, virulence and 'dose' of the infectious agents and the efficacy of potential hosts' immune responses are not the only determinants of infection, however. Unlike in the case of non-infectious diseases, the likelihood of an infection in an individual also depends on the presence of infections and immunity in other members of the population, as the following example shows. A microorganism such as polio virus can only cause an epidemic if it is able to circulate in a community of people who are not yet immune to it. Interventions against infectious diseases (e.g. polio vaccination) therefore have effects not only on the people receiving them but also on the other individuals in the population. Once 'herd immunity' has been developed (enough people are immune or have been immunized), the spread of the disease is halted. In order to assess the presence of infections we also need to take into account an inconvenient additional factor: it is not always clear who is infected, as people can spread the disease without being ill.

Thus infectious disease epidemiology differs from chronic disease epidemiology in certain important respects:

- It is important to differentiate between exposure to, colonization by, infection by and disease due to a potentially pathogenic microorganism (see also ▶ 8.3.4). The ultimate aim of infection control is to reduce the disease frequency, but the actions actually taken can be designed to reduce the number of cases or the number of infected or exposed people. An infected person (whether diseased or healthy) can be a source of, or a risk factor for, the exposure and infection of one or more other people. Thus the infectious disease risk depends on whether there are other people in the vicinity who have the infection, and the

network of contacts in the population is important when studying the transmission of infectious diseases.
- People can be immune to infection and/or disease to a greater or lesser extent.
- Microorganisms are not only a risk factor; they can sometimes have a protective effect. In a normal, healthy situation people live in symbiotic equilibrium with these microflora on the skin and mucosa such as in the throat and in the intestines, and they are an important line of natural defence against external threats.

From time to time a new disease will emerge somewhere, which was previously unknown (HIV/AIDS was unknown before 1983, for instance) or that did not occur (or had stopped occurring) in that particular area. A well known example is the rise of West Nile virus in the United States at the end of the twentieth century. A disease may emerge due to various causes:
- Changes in human behaviour (e.g. intravenous drug users sharing needles, increasing international tourism, or increasing contact between humans and animal carriers of pathogens). Since 2009, for instance, a new malaria species is found in Asia due to increasing contact between humans and monkeys
- Wars and natural disasters (e.g. the large number of rapes and resulting HIV infections in children during the Rwandan civil war)
- Changes in the environment (e.g. pollution or climate change)
- Changes in the agent itself (e.g. the development of resistance, or mutations in the microorganism's genetic material, as in the case of bird flu)
- Changes in healthcare (e.g. introducing treatments that have the side effect of severely impairing immune response, or relaxing preventive measures)
- Changes in food preparation or the food chain (e.g. the variant form of Creutzfeldt-Jakob disease that emerged in 1996 due to the contamination of food with prions which cause bovine spongiform encephalopathy (BSE or mad cow disease), or the use of growth-promoting antibiotics).

In addition to carrying out surveillance and developing disaster management plans, various preventive measures can be taken to reduce the risk of new diseases. These measures particularly target the causes listed above, e.g. improving blood, food and water hygiene, the responsible use of antibiotics, developing new vaccines, reducing human exposure to animal infections, protecting susceptible groups (children, the elderly, pregnant women) and combating bioterrorism.

Infectious disease epidemiology is very exciting because outbreaks usually develop much faster than in normal epidemiology, society is often far more involved, and we are increasingly confronted with emerging infections in our environment. This type of epidemiology can also be very satisfying for researchers, especially if the results can be used to prevent new cases of the disease. On the other hand, time pressure to produce results that are relevant to a current outbreak leaves less opportunity to design and execute studies very meticulously. Investigators have to be content with data of lower quality than they otherwise demand.

8.1.1 Commonly used terms in infectious disease epidemiology

Because of the specific nature of infectious disease epidemiology some additional concepts are needed to describe and explain the frequency and patterns of infectious diseases.

A person's infection often follows a characteristic pattern. When a person becomes infected, symptoms may develop and the disease manifest itself after a certain period (the incubation period). The symptoms of the disease usually disappear, with or without treatment, but the infection generally has consequences for immunity or carriership. The infected individual usually remains infectious for some time after the infection (the latency period) and can then pass it on to other people if it is human-to-human transmissible. The infectious period usually comes to an end too, but it is important to note that the start and end will not necessarily coincide with the period when the signs of the disease are manifest. For instance, an individual may be infectious before displaying symptoms of the disease (as

in the case of childhood chickenpox, influenza and HIV), whereas in other cases contagiousness may persist for years after the symptoms have disappeared (as in the case of hepatitis B). Individuals can therefore be infectious during the asymptomatic stage before the symptoms develop, the asymptomatic stage following the symptoms, or during an infection that runs a completely asymptomatic course.

$$\text{Attack rate} = \frac{\text{Number of diseased or infected individuals}}{\text{Total population at risk (non-immune)}}$$

$$\text{Secundary attack rate} = \frac{\text{Number of diseased or infected individuals}}{\text{Number of individuals exposed to an infected individual}}$$

The probability of a person becoming infected, given a particular source of infection, is referred to as the transmission rate (p). This will depend on the nature of the source (often an infected individual, animal or other infected source), any vector providing transmission (mosquito, food, airborne particles), how the contact takes place and for how long, the virulence and number of microorganisms to which the person is exposed, and the characteristics (genetic, immunological, behavioural) of the host being infected. The simplest ways of representing the transmission rate are (a) the attack rate (the number of cases or infected persons as a proportion of the total population) and (b) the secondary attack rate (the number of cases or infected persons among people who have been in contact with an infected individual divided by the total number of people who have been in contact). The basic reproduction number (R_o) is the average number of new infected individuals that one contagious individual can produce in a population in which everyone can contract the infection. This concept is elaborated in ▶ 8.6.2.

The events that can lead to the spread of an infectious disease (and its consequences) usually take place faster than in the case of a non-infectious disease. All sorts of assumptions that are standard in chronic disease epidemiology (e.g. independent disease risk) do not apply, so modified analytical methods need to be used in outbreak epidemiology.

8.2 Surveillance for early warning

Almost every outbreak begins with the discovery of an unusual disease-related finding in the population. Sometimes a single case of the disease provides sufficient warning (e.g. a case of botulism, paralysis due to shellfish poisoning or anthrax). In many cases, however, an outbreak will be suspected as a result of observing two or more cases in close chronological or geographical proximity. This is referred to as a disease cluster (see ▶ 8.6). It will usually be alert doctors – or in some cases alert citizens – who raise the red flag for a possible cluster of cases.

In order not to be dependent on the lucky vigilance of individual professionals and citizens we need systems designed to give early warning of unusual clusters of cases. As a result, various types of surveillance have been developed to provide ongoing collection, processing, analysis and reporting of disease data. Surveillance enables unusual shifts in reported numbers of diseases or infections in terms of time, place, people and pathogens to be identified. Surveillance has been going on for centuries and it has been found very useful in informing the authorities of public health trends. Back in the seventeenth century John Graunt (see ▶ fig. 1.6) published a weekly survey of causes of death in the London population, and William Farr (see ◘ fig. 8.3) analysed and published regular disease and mortality rates for England and Wales in the mid-nineteenth century. After World War II special bodies, such as the Center for Disease Control and Prevention (CDC) in the United States, were set up to carry out surveillance work.[1] There is also a European counterpart – the European Centre for Disease Prevention and Control (ECDC)[2] in Stockholm. This has a network of affiliated centres which organize national infectious disease surveillance in their respective countries.

It is partly thanks to active surveillance that we have succeeded in finally eliminating smallpox worldwide – an unparalleled achievement. During the smallpox campaign there was active surveillance

1 US Center for Disease Control and Prevention (website). ▶ http://1.usa.gov/1fgeJk1

2 EU Centre for Disease Prevention and Control (website) https://ecdc.europa.eu/en

Figure 8.3 William Farr (1807–1883), British epidemiologist and one of the founding fathers of today's medical statistics

for new cases throughout the world, and after each notification the patients were isolated and all their contacts vaccinated.

The principles of surveillance do not only apply to infectious diseases, of course; they also apply to cancer, congenital abnormalities and accident traumas. Surveillance is not designed primarily as a type of epidemiological research but as a tool for identifying changes in trends quickly, and to enable interventions to be carried out where necessary so as to nip an outbreak in the bud. Speed is usually more important than validity and precision here.

Many countries have statutory notification requirements for particular infectious diseases. If comprehensive notification is necessary, effective and feasible in order to control infectious diseases, the authorities can impose a notification requirement on e.g. doctors, microbiological laboratories and heads of institutions that come across cases. As a result of these notifications the regional and national authorities responsible for infectious disease control are informed quickly and can take timely action in the event of an outbreak. Up-to-date information on the notification requirements in the various countries can be found on the national public health institutes' websites.

National healthcare registries – e.g. cause of death registers, hospital registries and cancer registries (see ▶ 2.7) – can be used for this purpose, but they do not usually analyse the data fast enough to identify outbreaks before other authorities do so. GPs perform a very important warning function. In some countries this is done by a selection of GP practices, which report the number of new cases of a certain number of conditions once a week. This is an important source of information for e.g. influenza surveillance in those countries. This type of surveillance, in which data are collected from particular geographical locations, healthcare institutions or populations during defined time periods, is referred to as a sentinel system. Sentinel surveillance systems have also been developed for the surveillance of infectious diseases in nursing homes, for instance.

As the aim is to identify changes in disease frequency and clusters of disease in time and place as quickly as possible, a surveillance system needs to pick up most of the cases and therefore have high sensitivity, if necessary at the expense of specificity (see ▶ chap. 9). This can result in a large number of false positives being generated, and these need to be identified. In some cases (e.g. polio or SARS) even a mere suspicion is sufficient to warrant notification.

In order to understand the course of infectious diseases it is important to keep both cases and populations of microorganisms under surveillance. As part of influenza surveillance in various countries, for instance, a nose or throat swab is taken from patients with flu symptoms regularly (once a week) to see which, if any, microorganisms are present. The virology laboratories keep a record of the number of positive diagnoses of selected viruses. There are also computerized surveillance systems for local outbreaks involving resistance to antibiotics.[3]

A good example of the value of international surveillance is in detecting cases of Legionnaires'

3 ▶ Data from the ECDC Surveillance Atlas - Antimicrobial resistance (website) http://bit.ly/2yJxXeh

disease among tourists returning from holiday. This infection is often caused by showers or air conditioning systems in hotels, with patients only falling ill after returning home and looking like isolated cases. Pooling this data internationally enables outbreaks of Legionnaires' disease to be identified and possible sources tracked down and dealt with.

The value of a surveillance system stands or falls by the constancy and quality of the data. As these are usually supplied by healthcare providers it is important to involve them closely in the surveillance and furnish them with regular feedback to enable them to use the data and optimize its quality.

Surveillance systems can also be used to assess and monitor the efficacy of public health measures. Analysis of the numbers of reports of near misses provides hospitals and companies with information on the extent to which they comply with various quality and safety rules. Ongoing nationwide surveillance of infectious diseases is used to assess the impact of national vaccination programmes. An increase in the number of cases of childhood whooping cough, for instance, may be a reason to change the vaccination schedule, use a different vaccine and/or make special efforts to increase vaccination rates among particular sections of the population (see ▶ case 8.1).

Case 8.1 Surveillance of and vaccination against whooping cough

Whooping cough is an acute infectious lung disease usually caused by the bacterium *Bordetella pertussis*. The typical coughing fits are caused by the toxins that these bacteria produce. It is transmitted from person to person via the air. Patients are most infectious during the period preceding the coughing fits, but contagiousness continues for a few weeks thereafter. Whooping cough is highly contagious. On average, 90% of susceptible members of a family with a whooping cough patient will contract the infection. In countries with high vaccination rates adolescents and adults play a major role in spreading it. Having whooping cough does not confer lifelong immunity; immunity declines four to twenty years after a whooping cough infection. Nor does vaccination provide protection for life, only for four to twelve years.

Whooping cough occurs throughout the world, with an estimated 45 million cases and 400,000 deaths a year. Reported incidences differ widely, partly genuinely as a result of differences in prevention and control strategy, and partly as a result of different diagnosis and reporting criteria (see e.g. the atlas on the European Centre for Disease Prevention and Control website[4]). There was a sharp rise in the number of whooping cough cases in Europe around 2012, a year when 40,000 confirmed cases were recorded, as against 20,000 in previous years (▣ fig. 8.4). Hospitalization is needed in fewer than 10% of cases, mainly young infants. Deaths due to whooping cough are fortunately very rare, less than 0.2% of the cases.

Both cellular and non-cellular vaccines are available to prevent whooping cough. Cellular vaccines are made from dead bacteria. They contain large quantities of antigens and sometimes cause mild to severe side effects. Non-cellular vaccines contain combinations of protein components taken from the surface of the bacterium. These also, albeit to a lesser extent, can cause mild side effects, such as redness, pain and swelling at the injection site, as well as crying and listlessness.

Since 1950 infants in many European countries have been given a combined vaccine containing a cellular whooping cough vaccine and vaccines against e.g. diphtheria, tetanus and poliomyelitis. The efficacy of whooping cough vaccination in infancy declined around the millennium and there were repeated substantial epidemic waves of whooping cough in Europe, possibly due to a change in the *B. pertussis* bacteria in circulation. In response, most European countries decided to introduce an additional whooping cough vaccination at nursery school age. A non-cellular vaccine was chosen in most cases, as the cellular vaccine has too many side effects at that age.

4 ▶ Surveillance data on Pertussis (website) http://bit.ly/2lCa9qP

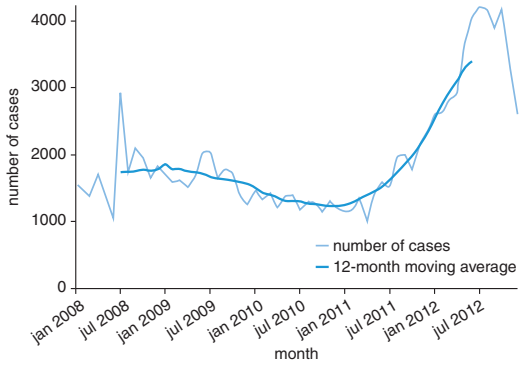


Figure 8.4 Overview of confirmed cases of reported whooping cough in Europe, 2008-2012

Many countries nowadays use a non-cellular whooping cough vaccine in the first year of life as well. Figure 8.4 shows a clear outbreak of whooping cough in Europe around 2012, which has again brought the efficacy of vaccination using the non-cellular vaccine into question. It is thanks to whooping cough surveillance that a debate of this kind is possible.

Following the epidemics of severe acute respiratory syndrome (SARS), H_5N_1 bird flu virus and several acts of bioterrorism around the millennium the need has grown to identify unusual rises in non-specific signs and symptoms (and combinations thereof). This syndrome surveillance involves studying the frequency of early non-specific signs or symptoms reported by e.g. GPs, hospitals, pharmacists (sales of medicines) or occupational health services (absenteeism figures). Syndrome definitions are based on combinations of symptoms without there being a precise single diagnosis. Increased incidence of a syndrome may then warrant further investigation or even measures.

An increase in the number of cases does not necessarily mean there has been an outbreak: the size or composition of the population may have changed, or there may have been changes in healthcare, the insurance system or the way cases are detected, diagnosed or recorded. Media attention usually results in a sharp increase in numbers of reported cases. For there to be an outbreak the reported cases must be interrelated in some way, involving the same characteristic disease parameters, the same microorganism, the same conditions, etc. It is less likely to be an outbreak if such shared parameters are absent.

A good case definition is extremely important in outbreak investigations and therefore in surveillance systems too. The clearer the definition, the better the quality of the data. If the level of misclassification is low, it will be easier to interpret the information and answer questions about the nature and severity of the outbreak. A case definition should always include the following four elements:

- Clinical symptoms
- Characteristics of the persons affected
- Geographical parameters (where the outbreak is taking place)
- Chronological parameters (when did the associated cases occur).

It is often worthwhile to differentiate between confirmed, probable and suspected cases (see ▶ case 8.5). Europe has uniform definitions of confirmed and probable cases for most infectious diseases of public health importance,[5] which enable international comparisons to be made.

Once we have established that there is an outbreak or a developing epidemic, we will want to survey the situation carefully in order to provide a basis for further investigation into the causes. At this point we need a whole range of information on each case, for example:

- Name, age, gender, ethnicity
- Place of residence
- Date and time when the disease first manifested itself
- Symptoms and their duration, and date of death if appropriate
- Possible transmission routes, high-risk behaviour
- Microbiological data (type/serotype, sequence, load)

5 EU case definitions (website) https://ecdc.europa.eu/en/infectious-diseases-public-health/surveillance-and-disease-data/eu-case-definitions

Figure 8.5 Surveillance systems are not only for infectious diseases

- Any other relevant information (vaccination status, occupation, etc.).

The more of these parameters a surveillance system has already recorded, the more effectively it can be used for the research to be done once it has identified an outbreak (fig. 8.5).

8.2.1 Rapid warning without false alarms

Timely identification of an outbreak or potential outbreak involves the ongoing comparison of numbers of cases observed with the numbers that would be expected from a chronological comparison or comparison with other regions. It goes without saying that we will want to avoid false alarms as much as possible. Before concluding that there is an outbreak we will therefore need to rule out other explanations, such as chance, incorrect estimation of the baseline level, changes in demography (e.g. a baby boom) or changes in diagnostic and recording procedures. There are statistical techniques that enable the number of warnings due solely to random variation to be reduced to an acceptable level, but this does not mean that we can rule out false alarms completely. The principle underlying these methods is that random variation over time and place is a less likely explanation if cases occur in successive periods and/or neighbouring areas.

Before using various statistical techniques we need to make a graphical representation of the disease frequencies, for example a graph of the cumulative numbers of cases observed and the numbers that would be expected based on the incidence of the disease remaining stable, plotted against time. This should ideally be based not on the date of notification or diagnosis but on the first day of the illness.

Local variations can be studied by making a

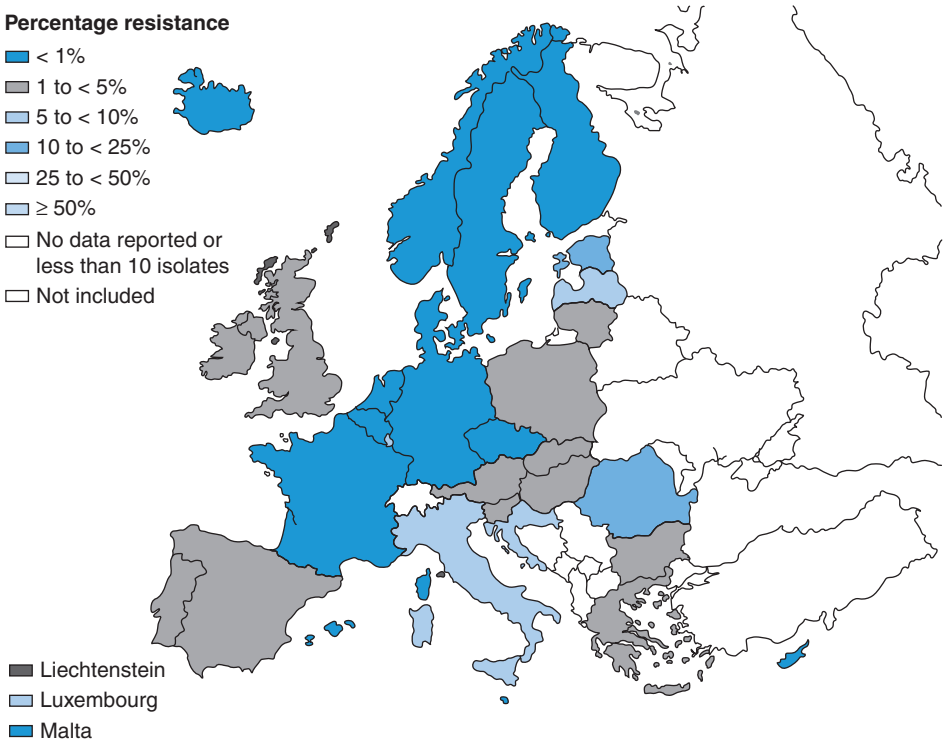


Figure 8.6 Prevalence of aminopenicillin resistance among Enterococcus faecalis isolates, Europe, 2014

map of the disease frequencies (incidence densities: see ▶ 2.5.2) showing higher frequency as increasing colour intensity (see fig. 8.6). Autocorrelation coefficients representing the correlations between disease frequencies in neighbouring areas can be used to show patterns in the area as a whole. The problem is that an outbreak is not the only possible explanation for a correlation between disease frequencies in neighbouring areas. Densely populated areas, for instance, will have more stable incidence densities and therefore stronger correlations. Differences in recording methods will also be smaller between neighbouring areas. Lastly, neighbouring areas will also display greater similarities in other determinants of the disease. To avoid being misled by random variation between areas we can 'smooth' the disease frequencies, i.e. calculate adjusted frequencies based on the information in the neighbouring areas. This technique is referred to as 'empirical Bayes estimation'. These methods enable us to detect underlying patterns in disease frequencies better than when using unsmoothed figures.

Geographical information systems (GIS) can also be used to plot cases on a map and analyse the data. A major problem with geographical methods is that these often have to rely on information from surveillance systems that only record the locations where the cases were diagnosed. Because of high mobility these often differ substantially from the location where the infection was contracted. The first international maps of the distribution of AIDS cases, for instance, showed the main airline routes rather than any epidemiological pattern. These geographical methods, however, are highly suitable for e.g. a local outbreak caused by a shared water supply (see ▶ case 1.5).

Special software (e.g. SaTScan) provides combined analysis of time and place data, enabling time and place-related outbreaks to be detected simultaneously, making it a very powerful tool in surveil-

lance. The software records the time and place of each case and calculates the numbers of observed and expected cases for each location at each time. Statistical procedures taking into account multiple testing enable the detection of clusters not due to random fluctuation. Technologies such as SaTScan are sensitive to certain user-defined parameters such as maximum cluster size or expected cluster shape (circular or elliptical).

8.3 Study designs for epidemiological research into outbreaks

An outbreak is a usually unexpected event and can only be studied once it has begun or even after it has stopped. This makes it difficult to design an adequate study. Usually there is substantial time pressure, and many outbreaks are of limited extent, making it difficult to differentiate between the effects of particular risk factors and random variation in the incidence of the disease in subgroups. In order to learn as much as possible from an outbreak, how prevention can be made more effective and what the most effective interventions are during and after an outbreak, we need to be properly prepared for outbreak investigations. Obtain, for example, ethical approval for a generic protocol, set up partnerships (e.g. between regional health authorities, hospitals, laboratories and GPs, but also internationally), and promote good surveillance and clinical vigilance to be able to identify and validate possible outbreaks quickly.

Case 8.2 Toxic shock syndrome

The first cases of toxic shock syndrome (TSS) were identified in seven adolescents by Todd et al. in 1978. Symptoms of TSS are sudden high fever, low blood pressure or shock and a bright red skin rash that persists for one or two weeks. TSS patients also feel miserable and often have gastrointestinal and muscular problems. The kidneys or liver may also be affected. TSS can be caused by Staphylococcal and Streptococcal bacteria.

In 1980 various young, otherwise healthy women were brought into Emergency Departments of hospitals in Wisconsin and Minnesota with the signs and symptoms described above. Remarkably, the disease had begun in all of these women at the start of an otherwise normal menstrual period. As the number of cases among menstruating women did not go down, it was decided to report women with these menstruation-linked symptoms to the Center for Disease Control (CDC) to enable the initiation of a case-control study. The first case-control study was carried out in Wisconsin on 35 patients with TSS during menstruation and 105 controls. For each patient three controls from the same area, of the same age and with a normal menstrual pattern were selected through GPs. Data on marital status, sexual habits, indications of sexually transmitted disease, length and intensity of menstrual periods, contraceptive use and use of tampons were collected through telephone interviews. It was found that patients were more likely to use tampons (97%) than controls (76%). There was also a difference in contraceptive use. No difference was found in the type or brand of tampons used. Both patients and controls often reported using 'highly absorbent' Rely brand tampons. A subsequent case-control study by the CDC in Atlanta confirmed the correlation with tampons (100% use among patients, as against 80% among controls) but not the correlation with contraception. The CDC carried out a further study among 50 patients with recent TSS and 150 controls recruited by the patients in summer 1980. In this study all the women were asked to find the box containing the tampons that they had recently used and report the brand and serial number. As 100% of the patients used tampons (83% of them Rely brand), as against 75% of the controls (26% Rely brand), the study concluded that the cause of the TSS in these young women probably lay in the use of this brand of tampons. Rely® tampons had been introduced in 1975 as 'super tampons' with high absorption and ease of use, as they only needed to be changed after a few days. Although they did indeed absorb more blood, they were also found to cause more mucosal irritation and higher growth of microorganisms

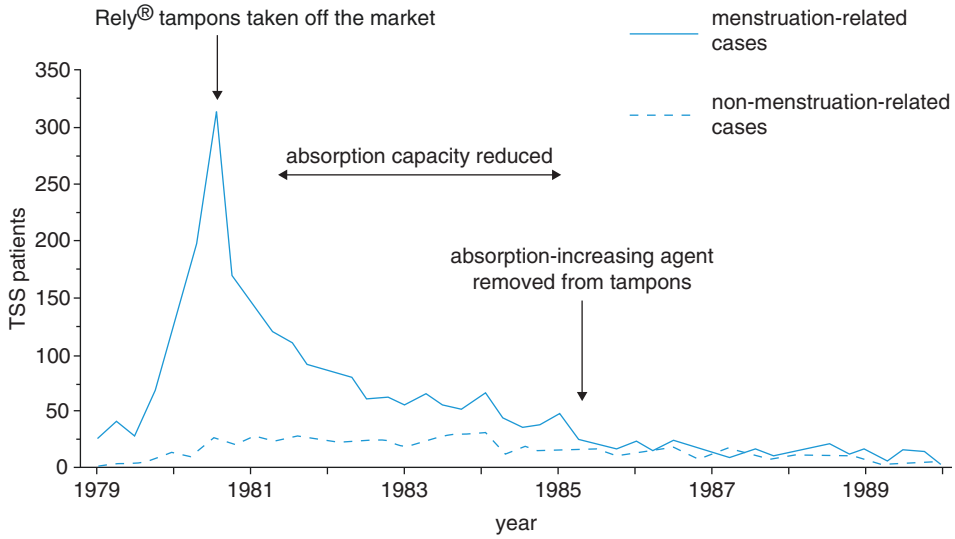


Figure 8.7 Numbers of TSS cases in the United States between 1979 and 1990

when used for a longer period. Women with *S. aureus* were indeed found to be at greater risk of TSS if they used these Rely brand tampons. The manufacturer took these tampons off the market on 22 September 1980, and the absorption capacity of tampons was reduced across the board. As fig. 8.7 shows, TSS disappeared almost entirely as a syndrome. Subsequent cases of TSS were usually due to wound contamination with *S. aureus*. As this case shows, three relatively small case-control studies can provide enough information to halt an epidemic.

8.3.1 The epidemic curve: learning from the time dimension

The epidemic curve is a standard element in the first descriptive stage of an outbreak investigation. It plots the number of cases detected against the time when the people start having symptoms of the disease. The scale used for the x-axis (the time when the symptoms began) depends on the nature of the disease, in particular the incubation period, and includes a period preceding the outbreak showing the normal incidence of the disease. As a rule of thumb, a quarter of the incubation period is often used as the unit of measurement for the x-axis. fig. 8.8 gives an example of an epidemic curve for an outbreak of fever caused by the Chikungunya virus (CHIKV) in a few villages in north east Italy in summer 2007. For each day it shows all the reported new cases of CHIKV fever in Castiglione di Cervia and Castiglione di Ravenna, in Cervia, and in other or unknown locations in the region during the July-September 2007 period (205 cases).

If the average incubation period of the disease is already known and all the cases were exposed at about the same time (as may be the case in an outbreak of food poisoning, for instance), the most likely time of exposure can be deduced from the figure. Note, however, that the shape of the epidemic curve can be influenced not only by case number, but also by delays in reporting, reporting fatigue, changes in the notification of new cases and changes in care-seeking by patients and potential patients.

The epidemic curve needs to be drawn at an early stage of the outbreak if it is to be useful for the evaluation of the measures taken. In point of fact it is one of the first analyses to be carried out once the

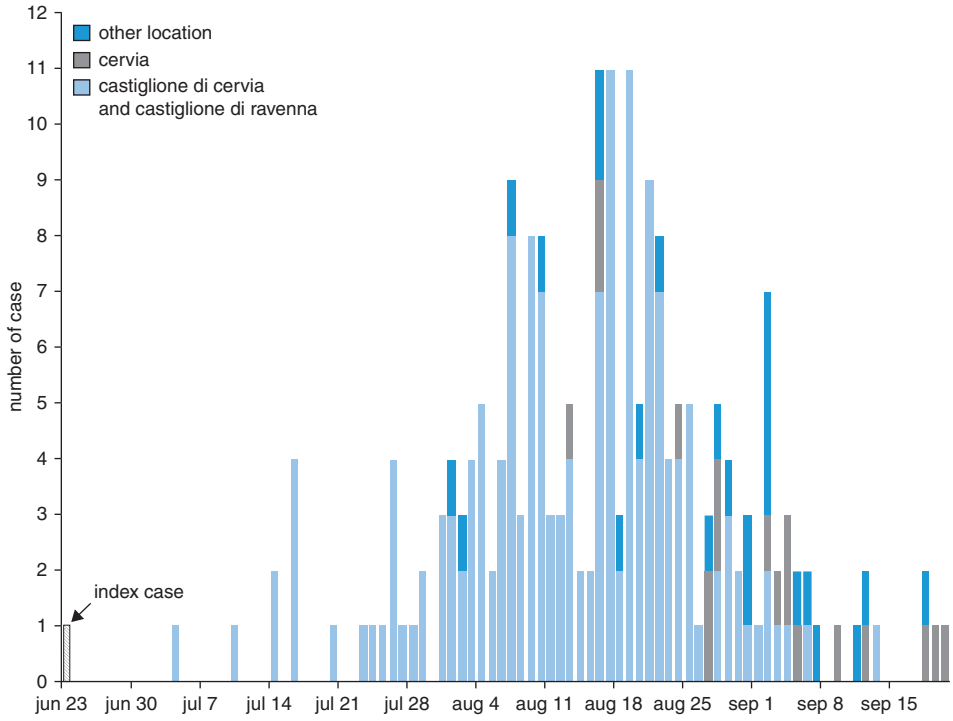


Figure 8.8 Example of an epidemic curve for an outbreak of Chikungunya fever in north east Italy

case definition has been decided upon and the cases have been identified.

The shape of the epidemic curve often provides information on the most likely mode of transmission. A point source outbreak in which all cases were infected at a particular time by the same source usually produces a sharply rising curve with a single early peak followed by a gradual decrease in the number of new cases. If the population has been exposed to the same source of infection over a lengthy period (e.g. where a food product contaminated with *Salmonella* has stayed on the supermarket shelf or been traded for a long time, or where people have been exposed to a cooling system containing *Legionella*), the curve will be much wider. An outbreak caused by person-to-person transmission (e.g. in the case of influenza or measles) produces an epidemic curve with several smaller peaks, the distance between which is the average incubation period. In practice, however, a characteristic pattern of this kind is often difficult to visualize because of variation in the latent period following infection, which blurs the pattern of peaks.

An epidemic curve also provides information on the course of the outbreak. As long as it remains in the rising part of the curve, transmission is likely to be continuing, whereas a falling line indicates that the outbreak may have already been halted. The epidemic curve is of course very important when deciding whether measures need to be taken, and if so, what measures. It can also furnish information for recommendations to administrators, healthcare providers, policy-makers and the public.

8.3.2 Cohort studies

As outbreaks usually span a short period of time, the effect of changes in the composition of the population at risk is negligible. Disease frequencies are therefore generally expressed in the form of attack rates. An attack rate is actually a cumulative incidence, i.e. a proportion, and consequently strictly

Table 8.1 Numbers of cases and attack rates for foods eaten by 226 attendees at a cold buffet

	eaten (attack rate)	not eaten (attack rate)	difference in attack rate	relative risk
Russian salad	116 (73%)	18 (27%)	+46%	2.7
kiwi fruit	12 (100%)	123 (57%)	+43%	1.8
prawn cocktail	34 (74%)	101 (56%)	+18%	1.3
French cheese	46 (53%)	89 (64%)	−11%	0.8
stuffed tomatoes	49 (82%)	86 (52%)	+30%	1.6

speaking not a rate in the sense that the incidence density is a rate (see ▶ 2.5.2). The difference in attack rates between people exposed and people not exposed to a particular determinant is therefore the attributable risk (▶ 3.2.1). Similarly, the ratio between attack rates is the relative risk (see ▶ 3.2.2).

By comparing attack rates in cohorts of exposed and non-exposed persons we can test hypotheses concerning potential causes of the outbreak. Table 8.1 gives an example of a salmonellosis outbreak among 226 people who attended a large buffet. From the table we can conclude that the Russian salad was the most likely source of the epidemic. Although in this example eating kiwi fruit also entailed an increased relative risk of 1.8, only 9% of the patients had eaten it, as against 90% who had eaten the Russian salad. The relative contribution of an exposure to the epidemic is expressed as the population attributable risk (see ▶ 3.2.6). We find that eating kiwi fruit was responsible for only $0.09 \times (1.8\text{-}1)/(0.09 \times (1.8\text{-}1)+1) = 0.07$, i.e. 7% of the cases, compared to 60% for the Russian salad. Once we have tracked down the most likely source, for the further investigation we need to know what microorganism caused the cases. Identifying the source is not always easy, even in the case of a point source outbreak, as there can be interaction between factors (human, environmental, microbiological).

8.3.3 Case-control studies

In the example shown in tab. 8.1 it was clearly possible to draw up a list of the attendees at a large buffet, on the basis of which a cohort of potential cases could be defined, but that was an exception. Usually it is not possible to assemble a cohort of potentially exposed individuals. If an outbreak of Q fever is diagnosed in a rural area, for instance, it will not be possible to identify everyone who has been exposed, so a case-control study will be needed. Another situation is when the incidence of the disease in an outbreak is low and cases are spread throughout the population. In these situations case-control studies are the most efficient design to study the causes of an outbreak. This type of design is discussed in detail in ▶ 4.3, but there are a few points of particular interest when carrying out outbreak investigations.

As an investigation into a particular outbreak will usually include only cases occurring in a particular period and a particular area (in other words, the case definition will be restricted in terms of time and place), the same restriction will need to be applied to the control group. Controls should ideally be selected from the population who would have been included in the case group if they had developed the disease (see ▶ 4.3.2). So in the case of an outbreak among visitors to a pop music festival the controls too should be sought among those visitors. However, their selection should be independent of exposure to the suspected determinant(s) of the disease. For example, if we want to find out whether playing on contaminated land could be the cause of an outbreak of poisoning among children, playing or not playing on contaminated land must not influence the likelihood of being included in the control group.

The longer we wait before investigating possible exposures, the more difficult it becomes to obtain

accurate data. If we are interviewing patients and controls, for instance, and controls remember less about possible exposures than patients, there is a risk of recall bias (differential misclassification: see ▶ 5.3.1). Unlike in epidemiological research into chronic diseases, in a case-control study of an outbreak we are in a hurry to collect the data, and that in itself poses a threat to the quality of the study. It is important, therefore, to prepare properly and use standard instruments and trained personnel for data collection, analysis and interpretation, so as to achieve adequate quality in studies of this kind.

Case 8.3 *Legionella* in Bovenkarspel

Legionnaires' disease was first described following a major outbreak of pneumonia among people who attended an American veterans' reunion in 1976 in Philadelphia. The study led to the discovery of a hitherto unknown organism in the patients' sputum, *Legionella pneumophila*. The bacteria had evidently spread via the conference hotel's air conditioning system. Following that first outbreak many more outbreaks of Legionnaires' disease were described. More than 40 different species of *Legionella* have now been identified, but *L. pneumophila* serogroup 1 has been found to be the main pathogen.

The north west of the Netherlands was hit by a very large Legionnaires disease outbreak in the town of Bovenkarspel in 1999. The first patient was admitted to the regional hospital in Hoorn with severe pneumonia on 6 March 1999, and a week later the number of cases had risen to twelve. Eight of them were in such a poor state that they needed ventilation at the intensive care unit. The doctors treating them initially thought it was a viral infection. The hospital alerted the regional health authority, and in view of the seriousness of the situation the national Institute for Public Health and the Environment (RIVM) immediately initiated urgent consultations with various medical specialists and public health representatives. Based on the symptoms, a fast urine test for *Legionella pneumophila* was carried out and yielded positive results in seven patients. The same bacterium was subsequently detected in other patients. As half of the patients were geographically clustered around Bovenkarspel a link was suspected with the flower show that had just ended there. An exploratory case-control study carried out by the regional public health authority showed that patients were more likely than other local residents to have visited the show. Given the large number of visitors (80,000) a major alert was sparked.

By the end of April 1999 233 patients had been notified as ill after visiting the show in Bovenkarspel. Legionnaires' disease was confirmed in 106 of them, while 48 had a probable and four a 'possible' diagnosis. Twenty-three of the patients died. A more detailed investigation of the exhibition sites where the patients had been led to a serious suspicion that the source of the outbreak were whirlpool baths on display which were contaminated with *Legionella*. This was confirmed by tests on water pipes, taps and equipment. The bacteria found in the whirlpool baths were of the same type as those obtained from patients. Since this outbreak, regulations for legionella prevention and inspection of public facilities where water is distributed have been intensified to prevent new outbreaks. It is inconceivable, however, that the bacterium itself can be eliminated from the environment.

8.3.4 Risks to individuals and populations

Infectious diseases are caused by microorganisms that are transmitted to a host where they can multiply. An individual can be infected by a source in the vicinity or by another infected person. Many microorganisms can be transmitted from person to person, thus creating transmission chains in the population – a typical characteristic of infectious diseases. Over time this person-to-person transmission of pathogens leads to dynamic changes in transmission rates in the process that causes disease. Because of these time-dependent events, which are unique to infectious diseases, we need to distinguish between different levels of disease risk.

An individual coming into contact with a source has a certain likelihood of exposure to the infectious agent. Once this exposure has taken place, the person has a certain likelihood of infection by the agent. The infected individual then has a certain likelihood of getting the disease, and finally a certain likelihood of dying from the disease. Knowing and reducing these risks is in the interest of the individual and his or her contacts, and this is what the prevention and treatment of cases focuses upon. Avoid an individual being exposed, ensure that once exposed the person cannot become infected, prevent the disease developing in infected persons by post-exposure prophylaxis and prevent them from transmitting the infection to others ('treatment as prevention'), and lastly ensure that manifest cases are cured as quickly as possible.

From the point of view of the population, of course, the likelihood of exposure and the likelihood of infection among exposed persons are also important, but the next most important thing for the population is the likelihood of infected persons in the population being infectious and the likelihood of them transmitting the pathogen to other people in the population during the infectious period. These factors determine the likelihood of an outbreak of a disease in the population and the likelihood of the epidemic increasing, decreasing or disappearing completely.

Epidemiologists study the epidemiological function at each of these levels: the likelihood of infection at various levels of exposure, the likelihood of disease given an infection, the likelihood of contagiousness, the actual transmission at various levels of infection, and so on. Given knowledge of these factors it is then possible to carry out targeted interventions with the aim of reducing these likelihoods for the entire population. Data might be used and interpreted differently, depending on whether one is concerned with the likelihood of disease in an individual with a particular risk profile in the context of clinical epidemiology or with the likelihood of an outbreak of the infectious disease in the population in a public health context.

Transmission by infectious persons, for instance, depends on how long cases are contagious, the level of contagiousness and the number of contacts they have with susceptible persons. So, given a particular degree of contagiousness, transmission depends on patients' behaviour and the degree of protection that their contacts have. Individual patients who have been treated for bacillary dysentery (shigellosis) will sometimes excrete *Shigella* bacteria in their faeces for several months, even after their symptoms have cleared up. An outbreak can therefore easily develop, for instance, among children at a nursery school where hygiene rules are not implemented sufficiently.

8.4 Stepwise approach to outbreak investigation

Although no two outbreaks are the same, the response to an outbreak or potential outbreak will nevertheless usually involve the same steps. Various international organizations such as the CDC, ECDC and WHO have developed roadmaps that can be used as a guide for situations where an outbreak is suspected.[6] A formal outbreak investigation will not always be needed: if it is immediately clear what is going on, what the source or cause is and which control measure suitable, the standard approach based on existing plans and protocols will suffice. If not, a systematic, structured outbreak investigation is needed to avoid unnecessary mistakes (e.g. time wasting, wrong measures, missed opportunities to gather unique information, loss of support among administrators, policy-makers and healthcare providers, and panic in the population). The roadmaps can be very useful in these situations.

It is important, however, for the person investigating the outbreak to have sufficient knowledge and expertise regarding diseases that can manifest themselves in an outbreak and the determinants that could be involved, as there is little time or opportunity to gain that knowledge once the outbreak has started. Bring in experts who have this knowledge at their fingertips if necessary. Ensure that proper administrative and logistics procedures are in place in advance, in the form of an outbreak protocol including the names and telephone numbers of relevant authorities and people. It must be clear from the

6 Outbreak investigations: 10 steps, 10 pitfalls (website). ▶ http://bit.ly/1yWfDwV

outset who is responsible for what, including who is in overall charge of the outbreak investigation and the control measures. In practice this will often be a physician expert in infectious diseases from the regional public health authority. If an outbreak covers more health regions the national centre for prevention and control of infectious disease will take charge. Ensure clarity about who is ultimately responsible for administration and communication with the press, the authorities and the public.

The various steps in an outbreak investigation are as follows:

Step 1: Establish that there is an outbreak

Determining that there actually is an outbreak involves comparing the observed number of cases with the expected number, the latter being based on the number of cases reported in other periods or at other sites. Even if the observed number of cases is higher than the expected number this does not necessarily mean that there actually is an outbreak. First we need to verify, for example, that the disease diagnosis and recording system remained unchanged, that the definition of a case is the same (i.e. no changes in the numerator of the epidemiological fraction), or that the population size was stable (i.e. no changes in the denominator). Random fluctuation also needs to be ruled out as much as possible.

Step 2: Verify the diagnosis

The aim of this step is to ensure that diagnostic or laboratory errors can be ruled out as an explanation for the increase in the number of cases. This involves studying the medical records and lab results for the reported cases. In order to control the outbreak the microbiological diagnosis needs to be ascertained quickly. Regional (or other) laboratories can provide fast diagnosis of many but not all infectious diseases using the polymerase chain reaction (PCR). The sooner the outbreak is detected and confirmed by microbiological diagnosis, the sooner steps can be taken to control it. It may also be wise at this stage to interview some patients with the disease to form your own hypotheses of the nature of the outbreak – taking the necessary health protection measures, of course.

Step 3: Decide upon the case definition

The aim of a good case definition is to include as many genuine cases of the outbreak as possible and as few unrelated cases as possible, as the latter group causes misclassification, thus reducing the likelihood of tracking down the cause of the outbreak. Use for a start a broad definition so as not to miss any cases, and collect as much information as possible on these patients so as to gain a good idea of what is going on. Soon afterwards we may apply a more restricted operational definition of which patients are part of the outbreak, including symptoms, signs, lab results, specific personal characteristics, places and times. If necessary, cases can be divided into 'probable' cases and cases 'confirmed' by laboratory results. National (or international) agreed definitions help to accomplish better comparability and easier interpretation.

Step 4: Identify the 'right' cases

Find the 'right' cases, in and outside any existing surveillance systems, that meet the definition decided upon in Step 3 and list them in a table. This should include at least data on the date of the first symptoms, the date of diagnosis, age, gender, geographical location, any specific symptoms and any proof of infection (from laboratory results).

Active detection of all cases is the basis for determining the extent of the outbreak. One way of achieving this is to actively ask local GPs, specialists and clinical laboratories to notify cases known to them and watch out for new cases. The public can also be asked to report new cases. Researchers must have a good relationship with the press so that requests of this kind can be made without creating panic.

Step 5: Describe the incidence of the cases in terms of time, place and personal characteristics

Initial descriptive analysis can be carried out using the table from Step 4. First draw and interpret the epidemic curve: this may provide indications of the nature of the outbreak, when exposure took place and about the course of the outbreak (see ▶ 8.3.1). If all the cases are plotted on a map of the area in question, this will show at a glance whether there are any geographical similarities that suggest a possible source of infection or transmission. ▶ Case 8.3 shows

how this can be used to develop hypotheses about the possible cause. It was in this way that John Snow discovered the source of contamination in the London cholera epidemic (see ▶ Case 1.5). In some outbreaks this descriptive step will already lead to hypotheses about factors that could explain the outbreak.

Step 6: Formulate hypotheses

Interviews with the patients may produce further ideas about the possible nature and source of the infection. If it is a 'standard' outbreak, where the pathogen is known but not the source, a standard questionnaire should be used for the interviews. If the pathogen is unknown, the interviews will need to be more exploratory. Experts and the medical literature may also provide hypotheses about the possible cause of the outbreak. Cases that do not quite fit in with the characteristic pattern of the epidemic curve can sometimes be a valuable source of information. An isolated case where the incubation period occurred earlier than in the other cases could identify the cook who unwittingly contaminated a food product. Many countries' national centres for prevention and control of infectious diseases carry out recurrent questionnaire surveys of a sample of the general population about possible risk factors. This sample can provide controls for patients notified as possibly having been exposed to sources. Hypotheses should ideally include the source, the mode of transmission and the determinant or determinants that is/are causing the disease. This step will often be the last one in a small-scale outbreak investigation (with a known microorganism and source). In other words, not all the steps will be required for every outbreak.

Step 7: Conduct an initial analytical study

The hypotheses on possible causes of the outbreak developed in this way are used in the preparation of an epidemiological study. This will often be a case-control study (see ▶ 8.3.3). It will indicate the greater or lesser likelihood of the possible explanations and enable the next steps to be organized.

Step 8: Adjust the hypotheses and carry out additional research

The hypotheses that remain intact and/or were adjusted during the previous step provide the basis for formal further research. This could take the form of epidemiological studies or studies with the aim of identifying exposure to the suspected cause. A visit to the location of the presumed pathogen (inspection of the kitchen, or a visit to the company or neighbourhood from which the patients come) can provide a lot of information and secure relevant 'evidence'. It is very important to take samples from patients, food, animals and the environment as soon as possible, as otherwise the causative agent may have disappeared, and because people are more likely to cooperate while the outbreak is still going on and the symptoms are still present. Incoming data should be checked for completeness and consistency and entered in a database. As outbreaks are usually short-lasting and the changes in the population at risk will therefore be negligible, the disease frequency is generally expressed in the form of attack rates (see ▶ 8.3.2).

Step 9: Make the findings known to the responsible people and the public

Effective, clear and timely information is vital in any investigation of an outbreak, for example:

- Informing the public of the nature of the disease and how it can be identified, treated and prevented. Press releases and social and other media can be used to disseminate this information. It goes without saying that the information should be phrased in language that is clear and easy to understand. It may be necessary to adapt it for particular subgroups or to set up special public information campaigns to target particular groups. Outbreaks also provide an opportunity to highlight important public health messages again.
- Informing the investigation team. All the staff involved should be given daily up-to-date information on the course of the outbreak and the investigation into the causes.
- Informing the responsible administrators. To enable them to fulfil their responsibilities, the people responsible for public health and the general administration of the population in question

should be given information on the progress of the investigation daily.
- Informing local GPs, specialists (insofar as relevant to the particular disease) and other healthcare workers who could come into contact with patients or people suspected of having the disease.
- Informing other authorities who could become involved with the outbreak: institutions or areas elsewhere with similar exposures, and also authorities indirectly involved, e.g. the Ministry of Infrastructure and the Environment, water companies, the Food and Consumer Product Safety Authority and environmental services.
- Informing fellow researchers. National infectious disease control centres will often have an electronic messaging network in place for this purpose.

Step 10: Take steps to control the outbreak and prevent new outbreaks

Investigations into outbreaks usually lead to a particular cause and/or source. What preventive measures can be taken to prevent future outbreaks will depend on the particular determinant and source and what is already known about the transmission etc. of that determinant. There are two types of interventions:
- Removing the source or agent (e.g. taking a product off the market, destroying an infectious animal or changing health protection measures).
- Protecting people who are neither ill nor infected, e.g. by means of vaccination, early treatment (post-exposure prophylaxis), screening of contacts, immunization, personal protection, personal hygiene and/or isolation of cases.

The sooner measures are taken, the greater the chances of bringing the outbreak under control and preventing new outbreaks. Surveillance systems can then be used to assess whether the measures taken have been effective, also in the long term.[7]

7 'Ebola: The Plague Fighters'. A moving documentary about how epidemiologists from the Center for Disease Control dealt with the advancing African Ebola epidemic in 1995-7. ▶ http://bit.ly/1lRawcp

8.5 Interpreting data on supposed outbreaks remains difficult

In practice there may be disagreement as to whether there is an outbreak and whether the correct source has been identified, as many kinds of bias can occur in outbreak investigations that make correct interpretation difficult:
- How is the affected population and/or outbreak area defined? False alarms may occur if definitions are based on accidental, biased or vague observations. There might be a tendency, for instance, to narrow the boundaries of the area or population down based on a selection of cases. A strict definition of the population will reduce the number of expected cases. As a consequence excess cases will become more apparent and might even produce significant results, even in situations of normal variation in disease incidence.
- How is the outbreak period defined? There may be seasonal variation, and this too can vary from one year to another, which makes it difficult to interpret a temporary increase in disease frequency. If the outbreak is observed over a longer period, random variation will be less of a problem, but important signals may be missed. Consider for instance the situation where an outbreak leads to earlier deaths among individuals who are at increased risk of mortality. Mean death rates covering a longer period might miss this phenomenon ('early harvest') because a temporary mortality peak are compensated by a period of lower mortality.
- What led to the hypotheses about possible causes? A poorly formulated hypothesis – or even having no hypothesis at all – is likely to result in false alarms or actual risk factors being missed.
- How good are the data? Inaccurate data on cases and missed cases make it difficult to identify the causes of outbreaks. Reported cases that emerge once a suspected cause of the outbreak has been identified can also cause bias. Lastly, the absence of good quantitative information on the nature and duration of exposure to risk factors will limit this kind of investigation. It will often no longer be possible to obtain these data, as the source of the exposure is gone and forgotten.

- Have all the confounders been taken into account? Confounding can cause serious bias, just as in other types of epidemiological research.
- Are the researchers being given enough time and opportunity to carry out the outbreak investigation? There is almost always public pressure to take action quickly, with the result that there is less time and attention available for systematic assessment of the problem and the impact of the measures taken.

Case study 8.4 Cancer cluster in Mountain View, California, USA

Trichloroethylene (TCE) is an industrial solvent associated with liver cancer, kidney cancer, and non-Hodgkin's lymphoma (NHL). Concerns in public media in 2014 about a possible excess of cancer diagnoses due to TCE exposure in the residential area of Mountain View, California prompted the Greater Bay Area Cancer Registry (GBACR) to examine the occurrence of TCE-associated cancers in this area. The US Environmental Protection Agency (US-EPA) has identified this region as an area where large volumes of TCE have been released for decades by several industrial, manufacturing and military operations. Continued attempts of US-EPA in this period to oversee site cleanup and monitoring had only limited effects on the reduction of the TCE burden in this area.

The GBACR evaluated incidence data for NHL, liver and kidney cancers in three time intervals: 1988-1995, 1996-2005 and 2006-2011 using population denominators obtained from the 1990, 2000 and 2010 US censuses respectively. To determine whether TCE-associated cancer occurrence was unusually high in the suspected area, the number of observed cases was compared to what would be expected based on the cancer rates in the entire Santa Clara Region, taking into consideration the race, gender, and ages of people who were diagnosed with cancer. Results were expressed as a standardized incidence ratio (SIR, see ▶ 2.5.3) with a corresponding 99% confidence interval.

There were no statistically significant differences between the neighbourhood of interest and the larger region for cancers of the liver or kidney. A statistically significant elevation was observed for NHL during the 1996-2005 period (SIR = 1.8, 99% CI 1.1-2.8) but not for the earlier 1988-1995 (SIR = 1.3, 99% CI 0.5-2.6) or later 2006-2011 (SIR = 1.3, 99% CI 0.6-2.4) periods.

The authors concluded that there was no evidence of a consistent, sustained or recent elevation of TCE-related cancer occurrence in this neighbourhood. The authors did not dispute that exposure to TCE is associated with adverse health effects. Because the time frame between exposure to a carcinogen and onset of disease can be many years, it is not possible to link an elevated incidence rate in a given time period to a specific cause or putative exposure with these data. The authors further acknowledged that cancer registry data do not allow firm conclusions about specific exposures incurred by past or current residents of this area, nor the effectiveness of ongoing cleanup efforts occurring in and around the neighbourhood of interest. However, if there were a major increase in cancer among residents who lived in this area for a long period of time, it could probably be detected in this type of investigation.

8.6 Special approaches are sometimes needed to study outbreaks and clusters

While the foregoing has been specifically concerned with outbreaks, even in situations of apparent equilibrium infections can have major effects on public health and epidemics can develop insidiously, due not only to human-to-human transmission but also to contamination of food or the environment or to the use of unsafe material in the health service. Epidemics of this kind (clusters) can go unnoticed, as the time dimension is different and the effects are less pronounced. Carefully designed epidemiological research should ensure that these signals are nevertheless discovered.

When analysing data from outbreak investigations classic methods of statistical analysis can gen-

erally be used: we translate the epidemiological problem into a logistic regression model and estimate the regression coefficients as a measure of the effect of the determinant on the occurrence of the disease, adjusting for confounders and indicating the precision of the estimate in terms of confidence intervals.

As the classic epidemiological methods are not always adequate when studying infectious diseases – and clusters in general – researchers sometimes have to apply special methods and techniques, the most common of which are briefly discussed below.

8.6.1 Models and simulations if the reality is too complex

Models can be very useful in both outbreak situations and situations of equilibrium. The great advantage of models is that problems due to random variability are non-existent. This can also be a disadvantage if the model becomes too simple and ignores all sorts of relevant variations (e.g. due to differences in dose, age, personal protection).

Infectious diseases behave dynamically in the population, and while these processes follow certain laws it cannot be captured in a simple model. ▶ Par. 8.6.2, for instance, looks at the basic reproduction number, a vital concept in explaining the rise and fall of outbreaks and infectious disease epidemics covering longer periods. As the description there shows, this process is influenced by so many factors that complex infectious disease models are needed to describe it fully. These models can be deterministic, i.e. describable in an algebraic equation, but more often dynamic transmission models (with random elements) provide better descriptions and better ways of estimating the effects of interventions. Take the average age at which an infection is contracted, for example. If many people are immunized against a particular infectious agent, the incidence of the disease will drop, thus increasing the average age at which infections occur among people who are susceptible. Some diseases (e.g. mumps, measles and rubella), however, are relatively serious when they occur in later life. A vaccination campaign can therefore reduce the total number of cases while increasing the number of severe cases. Complex processes of this kind are studied using dynamic transmission models, which can subsequently be used to evaluate various intervention scenarios. Additional models can then be used to estimate the cost-effectiveness of particular interventions before taking decisions on the matter.

Models are also needed to take the various contact patterns in a population into account in the calculations. Simple models are based on the assumption that the susceptible individuals in a population are distributed at random and therefore are at the same risk as anyone else of coming into contact with an infectious individual and becoming infected. This assumption is usually wrong, of course, so that more complex contact patterns in a population need to be described. In the case of virtually all infectious diseases there is considerable heterogeneity in the disease transmission rate in a population. As certain groups in the population come into close contact with one another to a greater or lesser extent, the mathematical models for the spread of a disease need to take different contact and transmission rates into account for different groups. Sexually active homosexual men played a major role in the initial spread of HIV/AIDS in Western countries, for instance. In the case of influenza outbreaks, on the other hand, nursery schools play a major role in spreading the disease, as they bring children from different neighbourhoods into close contact with one another. This heterogeneity in disease transmission can be taken into account by using different transmission rates in the models for the subgroups in question.

A model that is often used to study the course of a directly transmissible infectious disease (e.g. polio, mumps, measles, rubella, whooping cough) is the SLIR model. This is based on dividing the people in a population into four compartments, which represent the four possible stages that a person can occupy: susceptible (S), latent (L: infected but not yet infectious), infectious (I) and resistant (R: cured and immune). The SLIR model is a dynamic model for studying the likelihood of transitioning from one stage to the next. The model is described by a system of differential equations based on the particular characteristics of the infectious disease in question, the demographic characteristics of the population, the likelihood of contact between infectious and susceptible individuals and the microorganism's trans-

mission rate when contact takes place. The likelihood of transitioning from one stage to the next is modelled and estimated using empirical data from the literature or the researchers' own research. Various subgroups can of course be included in the models separately so as to reflect the heterogeneity in disease transmission.

In addition to the classic SLIR model, which is based strongly on population averages and various assumptions, there are other types of approach, such as microsimulation. This simulates an epidemic in a hypothetical population, with realistic parameters entered for the behaviour of microorganisms, people and so on. Given these assumptions, infectious disease risks can be estimated with great precision.

Case 8.5 A meningitis outbreak in Sudan

An epidemic of meningitis emerged in north west Sudan in January 1999. The first case was diagnosed in mid-January, and in the ensuing weeks the disease spread rapidly throughout the country and the neighbouring areas. The climatic conditions in Sudan are ideal for the spread of meningococcal meningitis for a large part of the year. Given the rapid spread and the country's experience of previous epidemics, Médecins sans Frontières (MSF) was soon called in to provide assistance. One of the MSF teams was sent to West Darfur, an area with a population of around 550,000 that had been hit badly by civil wars and was cut off from the rest of the country during a few months of the year because of poor infrastructure. The first cases were seen in this area at the beginning of February, and by Week 8 the number had risen to 15 per 10,000 per week. This area was therefore regarded as suffering an epidemic, in line with the WHO threshold of 15 per 100,000 per week. The MSF team arrived to carry out an exploratory mission that week, including taking lumbar punctures. *Neisseria meningitidis* group A was isolated from the samples, and preparations for a vaccination campaign began in Week 10. Within a few days eight vaccination teams had been trained and fully equipped with vaccination supplies. The public were informed via the radio, newspapers, mosques and public address systems. Meningitis patients were treated effectively with chloramphenicol in oil, administered intramuscularly. Vaccination started in Week 11, with up to 12,000 people a day being vaccinated in the urban areas.

Health workers in rural areas were trained in using the case definition and the medication and in providing information to the public. They also filled in surveillance forms, which were sent to the coordination centre once a week. This enabled the epidemic to be mapped out and further vaccination activities to be organized effectively.

The case definition was as follows:

- Suspected in the event of rapid-onset fever, stiff neck or rash of small red patches since February 1999 in the province of West Darfur (symptom in infants: bulging fontanelle)
- Probable if a suspected case has a cloudy lumbar puncture
- Confirmed if a suspected case has a positive culture or antigen detected in the lumbar puncture.

In spite of the poor infrastructure, this enabled an impression of the epidemic in the province to be gained quickly. The epidemic was brought under control within eight weeks of the start of the campaign, by which time 200,000 people had already been vaccinated. Of the 755 cases diagnosed, 106 died, representing an attack rate of 0.02% and a case fatality rate of 14%. Without vaccination the attack rate could have risen to 1 or 2%.

8.6.2 The basic reproduction number: a key concept when describing the infection level in the population

The basic reproduction number (R_o) is the average number of new infected individuals that one contagious individual can produce in a population in which everyone is susceptible to the infection. Measles is an example of a highly contagious dis-

Table 8.2 The basic reproduction number for some familiar infectious diseases

disease	R_0
measles	12–18
whooping cough	12–17
diphtheria	6–7
smallpox	5–7
polio	5–7
rubella	5–7
mumps	4–7
HIV/AIDS	2–5
SARS	2–5
influenza	2–3
ebola	1–4

ease, with an R_o of approximately 16 (see tab. 8.2). This means that during the period of contagiousness a child who has measles can successfully infect sixteen new children in a population of children who are all still susceptible to the disease. Ebola, on the other hand, is a far less contagious disease, with an R_o of 2. To be able to cause an outbreak R_o must be greater than 1. R_o is the product of the transmission rate (p), the number of contacts between infected and uninfected people per unit of time (c) and the duration of the contagiousness (D).

$$R_0 = p \times c \times D$$

The basic reproduction number is difficult to interpret as such, as R_o is based on the assumption that all contacts will produce potential patients, whereas in reality some contacts will be with people who are immune, or partially immune, to the pathogen. This leads us to the effective reproduction number (R):

$$R = R_0 \times x$$

where x is the proportion of people susceptible to the disease in a well-mixed population. If the basic reproduction number for measles is 16 and half of the children in the population are immune, a child who has measles will only be able to infect eight new (contagious) cases of the disease.

It is not so easy to estimate the basic reproduction number directly, so R_o is usually estimated indirectly. In a stable situation where the incidence and prevalence of the infectious disease do not change, for instance, each case will produce one new case on average. In this case the basic reproduction number is the reciprocal of the proportion of people susceptible to the disease. We can also take advantage of the fact that in a stable situation the average age at which the infection occurs depends on the basic reproduction number: the lower R_o, the higher the average age. The reverse is true of life expectancy: the higher the average life expectancy, the higher R_o. A rough indication of R_o can be gained by these indirect methods.

The basic reproduction number is a complex concept that includes various important characteristics of the behaviour of an infectious agent in a population, and which helps to understand many important features of the prevention and control of epidemics. To halt an outbreak of an infectious disease the reproduction number needs to be brought below 1. Suppose $R_o = 5$ for HIV infection in a particular population and use of condoms reduces the transmission rate (p) by 90%, the basic reproduction number could be reduced to e.g. 0.5 if condom use was frequent. An outbreak will also be halted if the fraction of people who have become immune is large enough. The size of the immunized fraction (f) needed to prevent an epidemic can be calculated in this way. As f=1−x:

$$R = R_0 \times (1 - f)$$

To prevent an outbreak R must be less than 1. This means that:

$$f > 1 - \frac{1}{R_0}$$

In the measles example with $R_o \approx 16$ this yields 1−(1/16) = 0.94: in other words, 94% of the children at risk of initial infection must have been immunized (naturally or by vaccination) to prevent a measles outbreak. This is only true, of course, if the immunized children are evenly distributed in the population. The foregoing leads us to an important concept in infectious disease control, namely group protec-

tion ('herd immunity'). This describes the protection status of the population (as opposed to that of an individual) with regard to a particular infectious agent. A population is 'immune' if the reproduction number R is less than 1. This does not necessary require all the individuals in the population to be immune: the more people have been vaccinated against the agent in question or have been become immune as a result of infection, the higher the population protection.

Recommended reading

Bonita R, Beaglehole R, Kjellstrom T. Basic epidemiology. 2nd ed. Geneva: World Health Organization; 2006.

Elliott P, Wakefield J.C, Best N.G, Briggs D.J. Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press; 2000.

Giesecke J. Modern infectious disease epidemiology. 2nd ed. London: Arnold Publishers; 2002.

Gregg M.B. Field epidemiology. 3rd ed. New York: Oxford University Press; 2008.

Koepsell T.D, Weiss N.S. Epidemiologic methods: studying the occurrence of illness. New York: Oxford University Press; 2003.

Morabia A. A History of Epidemiologic methods and concepts. Basel: Birkhäuser Verlag; 2004.

Murray C.J, Lopez A.D. The global epidemiology of infectious diseases. Cambridge: Harvard University Press; 2000.

Nelson K.E, Masters Williams C. Infectious disease epidemiology: theory and practice. 3rd ed. Boston: Jones and Bartlett Learning; 2013.

Rothman K.J, Greenland S, Lash T.L. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2012.

Rezza G, Nicoletti L, Angelini R, Romi R, Finarelli A.C, Panning M, Cordioli P, Fortuna C, Boros S, Magurano F, Silvi G, Angelini P, Dottori M, Ciufolini M.G, Majori G.C, Cassone A, for the CHIKV study group. Infection with chikungunya virus in Italy: an outbreak in a temperate region. Lancet 2007;370:1840–46.

Thomas J.C, Weber D.J, eds. Epidemiologic methods for the study of infectious diseases. New York: Oxford University Press; 2001.

Webb P, Bain C. Essential epidemiology: an introduction for students and health professionals. 2nd ed. Cambridge: Cambridge University Press; 2011.

Source references (cases)

European Centre for Disease Prevention and Control. Annual epidemiological report 2014 – vaccine-preventable diseases; 2014 (Case 8.1).

Holland W.W, Olsen J, du V. Florey C. The development of modern epidemiology: personal reports from those who were there. Oxford: Oxford University Press; 2007 (Case 8.2).

Den Boer J.W, Yzerman E.P.F, Schellekens J, Lettinga K.D, Boshuizen H.C. A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. Emerg Infect Dis. 2002;8:37–43 (Case 8.3).

Press D.J, McKinley M, Deapen D, Clarke C.A, Gomez, S.L. Residential cancer cluster investigation nearby a superfund study area with trichloroethylene contamination. Cancer Causes Control 2016;27:607–13 (Case 8.4).

Personal communication from R. Appels, University Medical Center Groningen (Case 8.5).

Diagnostic and Prognostic Research

9.1 Introduction

This chapter looks at epidemiological research for diagnostic and prognostic purposes. Diagnosis and prognosis are two of the basic elements of medicine and paramedicine. The Greek word 'diagnosis' means 'distinction', and the aim of the diagnostic process is indeed to distinguish between health and disease in individuals, between different diseases that appear similar, or between different stages of a particular condition in patients. The aim of prognosis is to predict the course or outcome of a disease process after the patient has been diagnosed: it is concerned with such questions as the likelihood of cure, of permanent disability or of death within a particular time frame. Both the diagnostic process and prognosis are descriptive in nature and relate to individuals. This is the main difference between this chapter and the other chapters: here we are not interested in explaining things in terms of cause and effect. On that topic we would refer to the chapters on observational and experimental research. This chapter is about describing the likelihood that a particular disease or health outcome is present (diagnosis) or will occur (prognosis) in an individual. Taking it to extremes, if we were to find that the length of the big toe shows whether a person will recover completely from a cerebral infarction, the length of the big toe would have major prognostic value. Whether the length of the big toe is also a causal factor is irrelevant.

9.1.1 Diagnostic and prognostic research are needed for clinical decision-making

Diagnostics and prognostics are not an end in themselves; they provide information on the basis of which patients and professionals take decisions. Diagnostic and prognostic information is vital when deciding whether to initiate treatment, and if so, what kind. Even if there is no suitable or acceptable treatment, diagnostic and prognostic information will often be useful to an individual patient, for example to assist with coping processes or behaviour modification. Diagnostic or prognostic information can also be useful when deciding whether the patient should undergo further diagnostic testing. The results of diagnostic and prognostic research can have far-reaching consequences, so it is essential that the findings are correct. We do not want to miss serious diseases or undesirable outcomes, but nor do we want diagnostic and prognostic procedures to give rise to incorrect suspicions that the patient has a condition or will have an undesirable outcome (fig. 9.1). The quality of the diagnostic and prognostic process is therefore also an object of epidemiological research.

When used in healthcare, diagnostic and prognostic processes must not only be of good quality, they must also be efficient: stressful tests must be minimized, results must be produced quickly and the cost to the individual and the community must be kept down. The requirement of good quality and the desire to reduce burden and cost are usually incompatible, as improving quality often means increased burden or cost. Epidemiological research into diagnostic or prognostic processes aims to find the minimum set of diagnostic (or prognostic) factors by which to distinguish individuals who have/will develop (or are likely to have/develop) a particular disease or outcome from individuals who do not have/will not develop (or are not likely to have/develop) that disease or outcome. In other words, diagnostic and prognostic research is research into the probability of a disease or outcome and the minimum set of determinants needed to calculate that probability.

Diagnostic and prognostic determinants generally fall into three categories:

1. Symptoms and signs
2. Results of laboratory and other tests
3. Patient characteristics (e.g. age and gender).

The purpose of history-taking (interviewing the patient) is to identify the symptoms observed by the patient. The practitioner carries out a physical examination to look for signs. This clinical investigation – history-taking and physical examination – can be supplemented with observations using imaging techniques such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI) or ultrasound, and lab tests such as haematological tests of blood parameters (e.g. haemoglobin), biochemical tests of the molecular composition of

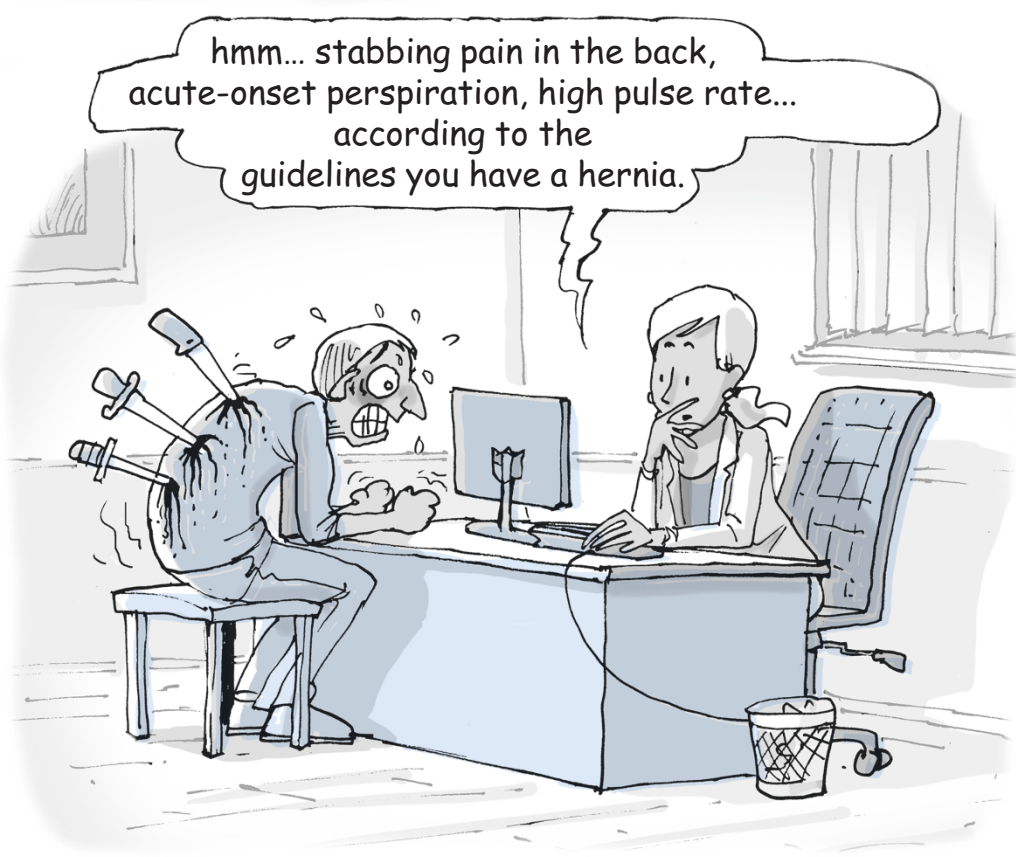


Figure 9.1 The practice of diagnosis

body fluids, tissues and excreta (e.g. serum, urine, faeces, fatty tissue, liver tissue, nails, hair, saliva and cerebrospinal fluid).

9.1.2 Diagnostic research is carried out in phases

Just as research into the effects of pharmaceuticals is not a single-step process, careful assessment of the quality and clinical utility of a diagnostic test requires a series of studies. The research required when developing a diagnostic test falls into four phases:

- Phase I: comparing the test results of patients with a clearly developed stage of the disease being diagnosed with healthy controls (a case-control study with contrasting disease stages).
- Phase II: comparing the test results of patients with the disease and controls with a variety of other diseases (a case-control study with related conditions).
- Phase III: comparing test results over the entire spectrum of the disease being diagnosed, with or without symptoms, mild and severe, with and without comorbidity, with different anatomical, microscopic and etiological characteristics (a full-spectrum case-control study).
- Phase IV: carrying out the diagnostic test on a large series of successive patients whose clinical presentation warrants it (a prospective cohort study of patients representative of the target group).

Table 9.1 Sources of variability in a measured diagnostic parameter

source		notes
actual (biological) variability	within individuals	changes in individuals in relation to time and test situation
	between individuals	biological differences between individuals
apparent variability (measurement errors)	measuring device	measuring device not working correctly
	observer	errors on the part of the person using the measuring device or interpreting the test results

This classification shows how important it is to select the study population for diagnostic research carefully.

9.1.3 Diagnostic research always concerns prevalence

Research questions for diagnostic studies relate to the likelihood of a disease or a particular stage of disease being present. As a result, diagnostic research always describes the prevalence of a disease outcome as a function of one or more determinants. Diagnostic research is therefor always cross-sectional (see ▶ chap. 4). In that respect diagnostic research differs substantially from etiological and therapeutic research, where the aim is to find causal determinants of the development or course of a disease or its consequences. These kinds of studies focus on incidences and therefore have a longitudinal design.

9.2 Validity and reproducibility of diagnostic tests

9.2.1 Many sources of diagnostic variability

The variability observed in the results of a diagnostic test is due to various sources of variability (see tab. 9.1), for example: instability of the parameter being measured in the person under consideration, inconsistency on the part of the observer, the number of measurements, the number of observation times, the number of observers and the number of persons under consideration (see fig. 9.2).

Test results thus reflect actual (biological) variability in the parameter being measured in the population as well as random or systematic errors (see ▶ chap. 5). As the aim of a diagnostic test is to distinguish individuals with abnormal actual values for a particular parameter from normal real variability, measuring errors need to be eliminated as far as possible. We need to allow for the fact, however, that many parameters can take on different values in a particular person over time and depending on the test situation. Most biological parameters have a normal distribution, with most values in the middle and relatively few extreme values.

The next question is which values of a diagnostic parameter should be regarded as normal and which as abnormal. When, for example, is creatine phosphokinase – the blood level of which is an indicator of myocardial infarction – too high or too low? This question implies that there is a cut-off point in the distribution curve of the diagnostic parameter that marks the transition from normal to abnormal. Ideally the distribution curve of a diagnostic parameter would be bimodal (i.e. there would be two clearly distinguishable subdivisions with a clear break between the values for people with and without the disease in question), but this is seldom the case. The distribution curve of the parameter for persons with the disease usually substantially overlaps the curve for those without it. This makes selecting a suitable cut-off point far from simple, especially since the distribution of a biological parameter often depends on other personal characteristics, e.g. age, gender, race and diet. To enable us to interpret the

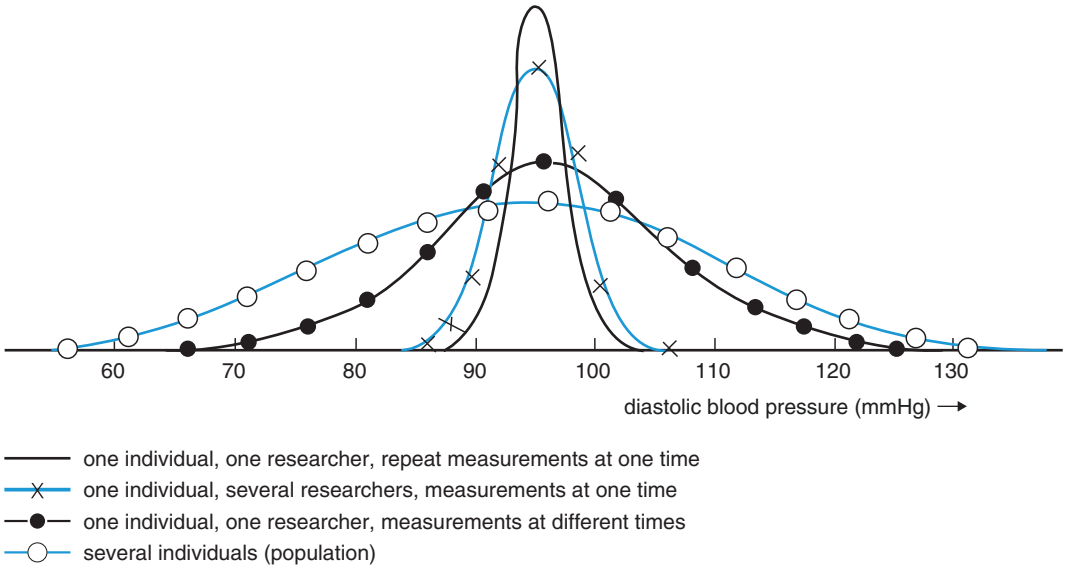


Figure 9.2 The cumulative effect of various sources of variability on the observed distribution of test results for a biological parameter (diastolic blood pressure)

result of a diagnostic test, then, we need to design a diagnostic function that includes that result along with other relevant diagnostic determinants. Only then can we investigate properly how the biological parameter in question relates to the presence of the disease. Abnormal values are thus those associated with the disease; normal test results are those where the disease is absent. Some biological parameters, however – e.g. serum cholesterol and blood pressure – display a continuous rise in morbidity and mortality risk over the entire range of possible test results, with each increase in the measured value being associated with additional risk. In such cases there are no normal values, strictly speaking. In practice we will often nevertheless select a cut-off point for deciding whether or not to treat in such a way that a therapeutic intervention in the case of an abnormal value has more benefits than disbenefits. We should also realize that these cut-off points may need to be adjusted later in the light of innovations in medical technology.

Measuring errors cloud our view of the true value of a diagnostic parameter and of biological variability in that parameter. Systematic measuring errors – due for example to incorrect calibration of the device – cause bias. If there are no systematic errors in the diagnostic measurement, it is valid. The validity of a measuring procedure thus expresses the extent to which the result corresponds to the true value of the measured parameter on average. We can correct for systematic errors if we know the magnitude and direction of the error. Random measuring errors that produce a random distribution of the test results around the true value of a diagnostic parameter are unavoidable. They affect reproducibility, i.e. the extent to which two measurements on the same individual produce the same result. Repeating the measuring procedure and taking the average of the test results reduces random measuring errors (and improves reproducibility). That is why a GP takes a patient's blood pressure more than once. This is illustrated in fig. 9.3, which shows the measurement of diastolic blood pressure in an individual with an actual low pressure (measured intra-arterially) of 92 mmHg.

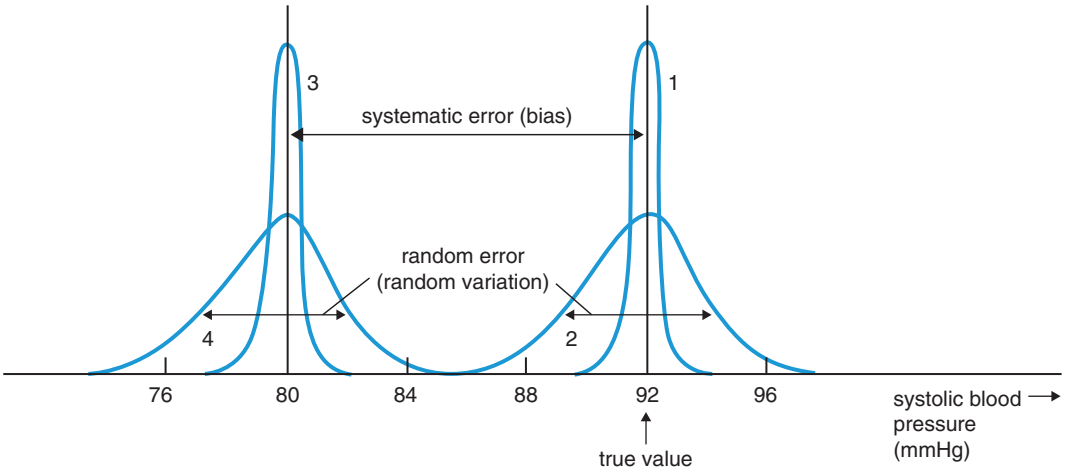


Figure 9.3 Systematic and random error when measuring a diagnostic parameter (blood pressure), *1* Measurement is valid and reproducible. *2* Measurement has poor reproducibility, but sufficient repeat measurements average out the random errors, giving a valid estimate of the actual blood pressure. *3* Measurement is reproducible but not valid. *4* Measurement is neither valid nor reproducible

9.2.2 Reproducibility indicates the precision of the test result

A gauge of the quality of a diagnostic test is the extent to which the results obtained when carrying a test out more than once on the same participants correspond. If the measuring conditions remain constant and the parameter being measured does not change, a repeat test will give consistent diagnostic results. Many different terms are in common use to describe the amount of agreement between repeat measurements and the absence of random errors in measurements. In ▶ chap. 5 we introduced the term reliability' to refer to the amount of agreement between the results of repeated cause-and-effect research based on samples of limited size. In the case of diagnostic tests the standard term is reproducibility, i.e. the correspondence between the results of repeat measurements by the same or different observers, or between repeat measurements under the same conditions. Use of the term 'accuracy' is not recommended, as there is no unequivocal definition of this concept and it is moreover associated with validity.

The reproducibility of a diagnostic test can relate to a situation where a single observer repeats the test on the same participants (intra-observer variability, stability, test-retest reliability). In research of this kind the successive observations must be independent of one another. If the time interval between them is too short, the precision of the diagnostic test could be overestimated because the observer remembers the previous result. If the interval is too long, on the other hand, the measurements could be affected by actual changes in the biological parameter and the precision could be underestimated. The reproducibility of a diagnostic test can also relate to a situation where two or more observers carry out the test on the same participants at about the same time (interobserver variability, interobserver agreement, inter-rater reliability).

9.2.3 Validity indicates whether the results are correct on average

Reliable test results are a necessary but insufficient precondition for the quality of a diagnostic test. The results should also give a valid representation of the disease status, or result in correct classification of the participants into the respective categories of severity or stage of the disease. The closer a diagnostic ob-

Table 9.2 Examples of conditions along with diagnostic tests that can be used to detect them and the respective external criteria (gold standards)

condition	diagnostic parameters or test	gold standard
gallstones	dyspepsia, pain	ultrasound
appendicitis	specific pain in the abdominal region	histological test after appendectomy
high blood pressure	external blood pressure measurement (sphygmomanometer)	intra-arterial blood pressure measurement
breast cancer	palpation (breast lump), mammogram, thermogram	histological test after biopsy
cervical cancer	contact bleeding, pain, cytological test after smear	histological test after biopsy
coronary heart disease	specific symptoms, exercise electrocardiogram, serum enzymes	coronary angiogram
fever	taking temperature using hand on forehead	taking temperature using a clinical thermometer

servation approximates to the actual clinical condition, the greater the validity of the test. Systematic discrepancies between the observed condition and the actual clinical condition (bias) result in the diagnostic test with imperfect validity. Random measuring errors (lack of precision, poor reproducibility) need not adversely affect validity if the test is repeated a sufficient number of times. In healthcare the preference is for diagnostic tests that are not only valid and precise but also quick, simple, inexpensive and not stressful.

In order to assess the validity of a diagnostic test we need to compare the results in a particular population with those from a different measuring tool (the external criterion). This is referred to as criterion validity. Ideally there will be a gold standard that can serve as the external criterion: this is a tool that measures the presence or stage of the particular condition objectively and independently and that is certain (or virtually certain) to give a correct picture of reality.

As the gold standard is usually expensive, labour-intensive, invasive or risky, and therefore not suitable for routine diagnostic testing, a good deal of diagnostic research focuses on making a good comparison between the results of a simple, quick, inexpensive and non-invasive test and those of a gold standard. The results of the comparison can then be used to justify using the simple test in future on patients similar to those on whom the validation study was carried out. tab. 9.2 shows examples of conditions along with the standard diagnostic tools and the respective gold standards.

For many conditions there is no suitable gold standard, e.g. complaints such as angina pectoris (chest pain), low back pain, migraine and most psychiatric conditions. Sometimes it is possible to draw a conclusion as to the validity of the test by waiting to see how the symptoms progress (without intervention) – assuming that the presence of the disease does not change substantially in the meantime. This is predictive validity, a special type of criterion validity.

In other cases the only ways of gaining some idea of validity are to check whether the measuring tool is seen to measure what it is supposed to measure (face validity), to check whether it takes account of all relevant aspects based on theoretical considerations (content validity), or to submit it to the scrutiny of one or more outside experts (expert validity).

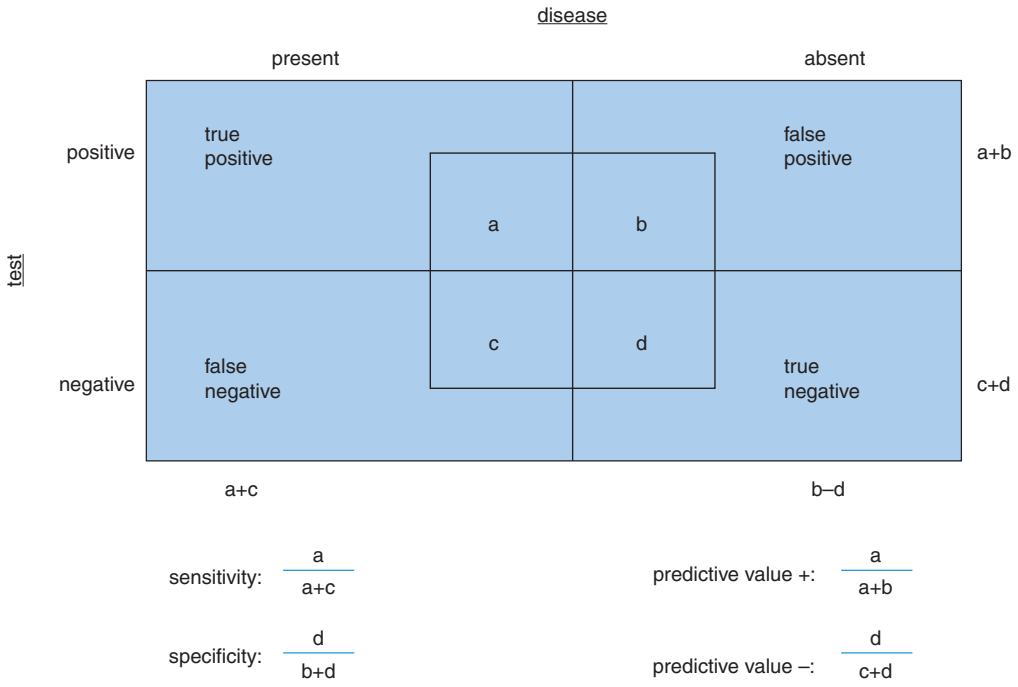


Figure 9.4 Calculating sensitivity, specificity and predictive value

9.3 Measures of validity of diagnostic tests

9.3.1 Sensitivity and specificity indicate the likelihood of a particular test result in persons with and without the disease

Two standard parameters for quantifying the validity of a diagnostic test are sensitivity and specificity. The sensitivity of a test indicates what percentage of the participants with a particular disease are correctly classified as diseased by the test (see fig. 9.4). Sensitivity therefore gives some indication of how 'sensitive' the test is in recognizing cases of the disease in question. If the test detects all the participants who have the disease – i.e. all of them test positive – it has maximum sensitivity, 100%, and we can say that the percentage of true positive test results is 100%. Note that the term 'positive' can be confusing: patients often get mixed up and think that it is a favourable result, and incorrectly conclude that they do not have the particular disease according to the test.

The specificity of a test indicates what percentage of a group of people without the disease being tested for are correctly classified as healthy by the test (see fig. 9.4). Specificity, then, gives some indication of the test's ability to correctly identify only people with the particular disease and no-one else. The more participants without the disease in question that the test identifies as 'healthy' – i.e. who test negative – the more specific the test is. In the optimum case the percentage of true negative test results will be 100%.

Using a diagnostic test with two possible outcomes – positive (test diagnosis = diseased) and negative (test diagnosis = healthy) – in a population containing both people with and without the particular disease produces a total of four categories:

- True positives (TP)
- False positives (FP)
- True negatives (TN)
- False negatives (FN)

Table 9.3 Relationship between the results of the GT and the OGTT in a thousand patients suspected of having diabetes

test	diabetes	no diabetes	
≥100 mg/ml	225	225	450
<100 mg/ml	25	525	550
	250	750	1,000

The two-by-two table in fig. 9.4 can only be filled in if we know with certainty which of the participants actually have the disease. This means that there must be a gold standard (see ▶ 9.2.3) and that the disease status of each person has been assessed using both the diagnostic test and the gold standard method. ▶ Case 9.1 illustrates the calculation of sensitivity and specificity based on hypothetical data from diabetes diagnostics.

Case 9.1 Diagnosing diabetes (hypothetical example)

Diabetes can be detected using the oral glucose tolerance test (OGTT). This involves orally administering a standard quantity of glucose in solution (50 g of glucose in approximately 200 ml of water) and then taking blood samples every half-hour for a few hours to measure the blood sugar level. The levels measured can be plotted on a graph. If the blood sugar curve does not fall sufficiently during the observation period the patient is diagnosed with diabetes. Diabetic patients have a hormone dysfunction (insulin deficiency) which makes it difficult for cells to absorb glucose from the blood. The question is whether diabetes could be diagnosed more simply, for example by looking at the blood glucose level two hours after eating a meal, i.e. a glucose test (GT).

To assess the quality of the GT method the results of the GT and OGTT methods for a thousand successive patients suspected of having diabetes based on certain symptoms are compared in a diagnostic study. The glucose tolerance test is taken to be the gold standard, although there is some doubt as to whether the OGTT measures the presence of diabetes correctly in every respect. For the purpose of the GT method a blood glucose level of ≥100 mg/ml two hours after a meal is regarded as indicative of diabetes. Based on the OGTT results 250 persons are classified as diabetic and 750 as non-diabetic. The distribution of the results of the quick test among them is shown in tab. 9.3.

The comparison between the two methods shows that the GT correctly detects 90% of the diabetics. Its sensitivity is therefore:

$$\frac{a}{a+c} = \frac{225}{250} \times 100\% = 90\%$$

The GT correctly classifies 70% of the non-diabetics (specificity):

$$\frac{d}{b+d} = \frac{525}{750} \times 100\% = 70\%$$

In total 75% (((225 + 525)/1,000) × 100%) of the participants are classified correctly. The GT produces a 'false negative' (FN) diagnosis in 25 cases and a 'false positive' (FP) diagnosis in 225 cases.

So what does this say about the value of the GT as a diagnostic tool in day-to-day practice? For this we need to look at the predictive value (PV). Prior to the diagnostic study 25% of the people in the practice population were thought to be diabetic (P(O)) and 75% non-diabetic (P($\overline{O}$)). After the diagnostic test we know that 50% of the people who test positive are indeed diabetic (positive predictive value, PV^+) and moreover that 95% of those who test negative are indeed non-diabetic (negative predictive value, PV^-). In the case of people who test positive, then, the probability of having the disease has gone up from 25% to 50%, whereas that probability has gone down from 25 to 5% in the case of people who test negative, as shown in fig. 9.5.

Thus the test has provided more information on the occurrence of diabetes in the practice population, but it has not been found capable of distinguishing diabetics from non-diabetics completely. The question is whether this increased knowledge is sufficient to base treat-

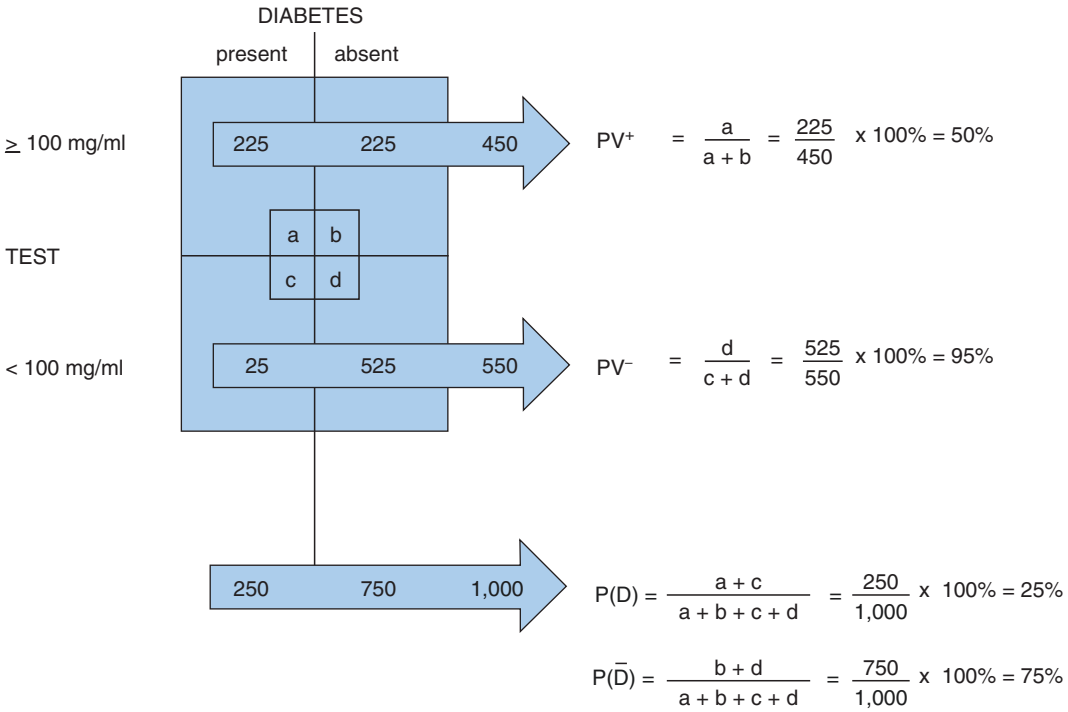


Figure 9.5 Predictive value of the blood glucose test in relation to the presence of diabetes in 1,000 patients suspected of having diabetes in a general practice situation. *P(O):* Prior probability of diabetes = prevalence of diabetes in the population being tested for the presence of the disease outcome.$P(\bar{O})$:Prior probability of the absence of diabetes = prevalence of non-diabetes in the population being tested for the presence of the disease. PV^+: Posterior probability of diabetes in persons who test positive = predictive value of a positive test result. PV^-: Posterior probability of the absence of diabetes in persons who test negative = predictive value of a negative test result

ment decisions or other follow-up activities on it.

Figure 9.5 shows that the predictive value of a test result depends on the sensitivity (a/(a + c)) and specificity (d/(b + d)) of the test. Higher sensitivity (a goes up and c goes down) is associated with a higher positive predictive value and a lower negative predictive value of the test. The prevalence (P(O)) of the disease outcome in the study population also affects the predictive value, however, as shown in fig. 9.6, where the same blood glucose test (with 90% sensitivity and 70% specificity) is used as a screening test on a predominantly healthy population of 100,000 in which the prevalence of diabetes is estimated to be 0.25%. The increase in the predictive values (PV^+ and PV^-) due to the screening test is minimal. The number of people with a false positive result is far higher than the number of diabetics detected. In a predominantly healthy population the test turns out to be worthless in detecting diabetes, in spite of the fact that the test conditions remain the same: hardly any of the people who test positive have diabetes. A negative test result is almost certainly indicative of the absence of diabetes, but that was already highly likely anyway.

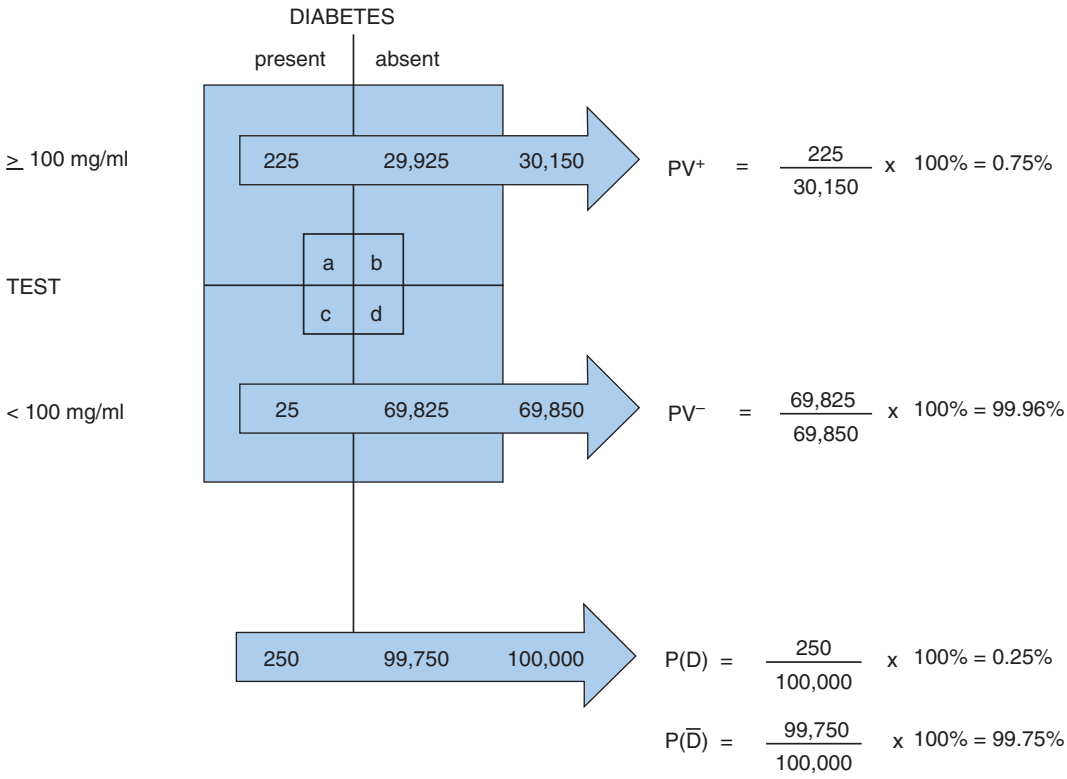


Figure 9.6 Predictive value of the GT for the presence of diabetes in 100,000 people in a predominantly healthy population

9.3.2 The ROC curve and selecting cut-off points

It is not difficult to develop a test that is 100% sensitive. If we blindly label everyone as diseased, i.e. select a very low cut-off point for a diagnostic test, everyone with the disease in question will be correctly classified as diseased. The price we pay for this is that everyone who does not have the disease will also be classified as having it, i.e. specificity is 0%. Conversely, we could develop a test that is 100% specific but does not detect a single patient with the disease. It is always important, therefore, to look at the test's sensitivity and specificity in combination, as they are 'communicating vessels'. Sensitivity can only be improved with specificity remaining constant – or vice versa – by using a different diagnostic test, or a different combination of diagnostic parameters.

The combinations of sensitivity and specificity for different cut-off points in a diagnostic test can be represented by an ROC (Receiver Operating Characteristic) curve, as exemplified by fig. 9.7.

Each dot in the graph represents the sensitivity and the complement of specificity for a different cut-off point. The discriminatory power of a test increases the closer the curve approaches the top left corner of the graph (where sensitivity and specificity are both 100%). This is expressed by calculating the area under the ROC curve and dividing it by the maximum area (the entire rectangle). The result – the proportion of the area under the ROC curve – is referred to as the area under the curve (AUC), the value of which is between 0 and 1. Note that an AUC of 0.5 means that the ROC coincides with the diagonal line in fig. 9.7. This is the 50/50 chance line, where the number of TPs is the same as that of FPs and the number of TNs the same as that of FNs. In other words, a test with an AUC of 0.5 has no discriminatory power; only at AUC values substantially higher than 0.5 does the test begin to have diagnostic

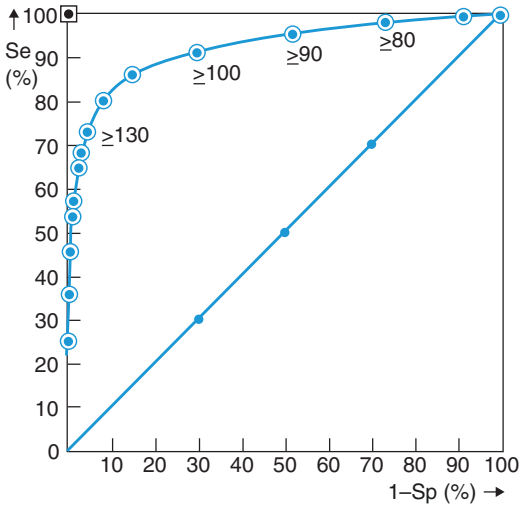


Figure 9.7 ROC curve for blood glucose tests with different sensitivity (Se) and specificity (Sp) to diagnose diabetes

value. The AUC – the discriminatory power of the diagnostic procedure – can only be improved, however, by using different tests or a combination of tests.

Many diagnostic parameters (e.g. blood pressure, lab results and function tests) are measured not as dichotomous variables (positive/negative, abnormal/normal, unhealthy/healthy) but on a scale (ordinal, interval or continuous). One or more cut-off points are then selected for measured values considered to be too high, too low or correct. But what are the rational grounds for selecting a cut-off point?

In the blood glucose test described in ▶ case 9.1 the cut-off point was set at 100 mg/100 ml. Based on that cut-off, 225 of the 250 diabetic patients were recognized as such (true positives, sensitivity = 90%) and 25 were missed (false negatives). Of the 750 patients without diabetes, 525 were classified correctly (true negatives, specificity = 70%) and 225 incorrectly (false positives). Was the correct cut-off selected?

Figure 9.8 shows the full range of the glucose test in the diabetics and non-diabetics. As the figure shows, each time the cut-off point is moved between a positive and negative test result the numbers of diabetics and non-diabetics detected change, illustrating the phenomenon mentioned earlier that the sensitivity and specificity of the test are communicating vessels. The figure therefore shows the corresponding sensitivity and specificity values for each possible cut-off point.

This distribution of blood sugar values among diabetics and non-diabetics can also be shown graphically in a histogram (see fig. 9.9) or frequency polygon (see fig. 9.10). The distributions for the diabetics and non-diabetics are projected 'back to back' here, showing that they overlap. If we want to use the diagnostic test only to detect cases of diabetes, the cut-off point should be set at a blood sugar level of 180 mg/100 ml, and if the test proves positive we will know for sure that the patient has the disease. We will miss quite a few diabetes cases, however. If we want to use the test only to exclude people without diabetes, we should set a cut-off point of 70 mg/100 ml, and only people who do not have the disease will test negative. A lot of non-diabetics will be identified as unhealthy, however. The optimum cut-off point lies somewhere between these two values.

Figure 9.10 shows that a cut-off point of 100 mg/100 ml yields approximately equal numbers of people with a true positive and a false positive test result. By increasing the cut-off point to 120 mg/100 ml we can reduce the number of false positives substantially, while the number of false negatives increases relatively little. At first sight this cut-off would therefore seem to be a better option.

In this example it is not possible to find a cut-off point that discriminates completely between diabetics and non-diabetics (no false positives and no false negatives, 100% sensitivity and 100% specificity). The perfect test situation is shown in fig. 9.11, where a cut-off point can be found that distinguishes perfectly between people who have and do not have the disease.

Given a particular type of test we can never increase sensitivity and specificity at the same time by moving the cut-off point. As we saw when introducing the ROC curve, we increase sensitivity at the expense of specificity and, conversely, increasing specificity means making sacrifices with regard to sensitivity (see fig. 9.12).

There is no general answer to the question of whether more weight should be attached to high sensitivity or to high specificity when selecting a cut-off point. Two considerations are important when weighing this up:

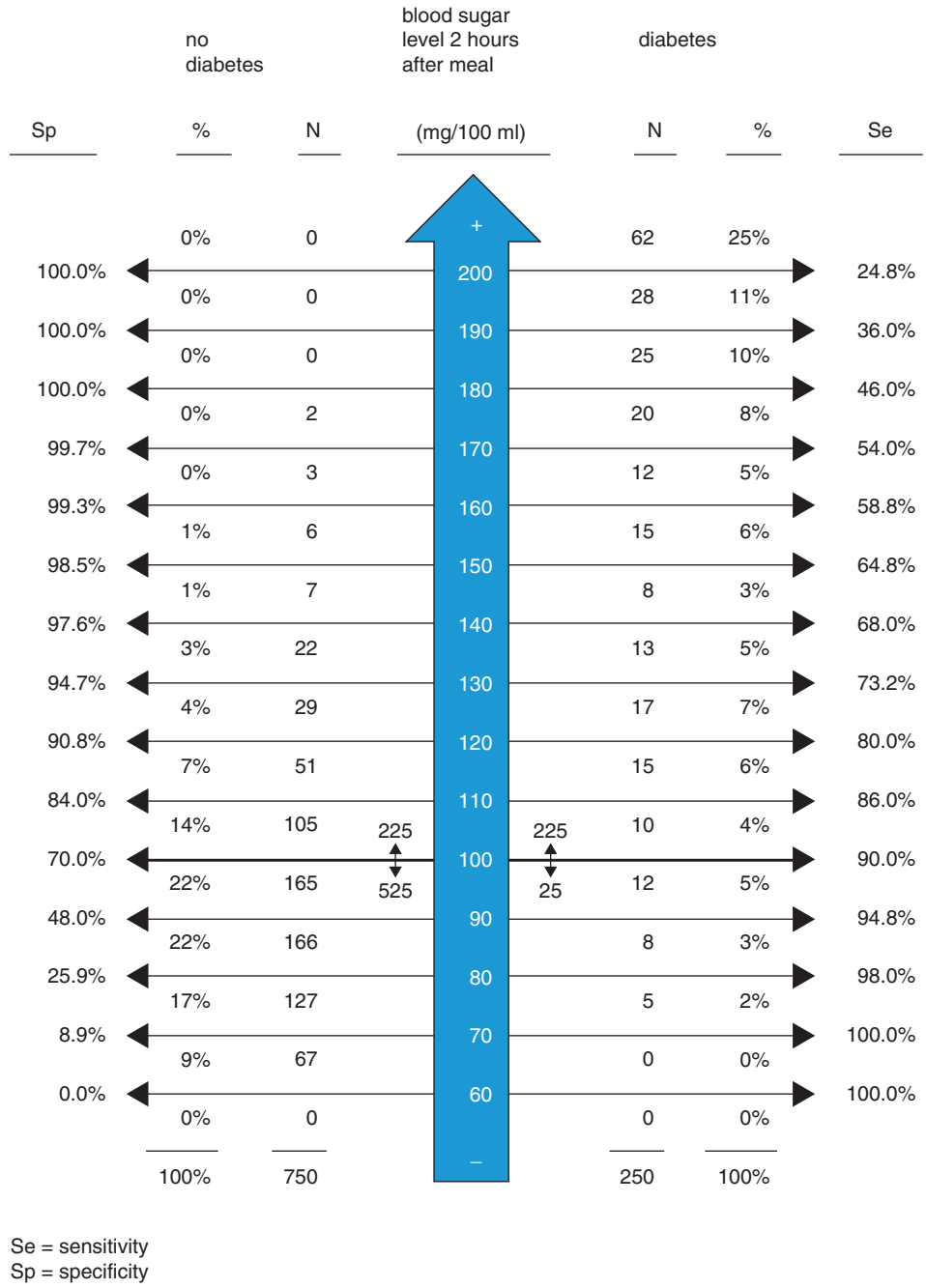


Figure 9.8 Distribution of the results of the glucose test in 250 diabetics and 750 non-diabetics. Sensitivity and specificity for various cut-off points for a positive/negative test result

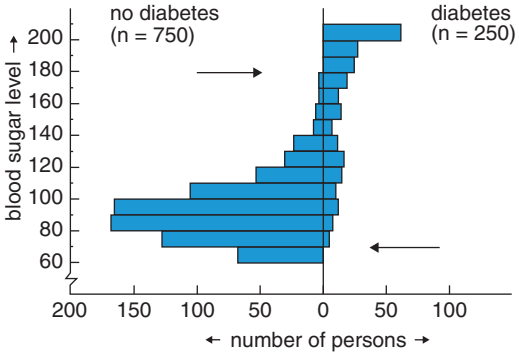


Figure 9.9 Frequency distribution of people with different results in the blood glucose test among diabetics and non-diabetics respectively

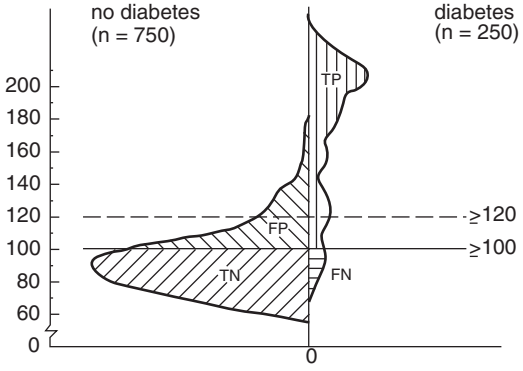


Figure 9.10 Distribution curves of the results of the blood glucose test in diabetics and non-diabetics (P(O) = 25%)

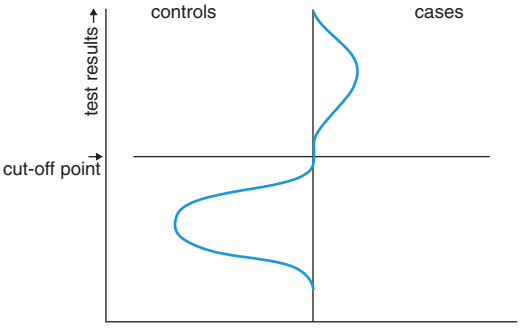


Figure 9.11 Example of a perfect diagnostic test

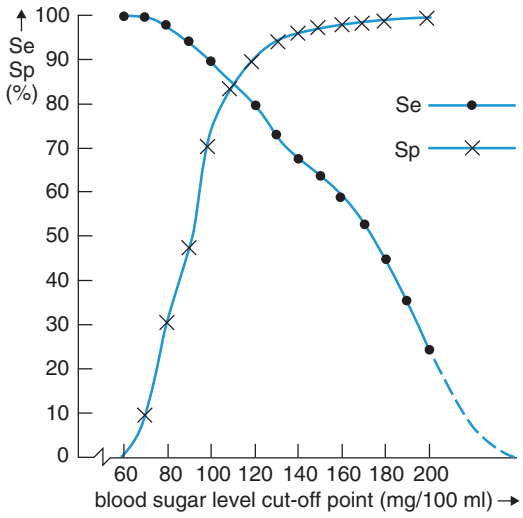


Figure 9.12 Relationship between sensitivity (Se) and specificity (Sp) for various cut-off points for a positive/negative result of the blood glucose test

- The consequences of incorrect diagnosis (FP, FN)
- The prevalence of the disease in the study population, as this is a factor in the numbers of false positive and false negative diagnoses.

A sensitive test is called for particularly if we do not want to miss a single case of the disease: for example, a disease that has a poor outcome if left untreated but that is treatable if discovered in time (e.g. Hodgkin's disease or phenylketonuria). A sensitive test can also be useful at the start of the diagnostic process in order to eliminate people who almost certainly do not have the disease in question (those who test negative), as it provides the greatest certainty if the result is negative. A sensitive test has the effect, however, of incorrectly labelling healthy people as 'unhealthy' (stigmatization). This is a problem particularly if further investigation or treat-

ment is stressful and risky and could result in physical, emotional or financial harm to the patients. An over-sensitive test results in over-consumption of healthcare.

9.3.3 The likelihood ratio combines information on the test in persons with and without the disease

The ROC curve (see ▶ 9.3.2) nicely illustrates how the sensitivity and sensitivity of a test depend on each other and on the cut-off point selected. The likelihood ratio is another way of combining information on those with and without the disease and is therefore used to compare various cut-off points. It is also useful in that it enables more than two possible test results to be used side by side.

The likelihood ratio of a positive test result (LR^+) is the ratio between the probability of a positive test result in people with the disease outcome and the probability of a positive test result in people without the disease outcome:

$$LR^+ = \frac{P(T^+|O)}{P(T^+|\bar{O})} = \frac{a/(a+c)}{b/(b+d)}$$
$$= \text{sensitivity}/(1-\text{specificity})$$

The likelihood ratio of a negative test result (LR^-) is the ratio between the probability of a negative test result in people with the disease and the probability of a negative test result in people without the disease:

$$LR^- = \frac{P(T^-|O)}{P(T^-|\bar{O})} = \frac{c/(a+c)}{d/(b+d)}$$
$$= (1-\text{sensitivity})/\text{specificity}$$

A likelihood ratio can be regarded as a yardstick of how much weight can be attached to a particular test in the diagnostic process (see fig. 9.13). A likelihood ratio of around 1 means that the test result in question does not add anything when gauging whether the disease is present or absent. The higher LR^+ is and the closer LR^- approximates to zero, the more information the diagnostic test provides. An LR^- below 0.3 means a reduction of at least 20–25% in the probability of the disease compared with the prior probability (prevalence), whereas an LR^+ higher than 3 means an increase of at least 20–25% in that probability. For example, if the prevalence of a disease in the general population is 4%, the estimated probability of the disease is 5% if the LR^+ of the test result in question is approximately 3. Diagnostic tests with LRs higher than 10 or below 0.1 are generally rated as 'good'.

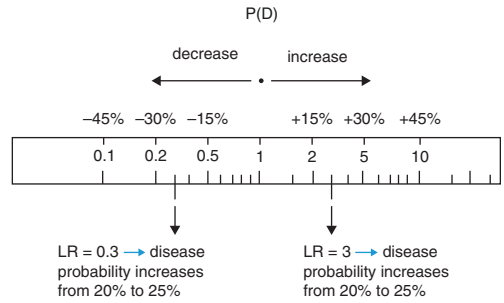


Figure 9.13 The likelihood ratio as a diagnostic yardstick

9.3.4 The predictive value indicates the posterior probability of the disease for a particular test result

The predictive value of an abnormal or positive test result (PV^+) indicates how likely it is that a person with this test result has the disease. The predictive value of a normal or negative test result (PV^-) indicates how likely it is that a person with this test result does not have the disease (see fig. 9.4). This is referred to as the posterior probability of the presence or absence of the disease, i.e. the likelihood once the test result is known. We can then compare this posterior probability with the prior probability of the disease, which is the same as the prevalence of the disease in the population in question (see ▶ 9.3.6). ▶ Case 9.1 illustrates this based on hypothetical data on how useful blood sugar levels are when diagnosing diabetes.

Generally speaking, higher prevalence of the disease in the study population results in a higher positive predictive value and a lower negative predictive value. Lower prevalence results in a lower positive predictive value and a higher negative predictive value. This point has important consequences for the interpretation of diagnostic tests (see ▶ 9.3.6).

Table 9.4 Positive and negative predictive values (percentages) of a diagnostic test with varying sensitivity (Se) and specificity (Sp) used in different populations

		P(O)	0.1%	0.5%	1.0%	5%	10%	25%	50%	75%	90%	95%	99%
Test 1	Se 90%	PV^+	0.30	1.49	2.9	13.6	25.0	50.0	75.0	90.0	96.4	98.3	99.7
	Sp 70%	PV^-	99.99	99.93	99.86	99.25	98.4	95.0	87.5	70.0	43.8	26.9	6.6
Test 2	Se 80%	PV^+	0.79	0.79	7.5	29.3	47.1	72.7	88.9	96.0	98.6	99.4	99.9
	Sp 90%	PV^-	99.98	99.98	99.78	98.84	97.6	93.1	81.8	60.0	33.3	19.2	4.4
Test 3	Se 70%	PV^+	0.17	0.87	1.7	8.4	16.3	36.8	63.7	84.0	94.0	97.1	99.4
	Sp 60%	PV^-	99.95	99.75	99.50	97.44	94.7	85.7	66.7	40.0	18.2	9.5	2.0

9.3.5 Predictive value depends on prevalence

As we saw in the previous section, the predictive value of a test is highly dependent on prevalence, and we shall quantify this relationship in this section. Table 9.4 shows the predictive values (PV^+ and PV^-) calculated for three different diagnostic tests (1: sensitivity = 0.90, specificity = 0.70; 2: sensitivity = 0.80, specificity = 0.90; 3: sensitivity = 0.70, specificity = 0.60) used in populations with different prevalences of the disease outcome (P(O)). You can also calculate these values for yourself, for example taking a population of 10,000.

Figure 9.14 shows the relationship between the prevalence of the disease, the sensitivity and specificity of the test and the predictive values of a positive/negative test result in graph form. The diagonal in the graph shows the situation for a test with both a sensitivity and a specificity of 50%. A test of this kind does not produce any information: we know just as much after it as we did before it. Clearly, the predictive value of a positive test result increases the more common the disease is and the higher the sensitivity and specificity are. Another interesting point is that the test provides the most additional information in situations where the disease is common (prevalence = 20–80%). A test that performs moderately is generally less suited to detecting a condition in populations where that condition is rare. Nor is there much point in using a test of that kind on a population where virtually everyone has the condition. Most diagnostic tests produce the most information when

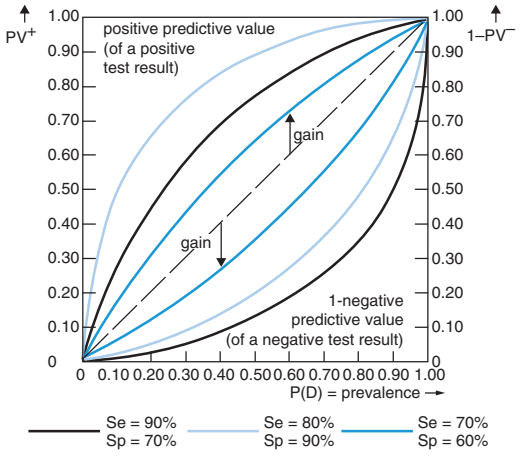


Figure 9.14 The effect of the prevalence of the disease and the sensitivity (Se) and specificity (Sp) of a test on the positive and negative predictive values of the test result

used if there are indications that the disease is present but they are not particularly strong. The degree of certainty needed regarding the presence of a disease to justify a treatment intervention based on the test result (the posterior probability) in fact differs from one disease to another, depending on such things as the severity of the condition and the risks associated with the treatment.

One problem is that the information on the prevalence of the disease in the group to which a person belongs is often inadequate, in which case a sensitivity analysis is called for. Broadly speaking, this involves calculating the PV^+ and PV^- for various pre-

valences (prior probabilities) within the range in which the actual prevalence is likely to be found. We then examine whether the results obtained lead to different diagnostic or treatment decisions. In this case, then, sensitivity analysis is used to find out how sensitive a diagnostic or treatment strategy is to changes in the assumptions regarding the occurrence of the disease in the study population. Sensitivity analysis can also be used to examine the consequences of different assumptions regarding the sensitivity and specificity of a diagnostic test.

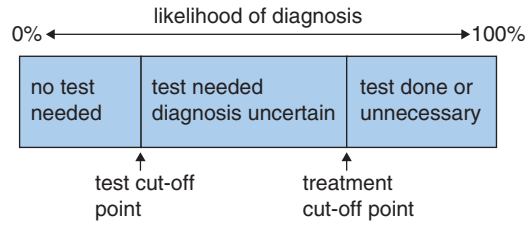


Figure 9.15 When is a diagnostic test required?

9.3.6 Bayes' theorem

In effect, diagnosis is nothing more nor less than the estimation of likelihoods (the prior and posterior probability of the presence or absence of the disease). The aim of diagnostic tests is to reduce uncertainty in these estimates. Figure 9.15 shows the spectrum of diagnostic uncertainty. If there is no suspicion of the disease (low prior probability) or it is patently obvious that it is present (high prior probability) there is no point in carrying out further investigation. In the grey area of diagnostic uncertainty, however, an additional test can have added value: this is the range of indications for the test. The result of this test can be used to calculate the posterior probability, enabling disease probabilities to be estimated better and with less uncertainty. This is Bayes' theorem in action (see Figure 9.16).

Figure 9.16 Thomas Bayes (1702–1761)

Bayes' theorem on conditional probabilities states that the posterior probability of disease outcome given a particular test result $P(O|T^x)$ can be calculated by multiplying the prior probability $P(O)$ by the additional evidence from the diagnostic test $P(T^x|O)/P(T^x)$.

As a formula:

$$P(O|T^x) = P(T^x|O)\frac{P(O)}{P(T^x)}$$

As we have already seen, the posterior probability is synonymous with the predictive value of a test. If we replace the general test result T^x in the formula with a positive or negative test result, this brings us back to our classic measures of validity (sensitivity = Se and specificity = Sp):

$$P(O^+|T^+) = P(T^+|O^+) \times \frac{P(O^+)}{P(T^+)}$$

$$\downarrow$$

$$PW^+ = \frac{Se \times prev.}{(Se \times prev) + [(1 - Sp) \times (1 - prev)]}$$

$$P(O^-|T^-) = P(T^-|O^-) \times \frac{P(O^-)}{P(T^-)}$$

$$\downarrow$$

$$PW^- = \frac{Sp \times (1 - prev)}{[Sp \times (1 - prev)] + [(1 - Se) \times prev]}$$

In these formulas too we notice that the predictive value (PV) of a test depends to a large extent on the prevalence of the disease.

Further application of Bayes' theorem shows the relationship between the prevalence of the disease

outcome (the prior probability P(O)), the likelihood ratio (LR) and the predictive values of a test result (the posterior probability PV):

$$\frac{P(O)}{1-P(O)} \times LR^{+} = \frac{PV^{+}}{1-PV^{+}}$$

and

$$\frac{P(O)}{1-P(O)} \times LR^{-} = \frac{1-PV^{-}}{PV^{-}}$$

Since (a similar derivation can be made for LR^{-}):

$$LR^{+} = \frac{a/(a+c)}{b/(b+d)} = \frac{a \times (b+d)}{b \times (a+c)}$$

$$= \frac{a/(a+b)}{b/(a+b)} \times \frac{(b+d)/(a+b+c+d)}{(a+c)/(a+b+c+d)}$$

$$= \frac{PV^{+}}{1-PV^{+}} \times \frac{1-P(O)}{P(O)}$$

Generally speaking, the following is true of test result x:

$$\frac{P(O)}{1-P(O)} \times LR^{x} = \frac{PV^{x}}{1-PV^{x}}$$

9.3.7 The diagnostic odds ratio as a product of the diagnostic function

Another way of examining diagnostic questions is to start from our familiar epidemiological function. The aim of epidemiological research into diagnostic questions is to estimate a function that describes as accurately as possible what diagnostic determinants are related to the likelihood of a particular disease being present. For practical purposes is it important to restrict the set of determinants to the minimum needed to predict the probability of the disease.

The diagnostic function indicates the likelihood of the presence of disease outcome O (prevalence P(O)) in relation to the set of diagnostic determinants D_i (symptoms, signs, tests, personal characteristics). Note that there are no confounders or effect modifiers in the diagnostic function, as it is a descriptive function, not an explanatory one.

Logistic functions are used mainly to describe the likelihood of the disease outcome dependent on the diagnostic parameters.

$$P(O) = \frac{1}{1+e^{-(b_0+b_1D_1+b_2D_2+\ldots+b_kD_k)}}$$

This function covers a range from 0 to 1 and produces an S-shaped curve, with a sharp increase in probability in the middle.

In the model P(O) represents the probability of the outcome being present given a particular combination of diagnostic parameters, b_o, b_1, b_2 etc. are weights that can be estimated based on the study data, and D_1, D_2 etc. are the values for a set of diagnostic parameters for a particular participant. From the function above we can derive the following:

$$\ln\left[\frac{P(O)}{1-P(O)}\right] = \ln \text{odds}(O)$$

$$= b_0 + b_1D_1 + b_2D_2 + \ldots + b_kD_k$$

And from this we can infer the following for people who test positive for parameter D_1 (value = 1) compared with people who test negative for parameter D_1 (value = 0), with the values of all the other parameters remaining constant:

$$\frac{\text{odds}(O)}{\text{odds}(\bar{O})} = OR_1 = e^{b_1}$$

The contribution made by each relevant diagnostic parameter can thus be estimated based on the logistic regression model. When diagnosing diabetes (▶ case 9.1), for example, we can use this approach to examine the unique contribution of fasting blood sugar level, having taken into account factors such as age, the pattern of symptoms and waist-hip ratio.

A new test parameter can easily be derived from the above, namely the diagnostic odds ratio (DOR). The DOR reflects the added value of a diagnostic parameter in a diagnostic function. It is in effect the diagnostic counterpart of the odds ratio that can be calculated in etiological research (see ▶ 3.2.4). There is a difference, however, in that the OR values regarded as relevant to a diagnostic factor are higher than for an etiological factor and the DOR does not permit interpretation in terms of causality. ◘ Table 9.5 shows the distribution of the test scores for a diagnostic test with regard to a hypothetical disease

Table 9.5 Distribution of test scores (positive/negative) among persons with and without the disease outcome

	disease outcome +	disease outcome –
test +	a	b
test –	c	d
total	*a + c*	*b + d*

as a contingency table. Here the DOR is defined as the ratio between the disease odds for a positive test result (a/b) and the disease odds (c/d) for a negative test result.

$$\mathrm{DOR} = \frac{\mathrm{a}/(\mathrm{a}+\mathrm{b})}{\mathrm{b}/(\mathrm{a}+\mathrm{b})} \Big/ \frac{\mathrm{c}/(\mathrm{c}+\mathrm{d})}{\mathrm{d}/(\mathrm{a}+\mathrm{d})} = \frac{\frac{\mathrm{a}}{\mathrm{b}}}{\frac{\mathrm{c}}{\mathrm{d}}} = \frac{\mathrm{ad}}{\mathrm{bc}}$$

Thus the DOR combines the information contained in both the abnormal (positive) and normal (negative) test result. In Bayesian terminology the disease odds are also referred to as the 'posterior odds', since they are the odds once the result (positive/negative) is known.

$$\mathrm{DOR} = \frac{\text{posterior odds}^{+}}{\text{posterior odds}^{-}}$$

The DOR also indicates the relationship between LR^+ and LR^-:

$$\mathrm{DOR} = \frac{\mathrm{TP}/\mathrm{FP}}{\mathrm{FN}/\mathrm{TN}} = \frac{\mathrm{Se}}{1-\mathrm{Se}} \Big/ \frac{1-\mathrm{Sp}}{\mathrm{Sp}}$$

$$= \frac{\mathrm{a}/(\mathrm{a}+\mathrm{c})}{\mathrm{c}/(\mathrm{a}+\mathrm{c})} \Big/ \frac{\mathrm{b}/(\mathrm{b}+\mathrm{d})}{\mathrm{d}/(\mathrm{b}+\mathrm{d})}$$

$$= \frac{(\mathrm{a}/\mathrm{a}+\mathrm{c})/(\mathrm{b}/\mathrm{b}+\mathrm{d})}{(\mathrm{c}/\mathrm{a}+\mathrm{c})/(\mathrm{d}/\mathrm{b}+\mathrm{d})} = \frac{\mathrm{LR}^{+}}{\mathrm{LR}^{-}}$$

A DOR of 1 indicates that the test does not discriminate between persons with and without the disease outcome; a DOR of less than 1 means that the test is useless (a negative test result is more likely among people with the disease); a DOR of more than 1 means that the test is usable (a positive test result is more likely among people with the disease). A useful test has a high DOR, but unfortunately it is not possible to indicate limits for this. For this reason the DOR is used mainly to pool the results of various diagnostic studies and used less in individual studies as to the validity of a diagnostic test.

9.3.8 Multiple tests

The above discussion enables us to say what strategies can be used to improve the efficiency of the diagnostic process. Broadly speaking, there are two options: ensuring that the prevalence is high in the study population, or using a combination of tests to detect the condition.

The prevalence of the disease in the study population (and hence the prior probability of the disease in the particular individual) can be influenced in various ways:

- Through referral policy: this increasingly concentrates cases of the disease in the healthcare system.
- By carrying out the diagnostic test selectively on people with particular demographic characteristics such as age, gender, race etc. (the likelihood of breast cancer is higher in older women than young women, for example).
- By targeting the diagnostic test on people with particular clinical characteristics (symptoms, signs, vague manifestations, exposure to known risk factors); these are referred to as 'indications' for carrying out a diagnostic test.

Combining two or more diagnostic tests is common in practice, based on the argument that combining a test with high sensitivity but low specificity and a test with the opposite properties can provide more certainty and thus a sufficient basis for a treatment strategy. The tests can be carried out in two ways, either in series or in parallel.

Serial testing

Serial testing involves carrying out independent tests successively. The second test is only carried out on people who score positive for the first test. The following decision rules are usually used: the overall result is positive if both of the tests are positive; in all other cases the result is negative. The net effect is a decrease in sensitivity and an increase in specificity.

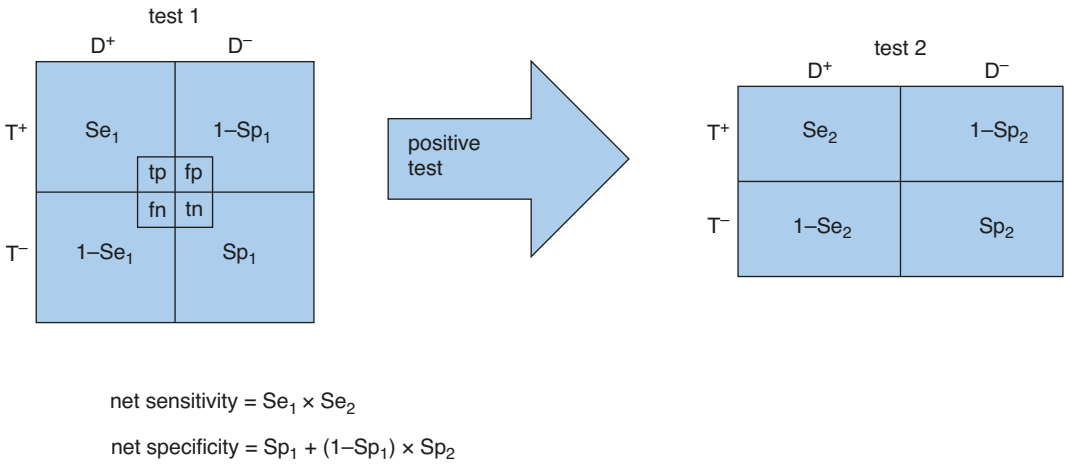


Figure 9.17 Formulas for calculating the combined sensitivity (Se) and specificity (Sp) of two diagnostic tests carried out in series

The predictive value of a positive test result goes up, whereas that of a negative test result goes down (see fig. 9.17). Serial testing is indicated where:

- A specific test is needed.
- There are two or more alternative tests available, each of which is insufficiently specific on its own.
- Having to wait a while for the final test result is not a problem.

The procedure is most efficient if the test with the highest specificity is carried out first. On the other hand, it may be decided to give priority to the least risky or least expensive procedure. Table 9.6 shows the hypothetical results of using two types of test that independently measure the presence of diabetes in series (Test 1: sensitivity = 90%, specificity = 70%; Test 2: sensitivity = 80%, specificity = 90%).

Parallel testing

Parallel tests are independent tests carried out at the same time on everyone in the study population. The following decision rules are often used: the overall result is positive if there is a positive result for at least one of the tests; the overall result is negative if the result of both tests is negative. If these rules are applied, the net effect is an increase in sensitivity and a decrease in specificity compared with those of each of the two separate tests. See fig. 9.18. The predictive value of a negative test result goes up, whereas that of a positive test result goes down. Parallel testing is indicated where:

- A sensitive test is required (i.e. where missing cases of the disease would have serious consequences).
- There are two or more alternative tests available, each of which is insufficiently sensitive on its own.
- The result of the test needs to be known quickly.

Parallel testing has the following disadvantages:

- Overdiagnosis (a larger number of false positives)
- Higher cost (everyone is subjected to two or more different tests, whereas in serial testing only part of the study population need to take a second test).

Table 9.7 illustrates the results of using two types of test in parallel to independently measure the presence of diabetes (Test 1: sensitivity = 90%, specificity = 70%; Test 2: sensitivity = 80%, specificity = 90%).

Table 9.6 Using two diagnostic tests in series to detect diabetes

A	first test 1, then test 2			B	first test 2, then test 1		
test 1	*diabetes*	*no diabetes*	*total*	*test 2*	*diabetes*	*no diabetes*	*total*
+	225	225	450	+	200	75	275
–	25	525	550	–	50	675	725
	250	750	1,000		250	750	1,000
test 2	*diabetes*	*no diabetes*	*total*	*test 1*	*diabetes*	*no diabetes*	*total*
+	180	22.5	202.5	+	180	22.5	202.5
–	45	202.5	247.5	–	20	52.5	72.5
	225	225	450		200	75	275
test 1 + test 2	*diabetes*	*no diabetes*	*total*	*test 2 + test 1*	*diabetes*	*no diabetes*	*total*
+	180	22.5	202.5	+	180	22.5	202.5
–	70	727.5	797.5	–	70	727.5	797.5
	250	750	1,000		250	750	1,000
sensitivity = 72% specificity = 97% PV^+ = 88.89% PV^- = 91.22% number of tests: 1,000 (test 1) + 450 (test 2) = 1,450				sensitivity = 72% specificity = 97% PV^+ = 88.89% PV^- = 91.22% number of tests: 1,000 (test 2) + 275 (test 1) = 1,275			

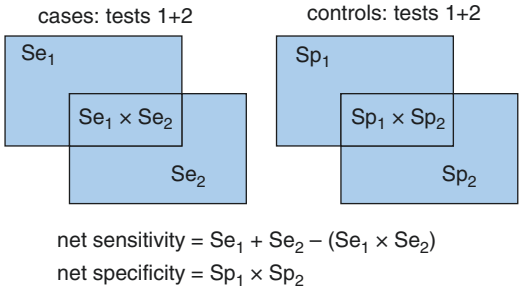


Figure 9.18 Formulas for calculating the sensitivity (Se) and specificity (Sp) of two diagnostic tests carried out in parallel

9.4 Measures of reproducibility of diagnostic tests

Research into the reproducibility of diagnostic tests has shown that professionals' opinions often differ, not only from one professional to another but also – albeit to a lesser extent – in a particular professional. These differences manifest themselves not only during history-taking and physical examination but also, contrary to expectation, in lab tests, which are often assumed to provide hard data. There are various measures to quantify the degree of consistency of a single rater or multiple raters. Which procedures and measures of association are suitable depends on various factors, including the scale on which the diagnostic parameter was measured.

Table 9.7 Using two diagnostic tests in parallel to detect diabetes

test 1	diabetes	no diabetes	total	test 2	diabetes	no diabetes	total
+	225	225	450	+	200	75	275
–	25	525	550	–	50	675	725
	250	750	1,000		250	750	1,000
sensitivity = 90% specificity = 70% PV^+ = 50% PV^- = 95.45% P(O) = 25%				sensitivity = 80% specificity = 90% PV^+ = 72.73% PV^- = 93.10% P(O) = 25%			
diabetics							
probability of positive result for both test 1 and test 2:				0.9 × 0.8 = 0.72			
probability of positive result for test 1 and negative result for test 2:				0.9 × 0.2 = 0.18			
probability of negative result for test 1 and positive result for test 2:				0.1 × 0.8 = 0.08			
probability of negative result for both test 1 and test 2:				0.1 × 0.2 = 0.02			
sensitivity of combined test:				(0.72 + 0.18 + 0.08)/1.00 = 0.98			
non-diabetics							
probability of positive result for both test 1 and test 2:				0.3 × 0.1 = 0.03			
probability of positive result for test 1 and negative result for test 2:				0.3 × 0.9 = 0.27			
probability of negative result for test 1 and positive result for test 2:				0.7 × 0.1 = 0.07			
probability of negative result for both test 1 and test 2:				0.7 × 0.9 = 0.63			
specificity of combined test:				0.63/1.0 = 0.63			
test 1 and test 2	diabetes	no diabetes	total	sensitivity = 98% specificity = 63% PV^+ = 46.89% PV^- = 98.95%			
+	245	277.5	522.5				
–	5	472.5	477.5				
	250	750	1,000				

9.4.1 The agreement rate for categorical test results

A commonly used measure is the agreement rate (between the first and second rating or the first and second rater). This is calculated mainly where there are two (or more) different test results (e.g. gallstones/no gallstones, no/mild/moderate/severe fatigue). Figure 9.19 illustrates the calculation of the agreement rate for a dichotomous test result, for example a situation where two radiologists independently assess (as positive or negative) the mammograms of a hundred women referred by their GP with suspected breast cancer.

Interpreting the calculated agreement rate (82% in this case) gives rise to several problems. First, the rate thus calculated is highly dependent on the expected frequency of abnormal characteristics in the study population. In fig. 9.20, for example, the same two radiologists independently assess the

		rater 2		
		normal test result	abnormal test result	
rater 1	normal test result	42 (a)	12 (b)	54
	abnormal test result	6 (c)	40 (d)	46
		48	52	100

$$\text{observed agreement} = \frac{\text{number of consistent observations}}{\text{total number of observations}} \times 100\%$$

$$= \frac{a + d}{a + b + c + d} \times 100\%$$

$$= \frac{42 + 40}{100} \times 100\% = 82\%$$

Figure 9.19 Calculating the observed inter-rater agreement on diagnostic tests (contingency table)

		rater 2		
		normal test result	abnormal test result	
rater 1	normal test result	932 (a)	22 (b)	954
	abnormal test result	16 (c)	30 (d)	46
		948	52	1.000

$$\text{observed agreement} = \frac{a + d}{a + b + c + d} \times 100\%$$

$$= \frac{932 + 30}{1{,}000} \times 100\% = 96.2\%$$

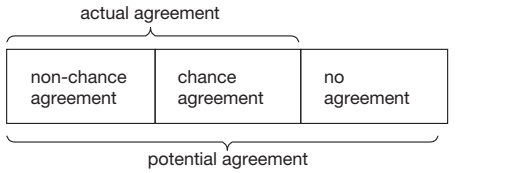
Figure 9.20 Calculating the observed inter-rater agreement on diagnostic tests (contingency table)

		rater 2		
		normal test result	abnormal test result	
rater 1	normal test result	$\frac{48}{100} \times 54 = 25.9$ (a)	$\frac{52}{100} \times 54 = 28.1$ (b)	54
	abnormal test result	$\frac{48}{100} \times 46 = 22.1$ (c)	$\frac{52}{100} \times 46 = 23.9$ (d)	46
		48	52	100

$$\text{expected agreement} = \frac{a+d}{a+b+c+d} \times 100\% = \frac{25.9+23.9}{100} \times 100\% = 49.8\%$$

Figure 9.21 Calculating the expected agreement rate based solely on chance between two ratings or raters on a diagnostic test (contingency table)

mammograms of a thousand women taking part in a breast cancer screening programme. The observed absolute agreement rate is higher in this situation, even though the raters' opinions differ more often in absolute terms; in particular, there is less consensus on whose test results should be regarded as abnormal. The difference between this situation and the one in fig. 9.19 is due to the prevalence of abnormal test results in the study population, which was approximately 50% in fig. 9.19 and approximately 5% in fig. 9.20.



$$\text{kappa} = \frac{\%\text{ actual (observed) agreement} - \%\text{ chance agreement}}{\%\text{ potential (full) agreement} - \%\text{ chance agreement}}$$

Figure 9.22 Kappa

9.4.2 Cohen's kappa: correcting the agreement rate for chance

When characterizing the reproducibility of a diagnostic test procedure we can allow for the fact that the agreement between the two ratings is due partly to chance. This is certainly a factor in the above examples (see figs. 9.19 and 9.20). If rater 1 concludes that there is an abnormal test result in 46 of the 100 cases and rater 2 finds that 52 of the 100 mammograms are abnormal, chance alone will cause rater 2 to classify as abnormal 52% of the 46 mammograms regarded as abnormal by rater 1. The remaining 48% will be diagnosed as normal. Likewise, rater 2 will record 52% of the 54 mammograms rated as normal by rater 1 as abnormal and 48% as normal (see fig. 9.21). In other words, both raters will agree in half of the cases purely on the basis of chance.

Cohen's kappa is a measure of inter-rater and intra-rater agreement that indicates actual agreement as a proportion of potential agreement after correcting for chance agreement (see fig. 9.22).

In the example, therefore, the kappa is:

$$\frac{82\% - 49.8\%}{100\% - 49.8\%} = 0.64$$

Table 9.8 Interpreting Cohen's kappa

value of cohen's kappa	agreement interpretation
<0.20	none
0.21–0.39	minimal
0.40–0.59	poor
0.60–0.79	moderate
0.80–0.90	strong
>0.90	almost perfect

Cohen's kappa normally has values between 0 (solely chance agreement) and 1 (perfect agreement). Agreement of 0.80 or more is regarded as strong inter-rater or intra-rater agreement, whereas agreement of 0.40 or less is regarded as low (see tab. 9.8). In clinical practice, however, diagnostic tools with kappas of the order of 0.40–0.70 are common. It could happen that two raters agree less often than expected based on chance, in which case the kappa is negative.

Although the kappa, as a measure of agreement between two raters or two rating sessions, takes into account the fact that the observed agreement is due partly to chance, there are still some difficulties when interpreting it, as:

- The kappa depends on the number of rating categories (strata): the more strata, the smaller the kappa (this is also true for the agreement rate).
- The kappa depends on the prevalence of the test results.

9.4.3 The correlation coefficient and the limits of agreement for continuous test results

In order to determine the reproducibility of a diagnostic test with a continuous result the first step is to sketch a scatter plot plotting the results of the two measurements for each participant against each other. If the measurement is reproducible, virtually all the points will be on a straight line passing through the zero point at an angle of 45°. The further the cloud is from that line, the less reproducible the measurement is. From this we can derive the following agreement rates for continuous variables:

- The correlation coefficient (r) is a measure of the density of the cloud (see also ▶ 3.3.2). A correlation coefficient of 1 is obtained if there is complete agreement. The lower the correlation coefficient, the poorer the reproducibility, with a value of zero indicating complete lack of correlation. The correlation coefficient does not take systematic differences between raters into account, however. If two raters agree perfectly, for example, and one rater systematically measures values twice as high as the other observer, the correlation coefficient remains constant (1 in this case). Also, Pearson's correlation coefficient is sensitive to the distribution of the values on the x and y-axes: an extreme value for both measurements causes a sharp increase in the correlation coefficient. A scatter plot should therefore always be made to look at the underlying distribution. A parameter-free version of the correlation coefficient (Spearman's) can be used if necessary. Note that the standard statistical tests for the correlation coefficient are irrelevant when testing for reproducibility, as they test the null hypothesis that the correlation coefficient is not zero, whereas we are interested in how close it approximates to 1.
- The distribution of differences between the two measurements for each individual: this will average zero if there is only random variability. The spread (standard deviation) of this distribution of differences provides a measure of reproducibility (see fig. 9.23). A Bland-Altman plot visualizes the agreement rate between two series of observations of a parameter measured on a continuous scale. This could be a comparison of the results for that parameter produced by two different measuring instruments (A and B), or a comparison of the results from repeat use of the same measuring instrument (A_1, A_2) on a number of participants. The Bland-Altman plot plots the difference between the two measurements (e.g. A–B) against the mean value of the two measurements ((A + B)/2) for every participant. The individual difference scores are expressed in

the same dimension as the measurements. Assuming a normal distribution, approximately 95% of the differences found will be between the mean difference score ± approximately 2× the standard deviation of the individual difference scores. These limits of agreement provide a good gauge of whether the individual differences in test results are clinically relevant and acceptable. As a way of determining the reproducibility and agreement rate of measurements of a continuous parameter the Bland-Altman plot does away with some of the limitations of Pearson's correlation coefficient (Pearson's r), as it shows up any systematic measuring errors (the mean difference score) and moreover provides a better picture of the size of intra-individual differences in results.

- The intra-class correlation coefficient (ICC) of agreement also overcomes the problem with Pearson's correlation coefficient (r) that it does not take systematic measuring differences into account: the ICC agreement only takes on the maximum value (i.e. 1) if the test results correspond precisely at the individual level. The other problems with Pearson's r – dependency on the spread of the test results and high sensitivity to outliers – also apply to the ICC agreement, however, and calculating the ICC is a very complex business.

9.5 Guidelines for diagnostic research

There are various checklists for assessing the quality of diagnostic research: these are used in systematic literature studies, but they also provide an extremely handy reference when designing new studies. The QUADAS-2 questionnaire is an example of a checklist of this kind that is in common use.[1]

There are STARD guidelines on the publication of diagnostic research, the latest version dated 2015.[2] Over 200 scientific journals require diagnostic studies to be reported in line with these guidelines. The STARD guidelines require a flow diagram showing the study design and the flow of patients. They also include a checklist of 25 items that a scientific article on diagnostic research must comply with: Title, Introduction (research question and aim), Method (participants, types of test and statistical methods), Results (participants, test results and estimates) and Discussion.

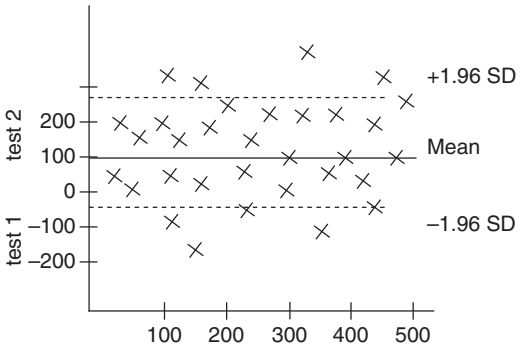


Figure 9.23 Bland-Altman plot

9.6 Prognostic research: describing the course of disease

Questions concerning the expected course of the condition are very different from diagnostic questions, but the aim is the same: to come up with information on the basis of which the practitioner can take treatment decisions or the patient can adjust his expectations or behaviour. Prognosis has a lot in common with diagnosis as described earlier in this chapter. The main difference is the time dimension. Prognosis is about changes over time, whereas diagnosis is about what is going on at a particular point in time. As stated in the introduction, in this chapter we only discuss prognostic prediction models, where causality or the effects of interventions are not considered. Prognostic research can of course also be concerned with the therapeutic and non-therapeutic causal determinants of the course of a disease, but for these we would refer to ▶ chap. 4 and 6 on observational studies and ▶ chap. 10 on intervention studies.

1 The QUADAS-2 questionnaire for assessing the quality of diagnostic research (website). http://bit.ly/1Hw3nGF
2 STARD guidelines for the publication of diagnostic research (website). ▶ http://bit.ly/1GeSymt

9.6.1 Prognostic research means following a cohort

In order to examine the prognosis for a particular condition or state of health in a particular patient, in terms of the likelihood of developing a particular outcome (death, complications, cure), a population of similar patients will need to be followed so as to estimate the incidence of that outcome. We can deduce from this that prognostic research always takes the form of a cohort study, with the cohort consisting of people with the condition or state of health for which we want to learn the prognosis. Such research only yields useful information, of course, if all sorts of data on the members of the cohort that can be linked to this incidence are collected at the start of the follow-up: these are the prognostic determinants. This brings us back to our familiar epidemiological function, which in this case describes the incidence of the feared or hoped-for future outcome as a function of a series of prognostic determinants. If the outcome variable (O) is a stage of disease whose likelihood of developing (incidence) we wish to know, the function will take the form of a logistic function with a linear set of determinants D_i:

$$P(O) = \frac{1}{1 + e^{-(b_0 + b_1 D_1 + b_2 D_2 + \ldots + b_k D_k)}}$$

Suitably modified models can be used for other types of outcome (continuous health parameters or length of survival).

► Case 9.2 provides an example of a prediction model developed to predict successful vaginal delivery after previous Caesarean section.

Like diagnostic functions, a prognostic model can include various determinants: symptoms and signs, results of lab and other tests, and characteristics of the people in question (age, gender, lifestyle). These parameters must relate to the situation at the start of the follow-up, i.e. when the question of the prognosis arises.

A historical cohort or case-control study can also be used for prognostic research, but this presupposes that all sorts of determinants that we want to include in the function have already been measured and recorded, or can still be ascertained. This will not be possible in the case of many determinants, in which case we shall have to fall back on a prospective cohort study.

For a good, correct description of the prognostic function all the determinants and the outcome must be measured correctly (validly and precisely) and independently of one another. In other words, we must not be influenced consciously or unconsciously by knowledge of the determinants when determining the outcome. The converse also applies when using historical data, i.e. we must not be guided by knowledge of the outcome when assessing the determinants.

Case 9.2 Prediction of successful vaginal delivery after previous Caesarean section

To help counsel pregnant women who had a previous Caesarean section (CS) to choose between an intended vaginal birth (i.e. a trial of labour, or TOL) and an elective repeat CS, a prediction model was developed to predict the probability of a successful vaginal delivery. Compared to a vaginal delivery, elective CS is associated with more complications and longer recovery. However, an unsuccessful TOL results in an emergency CS, which is associated with even higher risks of major complications, such as hysterectomy and operative injury.

The initial model was developed using logistic regression on a cohort of women recruited through U.S. academic medical centres. The predictor variables contained in the model are shown in ◘ tab. 9.9. The area under the receiver operating characteristic curve (AUC) of the model was 75.4%. The model was subsequently externally validated for usage in the U.S. It was unclear how this model would translate to Western European women because of significant differences in CS rates between the two regions and the lack of relevance of predictors related to ethnicity (being African-American or Hispanic). The model was therefore also externally validated in a Western European cohort. After external validation the AUC was reduced to 68%. This could be due to the aforementioned differences between regions, but also to overfitting of the initial model that may have resulted in optimistic measures of performance (see ► 9.6.3).

Table 9.9 Regression coefficients[a] for predictor variables in two prediction models for successful vaginal delivery in women with a previous Caesarean section (CS)

predictor variable	united states model (intercept: 3,766)	western european model (intercept: 1,876)
age (years)	−0.039	
prepregnancy body mass index (kg/m^2)	−0.060	−0.041
african-american (yes)	−0.671	
hispanic (yes)	−0.680	
any prior vaginal delivery (yes)	0.888	1.339
vaginal delivery after prior cs (yes)	1.003	
recurrent indication for cs (yes)	0.632	
white ethnicity (yes)		0.476
previous cs for non-progressive labour (yes)		−0.688
induction of labour		−0.660
expected foetal weight ≥ 90th percentile (yes)		−0.624

[a]Positive coefficients correspond to a higher probability of a successful vaginal delivery after a previous Caesarean section. The coefficients combined with the model intercept could be used to compute a woman's individual probability of successful vaginal delivery using the logistic function.

To overcome this problem, a new prediction model was developed on a Western European cohort. To reduce overfitting, predictor variables were selected based on previously published evidence and expert opinion instead of statistical significance. tab. 9.9 shows the predictor variables selected for the model and their regression coefficients. The AUC of the model was 70.8%, which was considerably lower than in the U.S. model in the derivation cohort. This could be due to large differences in patient management between regions which could have resulted in different selection mechanisms for women who were subjected to counselling.

9.6.2 Prognostic data to quantify the prognostic model

All sorts of practical problems arise when collecting data for a prognostic study: for example, a complete follow-up cannot be obtained for all the members of the study population (censoring), or useful information has not been obtained on all the potential prognostic determinants for all the members of the cohort (missing values). Assuming that this hurdle has been taken and all the data required to estimate the epidemiological function have been collected, creating the prognostic model is straightforward – similar to the procedure described for the diagnostic function. We first assess each potential prognostic determinant separately and then create a multivariate model containing all the promising prognostic factors. The final stage is to reduce the model to obtain one that is as efficient as possible but can still predict the outcome correctly. The criterion applied here is that a variable can be omitted from the model if the prognostic value of the model as a whole is not reduced to any noticeable extent by the omission. The prognostic value of the model as a whole can be seen – like that of a diagnostic study – from the AUC of the ROC curve (see ▶ 9.3.2).

As with diagnostic tests, the prognostic value of a parameter or test can be determined separately. Just

as with the DOR (▶ 9.3.7) we can calculate a prognostic odds ratio (POR) and a likelihood ratio (see ▶ 9.3.3). Given the nature of the question we are of course particularly interested in the prognostic value of a particular result, the PV (see ▶ 9.6.3). The calculations are similar, as are the problems that arise when interpreting these parameters.

9.6.3 Beware of overfitting when interpreting prognostic models

When interpreting data on prognostic value (PV^+ and PV^-), as with predictive values, we need to remember that the prognostic value is highly dependent on the composition of the population on which the prognostic test is used. Lower incidence of the outcome under consideration, with the validity of the prognostic testing procedure remaining the same, results in lower prognostic value of an abnormal test result.

If, as recommended in this chapter, we opt for an approach in which all the relevant prognostic information is assembled in a multivariate prognostic model, we need to remember that this model is mainly suited to the population for which it was estimated. The prognostic value of the model will usually be less if we apply the same model to a different, comparable population, including the patient population for which we ultimately wish to use it. This overestimation of the prognostic value is due to chance variability in the distribution of parameters in the population contributing to the shape of the original model. The regression coefficients in a different population will therefore be slightly different, and a lower prognostic value will be obtained if we apply the original model to a new population. This problem of overfitting of the model is worse the more prognostic determinants there are and the fewer participants we have at our disposal. It can be solved by testing ('validating') the original model on new populations and recalculating the prognostic value. If this corresponds closely to the original value, the model evidently has external validity and can be used with confidence on patients whose prognosis is not yet known. If the prognostic value of the model as a whole differs substantially from the original value, we can use the data about the test population to see whether the model can be improved. The new model will then need to be revalidated on new populations (see also ▶ case 9.2).

9.7 The examples show how relevant and how difficult diagnostic and prognostic research can be

To conclude this chapter and illustrate the theory discussed here, we now present a case providing an example of evaluation research in diagnostic epidemiology. ▶ Case 9.3 shows how a well-designed study can demonstrate the limited value of a diagnostic test with the aid of an uncontroversial gold standard.

Case 9.3 Diagnostic competence of the Alvarado score for the diagnosis of appendicitis in children

The Alvarado score is a diagnostic tool that combines signs, symptoms and laboratory results for the diagnosis of appendicitis. In children, appendicitis is the major cause of abdominal pain and the most common indication for emergency abdominal surgery. In cases of suspected appendicitis, imaging techniques such as ultrasound and computed tomography (CT) scanning can be used in addition to physical examination and laboratory testing. However, these imaging techniques are costly and may increase the time to final diagnosis due to variable availability, especially in low-resource countries. If accurate enough, the Alvarado score could guide patient management by helping physicians to decide which patients need surgery without the need for imaging.

A prospective observational study was performed among 588 children between 3 and 21 years of age (with a median age of 11.9 years) with abdominal pain suggestive of appendicitis. The final diagnosis was established either by the pathological report in case of surgery or by clinical review after two weeks of follow-up. This combination is regarded as the gold standard for both ruling in and ruling out appendicitis. The Alvarado scores of the participating children

Table 9.10 Distribution of Alvarado scores (test positive/test negative) among children with abdominal pain, prospectively reviewed for appendicitis

alvarado score	appendicitis		total
	yes	no	
≥7	142	75	217
<7	55	316	371
Total	*197*	*391*	*588*

Prevalence = 197/588 = 33.5%
Percentage with correct diagnosis = (142 + 316) / 588 = 77.9%
Sensitivity = 142/197 = 72.1%
Specificity = 316/391 = 80.8%
Positive predictive value = 142/217 = 65.4%
Negative predictive value = 316/371 = 85.2%
Positive likelihood ratio = (142/197)/(75/391) = 3.76
Negative likelihood ratio = (55/197)/(316/391) = 0.35

were not provided to the participating physicians, rendering them blind to the score's result. Of all the children suspected of appendicitis, 197 (33.5%) were ultimately diagnosed with this disease. An Alvarado score of 7 or higher has been published as the cut-off value to diagnose appendicitis and recommend surgery. Using this cut-off value yielded a sensitivity of 72.1% and a specificity of 80.8% (tab. 9.10). Considering that the treatment is invasive surgery, the number of false-positive test results (75 out of 217 test-positives) is high, resulting in a relatively low positive predictive value. Of all the test-positive children, only about 65.4% actually had appendicitis. If the Alvarado score alone were to be used for decision-making, 34.6% of test-positives would be overtreated and would undergo unnecessary surgery. The score performs better at ruling out appendicitis, as indicated by the higher negative predictive value. However, 55 out of 371 children that scored below 7 on the Alvarado score were still falsely classified as test-negative. When using the score to determine who is in need of surgical removal of the appendix, 15% of all test-negatives would not receive the necessary treatment in time. The Alvarado score alone cannot therefore be used to decide the need for surgery, emphasizing the importance of imaging techniques in diagnosing children with suspected appendicitis.

Recommended reading

Bland M. An introduction to medical statistics. 3rd ed. New York: Oxford University Press; 2000.
Fletcher RH, Fletcher SW, Fletcher GS. Clinical epidemiology: the essentials. 5th ed. Baltimore: Lippincott, Williams & Wilkins; 2012.
Grobbee DE, Hoes AW. clinical epidemiology: principles, methods, and applications for clinical research. 2nd ed. Burlington: Jones and Bartlett Learning; 2015.
Gordis L. Epidemiology. 5th edition. Philadelphia: Elsevier Saunders; 2014.
Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: how to do clinical practice research. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2006.
McDowell I. Measuring health: a guide to rating scales and questionnaires. New York: Oxford University Press; 2006.
Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–45.
Newman TB, Kohn MA. Evidence-based diagnosis. New York: Cambridge University Press; 2009.
Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. New York: Oxford University Press; 2008.
Vet HCW de, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.

Source reference (cases)

Schoorel ENC, Melman S, van Kuijk SMJ, et al. Predicting successful intended vaginal delivery after previous caesarean section: external validation of two predictive models in a Dutch nationwide registration-based cohort with a high intended vaginal delivery rate. BJOG; 2014. (Case 9.2)
Schneider C, Kharbanda A, Bachur R. Evaluating appendicitis scoring systems using a prospective pediatric cohort. Annals of Emergency Medicine; 2007. (Case 9.3)

Intervention

10.1 Introduction: research on intended effects differs from research on unintended effects

It will not always be certain whether the aim of a preventive, therapeutic or palliative intervention has actually been achieved. And if it has, was that because of or in spite of the intervention? Again, these are really questions of cause and effect (see ▶ chap. 6). This chapter discusses research into the outcomes of medical interventions that aim to have a beneficial effect on the occurrence or course of a disease (for example vaccinations, health information, rules on diet and lifestyle, reducing pollution, water fluoridation, medication, surgery and psychotherapy).

Intervention studies focus primarily on the intended effect of interventions. Preventive interventions, such as vaccination or health information, claim to be able to prevent a disease occurring. Whether that claim is justified needs to be proved by intervention studies. The aim of therapeutic interventions, such as medication, physiotherapy, surgery and radiotherapy, is to have a beneficial effect on the prognosis of the disease. To some extent this is also true of nursing and palliative or rehabilitation interventions, which are designed mainly to minimize the consequences of a disease. The intervention thus becomes one of the determinants in the epidemiological function. As we want to establish a causal relationship between the change in outcome and the intervention itself we will need to include all the other exposures that affect the health outcome in the study, and therefore in the epidemiological function:

$$P(O) = f(D_i)$$

where D_1 is the intervention and D_2, D_3 ... D_k represent the other determinants of the outcome (confounders).

At first sight it would seem that studying the effects of interventions is no different from etiological research as described in ▶ chap. 6. There is a substantial difference, however: etiological research is by definition concerned with unintended, usually unwanted, effects of specific determinants. Intervention studies, on the other hand, are concerned with estimating intended, usually desired, effects.

Research into the efficacy of medical interventions is usually experimental: participants – generally patients – undergo an intervention with the aim of answering the scientific question about the effect of the intervention. Such experiments are not always possible or necessary. In some cases the effect of medical interventions is better studied using an observational epidemiological design (in particular a cohort or case-control study), for example to investigate adverse effects of medication. The basic principle remains, however: the specific effect of an intervention can only be isolated from the effects of selection, confounding and information bias by means of a properly designed experimental study.

Research into the efficacy and safety of new drugs is strictly regulated. Human intervention studies – once all sorts of in vitro experiments and laboratory animal experiments have been concluded – are divided into four phases:

- Phase I research, where small doses of the product being tested are administered to a limited number of volunteers to study its pharmacokinetics, pharmacodynamics and any toxic effects, sometimes compared with placebo controls.
- Phase II research, which uses relatively small groups of patients (dozens for each dosage arm) who have the disease in question to determine the effective dose. The researchers also look at adverse and other biological effects. These trials are almost always blinded and randomized. Efficacy is not usually demonstrated convincingly in this phase, as the researchers generally focus on intermediate outcomes.
- Phase III research, a series of randomized controlled experiments to ascertain the clinical effects of the dosage selected in Phase II, based on strict protocols and involving far more participants than in Phase II. In these trials the group size is selected so that they will be able to demonstrate a clinically relevant effect with sufficient precision.
- Phase IV research, which comprises the systematic recording of adverse effects once the drug has marketed (post-marketing surveillance), supplemented by research – sometimes imposed on the manufacturer – into certain safety questions that have not yet been sufficiently investigated (long-term effects, high risk populations, etc.).

10.2 The question is always: among whom do we study what outcomes of what intervention against what comparison?

During the development of a medical intervention there is usually only a limited period available to test its efficacy properly. Sometimes it will be too early for a trial, as the optimum form of the intervention, the specific target group, the preferred dosage and/or timing is not yet clear. Sometimes concerns with respect to safety first need to be resolved. Sometimes, on the other hand, it will be too late for a trial, as the intervention is already being used on a large scale, and doctors and patients rightly or wrongly take the view that it would be unethical to withhold the intervention from patients for the purpose of a trial. Sometimes there will be methodological or practical obstacles to carrying out an intervention trial. Research into new drugs and new uses of existing drugs is subject to a large number of statutory provisions and requirements that the researchers and the trial must comply with. This is one reason why it usually takes many years before a new drug can be put on the market.

Once it has been decided to design and carry out a randomized controlled trial, the main question will usually be whether a particular new intervention is more effective than no treatment, a placebo or an existing treatment. It is not just a question of whether the new intervention works better than the reference treatment (effectiveness) but also whether the amount of improvement is worth the cost. This leads us to three questions:

1. What interventions will be compared? A new intervention will often be compared with the existing standard treatment. It goes without saying that it is important to describe precisely in the trial protocol and the final article what the interventions being compared comprise, so that they can be used in practice if so desired. Some complex interventions (in rehabilitation, for instance) are difficult to standardize, but have to be tailored to patients. Adapted designs are required to study the efficacy of such personalised treatments (e.g. an N = 1 trial: see ▶ par. 10.2.7.
2. To which patients are the interventions being given, i.e. what is the domain of the study? For what category of patients does the study aim to reach conclusions? This definition of the domain will determine the composition of the study population and the criteria for inclusion (inclusion criteria) or exclusion (exclusion criteria). For more information see ▶ chap. 4. Restricting the study population by age, gender or other prognostic variables improves homogeneity in the domain, i.e. will reduce the variation in the effects of the interventions in this domain. If the effects are likely to be substantially different in particular subdomains they will need to be reported separately for each subgroup. Specifying the effects for each subdomain may be highly desirable for the sake of generalization and broad application, but it requires such large trial groups that it is not often done in practice. Most intervention studies aim to estimate 'average' effects in 'standard' patients. The source from which the patients are taken also needs to be stated. In ▶ case 10.1, for instance, the study population was restricted to young breast cancer patients (under the age of 50) who entered the menopause early and reported subsequent menopausal problems as a result of cancer treatment.
3. What are the relevant outcome measures? The primary outcome measure follows the main claim for the intervention under consideration. The sample size for the trial is selected accordingly. The primary outcome is not necessarily the most objective or most precise parameter: quality of life may be a more important measure of the effect of a vascular procedure than the extent to which the arteries are actually widened, for instance. In ▶ case 10.1 menopausal problems were the primary outcome measure. Secondary outcomes were self-image and body image, sexual functioning, psychological well-being and quality of life.

Careful preparation and tight organization are required to carry out an intervention study. Special attention should be paid to the following points:

- A detailed trial protocol setting out the entire plan of the study. Important elements in the procedures need to be tested beforehand.

- A carefully thought-out strategy for recruiting participants. The principle underlying the informed-consent procedure is that patients can decide to participate after being informed adequately about the trial.
- A realistic estimate of the number of patients who will be included over a particular time period. As many researchers overestimate the number of inclusions that is feasible and experience problems later on, a very conservative estimate is recommended.
- The physicians and paramedics who will be recruiting the patients and carrying out the interventions should be involved at an early stage. The trial will only be successful if they make efforts to select the right patients and give them the treatment in line with the protocol.
- An ethics committee should assess the ethical and methodological aspects of the research proposal.
- The proposed trial should be registered. The major scientific journals require trials submitted for publication to be registered in an international trial register before they commence. The purpose of publishing the trial protocol in the register in advance is to prevent selective reporting.[1]
- The guidelines for trial reporting, e.g. the CONSORT guidelines, should be applied (see ▶ par. 4.2.2).

Case 10.1 The effect of cognitive behavioural therapy and exercise intervention on menopausal problems in young breast cancer patients

Every year approximately 1.7 in 1,000 women develop an invasive form of breast cancer, 30% of them under the age of 50. Premenopausal women with breast cancer who are treated with chemotherapy or hormone therapy can enter the menopause early. Oestrogen deficiency due to these treatments can cause vasomotor and urogenital problems and associated problems, e.g. insomnia, sexual problems, weight gain and psychological problems. Hormone replacement therapy (HRT) to relieve vasomotor and sexual symptoms cannot be given to patients with a history of breast cancer because of its possible tumour-promoting effects. As there are increasing indications that cognitive behavioural therapy (CBT), including relaxation techniques, is effective in reducing vasomotor symptoms during the natural menopause, and moderate exercise (60–80% of the maximum heart rate) for 2.5-3 hours a week reduces the risk of hot flushes and improves quality of life, a trial has been done into the effect of these interventions. Since each intervention has a different presumed working mechanism, a factorial design was adopted to enable the interventions to be assessed separately and in combination.

The women eligible for this trial were young breast cancer patients (under the age of 50) with menopausal problems caused by their treatment. A total of 422 women were randomized to the CBT group (N = 109), the 'physical activity' group (N = 104), the combined intervention group (N = 106) and the control group (N = 103). They answered questionnaires on inclusion, at the end of the intervention (after twelve weeks) and after six months follow-up. The primary outcome measure in this study was the occurrence of menopausal problems (e.g. hot flushes and night sweats). Other outcome measures (e.g. self-image and body image, sexual functioning, psychological well-being and quality of life) were studied as secondary outcomes. Scores on internationally approved standardized instruments were used as far as possible to measure the effects. The study showed that both CBT and physical activity have a beneficial effect on menopausal problems in young women with breast cancer. Much smaller was the effect found on sexual and physical functioning.

1 ClinicalTrials.gov, a registry of medical trials (website). ▶ http://1.usa.gov/1CN3biq

10.2.1 The observed effect is the sum of the specific effect, the placebo effect, the natural course, the external variables and the measurement error

Suppose a doctor is consulted by a patient who has been suffering for a number of days from certain symptoms and signs suggestive of the presence of disease X. The doctor makes a diagnosis and gives the patient treatment Y. If the patient is back to normal a week later, can the doctor conclude that Y is an effective treatment for disease X? No. Only if patients with an incurable disease recover fully and rapidly after a new treatment (e.g. when a patient with painful osteoarthritis of the hip is given a new hip) can the observed effect be attributed to that treatment with a high degree of certainty. However, such cases are exceptions to the rule. If a patient gets better after a treatment it may be due to the treatment, but there may be other explanations, as the intervention is only one of the determinants of the observed effect. As a rule, the observed effect (OE) of an intervention is made up of five different components:

- The specific effect (SE) of the intervention
- The non-specific effects produced by interaction with the therapist (the placebo effect, PE)
- The natural course (NC) of the disease
- The effect of the external variables (EV) that influence the effect, and
- The measurement errors (MEs) made when measuring the effects.

As a formula:

$$OE = SE + PE + NC + EV + ME$$

This breakdown applies both to the experimental group (e) being given the new treatment and the reference or comparison group (c) being given no treatment or the usual treatment. The difference in the observed effect between the two groups reflects the added value of the experimental treatment if the natural course (NC), the external variables (EV), the placebo effect (PE) and the measurement errors (MEs) are the same for both groups. As a formula:

$$\begin{aligned} OE_e &= SE_e + PE_e + NC_e + EV_e + ME \\ OE_c &= \underline{SE_c + PE_c + NC_c + EV_c + ME_c} \\ OE_e - OE_c &= SE_e - SE_c \end{aligned}$$

only if

$$\begin{aligned} NC_e &= NC_c \\ PE_e &= PE_c \\ EV_e &= EV_c \\ ME_e &= ME_c \end{aligned}$$

It is clear from the foregoing that it is not possible to make a valid estimate of the specific effect of an intervention without a suitable reference group. In a adequate study design we can ensure that the two groups are comparable in terms of the non-specific effects mentioned. Ultimately researchers aim to estimate the specific effect, i.e. the improvement in the course of the disease due to the actual treatment. Unfortunately this specific effect cannot usually be observed directly, but only by ruling out all other explanations.

The placebo effect (PE) of the intervention

In addition to any specific effect, every intervention will also have a non-specific effect, the placebo effect. This will manifest itself in its pure form in patients who are given a placebo drug or 'sham treatment'. This is an intervention that is credible to the patient, but with the presumed active ingredient left out without the patient's knowledge. As a general rule all interventions contain a non-specific (placebo) component as well as a specific component. The difference between specific and non-specific is based on our ideas regarding the working mechanism. If these change, the requirements for a suitable placebo treatment will change too. Although little is known as yet about the mechanism of the placebo effect, it should be emphasized that its influence, even on the 'hardest' effect parameters, is just as 'real' as that of the specific components in the intervention. The magnitude of the placebo effect depends among other things on the faith the patient has in the therapist and his intention to help the patient, and the faith the patient places in the therapy, but price, colour and taste can also be factors.

The natural course (NC) of the disease

Although patients, doctors and paramedics do not always fully realize it, an improvement can take place

without any intervention at all. A person who 'has thrown his back out', for example, will usually get better in a few days, even without treatment. A cold will last a week, with or without treatment. Even chronic patients, such as those with asthma, experience periods when the symptoms get worse (exacerbation) or improve (remission). After a patient has gone to the doctor with severe symptoms he will often enter a phase in which he feels better, even if he has not yet had any effective treatment (or any treatment at all). This phenomenon is referred to as regression to the mean. It goes without saying that if there is no intervention the condition can also get worse, with or without a fatal outcome or functional impairment. In a trial, differences in natural course between patient groups for which interventions are compared will of course have a major influence on the observed effects. In trials it is therefore crucial that the comparison groups of patients have the same average prognosis.

The external variables (EV) that influence the observed effect

An observed effect can be influenced by one or more determinants other than the intervention. These external variables can act as confounders of the effect of the intervention (see ▶ par. 5.4.1). Confounders will influence the magnitude of the observed effect, but if they are equally distributed among the trial groups they will not distort the difference in observed effect between the groups. External variables may already be present when the treatment starts. It is vital that the experimental and reference groups are comparable in terms of these factors. How this can be ensured in the design of the study or in the data analysis is explained in ▶ par. 10.2.2 and 10.2.3. Examples of external variables are age, gender, compliance, duration of symptoms, prior treatment, and any medication or therapy being given in addition to the intervention under consideration. For some diseases we know that lifestyle factors (e.g. smoking and alcohol consumption) may influence the course of the disease and hence the observed effect.

Some external variables do not develop until the treatment has already begun, for instance because they are influenced by it. These external factors, which develop 'along the course of the trial', are a much more complex issue. We must not adjust for intermediate effects of the therapy (see ▶ par. 5.4 and 5.5). For instance, suppose that participation in a fitness programme would rapidly improve cholesterol levels, eventually leading to a reduction in cardiovascular morbidity and mortality. Unintentional adjustment for differences in cholesterol levels that develop during the trial would at least partly mask the specific effect of the fitness programme.

Measurement errors (MEs) in outcome assessment

Random and systematic errors can creep in when measuring the outcome parameters. If the magnitude and direction of these errors are related to the intervention (differential misclassification: see ▶ par. 5.4.1) irreparable bias will occur. Patients who are given a new treatment may for example tell the doctor that they are feeling a lot better simply to please the doctor. Misclassification of this kind occurs mainly when the outcome measurement is subjective and can therefore be influenced by a preference for (or aversion to) one of the treatments being compared. We try to reduce these effects by making the outcome measurement as objective as possible and by standardization. With the aid of placebo treatments blinding becomes another option (see ▶ par. 10.2.4).

10.2.2 Comparing interventions: for explanation or decisions

Therapists and researchers generally approach the assessment of interventions from different angles. Therapists pose a pragmatic question: 'For which patients does this treatment have added value?' Researchers, on the other hand, want to find an explanation: 'What is the specific effect of this treatment?' An analytical study is concerned with the specific effect of the presumed active ingredient in the intervention – a vital piece of information when it comes to understanding the mechanism involved. To demonstrate the specific effect of the active ingredient we need to compare the experimental treatment with a placebo intervention that equals the experimental treatment in every respect except for that active ingredient. A pragmatic study, on the other hand, is concerned with the practical question

whether, for this indication, the new intervention is superior to the established treatment strategy. From a pragmatic point of view it makes no sense to compare the new intervention with a placebo; we want to know the added value of the treatment relative to the current standard of care. Moreover outcome measures usually differ in pragmatic and analytical studies. Pragmatic studies focus on outcomes relevant to patients (pain reduction, improving quality of life, etc.), whereas the interest in analytical studies is on relevant pathophysiological effects (tumour size, blood pressure, etc.). In evidence-based medicine these measures are often referred to as 'surrogate outcome measures', as they are regarded as precursors of the clinical endpoints relevant to patients.

10.2.3 Randomization to create comparability at baseline

Vital to the internal validity of an intervention study is how the interventions being compared are allocated to the patients in the trial. The allocation procedure needs to ensure that the groups of patients being exposed to the different interventions are comparable in terms of natural course (NC) and external variables (EV). If allocation is left up to the therapist the groups will almost certainly not be comparable (in terms of prognosis), as physicians and paramedics want to help patients as best they can and will therefore select whatever treatment has the best chances of success for them. Patients with a good prognosis will therefore be assigned a different intervention than patients with a poor prognosis. In the case of an intervention study involving these patients the results then become seriously biased: this is referred to as confounding by indication. Anticipated risks of a treatment for an individual patient could also play a role here, referred to as confounding by contraindication. Either way, differences in effect will not then be due solely to the intervention chosen but also to differences in prognosis between the patients allocated to the various interventions. There is confounding by indication whenever the prognosis influences the likelihood of being allocated to a particular intervention. In the example in ▶ case 10.1 this confounding could occur if the patients with severe menopausal problems were to be given cognitive behavioural therapy and/or an exercise intervention while the patients with mild problems were not given any intervention (or vice versa). Confounding by indication can also occur without the researchers realizing it. There is confounding by contraindication if a particular intervention is withheld from patients who supposedly run more risk of complications or adverse effects of the intervention.

The standard solution to this problem is to leave the allocation of the various interventions to chance, thus bypassing the preferences of patients, physicians and paramedics. This is the essence of randomization (see ▶ par. 4.2.1). The principle of randomization is simple, but in practice it is important to ensure that the allocation really is random and that at the time of randomization the intervention being assigned to the next patient is unknown and cannot be changed. This is referred to as allocation concealment. Correctly applied, randomization ensures equal distribution of, and hence control over, the exposures involved in the natural course (NC) as well as known and unknown external variables (EV), insofar as they are already present when the treatment is assigned. Randomization cannot of course prevent confounding by external factors that develop later (e.g. compliance and co-interventions), possibly influenced by the assigned treatment. Lastly, randomizing provides the possibility of blinding (see ▶ par. 10.2.4), which makes for comparability of measurement errors (MEs) and placebo effect (PE).

Randomization is regarded as the panacea for potential confounding and is thus the most essential element in a controlled experiment. It does not guarantee that the various exposures will actually be distributed equally among the various interventions, however, as unequal distributions can still occur (by chance). Especially in studies involving only a few patients an unequal distribution can easily occur, causing major problems when analysing the results. To ascertain that bad luck did not cause an unequal distribution on important prognostic variables a post randomization check on these distributions is needed. Of course statistical tests are of no use then, since we already know that any differences have been generated by chance. Such differences in one or more variables, if clinically relevant, need to be ad-

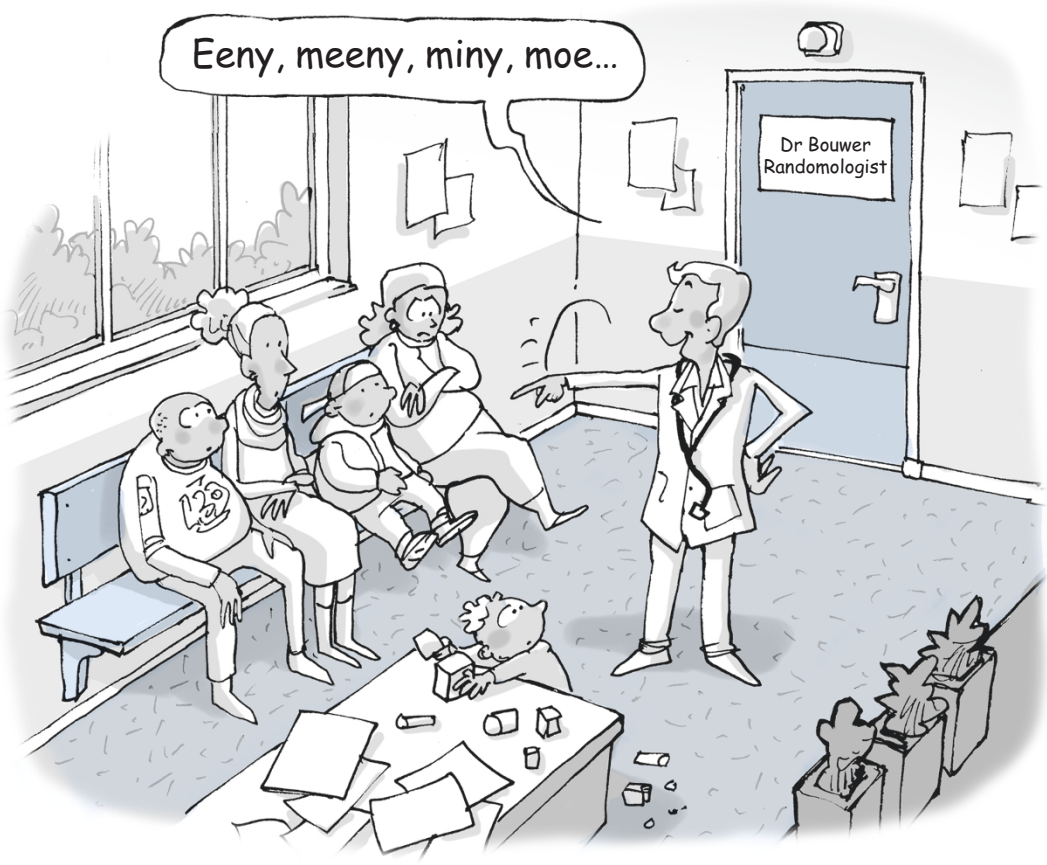


Figure 10.1 Dr Bouwer still needs some training in randomization and blinding

justed for during the analysis. Unfortunately this will partly undo the randomization, and can induce confounding on other variables (measured or unmeasured) that are associated with the variable for which the adjustment has been made (see fig. 10.1).

In the design of the experiment, when there is serious concern about ending up with unequal distribution on one or more strong prognostic factors in spite of randomization (e.g. because relatively few patients have been included), we can help the randomization process by pre-stratification. This involves separate 'stratified' randomization in each category (stratum) of the prognostic factor we are concerned about. This will ensure that that prognostic factor is distributed evenly among the intervention groups. In theory we could continue pre-stratifying for combinations of prognostic parameters until each stratum has only two individuals available for random selection (individual matching). An alternative is to use minimization: this involves manipulating the likelihood of each new patient being allocated to one group or the other to some extent (using a computer algorithm) so as to achieve an even distribution of important prognostic factors.

Pre-stratification can only be used on a few external variables, as the strata will otherwise be too small. It also entails the risk of the strata having different group sizes by chance, resulting in inefficiency in the statistical analyses. Pre-stratification is therefore often combined with block randomization. Block randomization involves random allocation of a fixed number of patients to all the groups in each stratum (see fig. 10.2). This makes the group

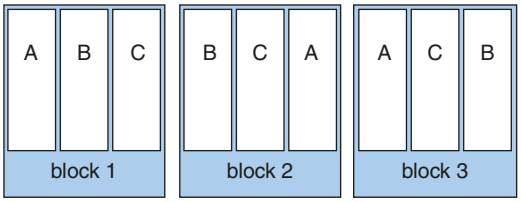


Figure 10.2 Block randomization with a block size of three and three interventions

sizes of the various strata comparable between the groups, both during the trial (in the interim analysis) and at the end of the study. When block randomization is used it is important that only the statistician drawing up the randomization scheme knows the block size, otherwise the researchers will know which treatment the last patient in the block will be allocated to, thus conflicting with the principle of allocation concealment. Block sizes can also be varied to improve allocation concealment.

A major advantage of randomization is that on average it also creates comparability for external confounders that we do not know or have not measured and cannot therefore stratify for (either before or after analysis). The larger the study population, the more this will be the case.

For a patient, an intervention study effectively starts at the time of randomization. The principle is that once patients have been randomized they are members of the study population. A randomized controlled experiment is in effect a cohort study: randomization is the event that defines cohort membership (see ▶ par. 2.3.1). This has major repercussions for the way in which the data, once collected, are analysed. It is vital for each patient – whatever happens – to be linked in the analysis to the intervention to which he was originally allocated during randomization (intention-to-treat analysis), regardless of the actual intervention carried out. This point is developed in ▶ par. 10.3.2.

10.2.4 Blinding to maintain comparability right to the end

In order to attribute a difference in outcomes between intervention groups to the effects of the interventions, the influence of external factors and measurement errors needs to be comparable for the comparison groups (see ▶ par. 10.2.1). This condition will not be met if knowledge about who received which intervention can influence external factors following randomization or measurement errors (MEs). Think for instance of differences in placebo effect or in the use of co-interventions (performance bias). If the patient knows he is being given a 'sham treatment' there will be less placebo effect and the specific effect of active treatment will be overestimated. The results will also be distorted if the rater has a biased view of the disease outcome and interprets the condition of patients treated with his preferred intervention in a more favourable light (observer bias).

The solution to these potential forms of bias is blinding, which masks which patient is receiving which treatment (see ▶ par. 4.2.1). Ideally everyone who could create bias (differential measurement errors) – patients, therapists, effect raters and those analysing the data – should be blinded, but such full blinding is really only possible with perfect placebos.

Full blinding is not possible in many cases. In ▶ case 10.1, for example, the physiotherapist or the person giving the CBT cannot be blinded. With some creativity (sham operations, sham radiotherapy and even sham psychotherapy), however, a host of credible placebo interventions can be constructed. Blinding the person analysing the data is not common, yet such blinding has demonstrated added value for bias reduction (fig. 10.3).

Opinions are divided on whether blinding is necessary for pragmatic studies (see ▶ par. 10.2.2). Advocates say that this is the only way of determining the added value of the experimental treatment (compared with the standard treatment). Opponents, on the other hand, say that external effects of the treatment (including 'faith' in the effect) are simply inherent in a medical intervention and as such should be included in the estimate of the treatment's added value. An oft-heard maxim is: 'A treatment doesn't have to work as long as it helps'.

10

Figure 10.3 This is a case of full blinding

10.2.5 The cross-over trial

So far we have been looking at a study design in which patients are divided (randomized) into two or more intervention groups. Each patient is allocated to only one intervention and thus contributes to the outcome assessment only once. This type of design is referred to as parallel (see fig. 10.4). The implicit assumption is that the difference in disease outcome between the two groups gives a valid estimate of the effect that would have occurred if we had been able to compare the effects of the two interventions on disease outcome in each of the patients. In large randomized groups this is a reasonable assumption, but in smaller trials incomparability of the trial groups can be a problem.

In some cases (where the effects are rapid, short-lived and reversible) it is possible to give patients first one and then the other treatment (serial design). Aside from the fact that this is not always possible, in a serial design we do not have proper control for the effect of the natural course of the disease.

A cross-over trial (fig. 10.5) combines the advantages of a parallel design and a serial design. Here randomization determines not who receives which treatment but in what order the interventions are carried out on each patient. Thus each patient provides his own reference for the effects of the various treatments, and we can adjust for natural course by comparing the different orders. As each patient is compared with himself in the statistical analysis, the standard deviation becomes much smaller and far fewer patients will be needed in a cross-over experiment than in a parallel design. Although the cross-over design is therefore very attractive, and highly popular in pharmacological research, it is rarely feasible in the non-pharmacological domain. Not only must the effect of the intervention be rapid and reversible, the effect of one intervention must also not continue during the period when the second intervention is being given (i.e. no carry-over effect or selective dropout), otherwise the patient groups will not be comparable in the second half of the experiment.

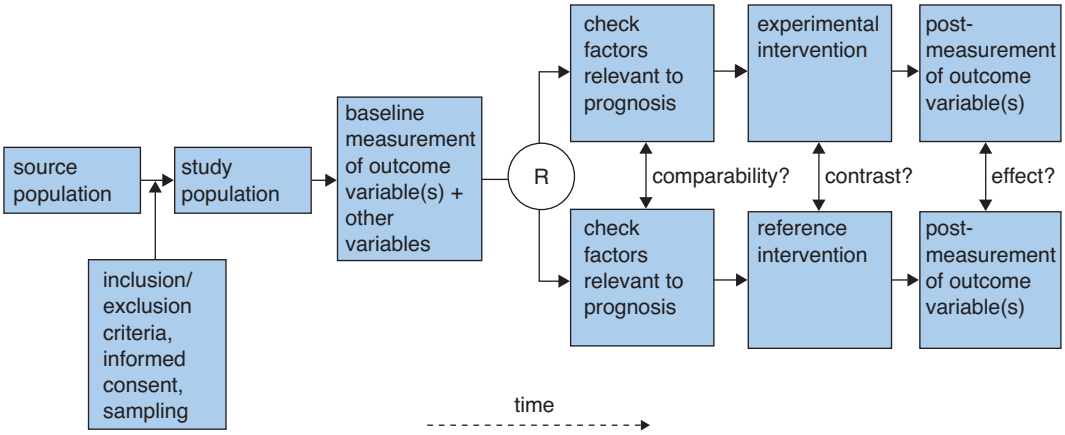


Figure 10.4 Basic structure of a parallel design trial

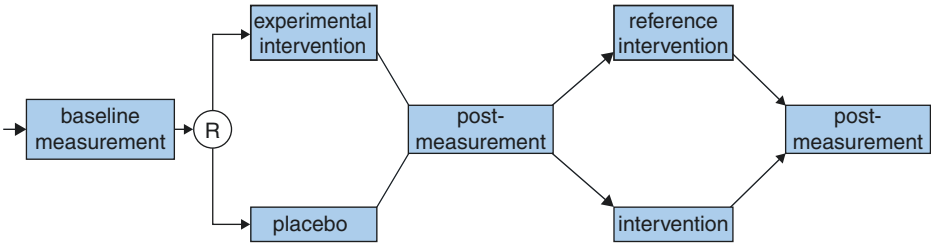


Figure 10.5 Basic structure of a cross-over trial

10.2.6 The factorial trial

A factorial trial enables the effect of two interventions (A and B) to be studied at the same time in a single patient group. This design may save costs, and also enables the interaction between the two treatments to be studied. In its simplest form, a two-by-two factorial design, the participants are randomized into four groups. Group 1 receives both treatments, Groups 2 and 3 receive one of the two, and Group 4 receives neither of them. This design was used in ▶ case 10.1. It is even possible to use placebos for both treatments, enforcing blind assessment: each participant receives two interventions, of which neither, one or both can produce a specific effect (tab. 10.1).

The added value of the combined intervention can be assessed by comparing the effect in the combination group (group 1) with the sum of the two separate effects (groups 2 and 3). If the combined effect is substantially different (higher or lower) from the sum of the two effects, there must be interaction. In this case interaction, also referred to as 'effect modification', means that the strength of the effect of one intervention depends on whether the other intervention is also given simultaneously (see ▶ par. 5.5). If the primary aim of the trial is to study interaction, each treatment group of course must include enough patients.

10.2.7 The N=1 experiment

Results of randomized controlled experiments indicate which of the treatments being compared are most successful on average. Sometimes we encounter a situation where it is desirable and feasible to test which treatment gives the best result in an individual

Table 10.1 Basic structure of a factorial trial

	intervention B	placebo (B)	
intervention A	group 1	group 2	groups 1+2
placebo (A)	group 3	group 4	groups 3+4
	groups 1+3	groups 2+4	

Effect of A: Groups 1+2 – Groups 3+4
Effect of B: Groups 1+3 – Groups 2+4
Effect of A and B: Group 1 – Group 4
Interaction effect: Group 1 – Groups 2+3

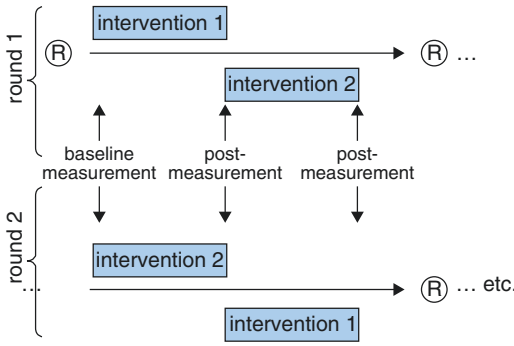


Figure 10.6 Basic structure of the *N* = 1 experiment

patient. Such situations occur e.g. when patients with particular characteristics have been excluded from the intervention study that has been carried out, or when the available results have not produced any clear conclusions. If the outcome is reversible and the treatments' effects show up relatively quickly, we can study the outcome in an individual patient who receives both treatment alternatives at least once but in random order. This design has the characteristics of a serial design but with a repeating, random allocation of the treatment sequence (see fig. 10.6).

It is important to repeat the treatment episodes several times to obtain enough outcome events and eliminate the role of chance, natural course and external effects. N=1 experiments are therefore only possible on stable patients who have a chronic condition with a treatment effect that comes and go relatively quickly.

Although informed consent is required, just as in a regular experiment, the great advantage of an N=1 experiment is that the patient benefits directly from the results. The scientific question is what the results of an N=1 experiment mean for similar patients. The answer is 'not very much', as there is no representative sample of the group to which we would like to extrapolate the results. The question can be answered better if we carry out a number of N=1 experiments with the same research question and design. Single N=1 experiments can produce hypotheses, prompt similar trials and provide the basis for larger trials on patient groups. Several analogous N=1 experiments can be summarized in a meta-analysis and thus produce results that can be generalized.

10.3 Data from experimental studies is analysed to produce a valid estimate of effect

Data from experiments can be used to arrive at valid and precise estimates of the relative effect of a treatment. It is important to retain the methodological strengths of an experimental design (the comparability of natural course, external variables and measurement errors) and even enhance them where possible. Data analysis is once again based on the epidemiological function, which calculates the likelihood of a particular outcome as a function of the intervention and any other determinants of the effect.

10.3.1 Successful randomization needs to be checked and adjusted if necessary

On average randomization creates equal distributions for all external variables (EV). It is still a good idea to check whether this has been accomplished for all external variables included in the study. Should an unequal distribution of extraneous determinants occur, the resulting confounding can be adjusted for in the data analysis: some methods are described in ▶ par. 5.4.2, referred to as 'stratified analysis' and 'multivariable regression'. In the example of ▶ case 10.1 randomization could have allocated more people who play sport to the exercise therapy, or more depressive women to the cognitive behavioural therapy, by chance. If we can assume that sports have a positive effect on menopausal problems and depression a negative one, there will be confounding. Provided that both variables have been assessed in all participants, confounder adjustment can be done using stratified analysis. To that end intervention effects are assessed separately for those who play sports and those who do not. Then the average (weighted) effects for these two strata are calculated. A similar analysis can be carried out on the depressive and non-depressive women. Note that stratified analysis is not a solution for unmeasured external variables (confounders). We just have to hope that randomization has done its job for these variables.

The issue of incomparability of groups on external variables (confounding) should not be confused with the issue of subgroup differences in effects. If these differences exist (effect modification: see ▶ par. 5.5), reports should show the effects for each subgroup separately together with the corresponding confidence intervals. Most intervention trials, however, do not have enough participants to distinguish such subgroup differences in effects from random variation.

10.3.2 What about patients who have not been treated in line with the protocol?

The experimental design assumes random assignment of the interventions to each patient. Yet some patients will not receive the intervention (or the full intervention) assigned to them. Some may be too ill for surgery, for example, or they may not take their medicine or fail to comply with the rules in some other way. In ▶ case 10.1 women were sometimes absent from the group therapy (CBT) sessions, and not all participants complied with the exercise programme. In such situations there are two options for data analysis:

- In an intention-to-treat analysis every patient is included in the analysis as originally planned, regardless of whether or not he or she actually received the treatment. Such 'contamination' might produce underestimation of the effect, but the advantage of this approach is that it keeps randomization intact.
- In a per-protocol analysis patient groups are compared in the analysis according to the actual treatment they received. Obviously the breakdown of the trial groups is then a better reflection of what the patients actually received, but the original comparability of the groups accomplished by the randomization is lost. The experiment then effectively becomes an observational study, with all the disadvantages that that entails. Suppose compliance with treatment A is lower than with treatment B because treatment A does relieve patients' symptoms sufficiently. Not including patients who discontinue their allocated treatment in the analysis would produce biased results (the effect of treatment A will be overestimated).

10.3.3 Sequential analyses to overcome researchers' impatience

It is tempting to analyse the available data while a randomized controlled experiment is in progress, in particular to see whether it is already clear which intervention is superior. If so, the researchers will want to halt the experiment and offer all subsequent patients the more effective treatment. It would indeed be unethical to include more patients in an experiment than needed to answer the original research question. This argument is fair, yet it carries a major risk that we will never find out whether this initial positive effect was merely due to random fluc-

tuation. Special sequential analysis procedures have been developed to deal with this dilemma. These involve prior formulation of the statistical limits that need to be exceeded before a decision is taken to halt the experiment while in progress. These limits are set to give chance only a limited role. A variant of this method is to plan in advance that an interim analysis will be done, for example when outcome has been measured in half of the study population. Part of that prior plan is a decision rule about what results of the interim analysis will lead to the trial being halted in the event of an unexpectedly large (or negligible) effect.

10.4 Randomized trials are not needed to investigate unintended effects

As we have seen in this chapter, the randomized controlled experiment is a superior study design to estimate intended effects of interventions, mainly because it enables the specific effects of an intervention to be differentiated from the effects of natural course, external variables and measurement errors. For the study of unintended effects the experiment will often be impractical. A study of rare severe adverse effects, for instance, would require gigantic numbers of participants. Lack of severe adverse effects in a randomized controlled trial of regular size does not provide much evidence about the safety of the experimental intervention.

Fortunately an experimental design is not needed to study unintended effects, as confounding by indication (see ▶ par. 10.2.3) is not really a problem in research questions of this kind: the indication for a particular intervention is based primarily on the possibility of having a beneficial effect on the prognosis, not on producing a particular adverse effect. There are exceptions to this rule – for example where a particular treatment is deliberately being avoided because of the risk of an allergic reaction – but in those cases the adverse effects are usually known and no longer a topic for research.

Research into adverse effects of medical interventions is the domain of observational research, as described in ▶ chap. 6. These studies thus obviously have to contend with the same problems as etiologic studies (in particular bias due to confounding, selection bias and information bias), but these problems, unlike that of confounding by indication, can be resolved both theoretically and practically in observational research. This section can therefore be brief, as virtually all the relevant information has already been provided in ▶ chap. 4-6. We shall briefly consider some specific aspects of research into adverse effects below.

10.4.1 Adverse effects come in all shapes and sizes

Drugs are by definition toxic substances designed to interfere with particular aspects of metabolism. The whole development of drugs aims, of course, to select those that act selectively on the pathological process without affecting other, vital functions. This is only partly successful: there will always be a risk of unintentional, usually harmful, adverse effects (■ fig. 10.7). Although avoiding adverse effects has always been an inherent part of the development and evaluation of drugs, efforts to identify adverse effects were greatly stepped up after the thalidomide scandal, when mothers who took this drug during pregnancy gave birth to children with no limbs.

There are various types of adverse effects:

- Type A reactions are predictable based on the mechanism of the drug. They are generally relatively common, especially at higher doses, and are therefore usually discovered soon, in some cases even before the drug is tested on trial participants. Hair loss due to chemotherapy is an example of a Type A reaction.
- Type B reactions are unpredictable and rare, and are therefore generally only discovered after the drug has been placed on the market. These are the subject of Phase IV studies (post-marketing surveillance). Vaginal cancer in daughters of women who took diethylstilbestrol (DES) as a contraction inhibitor at the end of pregnancy is a well known example.
- Type C reactions are usually also rare but more or less predictable. They are caused by drugs exacerbating rather than reducing the severity of the condition for which they are prescribed. They are very difficult to detect, because it is difficult to distinguish between the effect of the

Figure 10.7 Nurse, I stopped taking the tablets. The cure is worse than the disease

underlying disease and that of the drug. They usually only come to light in large-scale Phase IV studies. An example of a Type C reaction is increased risk of severe asthma and asthma mortality in patients given certain bronchodilators (fenoterol) for their asthma.

It is clear from the foregoing that different approaches are needed to detect and demonstrate the various types of adverse effects. Type A and C reactions can be predicted rationally to a large extent, leading to specific hypotheses that can be tested. Detecting Type B reactions is mainly a question of vigilance on the part of patients and therapists and the untargeted, systematic monitoring of patients who have been treated.

10.4.2 Post-marketing surveillance to detect adverse effects; cohort and case-control studies to test hypotheses

Phase IV research, also referred to as post-marketing surveillance (PMS), is legally required for every drug in the first few years after its approval. The most common adverse effects are usually discovered early on, long before the drugs are regularly prescribed to patients. Occasionally, however, a drug is taken off the market as a result of findings from post-marketing surveillance. It is often vigilant patients and therapists who first express suspicion of a relationship between a drug and the symptoms and signs it has produced. Therapists are expected to report the adverse effect (or suspicion) to the national pharmacovigilance centre and/or the drug manufacturer. These reports are evaluated systematically (also at the European level, by the European Medicines

Agency, EMA)[2] and compared with other reports worldwide. If a report gives rise to serious suspicions the drugs inspectorate may intervene directly. It goes without saying that spontaneous reports of suspected adverse effects can only be systematically collected and organized after drugs have been approved and allowed onto the market.

If there are suspicions regarding possible causal relationships between the use of a particular drug and the occurrence of particular health problems, it makes sense to test them using one of the appropriate types of etiological research. Because of the rarity of most adverse effects the obvious choice is a case-control study, but a cohort study will sometimes be an option. Some regions in the Netherlands, for instance, keep records of all drugs issued by pharmacies (ordered by person and by date) and link these to hospital admission records of the same population. This enables testing of particular hypotheses on increased risk for adverse effects in patients using particular drugs. From the database cohorts of users of specific drugs (ideally for the same indication) are compared for the occurrence of the adverse condition in question. As with any cohort study, the validity of the results of such a study will depend on the availability of data on the determinants and the possibility of overcoming issues of confounding and effect modification. Analyses done on PMS databases will also usually have a cohort study approach.

All the possibilities and limitations of cohort and case-control studies mentioned in ▶ chap. 4 apply here too. For the sake of brevity we would refer to that chapter.

10.5 Examples show the broad applicability of randomized controlled trials

We shall close the book, and this chapter, with some examples of experimental intervention trials. ▶ Case 10.1 has already provided an example of an evaluation of an exercise intervention and CBT treatment. The other cases describe other types of intervention: pharmaceutical (▶ case 10.2), surgery (▶ cases 10.3 and 10.4), artificially induced labour (▶ case 10.5) and preventive measures (▶ cases 10.6, 10.7 and 10.8). These cases will give readers some idea of the broad applicability of randomized controlled trials.

Case 10.2 Use of lidocaine in suspected heart attacks

The first few hours after a heart attack are particularly risky. Arrhythmia in particular can cause rapid and irregular contraction of the heart ventricles (ventricular fibrillation) and ultimately cardiac arrest. If the patient is not resuscitated using cardiac massage or electrical stimulation he will die. Lidocaine may be able to prevent the arrhythmia. A large-scale study (the Amsterdam Lidocaine Intervention Trial) therefore investigated the effect of its intramuscular administration while the patient is being transported to hospital. The trial was conducted by the ambulance services that transport patients in the Amsterdam region to fifteen hospitals. Over a period of just under three years over 6,000 patients with a suspected heart attack were transported. Following randomization half of them were injected with lidocaine. Some patients were excluded, including those whose GP had already done this, those whose heart rate was lower than 45 bpm and those who had already been resuscitated because of ventricular fibrillation before randomization. No prior informed consent was obtained because of time pressure and the patients' condition.
The patients were injected using an automatic syringe, omitting the needle and the liquid for the control group. The ambulance team was not aware of whether the injection or sham injection was given until after the procedure was done. Only 32% of the randomized patients were subsequently diagnosed with acute myocardial infarction. There was a slight difference between the lidocaine group (34%) and the control group (31%) in this respect. The authors argued that this could be due to lidocaine

2 European Medicines Agency (website). ▶ http://bit.ly/LZ0yUm

Table 10.2 Ventricular fibrillation (VF) within 15 minutes and between 15 and 60 minutes after randomization and numbers of actual or potential deaths

	lidocaine (N = 2,987)	controls (N = 3,037)	*p*-value
VF < 15 min.	6	5	NS
15 min. < VF < 60 min.	2	12	< 0.01
Total	8	17	0.08
VF, survived	6	15	0.05
VF, died	2	2	NS
No VF, died	17	19	NS
total[a]	25	36	NS

[a] Number of patients who died plus VF patients who were successfully resuscitated

affecting the level of the enzyme creatine kinase in the blood plasma, on which the diagnosis was usually based. A blinded analysis of the electrocardiograms (ECGs) carried out on admission and after 24 hours in hospital did not show any difference between the trial groups in the frequency of acute myocardial infarction. The main outcome measure was the occurrence of ventricular fibrillation in the first hour after randomization. This was measured by recording each patient's ECG for one hour from the time he was put in the ambulance (half a minute before randomization). Ventricular fibrillation occurred in 8 patients in the lidocaine group, as against 17 in the control group. When this was only checked from 15 minutes onwards, once the level of lidocaine in the blood was adequate, the figures were 2 and 12 respectively. As these patients were generally successfully resuscitated in the ambulance using electrical stimulation, no difference in mortality between the trial groups was found. The figures are shown in tab. 10.2. The results also showed that lidocaine can successfully reduce excessive heart rate (tachycardia). The authors concluded that lidocaine administered soon after a suspected myocardial infarction can prevent ventricular fibrillation. They recommend rapid injection of lidocaine by the patient himself, the GP or a paramedic, especially in cases where electrical defibrillation cannot be carried out.

Case 10.3 Removing wisdom teeth

Removing wisdom teeth can be quite unpleasant, mainly because of the pain that develops once the anaesthetic has worn off. Paracetamol (1,000 mg) has traditionally been prescribed to relieve the pain, but other painkillers (NSAIDs, non-steroidal anti-inflammatory drugs) are also popular nowadays. As it was not entirely clear whether NSAIDs do actually provide better pain relief than paracetamol, 40 patients who were to have their lower wisdom teeth on both sides removed were asked to take part in a study of the effects of the two drugs. As these procedures are often divided over two sessions (one for each side), the researchers opted for a cross-over design. Another reason is that pain is subjective and varies widely among patients. As a cross-over design was chosen, the patients taking part formed their own control group, as it were, and possible differences between patients were tackled. After the first procedure the participants received pain relief in the form of either an NSAID or paracetamol, depending on the randomization. Following the second procedure the group who had been given an NSAID

received paracetamol, and the paracetamol group received an NSAID. Four of the 40 participants dropped out: two did not show up for the second procedure, one developed a wound infection and one did not comply with the medication protocol.
The outcome measures were pain, measured on a 100 mm visual analogue scale, and swelling, measured using a face bow capable of detecting volumetric changes in the face. No clinically relevant or statistically significant difference was found between paracetamol and NSAID in terms of acute postoperative pain and postoperative swelling.

Case 10.4 Surgical treatment for lumbar spinal stenosis

Lumbar spinal stenosis is a degenerative condition that is characterized by low back pain and leg pain caused by a gradual narrowing of the spinal canal. In addition to pain, patients often report a decrease in quality of life and walking ability. One mode of treatment is removing bone tissue to alleviate the nerve root impingement. In addition to decompression surgery, many patients also undergo fusion surgery as an adjunct. However, the evidence in favour of the addition of lumbar fusion was weak, and no randomized controlled trials had been performed.
To assess the efficacy of fusion surgery in addition to decompression surgery alone, a randomized multicentre trial was performed in Sweden. The authors compared decompression surgery alone to compression surgery plus lumbar fusion surgery. Neither the patients nor the surgeons were blinded, but the surgeons were not involved in the assessment of the outcomes. All the patients had been diagnosed with lumbar spinal stenosis and had to have had symptoms for over six months, pseudoclaudication in one or both legs, leg pain and one or two adjacent stenotic segments. In total, 247 patients were randomized using computerized randomization stratified for the presence of degenerative spondylolisthesis, a condition in which one vertebra slips forward over the one below it. The primary outcome measured was the Oswestry Disability Index (ODI).
Patients in the group receiving both decompression and lumbar fusion were operated upon for a significantly longer time and lost more blood during the procedure compared with those receiving only decompression. After two years of follow-up, patients in the decompression group improved even more on the ODI than patients in the combination group, but the difference was not statistically significant ($p = 0.36$). In addition, none of the secondary outcomes differed significantly in favour of the combination procedure. The researchers concluded that adding fusion to decompression surgery has no clinical benefit.

Case 10.5 Immediate delivery to treat hypertensive disorders of pregnancy

Hypertensive complications of the pregnancy, which occur in approximately 10% of all pregnancies, can only be resolved by delivery of the placenta. After delivery, progression is halted. Artificially induced delivery is not always an option, as preterm delivery is associated with an increased risk of neonatal complications. Only a few studies have focused on the way women with hypertensive disorders in late preterm pregnancy (i.e. between 34 and 37 weeks of gestational age) should be managed.
To evaluate the difference between immediate delivery and expectant monitoring aimed at prolonging pregnancy until 37 weeks (i.e. early-term pregnancy) a non-blinded randomized controlled trial was performed in 51 hospitals. A total of 897 women were asked to participate in the study if they had gestational hypertension, pre-eclampsia or deteriorating pre-existing hypertension and had a gestational age between 34 and 37 weeks. In the end, 704 women were randomly assigned to immediate delivery by induction or Caesarean section starting within 24 hours of randomization, or

expectant monitoring. The expectant monitoring group consisted of women monitored as outpatients, through a home care programme, or as inpatients, depending on their condition. The primary maternal outcome selected was a composite measure consisting of thromboembolic complications, pulmonary oedema, the syndrome of haemolysis, elevated liver enzymes and low platelet count (HELLP), eclampsia, placental abruption and maternal death. The researchers defined the primary neonatal outcome as the need for supplemental oxygen for more than 24 hours and radiographic findings indicative of respiratory distress syndrome.
Only in 4 women (1.1%) in the immediate delivery group did the primary outcome occur, compared to 11 women (3.1%) who were assigned to the expectant monitoring group. The relative risk was not statistically significant (RR = 0.36, 95% confidence interval (CI) 0.12–1.11). However, a clinically relevant and statistically significant difference was observed with respect to the primary neonatal outcome, in favour of the expectant group. In the immediate delivery group 20 neonates (5.7%) experienced the primary outcome, compared with 6 (1.7%) in the expectant monitoring group. The relative risk was 3.3 (95% CI 1.4–8.2, p = 0.005). The authors concluded that it is not justified to immediately induce delivery in women with hypertensive complications of pregnancy between 34 and 37 weeks of gestation, because of the increased risk of neonatal respiratory distress.

Case 10.6 Stopping smoking

Stopping smoking is highly cost-effective, especially from the medical point of view. Unfortunately, not everyone manages to stop smoking in the long term, and many smokers fall back into their old habit. Using nicotine replacements doubles the likelihood of an attempt to stop being successful, but there are other products on the market that are used as aids to stopping, such as varenicline (a nicotinic agonist), bupropion and nortriptyline (an antidepressant). The mechanism of these drugs is thought to be based on reducing the psychological and somatic symptoms that occur when stopping smoking. There are indications that the simultaneous use of nicotine replacements and the antidepressant nortriptyline could produce a complementary effect.
A total of 901 patients was recruited from a number of anti-smoking clinics and randomized into two groups. One group was given nicotine replacements and a placebo for twelve months, while the other was given nortriptyline and nicotine replacements for twelve months. The medication for both groups looked identical, and neither the patients nor the trial team knew which patients received which medication (a double-blind design). All the patients were allowed to attend as many Stop Smoking groups as they thought necessary. The main (primary) outcome measure was a successful stopping attempt within six months of the start of the study. Secondary outcome measures included a successful attempt after twelve months, the urge to smoke and mood changes. Whether the participants had actually stopped was checked by analysing a saliva sample. The nortriptyline group experienced some adverse effects (dry mouth and constipation). The combined use of nortriptyline and nicotine replacements did not lead to significantly more successful attempts than the use of nicotine replacements only. The authors concluded that, given the side effects, it is not a good idea to routinely prescribe nortriptyline in addition to nicotine replacements.

Case 10.7 Web-based intervention to prevent major depression

Major depressive disorders (MDD) are expected to become the leading cause of premature mortality and disability in high-income countries by 2030. As evidence-based treatments are not very successful in improving functional and health outcomes, many studies focus on prevention.

German researchers developed an online guided self-help intervention to prevent MDD for individuals who have sub-threshold depression. They recruited participants through a large German health insurer and by using newspaper articles, on-air media and related websites. All the participants were screened online to assess whether they were experiencing sub-threshold depression (using the Centre for Epidemiological Studies Depression Scale, or CES-D) and were 18 years of age or older. If they were already receiving psychotherapy, were on a waiting list for it, or had been treated with psychotherapy in the past six months, or were at notable risk of suicide (assessed using the Beck Depression Inventory), they were excluded from the study. All 406 participants had unrestricted access to usual care and were randomly allocated to receive either the web-based guided self-help intervention or enhanced usual care. The web-based intervention consisted of six sessions based on psychoeducation, behaviour therapy and problem-solving therapy. The primary outcome of the study was the time to onset of MDD assessed using a telephone-administered structured SCID at six and twelve months follow-up. In the web-based intervention group 27% progressed to MDD, compared with 41% of the participants in the comparison group. The incidence rate ratio was 0.60 (95% confidence interval (CI) 0.42–0.84). The log-rank test, used to test for differences between the Kaplan-Meier curves of both groups, yielded a P-value of 0.002, indicating a difference in incidence rates over time. The number needed to treat to avoid one incident case of MDD was 5.9 (95% CI 3.9–14.6). The authors concluded that a web-based guided self-help intervention could reduce the incidence of MDD over twelve months, compared with enhanced usual care.

Case 10.8 Lifestyle intervention for obese primary care patients

Primary care physicians play a vital role in dealing with obesity in their patient populations. Losing weight and increasing physical activity are paramount, as obesity is associated with increased morbidity and mortality. For instance, obese people are more prone to hypertension, diabetes, coronary heart disease, stroke and cancer. Lifestyle interventions specifically tailored towards the individual are hypothesized to reduce weight more effectively and have more long-lasting results, compared with generic lifestyle interventions.

Researchers conducted a randomized controlled trial among 211 obese patients from the practices of 24 primary care physicians in Rhode Island (USA). Participants had to have a body mass index of at least 25 kg/m^2 to be included. All the patients received twelve months of treatment by dieticians trained as lifestyle counsellors, aiming at weight loss and lifestyle changes. It consisted of three face-to-face meetings over a year-long period to discuss their weight loss goals and structured meal plans. The intervention group received monthly phone calls from the lifestyle counsellor during the first six months in addition to the meetings, and weekly mailings for a whole year. The mailings consisted of printed materials, feedback on food and exercise logs and exercise-related DVDs. The control group only received pamphlets on weight loss, physical activity and healthy diet.

The proportion of patients who lost 5% or more of their baseline weight was significantly higher in the tailored intervention group during the first year. At six months 37% of the intervention group experienced this outcome, compared with just 12.9% of the comparison group ($P < 0.01$). At twelve months the contrast was even larger. In the intervention group 47.8% of participants lost 5% or more of their weight, compared with 11.6% ($P < 0.01$) in the comparison group. However, in the year after the intervention the differences decreased substantially. At 18 and 24 months the proportions of patients

having lost 5% or more compared with baseline were 31.4% versus 26.7% (P = 0.64) and 33.3% versus 24.6% (P = 0.39) respectively. The authors concluded that, although the intervention was effective, it needs to be better understood how much continued contact through which channel is needed to maintain these differences over a longer period of time.

Recommended reading

Gordis L. Epidemiology. 5th ed. Philadelphia: Elsevier Saunders; 2014.

Grobbee DE, Hoes AW. Clinical epidemiology: principles, methods, and applications for clinical research. 2nd ed. Burlington: Jones and Bartlett Learning; 2015.

Jadad AR, Enkin MW. randomized controlled trials: questions, answers and musings. 2nd ed. London: BMJ Books; 2007.

Piantadosi S. Clinical trials: a methodologic perspective. 2nd ed. New York: Wiley; 2005.

Straus SE, Richardson WS, Glaszlou P, Haynes RB. Evidence-based medicine: how to practice and teach It. 4th ed. New York: Churchill Livingstone; 2010.

Weiss NS. Clinical epidemiology: the study of the outcome of illness. 3rd ed. New York: Oxford University Press; 2006.

Source references (cases)

Duijts SFA, Beurden M van, Oldenburg HSA, Hunter MS, Kieffer JM, Stuiver MM, et al. Efficacy of cognitive behavioral therapy and physical exercise in alleviating treatment-induced menopausal symptoms in patients with breast cancer: results of a randomized, controlled, multicenter trial. J Clin Oncol. 2012;30(33):4124–33 (Case 10.1).

Koster RW, Dunning AJ. Intramuscular lidocaine for prevention of lethal arrhythmias in the prehospitalization phase of acute myocardial infarction. N Engl J Med. 1985;313:1105–10 (Case 10.2).

Bjørnsson GA, Haanæs HR, Skoglund LA. A randomized, double-blind crossover trial of paracetamol 1000 mg four times daily vs ibuprofen 600 mg: effect on swelling and other postoperative events after third molar surgery. Br J Clin Pharmacol. 2003;55(4):405–12 (Case 10.3).

Peter Försth GO, Carlsson T, Frost A, Borgström F, Fritzell P, Öhagen P, Michaëlsson K, Sandén B. A randomized, controlled trial of fusion surgery for lumbar spinal stenosis. N Engl J Med 2016;374:1413–23 (Case 10.4).

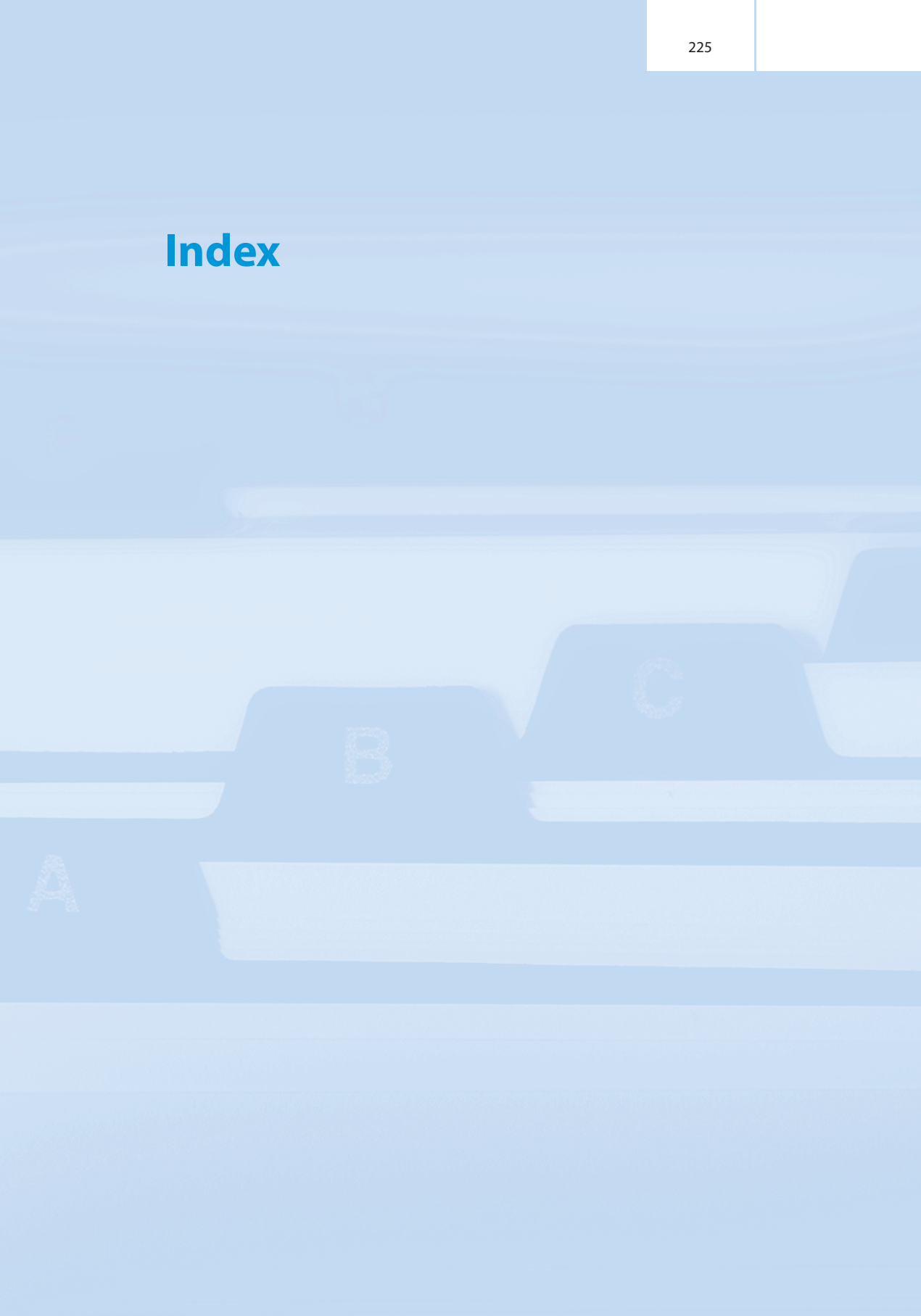
Broekhuijsen K, van Baaren GJ, van Pampus MG, Ganzevoort W, Sikkema JM, Woiski MD, Oudijk MA, Bloemenkamp KW, Scheepers HC, Bremer HA, Rijnders RJ, van Loon AJ, Perquin DA, Sporken JM, Papatsonis DN, van Huizen ME, Vredevoogd CB, Brons JT, Kaplan M, van Kaam AH, Groen H, Porath MM, van den Berg PP, Mol BW, Franssen MT, Langenveld J; HYPITAT-II study group. Immediate delivery versus expectant monitoring for hypertensive disorders of pregnancy between 34 and 37 weeks of gestation (HYPITAT-II): an open-label, randomised controlled trial. Lancet. 2015;385:2492–501 (Case 10.5).

Aveyard P, Johnson C, Fillingham S, Parsons A, Murphy M. Nortriptyline plus nicotine replacement versus placebo plus nicotine replacement for smoking cessation: pragmatic randomised controlled trial. BMJ. 2008;336:1223–7 (Case 10.6).

Buntrock C, Ebert DD, Lehr D, Smit F, Riper H, Berking M, Cuijpers P. Effect of a web-based guided self-help intervention for prevention of major depression in adults with subthreshold depression: a randomized clinical trial. JAMA 2016 May 3;315:1854–63 (Case 10.7).

Eaton CB, Hartman SJ, Perzanowski E, Pan G, Roberts MB, Risica PM, Gans KM, Jakicic JM, Marcus BH. A randomized clinical trial of a tailored lifestyle Intervention for obese, sedentary, primary care patients. Ann Fam Med. 2016;14:311–9 (Case 10.8).

Index



L

M

N

O

P

Q

R

S

T

U

V

W