1

Epidemiology is …

Box 1.1 Epidemiology is …

'The science of epidemics' (*Concise Oxford Dictionary*, 1964)
'The science of the occurrence of illness' (Miettinen, 1978)
'The study of the **distribution** and **determinants** of disease in humans' (MacMahon and Pugh, 1970)
'The study of the distribution and determinants of **health-related states or events** in specified populations, and the **application of this study to control of health problems**' (Porta, 2008)
'The study of the occurrence and distribution of health-related events, states and processes in specified populations, including the study of the determinants influencing such processes, and the application of this knowledge to control relevant health problems' (Porta, 2014)

So what is epidemiology anyway? As shown in Box 1.1, the *Concise Oxford Dictionary* (1964) defined it accurately, but not very helpfully, as 'the science of epidemics'. In 1970, MacMahon and Pugh came up with something a bit more concrete: 'the study of the *distribution* and *determinants* of disease'. Their definition succinctly identifies the two core strands of traditional epidemiology: *who* is developing disease (and *where* and *when*), and *why* are they developing it? The next definition, from the 2008 edition of the *Dictionary of Epidemiology* (Porta, 2008), takes things two steps further by broadening the scope to include health in general, not just disease, as well as highlighting the essential role of epidemiology in translating research findings into health policy and medical practice to control disease. The most recent definition (Porta, 2014) elaborates further still but, in doing so, loses some of the elegance of the earlier versions.

Epidemiology, therefore, is about measuring disease or other aspects of health, identifying the causes of ill-health and intervening to improve health; but what do we mean by 'health'? Back in 1948, the World Health Organization (WHO, 1948) defined it as '… a state of physical, mental and social well-being'. In practice, what we usually measure is physical health, and this focus is reflected in the content of most routine reports of health data and in many of the health measures that we will consider here. However, methods that attempt to capture the more elusive components of mental and social well-being are now emerging. Instead of simply measuring 'life expectancy', WHO introduced the concepts of 'health-adjusted life expectancy' (HALE) and subsequently 'disability-adjusted life years' (DALYs) to allow better international comparisons of the effectiveness of health systems. In doing so, they recognised that it is not longevity per se that we seek, but a long and healthy life. We will discuss these and other measures in more detail in Chapter 2.

Perhaps epidemiology's most fundamental role is to provide a logic and structure for the analysis of health problems both large and small or, as described by Wade Hampton Frost, epidemiology involves the 'orderly arrangement of [established facts] into chains of inference which extend more or less beyond the bounds of direct observation' (Frost, 1927). It emphasises the sound use of numbers – we have to count and we have to think. We have to think about what is worth counting and how best to count it, about what is practical and, importantly, about how well we (or others) finally measured whatever it was we set out to measure, and what it all means. Accurate measurement of health is clearly the cornerstone of the discipline, but we believe the special value of epidemiology flows from a way of thought that is open, alert to the potential for error, willing to consider alternative explanations and, finally, constructively critical and pragmatic.

We offer this book as an aid to such thought. It does not aim to turn you into a practising epidemiologist overnight, but will give clear directions if that is where you decide to go. Its primary goal is to help you interpret the mass of

Table 1.1 Numbers of people who became ill after eating various foods at a youth camp.

	People who ate the food		People who didn't eat the food	
Food	Total	Number ill	Total	Number ill
Friday dinner:				
Hot chicken	343	156	231	74
Peas	390	175	184	55
Potato fries	422	184	152	46
Saturday lunch:				
Cold chicken	202	155	372	75
Salad	385	171	189	59
Saturday dinner:				
Fruit salad	324	146	250	84

(Adapted from Hook *et al.*, 1996, with permission from John Wiley and Sons. © 1996 The Public Health Association of Australia Inc.)

epidemiological literature and the various types of health data that you may come across. We hope that you will see, by reading and by doing, that the fundamental concepts and tools of epidemiology are relatively simple, although the tasks of integrating, synthesising and interpreting health information are more challenging. But before we go any further, let us do some public health epidemiology.

A case of food poisoning

Epidemiology is a bit like detective work in that we try to find out why and how disease occurs. Our first example illustrates this. After an outbreak of food poisoning at a youth camp, the local public health unit was called in to identify the cause (Hook *et al.*, 1996). They first asked everyone at the camp what they had eaten prior to the outbreak and some results of this investigation are shown in Table 1.1.

Looking at the numbers in Table 1.1, it is difficult to see which of the foods might have been responsible for the outbreak. (Note that everyone is recorded as either having eaten or not eaten each food; and that most people will have eaten more than one of the foods.) More people became ill after eating potato fries than after eating cold chicken (184 versus 155) – but then more people ate the fries (422 versus 202). How then can we best compare the two foods? One simple way to do this is to calculate the *percentage* of people who became ill among those who ate (or did not eat) each type of food. For example, 156 out of 343 people who ate hot chicken became ill and

$$156 \div 343 = 0.45 = 45\%$$

So 45% of people who ate hot chicken became sick. This is known as the **attack rate** for hot chicken, i.e. 45% of hot chicken eaters were 'attacked' by food poisoning.

Calculate the attack rates for the other foods. Which food has the highest attack rate?

Although cold chicken has the highest attack rate (77%), not everyone who ate it (or, more precisely, who *reported* eating it) became ill and 20% or one in five people who did *not* eat cold chicken still became ill. This is to be expected; no matter what the cause of concern, it is rare that everyone who is exposed to it will show the effects (in this case, become ill). What can help here is to work out how much *more likely* people who ate a particular food were to become ill than those who did not eat it. For example, 45% of people who ate hot chicken became ill, compared with 32% of people who did not eat hot chicken. Hot-chicken eaters were therefore 1.4 times (45% ÷ 32% = 1.4) more likely to become ill than people who did not eat hot chicken. This measure gives us the risk of sickness in hot-chicken eaters *relative* to non-eaters, hence its name – **relative risk**.

Calculate the relative risk of developing food poisoning associated with each of the other food items. Which food is associated with the highest relative risk of sickness?

We can now conclude that the food item most likely to have been responsible for the outbreak was the cold chicken – people who ate this were almost four times as likely to become ill as those who did not. This is quite a strong relative risk; in comparison, eating any of the other foods was associated with no more than one and a half times the risk of disease. The relevant data, including the attack rates and relative risks, are summarised in Table 1.2, which is much more informative than the raw numbers of Table 1.1.

In identifying the cause of the outbreak you have just solved an epidemiological problem. The 'attack rates' and 'relative risks' that you used are simple to calculate and are two very useful epidemiological measures. We will discuss them further in Chapters 2 and 5 and they will appear throughout the book.

Subdisciplines of epidemiology

The outbreak investigation above is an example of what might be called *public health epidemiology*, or *infectious disease epidemiology*, with the first name reflecting the broad field of application and the second the nature of both the aetiological (causal) agent and the disease. It is quite common now to specify such subfields of epidemiology, which range on the one hand from *nutritional* through *social* to *environmental* and *eco-epidemiology*, and on the other from

Table 1.2 Numbers of people who became ill after eating various foods at a youth camp and attack rates and relative risks for each food.

	People who ate the food			People who didn't eat the food			
Food	Total	Number ill	Attack rate	Total	Number ill	Attack rate	Relative risk[a]
Friday dinner:							
Hot chicken	343	156	45%	231	74	32%	1.4
Peas	390	175	45%	184	55	30%	1.5
Potato fries	422	184	44%	152	46	30%	1.4
Saturday lunch:							
Cold chicken	202	155	77%	372	75	20%	3.8
Salad	385	171	44%	189	59	31%	1.4
Saturday dinner:							
Fruit salad	324	146	45%	250	84	34%	1.3

[a] Note, relative risks are calculated using the exact percentages and not the rounded values shown.
(Adapted from Hook *et al.*, 1996, with permission from John Wiley and Sons. © 1996 The Public Health Association of Australia Inc.)

cancer to *injury* or *perinatal epidemiology*: the former grouping being exposure-oriented and the latter focused on the particular disease or outcome. Nonetheless, the core methods and techniques of epidemiology remain common to all subdisciplines, so the contents of this book are relevant to all. Setting subspeciality boundaries largely reflects the explosion of knowledge in these areas, although some areas do present special challenges. For example, capturing a person's usual diet is remarkably challenging and the subsequent data analysis equally so; epidemiologists coming fresh to the field of nutritional epidemiology will need to develop experience and expertise in that specific area. You will meet examples from a wide cross-section of health research as you read on, and the common threads of logic, study design and interpretation will, we trust, become apparent.

It is of some interest to know a bit more about a few of the special epidemiologies. *Occupational epidemiology* has the longest history of all, with influential early observations of diseases linked to occupations such as mining appearing in the sixteenth century, and a systematic treatise on occupational diseases was published by Ramazzini back in 1700 (Rosen, 1958). Occupational health research in general, and epidemiology in particular, continue to contribute to enhancing workplace health today. Seminal contributions in the field include identification of the pulmonary (lung) hazards of asbestos for miners and construction workers (Selikoff *et al.*, 1965) and the work practices that led to an epidemic of a rare fatal cancer in workers in the polyvinyl chloride industry (Makk *et al.*, 1974). Company records of job tasks can

provide measures of past exposure among employees, allowing researchers to look back in time and link, for example, past asbestos exposure to subsequent deaths in the workforce. (This type of study is a *historical cohort design* – see Chapter 4. It is only possible when there are good records of both exposure and outcome, usually death, and for this reason has proved particularly useful in occupational studies where such records often do exist.)

Far more modern are the subdisciplines of *molecular* and *clinical epidemiology*. The former aims to weld the population perspective of epidemiology with our rapidly increasing understanding of how variations in genes and their products affect the growth, form and function of cells and tissues. It thus has the potential to define genetic contributions to disease risk and can also provide biological markers of some exposures (e.g. changes to DNA following exposure to tobacco smoke). In contrast, clinical epidemiology differs from other branches of epidemiology in its focus on enhancing clinical decisions to benefit *individual patients*, rather than improving the health of *populations*. For this reason, clinical epidemiology is sometimes regarded as a separate discipline, a view encouraged by the fact that it has developed its own names for many standard epidemiological measures. The foundations are, however, identical to those of public health epidemiology and when appropriate we will discuss the two in parallel, highlighting any differences in language or approach along the way. There is also increasing interest in *lifecourse epidemiology*, which attempts to integrate events across the lifetime, often going right back to conception and sometimes to previous generations, to understand disease risk.

On epidemics

If we take the word 'epidemiology' itself, its origins from '*epidemic*' are clear. If we talk about an epidemic we immediately conjure up pictures of an acute outbreak of infectious disease but, both for practical and for etymological reasons, it seems reasonable to use the term to describe a notable excess of any disease over time. Many developed countries could, for example, be described as undergoing an epidemic of lung cancer over the last few decades (Figure 1.1). Notably the pattern of lung cancer over time differs for men and women; rates in men rose sharply between 1950 and 1980 but have been falling for some years now, while those in women rose later and started to fall more recently – a consequence of the fact that, as a group, women took up smoking more recently than men. To describe this excessive occurrence of disease (or death) as an 'epidemic' captures some of the urgency the numbers demand.

The derivation of the word 'epidemiology' itself is from the Greek *epi*, upon, *demos*, the people, and *logia*, study. Literally, therefore, it means the 'study

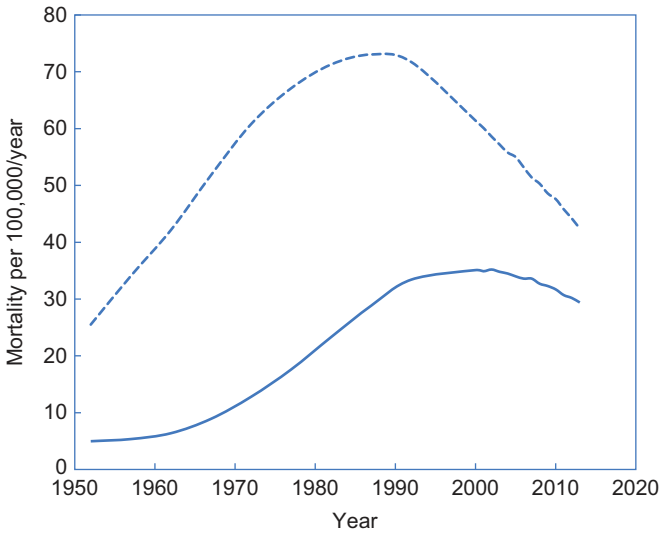

Figure 1.1 Time trends in lung cancer mortality rates in the USA (age-standardised to the 1970 US population) for white men (––) and women (____). (*Drawn from*: CDC Wonder Database (CDC), accessed 26 February 2015.)

(of what is) upon the people'. Such study suggests a simple set of questions that have long lain at the heart of epidemiology.

- **What** disease/condition is present in excess?
- **Who** is ill?
- **Where** do they live?
- **When** did they become ill?
- **Why** did they become ill?

The first question reflects the need for a sound, common definition of a disease so that like is compared with like. Epidemiology is all about comparison – without some reference to what is usual, how can we identify excess? The next three questions form the mantra of descriptive epidemiology: '*person, place and time*'. As Figure 1.2 shows, an 'epidemic of premature mortality' occurred among young and middle-aged men in Russia in the mid-1990s and again in the early 2000s. This description captures the essence of the problem and prompts the next questions: what caused these epidemics? What changed in the circumstances of younger Russian men to reverse the pattern of falling mortality in the early 1980s and then cause it to almost double in less than 10 years? And why did this happen again in the late 1990s? Other data show that there were no such mortality changes in Western Europe, or among older Russian men or infants, or (to the same extent) in Russian women. This simple graph captures a public health disaster for Russia and prompts urgent causal speculation: *Why did this happen?* Solving and responding to this final question is critical for public health progress, but there is clearly no simple solution. In this case, a high proportion of the deaths were

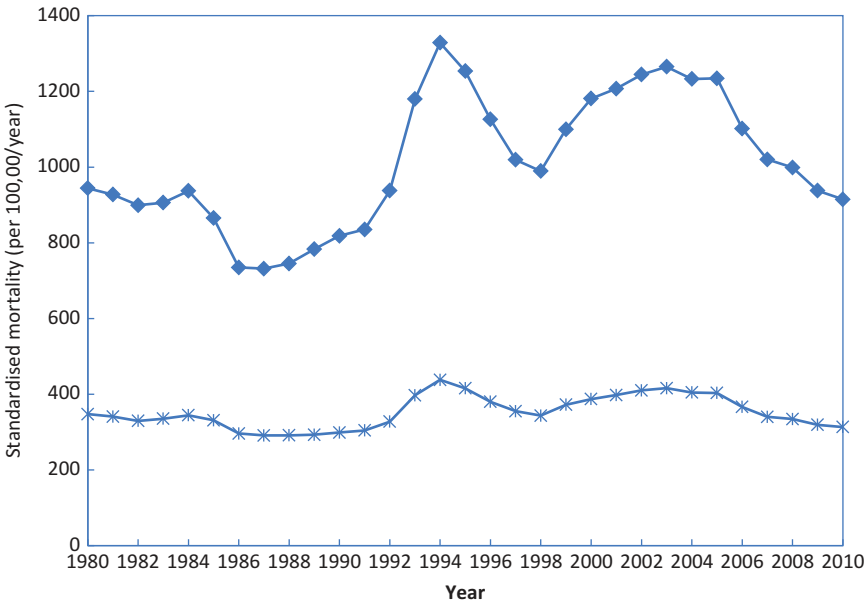

Figure 1.2 Changes in all cause mortality at ages 0–64 years in the Russian Federation from 1980 to 2010 (◆ men, ✳ women). (*Data from*: the European Health for All Database. WHO Regional Office for Europe, Copenhagen, Denmark, http://data.euro.who.int/hfadb/, accessed 27 February 2015.)

linked to excess consumption of alcohol during the 1990s: increases in mortality coincided with periods of economic and societal crisis, and rates fell when the economic situation improved (Zaridze *et al.*, 2009). The earlier decline during the 1980s coincided with an anti-alcohol campaign involving higher taxes and reduced production which led to sharp decreases in alcohol consumption in the short term, and lower rates of alcohol-related mortality and suicide (Pridemore and Spivak, 2003). This example highlights the importance of paying close attention to descriptive data that provide a 'community diagnosis' or take the public health 'pulse' of a nation. Much can be gleaned from apparently simple data to give a quite precise description of the overall health of a population or a more specific health event, as the following exercise shows.

An historical epidemic

Table 1.3 shows some data that relate to an actual human experience. It tells you how many people there were in various age, sex and socioeconomic groups and what percentage of these people died during the 'epidemic'. The challenge is to use these data to describe the event systematically in terms of **whom** this happened to (we have no data on place or time) and then to think about the sort of event that might have induced such a pattern.

Table 1.3 An historical event.

	Adult males		Adult females		Children (both sexes)		Total population	
SES[a]	Total	% Dead	Total	% Dead	Total	% Dead	Total	% Dead
High	175	67.4	144	2.8	6	–	325	37.5
Medium	168	91.7	93	14.0	24	–	285	58.6
Low	462	83.8	165	53.9	79	65.8	706	74.8
Other	885	78.3	23	13.0	0	–	908	76.7
Total	1690	80.0	425	25.6	109	47.7	2224	68.0

[a] SES, socioeconomic status.
(*Source*: www.anesi.com/titanic.htm, The *Titanic* casualty figures (and what they mean), accessed 29 April 2015.)

The following questions are designed to help you identify key features of the data.

1. What is distinctive about this isolated population with regard to:
   - the numbers of men and women (sex distribution),
   - the numbers of adults and children (age distribution), and
   - the numbers in each socioeconomic group (socioeconomic distribution)?
2. What strikes you about the percentage of people who died (the 'death rate')? Is this different for (a) adults and children, (b) men and women, (c) high and low socioeconomic status (SES) and (d) any particular combinations of the above?
3. How many times more likely were:
   - men to die than women, and
   - those of low SES to die than those of high SES?
4. To what historical event might these data refer?

Table 1.3 displays more complicated data than Table 1.2 because you had to consider the joint effects of three factors (sex, SES, and age) on mortality. The sequence of questions above underlines a general principle in describing such tables – i.e. to look at overall patterns first, then move on to more detail. We all see things in different ways, but until you develop your own style the approach shown in Box 1.2 may help you avoid getting lost in the array of possible relationships. You need first to grasp the size of the *whole group* under study and how many died; then check the overall patterns (the numbers and death rates[1]) across each 'exposure' separately (sex, SES, age).

[1] As you will see in Chapter 2, these are not technically 'rates' in the true sense of the word but it is convenient to call them rates as they are essentially identical in form to the attack rates in Table 1.2.

Box 1.2 An historical event

Things to note about the population include:

- the predominance of adult males (1690 ÷ 2224 = 76%), the much smaller proportion of adult females (19%), and the very few children;
- the substantial excess of persons of low SES (men and children in particular); and
- the total population (2224) is quite large – a village, small town, an army barracks . . . ?

Things to note about the 'death rates' include the following.

- The overall death rate is very high – more than two-thirds died.
- Overall, death rates increased with decreasing SES.
- The death rate in men (80.0%) was much higher than that in women (25.6%); the death rate in children was between these two.
- In men, the death rate was high in all socioeconomic classes, although those of high SES fared better than the rest; in women, the death rate was always lower than that for males of equivalent SES, but it increased strikingly from high to medium to low SES.
- The only children to die were of low SES.

Overall, the relative risk (RR) for men versus women is 80.0 ÷ 25.6 = 3.1
The RR for low versus high SES is 74.8 ÷ 37.5 = 2.0
The RR for women of low SES versus women of high SES is 53.9 ÷ 2.8 = 19.3
The RR for men of low SES versus women of high SES is 83.8 ÷ 2.8 = 29.9

A disaster has occurred, causing a high death rate that predominantly affected men (of all social classes) and, to a lesser extent, women and children of low social class. Overall there is a modest benefit of belonging to a higher social stratum, and among women this protection was exceptionally strong (a 19-fold higher risk of dying for low versus high SES).

Such substantial differences in risk reflect powerful preventive effects and in this instance it was a mix of social custom and the physical consequences of social stratification. The event was the sinking of the *Titanic*, where those of higher SES (the first-class passengers) were situated on the upper decks and were therefore closer to the lifeboats than those of medium and low SES (those travelling second and third class, respectively). The males gallantly helped the females and children into the lifeboats first. Those of 'other' SES were the crew.