Chapter 5

Cause and effect: The epidemiological approach

Objectives

On completion of the chapter you should understand:

- that the purpose of studying cause and effect in epidemiology is to generate knowledge to prevent, cure, treat, and control disease;
- that cause and effect understanding is particularly difficult to achieve in epidemiology because of the complex development of diseases over long timescales and because of ethical restraints on human experimentation;
- how causal thinking in epidemiology depends on and contributes to other domains of knowledge, both scientific and non-scientific;
- the importance of setting out a causal framework in advance of designing a study;
- the potential role of causal graphs in both concepts (e.g. web of causation) and analytical aids (e.g. directed acyclic graphs);
- the potential contributions of epidemiological study designs for contributing to causal knowledge;
- how to use a systematic approach, which checks for error, chance, bias, and confounding, before reaching judgements on cause and effect;
- the distinction between confounding, moderator, and mediator variables;
- the meaning of interaction and effect modification;
- that epidemiological approaches to, and guidelines for, causality are not a checklist and therefore conclusions must be carefully judged and tentative;
- the value of synthesizing data from epidemiological studies and other disciplines, including laboratory experiments, before reaching conclusions.

5.1 Introduction: causality in science and philosophy and its relevance to epidemiology

Cause and effect understanding is the highest achievement (the jewel in the crown) of scientific knowledge, including epidemiology. Causal knowledge points to actions to break the links between the factors causing disease, and disease itself. As such, it underpins prevention of disease. It also helps to predict the outcome of an intervention and helps to treat disease. To quote Hippocrates writing about 2000 years ago, 'To know the causes of a disease and to understand the use of the various methods by which the disease may be prevented amounts to the same thing as being able to cure the disease' (see Chadwick and Mann 1950). This is both a modest exaggeration

and an understatement. Sometimes we know the exact cause of a disease but we can neither prevent it nor cure it (e.g. serious genetic disorders); but where we can prevent it, the result is far superior to curing the disease (e.g. it is better to prevent lung cancer than cure it through surgery, chemotherapy, and radiotherapy).

Epidemiology enjoys the status of a science. As in all sciences including physics and chemistry, epidemiological understanding of cause and effect does not have to be 100 per cent complete or accurate to permit useful application. After all, gravity is still a mystery but scientists have learned enough about it to land a rocket on a comet. Arguably, more so than in other sciences, in epidemiology even partial understanding must be applied as quickly as possible, for it may be a life and death matter. There is, therefore, an ethical responsibility to apply knowledge even when, from a scientific point of view, further research is advised. Yet, this ethical imperative may be perilous.

Early application of knowledge sometimes has devastating effects and sometimes beneficial effects. Sylvia Tesh (1988) gives two examples of this. The public health endeavours of the nineteenth century, including the building of sewers, the delivery of clean water, and the improvement of the sanitary conditions of the home and workplace, were driven by the 'miasma' theory of health and disease. This theory presumed that noxious air carrying 'miasma' released from filth in the environment was the cause of most of the prevalent diseases, including cholera. Though wrong, the miasma theory worked and the benefits of the applications of this simplistic theory have been immense and remain the bedrock of public health everywhere. Time has shown the 'filth' allowed microbes (and the insects and other creatures that spread them) to flourish. Many of these diseases were contagious—that is, they spread by contact. There was no 'miasma'.

By contrast, according to Tesh, the contagion theory was both correct and dominant in explaining the occurrence of plague. Jews were, however, incriminated in a poorly understood causal pathway of contagion and thousands were executed in a vain attempt to control plague. Tesh gives a figure of 16 000 Jews killed in Strasbourg alone. (Roy Porter gives a figure of 2000 Jews slaughtered in Strasbourg and 12 000 in Mainz.) The contagion theory was ineffective in this application, which was outrageous, and surely underpinned by the anti-Semitism of those times. Nonetheless, this history has contemporary lessons as some, usually poor or minority, populations are still scapegoated when there are threats of contagious diseases.

The effective application of incomplete knowledge requires art and science. Epidemiology is one of the principle sciences that public health policy draws upon. Recent examples of major public health policy decisions requiring the application of incomplete data include: whether to ban consumption of beef products given the epidemic of bovine spongiform encephalopathy in cattle; what action to take given evidence that living near a nuclear power plant is associated with raised risk of childhood leukaemia; what proportion of daily energy intake should be consumed as fat; what is the recommended daily salt intake; whether we should tax sugar to help reduce obesity; whether we should strive to increase the average birthweight of newborn babies; and whether women should or should not take oestrogen-based hormonal therapy to reduce post-menopausal problems and chronic diseases. These kinds of decisions generate massive controversy that hinges around causality.

To the study of causality, epidemiology has contributed:

- a philosophy of health and disease
- models that illustrate that philosophy
- study designs to produce quantitative evidence
- information from quantitative studies on the relationships between numerous factors and diseases
- frameworks for interpreting and applying the accumulated evidence.

Box 5.1 Some simple questions in causality in epidemiology

- What is a cause?
- What will be the result of a cause in epidemiology?
- How might we measure the result of a cause in epidemiology?

Epidemiology has absorbed a great deal of causal thinking from other disciplines and the occasionally heard and read proposition that epidemiology is *the* science of causal thinking (whether confined to populations or not) is immodest and false.

Scientific thinking encourages turning empirical observations into theories and hypotheses that permit tests of generalizable cause and effect judgements. Epidemiological reasoning on cause and effect is embedded in observations of disease variation, the discovery of associations between putative causes and the disease, and ways of testing hypotheses so we can use associations to progress towards causation. Epidemiology draws upon the reasoning of other disciplines, including philosophy and microbiology. Epidemiology shares similar problems of disentangling cause and effect relationships with other disciplines (particularly those mainly reliant on observation of naturally occurring events). Solutions to problems are likely to arise from the sharing of ideas among such disciplines.

This background understanding of how science works is necessary to counter the criticism that epidemiological reasoning on cause and effect is merely empirical and atheoretical. On a pragmatic note, epidemiological debates on cause and effect are often in the public eye and, more so than most other sciences—so non-epidemiologists become involved in the interpretation of data and making judgements on their meaning (see also Chapters 4 and 10). This requires that epidemiological approaches to analysis of cause and effect are easy to understand and apply. Before reading on, do the exercise in Box 5.1.

A cause is something which has an effect, that is, it brings about or produces something. In epidemiology, a cause can be considered something that alters the frequency of disease, health status, or associated causal factors in a population. We measure these effects by defining changes in the incidence (preferably), prevalence, and other outcomes due to changes in the presumed causal factors (i.e. exposures or risk factors). We will also measure how other (moderating and mediating) variables alter the relationship between postulated causal factors and outcomes. These are pragmatic definitions, but it is worth knowing more about the broader debates and controversies on cause, and where such simple ideas fit. This is important so that epidemiologists can converse about cause and effect in multidisciplinary settings, where pragmatic definitions may be questioned, or even derided.

5.1.1 Some philosophy

Philosophers have grappled with the nature of causality for thousands of years (Cottingham 1996). Aristotle, for example, held a broad view that there were four elements to cause, which have been reconsidered in the context of a house by John Dreker, as extracted in Cottingham. The causes of a house are the material (the stone, brick, or wood), the formal (the plan), the efficient (the thing which puts it into effect, here the builder), and the final (the purpose, being to create a comfortable home). Aristotle foresaw one effect could have several causes. The cause of Legionnaires' disease is, at its simplest, exposure to the causal bacteria. From an Aristotelian point of view, the four causes would be: the existence of living bacteria (material); the essence of the nature of the relationship between bacteria and humans (the formal); the delivery of an infective dose by some

mechanism, such as a cooling tower (the efficient); and the quest for processes to increase human comfort that leads to complex water systems such as cooling towers (final).

David Hume's philosophy has also been influential (Cottingham 1996). Hume's view that a cause cannot be deduced logically from the fact that two events are linked, but needs to be experienced or perceived at a deep level, is crucially important to epidemiology. Just because thunder follows lightning does not mean thunder is caused by lightning (indeed, it is not as we discuss later). When we flick a light switch the light may go on, but this does not prove that the one act causes the other. To stretch your imagination, can you think of alternative explanations, no matter how absurd they seem?

One explanation is that there is someone observing and as soon as you flick the switch, he or she puts the light on. This is, indeed, absurd, but it is possible. To someone who had no understanding of electrical circuits it might seem more plausible than the truth. When we understand the mechanism of electrical circuits, however, we accept that there is cause and effect.

This perspective is echoed in the axiom, 'association is not causation'. Cause and effect deductions need more than linkage; they need understanding. Hume's thoughts are relevant to the debate on black box epidemiology. The black box metaphor comes from the increasing availability of technology as a closed unit, not amenable to easy opening and exploration, for example, a DVD player, modem, or mobile telephone. The unit works, or if it does not it is discarded and replaced, without regard to what the problem is. This has become an apt metaphor for epidemiological research based on the study of associations (risk factor epidemiology) and the evaluation of complex interventions. The late Petr Skrabanek (1994) described it as epidemiology where the causal mechanism behind an association remained unknown but hidden (black), but with the inference that the causal mechanism was within the association (box). Skrabanek argued that the purpose of science is to open and understand the black box, which epidemiology too often failed to do. While failing to understand the association (or the effective components of a complex intervention) is a limitation, the greater problem occurs when the true causes lie outside the putative association (i.e. outside the black box)—and that happens.

The contribution of another philosopher, John Stuart Mill, captured in his 'canons', is so similar to the modern ideas of epidemiology that it is discussed in the section on guidelines for causality (section 5.6.1). Philosophical discussion on the nature of causality, questioning whether causes can be stated definitively or only as a matter of probability is of importance to epidemiology. Epidemiology tends to work closely with statistics, which deals with probability, and tends to side with this approach.

5.2 Epidemiological causal strategy and reasoning: the example of Semmelweis

The epidemiological strategy is simple but its successful execution is not. To reiterate, at the population level diseases form patterns, which are ever-changing. Over short time periods the changes are largely, but not exclusively, caused by environmental changes. The exception is that genetic changes in microbes can be rapid and hence the pattern of human microbial diseases can change fast. Over long time periods, that is, over many generations, genetic variation also changes the population pattern of human disease. Clues to the causes of disease are inherent within these patterns. These patterns, therefore, can be studied both to generate and test ideas on causation and to test out ideas developed in other fields of enquiry. The combination of epidemiological and other types of observation is particularly valuable.

The epidemiological mode of reasoning combined with other observations is illustrated by the discovery by Ignaz Semmelweis of the general cause of puerperal fever. Semmelweis (1818–1865)

Table 5.1 Births, deaths, and mortality rates (%) for all patients at the two clinics of the Vienna maternity hospital from 1841 to 1846

First clinic (doctors)			Second clinic (midwives)		
Births	**Deaths**	**Rate (%)**	**Births**	**Deaths**	**Rate (%)**
20 042	1989	9.92	17 791	691	3.38

Source: data from Semmelweis I. The etiology, concept and prophylaxis of childbed fever, 1983 [excerpts] In: Buck C, Llopis A, Najera E, Terris M (eds.). *The challenge of epidemiology–issues and selected readings*. Washington: Pan American Health Organization (PAHO) Scientific Publication, Copyright © 1988 PAHO. pp. 46–59.

was training in obstetrics in Vienna when he observed that the mortality from childbed fever (now known as puerperal fever) was lower in women attending clinic 2, run by midwives, than in those attending clinic 1, run by doctors. He also noted that women who gave birth in the street, or prematurely, had a lower mortality than those in clinic 1. The statistics he collected are given in Table 5.1. Do the exercise in Box 5.2 before reading on.

Semmelweis also noted that while the cases in clinic 2 were sporadic, in clinic 1 a whole row of patients might be sick at the same time. Semmelweis was perplexed but saw that the pattern he observed meant an endemic cause, that is, the cause lay within the clinic itself. He tried, unsuccessfully, to solve the problem by delivering the mothers by laying them on their sides rather than on their backs. At this stage, based on a case series study (see Chapter 9 for details on this design) Semmelweis has observed a pattern, come to a general conclusion (internal, not external cause), developed a hypothesis, and tested an intervention which was unsuccessful (delivery position). It could be said that he has observed an association but the explanation remains hidden. This makes unravelling the mystery even more, not less, important. At the time there was no developed suite of methods in epidemiology—if there had been a case–control study would have been the next step (see Chapter 9 for this study design). Equally importantly, there was no field of medical microbiology either.

A year or so later, in 1847, his colleague and friend Professor Kolletschka died following a fingerprick with a knife used at an autopsy. Kolletschka's own autopsy showed inflammation to be widespread in his corpse, with peritonitis and meningitis. Semmelweis's mind was alert and he connected the childbed fever disease in women with that of his friend. He wrote:

> Day and night I was haunted by the image of Kolletschka's disease and was forced to recognize, ever more decisively that the disease from which Kolletschka died was identical to that from which so many maternity patients died.
>
> Semmelweis (reprinted 1988, p. 52)

Semmelweis was 'compelled to ask' whether cadaverous particles, that is, the substance passed from the cadaver (corpse) Kolletschka dissected, had been introduced into the vascular systems of maternity patients, as seemed to have happened in the case of his friend.

Box 5.2 Causal question from statistics presented in Table 5.1

Do these figures spark off any ideas of causation in your mind? What explanations can you generate? Reflect on this question before reading on.

Semmelweis's inspired idea was that particles had been transferred from the scalpel to the vascular system of his friend and that the same kind of particles were killing maternity patients. He foresaw that the particles could be transferred from the hands of medical students and doctors to the women during pelvic examinations. If so, something stronger than ordinary soap was needed for handwashing. He introduced chlorina liquida, and then for economy, chlorinated lime. These substances, unlike ordinary soap, have antiseptic properties. The maternal mortality rate plummeted, reaching the level of the midwives' clinic.

Although Semmelweis was not the first to link puerperal fever to lack of hygiene, his contribution was huge; particularly because of the systematic, epidemiological evidence he accumulated and the way he tested his ideas (hypotheses). The epidemiological observations outlined the problem and prepared the mind to seek a solution, itself inspired by clinical and autopsy observation, and tested by experimentation and epidemiological monitoring of outcomes.

Two great principles are illustrated by this work. First, deep and generalizable knowledge lies in the explanation of disease patterns, rather than in their description, which is just the first step. The questioning (and persistent) mind may solve the riddle inherent in the pattern. Second, inspiration is needed, and may come from unexpected sources, as here from Kolletschka's autopsy. Such inspiration needs to be converted into a scientific hypothesis so it can be tested by scientific observation or experiment, as by Semmelweis's interventions, first in the way labour was conducted (unsuccessful), and then of handwashing with chlorinated lime (successful).

Most disease patterns remain unexplained despite lengthy study and others are never explored fully (so-called cul-de-sac epidemiology). Those that are explained usually lead to profound insights. Epidemiology does not, however, have the tools to demonstrate biological disease mechanisms. Whether the cause is biochemical—as in scurvy, or social, as in the rise of suicide in populations hit by unemployment—epidemiologists are reliant on other sciences to be equal partners in pursuit of the mechanisms. Action cannot always, however, await understanding of the mechanism. A contemporary example of this is the use of epidemiological data showing that laying an infant on its front (prone position) to sleep raises the risk of 'cot death', or in medical terms sudden infant death syndrome. Yet, the prone position was long and wrongly advocated as a means of avoiding the potential danger of infants inhaling their own vomit. A campaign to persuade parents to lay their infants on their backs has halved the incidence of cot death. In countries where the evidence has not been implemented, we have seen no such change. The mechanism is yet to be fully explained but the association is agreed as causal.

It would be several decades after Semmelweis that, among others, Louis Pasteur firmly established an old idea—the germ theory of diseases—to be true and the bacterial nature of the cadaverous particles was established. Then Semmelweis' discovery could be understood. Even in this straightforward disease we see complex layers of causation: the organization of health care; the behaviour of doctors in relation to autopsies; hygiene practices; cadaverous materials; and transfer of bacteria from doctors to mothers in labour. This complexity is the topic we cover next, where we consider models of causation.

5.3 Models of cause in epidemiology

5.3.1 Interplay of host, agent, and environment

The idea that disease is *virtually always* a result of the interplay of the environment, the genetic and physical make-up of the individual, and the agent of disease, is one of the most important of the cause and effect ideas underpinned by epidemiology. This theory applies both to diseases said

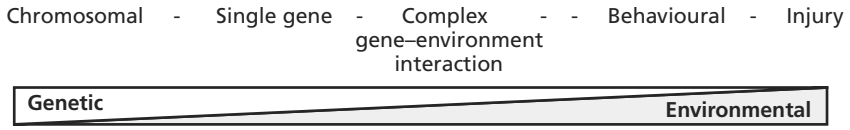


Fig. 5.1 Line of causation—a simple model for considering genetic and environmental factors.

to be multifactorial (e.g. cancers or heart disease) and to diseases which are by their definition a result of a single cause, such as tuberculosis, a drug side effect, or an overdose. This way of thinking, shared with social scientists, contrasts with the strong focus on very specific and narrowly defined causes (reductionist approaches) of most sciences, including medical sciences.

Diseases attributed to single causes are invariably so by definition. For example, tuberculosis is a disease that has many manifestations. It is characterized by a multiplicity of diffuse signs and symptoms, which affect nearly every part of the body. Some diseases, for example, sarcoidosis, are often indistinguishable from tuberculosis clinically, while the microscopic findings in Crohn's disease look very similar to tuberculosis. In some ways tuberculosis is a number of diseases (e.g. pulmonary tuberculosis, cutaneous tuberculosis, tuberculous meningitis), some of which are indistinguishable from other diseases. The fact that tuberculosis is caused by the tubercle bacillus is a matter of redefinition following Koch's discovery of the causal bacterium in 1882. Another perspective is that the causes of tuberculosis are many, including malnutrition and overcrowding.

This idea is captured by several well-known disease causation models such as the line, the triangle, the wheel, the web, and the pie. These models help to organize ideas about causes and strategies to prevent and control disease. In analysing causes, it is advisable to move from simple to complex models.

Figure 5.1 illustrates the idea of the line of causation. First, an arbitrary division is made between genetic and all other causes, categorized by convention as the environment. The line conceptualizes causes as lying on a spectrum from being wholly caused by genetic factors, or by environmental ones. Although the interaction of the genome and the environment is the key to understanding causation, the gene–environment division, though artificial, is widely used as a simple first step in analysing causes. At one extreme lie disorders which are almost entirely genetic, such as Down's syndrome (trisomy 21). At the other extreme lies injury arising from a road traffic accident. Most disorders lie in between. One of the early judgements required on diseases of unknown cause is the likely relative importance of genetic and environmental factors, for the preventive or control strategy will be fundamentally different. Try the exercise in Box 5.3 before reading on.

Figure 5.2 shows how epidemiology can help to make judgements on the question in Box 5.2. Diseases where the incidence varies rapidly over time or is different in genetically similar groups are clearly strongly influenced by environmental factors, while diseases which have a stable

Box 5.3 Exercise on gene and environment in causation

Think about the causes of three or four health problems or diseases that you or your friends or relatives have had. Place them on the line of causation. (Use these diseases for the following exercises too.)

Think through the cause of disease X using this model (Box 1.7, Chapter 1). What is your judgement? Is disease X likely to be genetic or environmental? Why? What makes you favour genetic factors over environmental ones?

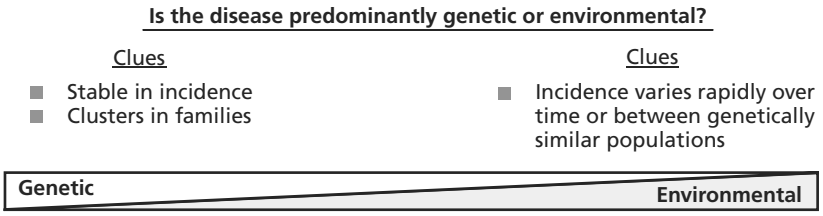


Fig. 5.2 Line of causation: epidemiological clues to environment or genetic causation.

incidence or are clustered in blood relatives are more likely to have strong genetic influences. Figure 5.3 places some diseases on this spectrum.

The triangle, wheel, and the web are more complex versions of the same concept as the epidemiological line. Each model has its strengths and limitations for helping to clarify causal thinking. Each model is, however, a simplification—that is the value of a model. The categories of host, agent, and environment (Fig. 5.4) are arbitrary. While the meaning of the words host and agent of disease are self-evident, or can be illustrated with simple examples (Boxes 5.4 and 5.5), this is not the case for the environment, which has an immensely broad meaning (Box 5.6). The host and agent are, of course, both part of the environment. The environment, in this context, is arbitrarily defined to mean factors other than the host and the agent of disease. The environment, in particular, can be split to some benefit into several categories, such as the social, chemical, or physical environment.

Boxes 5.4, 5.5, and 5.6, list some of the many host, agent, and environmental factors which are generally important causes of population level variations in human disease. Of the factors listed in Table 5.2, age is the most powerful, and for many diseases, particularly of the reproductive tract, sex equally or even more so. Obviously, neither age nor sex are direct causes of disease outcomes, but they are on the causal pathway. Age is a measure of time, and all causes need time to have their effects, some needing many decades. Paradoxically, however, these two factors are seldom studied as causal factors but as confounders. Before reading on, reflect on why this might be so.

In using epidemiological comparisons to spark *new* understanding of disease causation, it is essential that the populations compared are alike in *known* causal factors, of which age and sex are the most important. Hence, we see the almost routine use of age and sex matching or adjustment techniques in causal epidemiology (Chapters 4 and 8). This said, even for variables such as age and sex, the causal effects and mechanisms are complex and cannot usually be

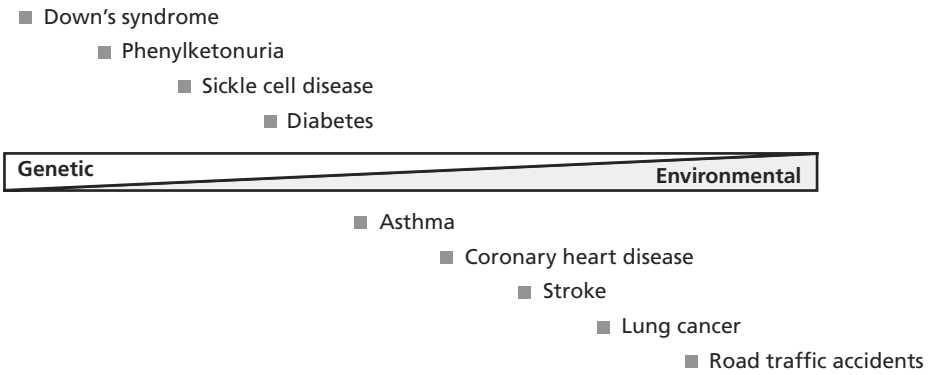
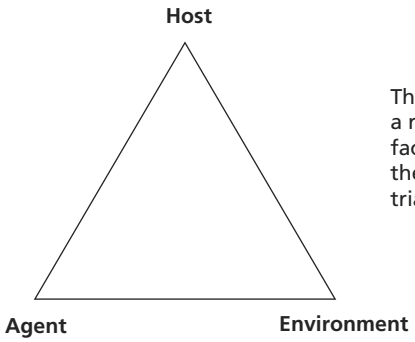


Fig. 5.3 Line of causation: some examples of where disease/health problems lie.



The underlying cause of the disease is a result of the interaction of several factors, which can be analysed using the components of the epidemiological triangle

Fig. 5.4 Triangle of causation.

Reproduced from Centers of Disease Control and Prevention (2012), '*Principles of Epidemiology in Public Health Practice, Third Edition: An Introduction to Applied Epidemiology and Biostatistics*', available from http://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson1/section8.html, 29th Feb. 2016.

specified as biosocial mechanisms. For example, at any age, women have a lower incidence of cardiovascular diseases such as myocardial infarction (heart attack). This sex difference is well characterized, but the mechanisms have not been specified and are likely to involve a mix of genetic, behavioural, and social factors. The human genome project may lead to such mechanistic understanding.

Before reading on, do the exercise in Box 5.7.

The triangle is a useful model for analysing causal relationships and to derive public health strategies, as shown in Figures 5.5 and 5.6, for example, for the control of Legionnaires' disease. In this and other infectious diseases, the concept of the disease agent is central to causation, and usually a specific agent can be identified or assumed.

In explaining population differences in the pattern of disease, agent factors, examples of which are in Box 5.5, receive less attention than they deserve. In infectious disease epidemiology, characterizing the virulence of organisms is difficult and sometimes impossible, and in other diseases conceptualizing the cause as an agent is not easy. The issue of agent virulence is likely to be considered more carefully in future. The reason is that the genome of most pathogenic microorganisms is being mapped and understanding of gene variants associated with virulence is growing fast. The bacterium *Helicobacter pylori*, for example, is associated with severe inflammation and duodenal ulceration in 89 per cent of infections with the VacAsla strains and 20 per cent of infections with the vac s2 strain. Virulence genes can be identified and potentially can be removed to create organisms that are less pathogenic to humans.

Box 5.4 Causes of diseases: examples of host factors

Genetic inheritance
Age
Sex
Previous disability
Behaviours (such as smoking)
Height and weight
Cholesterol in blood

Box 5.5 Causes of diseases: examples of agent factors

Virulence of a microorganism
Serotype of microorganism
Antibiotic resistance of microorganism
Cigarette—tar content
Type of glass in motor car windscreen

Box 5.6 Causes of diseases: examples of environmental factors

Home overcrowding
Air composition
Workplace hygiene
Weather
Water composition
Food contamination
Animal/human contact
Cooling tower use
Radiation

Table 5.2 Control of Legionnaires' disease: triangle and levels of prevention

	Agent	Host	Environment
Primary	Design and hygiene of water systems to prevent growth of bacteria	Smoking cessation and general health improvement	Use and location of cooling towers to be regulated
Secondary	Hygiene to keep bacterial growth controlled	Nil—there is no way to pick up cases at the pre-symptom stage	Separate people from source once outbreak has occurred, e.g. if in a hospital ward
Tertiary	Once an outbreak has occurred, decontaminate the water system	Medical therapy	Close damaged cooling towers or water systems; or repair them

Box 5.7 Analysing disease using the triangle of causation

Reconsider your chosen health problems (Box 5.3) using the triangle of causation (Fig. 5.4). Also, think through the cause of disease X (Chapter 1, Box 1.7) using this model. Finally, think through how these problems could be controlled by actions targeted at the host, agent, and environment.

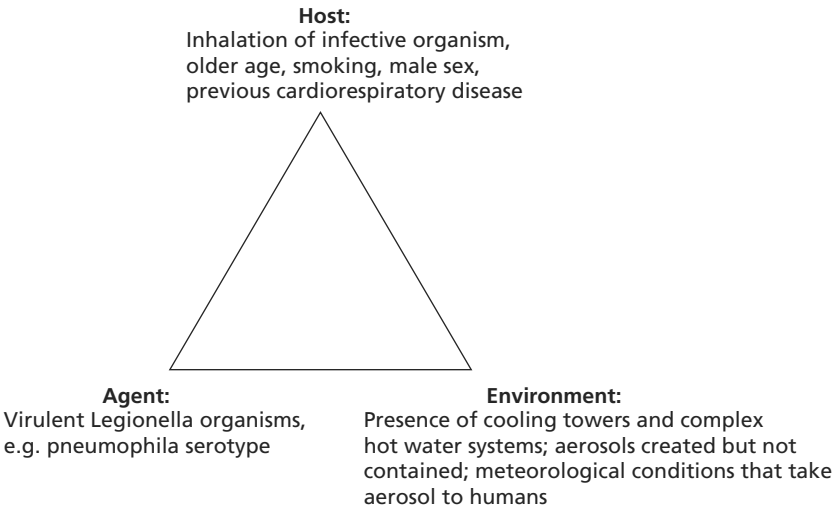


Fig. 5.5 Analysis of the causes of Legionnaires' disease: triangle of causation.

The concept of the disease agent also works with many non-infectious agents, for example, cigarettes, motor cars, and alcohol can be considered as the agents of disease and injury. A reduction of the tar content of cigarettes, and hence their virulence (in the literal sense of being toxic or hostile to health) could be responsible for some of the recent reduction of lung cancer incidence.

The interaction of the host, agent, and environment is rarely understood. For example, the effect of cigarette smoking is substantially greater in poor people than in rich people. The reason is unclear. It may be that there is an interaction between the agent (cigarettes), host factors such as nutritional status, and environmental factors such as air quality. These ideas are illustrated as follows in the simpler context of Legionnaires' disease.

Legionnaires' disease is a pneumonia (an inflammation of the lungs) that presents with some atypical features. It results from the inhalation, by susceptible people, of virulent organisms belonging to the genus *Legionellaceae* (legionellas for short). The organisms that cause Legionnaires'

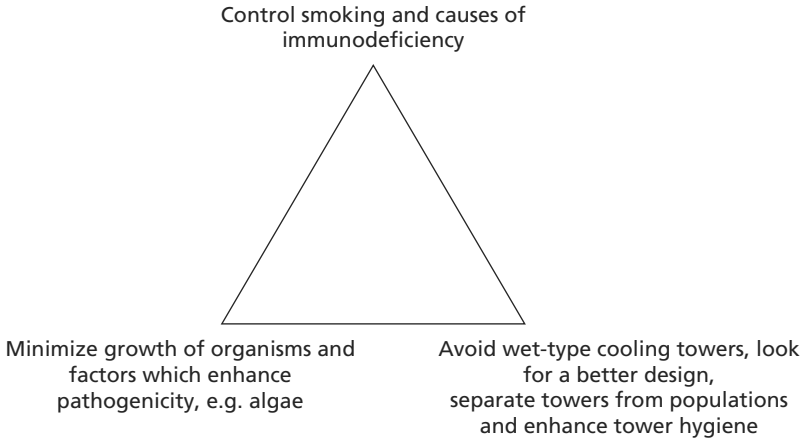


Fig. 5.6 Analysis of the control of Legionnaires' disease: triangle of causation.

Box 5.8 Reflection on the value of models

Consider how your thinking on the cause of Legionnaires' disease has changed because of the analysis in Figures 5.5 and 5.6, and the analysis you conducted on the diseases you chose in Box 5.3. What has been the additional value of employing this kind of model?

disease are environmentally acquired. The causal microorganism is found in most natural waters and is usually harmless. It is, therefore, a simplification to say that this normally harmless bacterium is *the* cause of Legionnaires' disease. Such a view could lead to erroneous, costly, and ineffective action to control this disease through futile attempts to eliminate this widely distributed organism from water.

The underlying cause of Legionnaires' disease lies in the creation by humans of water systems which permit the organism to thrive and be aerosolized at sufficient concentration to cause human disease. The ageing of the population, the presence of immunocompromized people, and people who impair their lung's defence mechanisms by smoking are also important causal factors. The bacterium, which is not normally a human pathogen, finds itself interacting with humans in this environment. The triangle of causality provides a framework for this type of reasoning, as illustrated in Figure 5.5. An understanding of the range of causes permits the development of a rational preventive strategy as shown in Figure 5.6. Before reading on, do the exercise in Box 5.8.

In a systematic analysis based on a model, as shown in Figures 5.5 and 5.6, attention is deflected from the microorganism as a specific cause, to the environment, host, and agent as interacting causes. This thinking broadens the control strategy. On current thinking the most effective approaches are to design better complex water systems, and to use hygiene and chemical measures to inhibit bacterial growth.

Table 5.2 shows how the epidemiological triangle can be combined with the schema of the levels of prevention to devise a comprehensive framework for thinking about possible preventive actions. Primary prevention is action to prevent the disease or problem from actually arising; secondary prevention is the early detection of the problem to prevent its damaging effects; and tertiary prevention is to contain, and if possible reverse, the damage already done. (As an aside, most clinicians and policy-makers working in clinical settings combine tertiary prevention with the secondary prevention category, calling it all secondary prevention. Epidemiologists and public health practitioners, in this context, usually conform to this simpler schema.) It is worth re-emphasizing that these frameworks aid systematic thinking by simplifying the problem into logical constituent parts. Before reading on, do the exercise in Box 5.9.

Figure 5.7 shows the wheel of causation. The principles behind this model are as for the triangle, but it emphasizes the unity of the interacting factors. The genetic make-up of the individual and its expression in the body (called the phenotype) is shown as the hub of the wheel, but enveloped

Box 5.9 Combining causal models and the levels of prevention

Think about the control of the three or four health problems you picked and disease X (Chapter 1, Box 1.7) using the triangle and the levels of prevention. Create a table from the information in Box 5.4 and a figure like Figure 5.6.

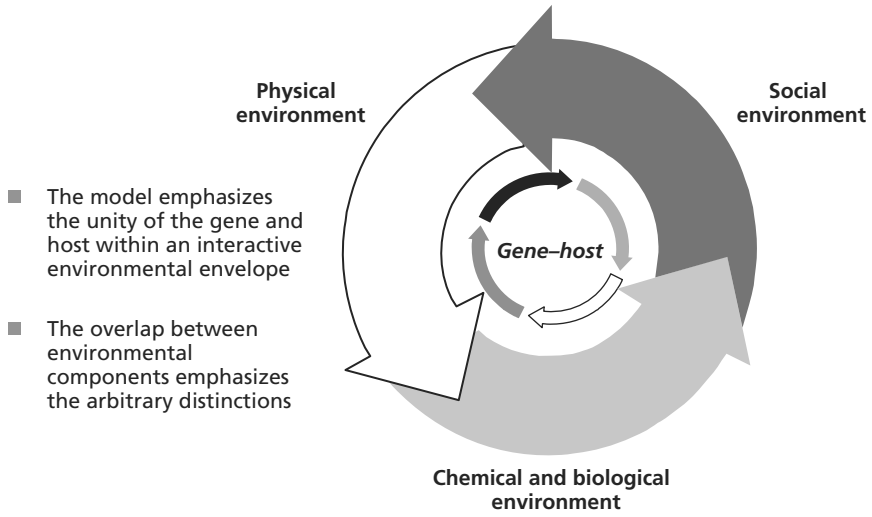


Fig. 5.7 Wheel of causation.

Reprinted, with minor adaptation, from *Epidemiology: An Introductory Text*, 2nd edition, Mausner JS and Kramer S, Philadelphia, PA: WB Saunders, Copyright © 1985, with permission from Elsevier.

within an interacting environment. This version of the model emphasizes the fact that the division of the environment into components is somewhat arbitrary.

In Figure 5.8 the wheel model is applied to phenylketonuria, the classic genetic disorder. Phenylketonuria is an autosomal single gene disease (autosomal means it is not on the sex chromosomes). As a result, an enzyme required to metabolize the dietary amino acid phenylalanine and turn it into tyrosine is deficient, and so phenylalanine accumulates in the blood, causing brain damage. Early diagnosis, usually through screening, and then following a diet low in phenylalanine can prevent the disease. The cause of this disease could be said to be a faulty gene. The cause

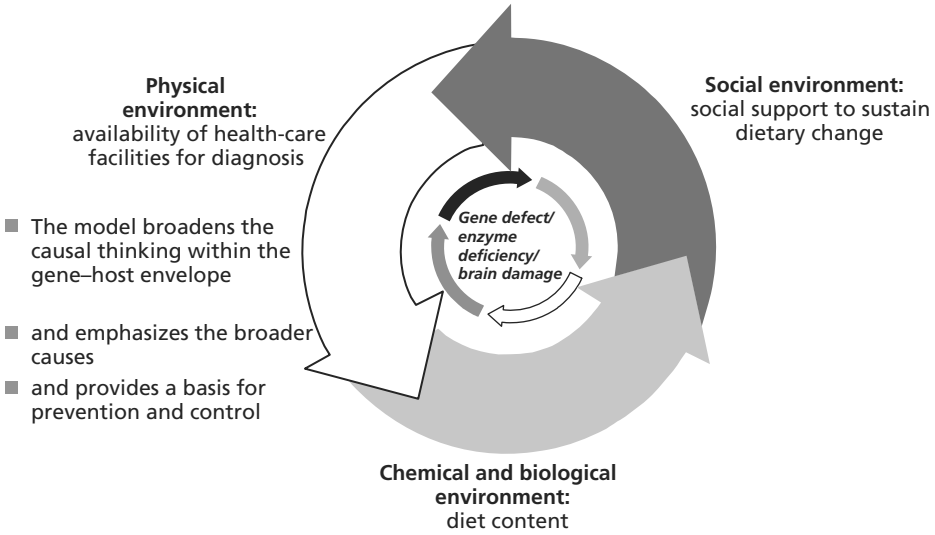


Fig. 5.8 Wheel of causation applied to phenylketonuria.

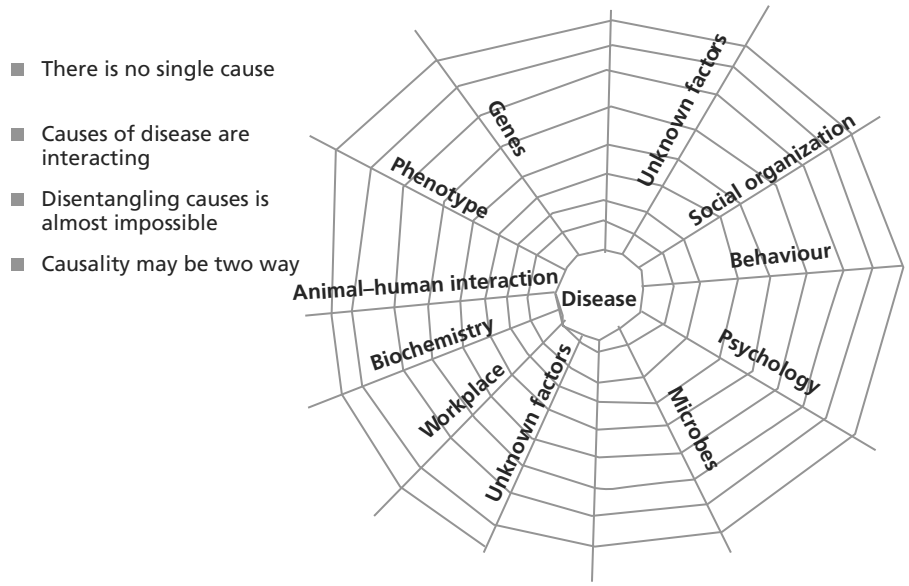


Fig. 5.9 Web of causation.

of the disease is, in reality, a combination of a faulty gene, exposure to a diet containing a high amount of phenylalanine (about 15% of the protein of most natural foods), and in the case of failure of diagnosis and dietary advice, a social environment unable to protect the child.

For many disorders such as coronary heart disease, and many cancers, our understanding of the causes is highly complex. Either the causes are truly complex, or equally likely, our understanding is too poor to permit clarity. These disorders are referred to as multifactorial or polyfactorial disorders. As discussed earlier, all disorders have several causes and where that is not the case, it is simply a matter of our definition. In disorders with multifactorial causation often no specific causes are known, many factors appear to be important, and mechanisms of causation are not apparent. (Usually, greater knowledge of causes brings simplification. For example, until recently gastric ulcer was thought of as a complex, multifactorial problem associated with stress. Now, we know the bacterium *Helicobacter pylori* lies at the heart of causation. Gastric ulcers are no longer thought of as multifactorial disorders.)

The complexity of these multifactorial diseases is not captured by the line, wheel, and triangle concepts (which remain useful nonetheless) and is better portrayed by the metaphor of the spider's web. In some portrayals, the web is shown as a highly schematized diagram, more like an electronic circuit or an underground transport map. Such portrayals tend to underestimate the complexity and overestimate the state of understanding. The web, as shown in Figure 5.9, emphasizes the interconnections among the postulated causes. This model, more than the others, indicates the potential for the disease to influence the causes and not just the other way around. For example, lack of exercise may be one of the causes of heart disease but the disease can also cause people to stop exercising (called reverse causality). The metaphor of the web permits the still broader causal question: where is the spider that spun the web? (after Krieger 1994). The question can be answered at a number of levels, for example, evolutionary biology, social structures, economics, and role of industries. This question chimes with Rose's concept of the causes of the causes (Chapter 2). The analysis of heart disease causation using the web of causation begins to illustrate the great complexity of this disease (see Fig. 5.10).

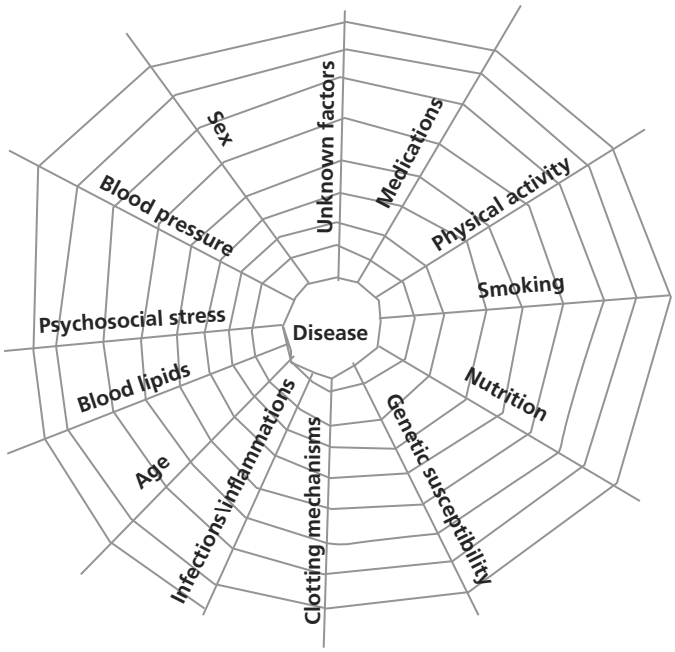


Fig. 5.10 Web of causation and coronary heart disease.

The purpose of models is to simplify reality and promote understanding. The web permits us to grasp the complexity of multifactorial diseases but the line, triangle, and the wheel help us to focus on their essentials and think through practical actions. Before reading on, do the exercise in Box 5.10.

Models provide a means of analysing causal pathways and a foundation for the application of epidemiological knowledge to public health action. Narrow causal thinking based on single causes, in contrast, can mislead epidemiologists into prematurely believing that a problem has been resolved and can seriously distort public health action. Models also help to lay out what is, and what is not, known. Causal models also help us to understand the ideas of necessary or sufficient causes, ideas that have been formalized in the interacting causal components model (causal pies) considered below.

5.3.2 Necessary and sufficient cause, proximal and distal cause, and the interacting component causes, models: individuals and populations

Epidemiological thinking on causality has been influenced by the concepts of necessary and sufficient cause. Mostly, this thinking relates to causation at the individual level (I will consider the

Box 5.10 Analysing disease using the wheel and web models

Review the health problems or diseases that you picked and disease X (Chapter 1, Box 1.7) using the wheel and web models.

implications for populations at the end of this section). *A Dictionary of Epidemiology* tells us that a necessary cause is a factor whose presence is required for the occurrence of the effect. It defines sufficient cause as a set of conditions, factors, or events sufficient to produce a given outcome. A factor, or a group of factors, whose presence leads to an effect is a sufficient cause, so some causes of diseases are said to be sufficient in themselves to induce disease while others are said to be necessary components in a larger jigsaw of causes. For example, the tubercle bacillus is required to cause tuberculosis but, alone, does not always cause it, so it is a necessary, not a sufficient, cause. In other words, a single factor does not cause this disease. This is, of course, the key message of the causal models discussed here.

The problem in practice is that a cause on its own rarely induces a disease in an individual except for extremely serious genetic defects. The necessary and sufficient causes model has theoretical value for analysing causes, but in epidemiology, as Susser (1977) points out, most causal factors are neither necessary nor sufficient, but contributory. Try the exercise in Box 5.11 before reading on.

Down's syndrome is the name given to a disorder where a person has a highly characteristic appearance (leading to the previous name, mongolism), and who will inevitably be mentally retarded because they have three chromosomes at the position of chromosome 21 instead of the normal two (trisomy 21). This genetic feature is a sufficient cause of Down's syndrome. In other words, this chromosome abnormality alone will lead to the characteristics that define Down's syndrome.

Sickle cell disease (two sickle cell gene alleles per cell) is a genetically inherited condition. The position is not quite the same as for Down's syndrome because the word disease leads to an expectation that the person has, or will develop, a health problem. The presence of sickle cell genes is a necessary cause of sickle cell disease. In milder cases especially, external stimuli such as infections are required to cause clinical disease. Here we have another example (phenylketonuria was discussed earlier) of genes being necessary, but not always sufficient causes of genetic diseases.

Scurvy occurs when there is insufficient vitamin C in the diet to maintain health, usually due to lack of fruit and vegetables. This does not occur in natural circumstances, but does when a restricted diet is taken, as in the past by sailors on long voyages in sailing ships, and nowadays by food-related problems or in the mentally disturbed. Vitamin C insufficiency is both a necessary and sufficient cause of scurvy. By definition, other diseases, several of which look like scurvy, are not scurvy unless there is a lack of vitamin C. Yet, dietary insufficiency of vitamin C is unnatural, so other factors, in practice, come into play.

For tuberculosis, exposure to the bacillus is necessary but not sufficient in most people, and in many people the organism is controlled in the host. For both tuberculosis and scurvy, contributory causes include poor socio-economic conditions. These increase both the risk of exposure to the necessary cause and, for tuberculosis, the likelihood of the organism establishing a clinically important infection.

Box 5.11 Necessary and sufficient causes for some disorders

Consider the causes of one or all of Down's syndrome (trisomy 21), sickle cell disease, tuberculosis, scurvy, phenylketonuria, and lung cancer. If the cause is sufficient, its presence alone would induce the disease and if it is necessary, the disease would not occur in its absence. What do you think: are the causes sufficient and are they necessary?

For phenylketonuria, the necessary cause is a genetic defect and that together with a diet containing normal amounts of phenylalanine is sufficient (Fig. 5.8).

For lung cancer tobacco smoke, by far and away the most important causal factor is neither necessary nor sufficient, for there are many other causes. Some smokers do not develop the disease and some non-smokers do.

This analysis shows the strengths and weaknesses of the necessary/sufficient cause concept. When a specific cause of disease is known it can be incorporated into its definition. The specific cause becomes necessary by definition. For multifactorial diseases, at least at present, there are no known necessary causes. The example of lung cancer illustrates this well. In practice, except for unusual or unhelpful scenarios (e.g. a bolt of lightning, or falling off a cliff), there are no single sufficient factors that inevitably lead to chronic diseases or death. Ageing (and birth!) is probably the only sufficient cause of death. The concept of sufficient causes has, therefore, veered from single causes to groups of causes.

Rothman's interacting component causes model (Fig. 5.11) has emphasized that the causes of disease comprise a constellation of factors. It has broadened the sufficient cause concept to be a minimal set of conditions which together inevitably produce the disease. Different combinations of these factors may cause the disease. Figure 5.11 is a simplified version of Rothman's ideas. Three combinations of factors (ABC, BED, AEC) are shown here as sufficient causes of the disease. Each of the constituents of the causal 'pie' is necessary, and hence contributes to 100 per cent of the risk of disease attributed to that particular combination of causes. The factors are conceived to act in a biological sequence, which determines the period between the beginning of causal action and the initiation of disease. It follows that control of the disease could be achieved by removing one of the components in each 'pie'. If there were a factor common to all 'pies', the disease would be eliminated by removing that factor alone. In this case removing factor A would remove all the disease caused by the first and third constellation of causes. This mode of reasoning, and model, is hard to apply to specific diseases but has considerable theoretical value.

A sequence of causes can be considered in terms of time, and also in terms of space. Causes that are close to the individual—in terms of time or space—are sometimes referred to as proximal causes, in the sense of 'near to'. Those that are distant are called distal causes in the sense of 'away from'. To give an example, proximal causes of lung cancer would be smoking cigarettes and exposure to radiation. Distal causes would be poor diet, poverty, or tobacco farming. Distal causes are sometimes referred to as upstream.

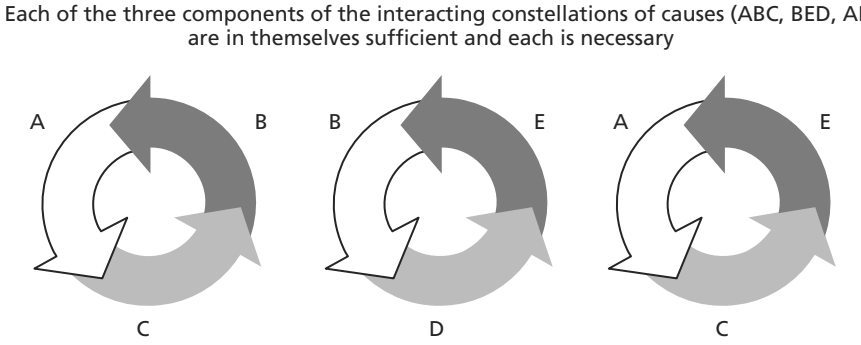


Fig. 5.11 Interacting component causes.

The problem with these complex models, in practice, is that we do not have the knowledge to define even a sufficient constellation of causes, that is, we cannot define the 'pies' in Figure 5.11. The components of the pies, as sufficient causes, must be acting differently than if they were separate (when they do not cause disease). There is, therefore, a modification of the biological effect (i.e. biological interaction that leads to disease). The model is, therefore, useful in thinking about the issues discussed in section 5.4.

The model is also pertinent to thinking about the strength of the association (section 5.6.2) and relative and attributable risk (sections 8.4 and 8.6). Interested readers should consult Rothman's writings.

Epidemiology is a population science but the work has to have meaning and applications for individuals, otherwise it would be of little or no use to clinicians and others, including health promoters, when dealing with patients. Clinicians apply probabilities in the management of the individual. It is true that sufficient causes can rarely be defined at the individual level. This is not true, though it is seldom discussed, at the population level. We know many single causes are sufficient to alter the incidence of disease at the population level and we can predict the consequence with great certainty. So, at the individual level smoking cigarettes is not a necessary or sufficient cause of lung cancer. At the population level, it is not a necessary cause of a rise in lung cancer, but alone, it is a sufficient cause of a rise in the incidence of lung cancer. Therefore, the mostly theoretical concepts of necessary and sufficient causes can find practical and firmly grounded applications at the population level.

5.4 Susceptibility, risk/effect modification, and interaction

We have already considered, briefly, the idea of disease susceptibility, for example, in the extreme case where some diseases rarely, or never, occur because humans have no susceptibility to them. Even for human diseases, exposure to the causes is often common but the disease is rare. The fact that some people who are exposed to the known causes get disease, and others do not, implies there are other factors that determine the outcome, as emphasized by all the models we discussed.

Whether we consider trivial problems such as the common cold, or serious diseases such as coronary heart disease (CHD), it is clear that some individuals are more susceptible than others. Epidemiology cannot help with understanding such individual differences. The same variable susceptibility, however, applies to populations, where epidemiology can shed light. In the case of the common cold the reasons tend to be fairly obvious: poor populations living in overcrowded conditions are more susceptible to being exposed to the common cold virus, may get a larger dose, and may have lower resistance and therefore are more likely to get ill.

For CHD it is not at all clear why, to take one of many potential comparisons, Indian-born women in England and Wales should have more CHD than predominantly White European-origin women born in England and Wales. This observation seems real, and not a data artefact. In fact, on first principles we would expect the opposite because of less smoking in Indian-born women for example. Just as with the common cold, there must be one or more factors that increase Indian-born women's risk. Typically, in studies of ethnic variations in CHD, the known risk factors do not appear to explain such differences, although at least part of this may be due to difficulties in measuring risk factors precisely. There may be unknown factors that cause these differences in risk of CHD. Such factors may even modify the risk of CHD through known risk factors; for example, hypertension may be a more potent cause of CHD in one ethnic group than in another. If that were true, the question of why this is remains. At this point, speculation is usually required, for example, unknown genetic variations, or life-course effects, possibly related to fetal development that alter susceptibility to the effects of high blood pressure.

The factor that influences susceptibility is known as a risk modifier or effect modifier, and there is said to be an interaction between the study exposure and the effect modifier on outcome. Strictly speaking, it is an association modifier but the word *effect* is now embedded. Interaction is a word from statistics and is used, mostly, loosely and synonymously with effect modification, but care is needed in the use of these terms as will be considered later. Risk and effect is best judged in epidemiology by how the incidence of disease is changed by exposure to the risk factor (Chapter 7), and it is quantified using either absolute or relative measures of risk (Chapters 7 and 8).

Risk or effect modification is different from confounding. It occurs when two factors reduce or increase risk. Before reading on, do the exercise in Box 5.12.

The possible combined effects are that the coexistence of these factors can lead to the addition of the individual cumulative incidence or relative risks, less than the addition of the two (antagonism, or negative interaction), or more than the addition of the two (synergy, or positive interaction). Our default situation, and expectation on first principles, is that at least the two risks will be combined. We call this the additive model of effect modification. There are good reasons for basing biomedical and public health research and practice on this model (readers should consult Kenneth Rothman (2012) for the arguments but be aware that this is still a controversial area). In this case, if there is no effect modification or interaction, the cumulative incidence in this population is 100 (baseline) plus 200 (excess from smoking), plus 200 (excess from alcohol)—that is, 500 per 10 000. This gives a relative risk of five. If our research shows a different result—say 800 per 10 000 or 300 per 10 000, and it is not a result of chance, error, or bias (including confounding), then we have risk modification on the additive scale. The two risk factors, smoking and alcohol, are then combining to give a result other than that expected on the additive effect. We can say that the presence of one factor has modified the effect of the other.

A simple definition of effect modification is this: the association between a risk factor and an outcome differs in subgroups of the population. In this example, the association between smoking and cancer X has been altered by the presence of alcohol, and vice versa, and this alteration occurs on the additive scale in a subgroup of the population, where both risk factors are present.

Box 5.12 Potential effects of smoking cigarettes and drinking alcohol on a hypothetical cancer X

When two (or more) causal risk factors coexist, what is the likely effect on the outcome? For example, say that smoking cigarettes triples the risk of cancer X, and drinking alcohol also triples the risk. What is the risk of cancer X in a population where people both smoke cigarettes and drink alcohol? Let us say that the cumulative incidence of disease in people who do not drink or smoke is 100 per 10 000 people. In those that smoke, the risk is 300 per 10 000 people, 200 extra cases per 10 000 people being contributed by smoking (relative risk 3, excess relative risk 2, where excess relative risk is relative risk minus the baseline risk, which is by definition 1), and similarly for those that drink alcohol (200 extra cases per 10 000 people, relative risk 3, excess relative risk 2). Assume the four groups of people—those who do not smoke or drink, those who smoke, those who drink, and those who both smoke and drink—are identical in every other respect, that is, there is no confounding. Now imagine we have 10 000 people who all drink alcohol and smoke cigarettes. What are the possible combined effects in this group? You may wish to prepare an appropriate table including the cumulative incidences and relative risks to test your thinking, before looking at Table 5.3(a), which summarizes the text that follows.

The two risks might have large modifying effects, perhaps even so much that the effects are multiplicative or more; that is, the two factors of smoking and alcohol together increase relative risk by ninefold (1 · 3 · 3; the 1 signifying baseline risk, in the absence of the two risk factors) and not by fivefold (1 + 2 + 2). If a third factor (say a gene variant) also tripled relative risk, people with the three factors would have a 27-fold risk (1 · 3 · 3 · 3) and not a sevenfold risk as for the addition of baseline and excess risks (1 + 2 + 2 + 2).

In fact, this multiplication effect is often seen. As a result, sometimes statistical interaction is defined as a departure from the expectation of multiplication of risks. Investigators can conclude, sometimes without meaning to because as we will see the claim is not tenable, that there is no effect modification, because there is no departure from multiplication of risks. Logically, if there is no interaction on the additive scale then there must be interaction on the multiplicative scale (negative interaction), and if there is positive interaction on the multiplicative scale, then there must be interaction on the additive scale also.

The combined effects might be even bigger than 27. Then we would have effect modification demonstrable even using statistical methods that are based on the multiplication of risks. It is, however, exceptional to demonstrate such large interactive effects in epidemiology.

These additive and multiplicative considerations in relation to risk are important to the development of research questions, study design, analysis of data, the choice of statistical methods and analysis programmes, and interpretation of data. Statistical models sometimes assume a multiplicative model of risks. They may use logarithms, where addition is equal to multiplication, for example, for logarithms on base 10, where 0 is 1 in ordinary numbers, and 1 is 10 in ordinary numbers, the addition of the logarithms 1 plus 1 is equivalent to 100 in ordinary numbers. (The multiple logistic regression model and the Poisson regression model are examples. The multiple logistic regression model uses natural not decimal, logarithms.) Other models, for example, multiple linear regression, do not use logarithms.

In epidemiology and statistics, the concept of effect modification is increasingly discussed as interaction. Interaction is often demonstrable on the additive scale (relating to the biological and public health concept of interaction) but rarely on the multiplicative scale.

In the classic and most commonly cited example of effect modification (interaction) between smoking (tenfold increased risk) and asbestos (fivefold increased risk), the combined effect on lung cancer is multiplicative with a combined greater than 50-fold increased relative risk (compared with 15.1, i.e. 1 plus 9.9 plus 4.2, on the additive model). This result, summarized in Table 5.3 (b), has been shown by Hammond *et al*. (1979). If this interaction had been tested out using statistical analysis based on the multiplication of risks as the standard approach, investigators could have concluded that there was no effect modification. Such a general conclusion would be wrong because there is interaction on the additive model. If the combined effect had been substantially more than 50, there would also have been interaction on a multiplicative scale. The important point is that the risk of lung cancer in smokers who are exposed to asbestos is much higher than in those not so exposed.

One difference between confounding and effect modification is that the exposure–outcome association would be similar in all levels (strata) of a confounder but that the exposure–outcome association differs in different levels (strata) of an effect modifier. Confounding is an obstacle to proper interpretation that should be controlled for but effect modification is of causal and public health interest and should not be controlled. (A variable can act as both a confounder and an effect modifier in different circumstances, e.g. age.)

Table 5.3 Effect modification for (a) cancer X (imaginary) and (b) lung cancer (real)

Exposure	**Cancer X cumulative incidence/10000**		**Relative risk**	
(a) Imaginary data	**Actual**	**Excess***	**Actual**	**Excess**
No smoking, no alcohol**	100	0	1	0
Smoking, no alcohol	300	200	3	2
Alcohol, no smoking	300	200	3	2
Smoking and alcohol				
No effect modification on the additive model	500	400	5	4
Effect modification on the additive model (example)	800	700	8	7
(b) Based on classic example (Hammond *et al*.)	**Lung cancer Death rates/100 000**		**Relative risk**	
	Actual	**Excess**	**Actual**	**Excess**
No smoking or asbestos	11.3	0	1	0
Smoking, no asbestos	58.4	47.1	5.2	4.2
Asbestos, no smoking	122.6	111.3	10.9	9.9
Asbestos and smoking				
- no effect modification (as expected) if no additive model interaction	169.7***	158.4	15.1	14.1
- with effect modification on additive model (as found)	601.6	590.3	53.2	52.2

*Over baseline

**Baseline

*** 169.7 = 11.3 + 47.1 + 111.3

As we will discuss in Chapter 8, there are two main ways of presenting risk data—by absolute/actual risk, and relative/comparative risk. In discussing interactions, we should clarify which approach and analysis we intend to pursue, our prior definition of interaction, and the causal model we are assuming. In the classic example discussed just now involving smoking and asbestos in lung cancer, the interaction is clear if the causal model is an additive one. If, however, the chosen causal model presumes a multiplication of risks, then there is no important departure from that. The latter is an awkward interpretation, as this is the prime example of interaction, in the sense of effect modification, in epidemiology.

The surprising conclusion of this is there is always interaction from a statistical perspective. It is vital, therefore, for researchers to tell us how they examined interaction, why they did so, and on what scale there was interaction. Unfortunately, this essential and simple information is only exceptionally provided (this can sometimes be inferred from the kind of statistical analysis done). My recommendation to readers is to think of interaction as a departure from additive risks (following Rothman's advice (Rothman, 2012)): Rothman recommends stratified analysis to spot effect modification, prior to statistical analysis of interactions. That way examination of effect modification is not just a by-product of analysis.

Interaction can also occur when the outcome is a continuous variable, for example, blood pressure or cholesterol levels. The current trend is for very large studies and this is driven by the large samples required for studying gene–environment interactions.

The promise of the new genetics is to clarify the nature of individual and group-level susceptibility to diseases, and the variable response to treatments. Gene variants can act as effect modifiers/interactive factors of the relationship between an environmental exposure and an outcome. The problem is that studies that can accurately assess interaction need to be very large, especially when effects are small (as with most gene variants). Mostly, studies that report there is no interaction are too small to reach such a conclusion, and many others have applied the wrong conceptual approach. If there is heterogeneity in the effects of a risk factor within populations, as is highly likely, there must be effect modification/interaction and vice versa. Failure to seek or notice risk modification can lead to a false measure of population risk and the possibility of missing an important finding, at least for population subgroups. The importance, in practice, of effect modification/interactions is under debate.

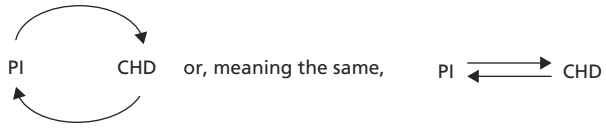
Causal graph methods that are helpful in data gathering, analysis, and interpretation have become available in epidemiology (mainly developed and advanced in other disciplines), and these are considered next. At present, these methods do not incorporate effect modification, and are being developed for this.

5.5 Causal graphs: introducing the directed acyclic graph

A diagram (or graph) that makes explicit the postulated relations between variables is recommended. Producing such a diagram is difficult because it requires considerable understanding, including of the biology, of the topic under study. This step is a component of a disciplined approach to research that includes prior statement of a fully articulated scientific hypothesis (not merely the statistical null hypothesis) and a detailed analysis plan prepared in advance. With a hypothesis, causal diagram, and an analysis plan, we can assess whether the results of the research fit with our prior presumed causal understanding. If yes, that increases confidence that our prior understanding was correct. If not, then it motivates us to alter and improve our hypothesis and causal diagram. This can then be checked with new research. The field of causal diagrams in epidemiology is large, complex, and is developing fast. One important and increasingly used form of causal diagram is called the directed acyclic graph (DAG). The field has been advanced by, among others, Judea Pearl, who is a professor of computing sciences and statistics. Pearl has claimed in his book *Causality* (second edition) that causality has been mathematized. While this claim is not (yet) true for epidemiology and related population health and medical sciences, it merits examination.

The DAG is based on graph theory in mathematics and can be written as an algebraic equation (in a form known as a structural equation model). Equally, algebraic equations can be expressed as DAGs. This property allows DAGs to be used both for expressing potential causal pathways, and for guiding and interpreting data analysis. The field of DAGs has vocabulary that is different but related to that of epidemiology. Some of this vocabulary is in Table 5.4, which also re-expresses it in similar epidemiological terms. The DAG uses lines and arrows (Table 5.5) that follow formal rules.

Assume we have a relationship between two variables (A and B) that we believe is potentially causal. Before examining the DAG approach, we will look at this relationship in general terms. To make this less abstract let us name A as the variable physical inactivity (PI) and the outcome B as coronary heart disease (CHD). It is known that physical inactivity can cause CHD, but also that CHD can cause inactivity. This complex relationship can be designated as circular:



Epidemiology simplifies this kind of problem by examining the possibilities separately. Probably the best way to do this is by studying young people long before they get the outcome, here CHD, which occurs in later life. Therefore, we could, for example, consider the relationship between physical inactivity in early life, childhood, adolescence, and early adulthood and the outcome

Table 5.4 Terminology commonly used in DAGs in relation to similar or related terminology in epidemiology

DAG terminology	Epidemiology terminology for similar ideas
Ancestors	Distal causes
Back door pathway	Confounding variable(s) creating the association
Block/blocking	Eliminating an association through, for example, adjusting, stratifying, etc.
Blocked path	An association that has been eliminated as it has been controlled for, e.g. by adjusting for confounders
Blocking	Presence of confounding or selection bias
Child	Effect, outcome
Collapsibility	The measure of the association is not affected whether examining stratified or overall, actual (crude) rates
Collider	A variable that is caused by both the exposure and outcome under study
Collider stratification bias	See M-bias
Conditioning	General term to include stratification, standardization, and adjustment in a model (conditioning means holding a variable constant)
Descendants	Effects on potential causal path including mediators
D-connected	The postulated causal path is open (see open path and path)
D-separation or D-unconnected (directional separation)	The postulated causal path is closed
Endogenous selection	See M-bias
Identification	Analysis of associations to separate error/bias/confounding from causal effects
M-bias	Berkson's bias/selection bias. It arises from, for example, adjusting for a collider
Nodes	Variables
Open path	Potential causal relationship, i.e. association
Parent	Proximal cause
Path	The route to potential causality, i.e. from A to B
Vertices	Variables

Table 5.5 Symbols used in the construction of directed acyclic graphs (DAGs)

Symbol	Name of symbol
———	Edge. This edge is undirected, i.e. there is no arrow to give the direction
>	The direction indicator for an edge
⟶	Directed edge (arrow) indicating association (arrow can be thickened to denote stronger association)
X	Variable or in some notations variables conditioned on X (the practice of putting variables in a box is not always followed in epidemiology)
. . . . .	Association induced by collider bias
< ····· >	Bidirected edge, denoting an association induced by confounding (the line can be solid)

CHD in middle age and beyond. In this case we postulate and study that physical activity leads to CHD i.e. PI → CHD. It is not plausible that CHD in middle age leads to inactivity earlier. So in doing this, we have simplified our research.

In contrast to this example from a chronic disease, nearly all infectious diseases have the simple A → B relationship where A is the infectious agent and B is the outcome; for example, A is the measles virus and B is the illness. It is not plausible that measles illness causes the acquisition of the measles virus.

So, for causal analysis we usually start by simplifying the matter under study. Simplifying means the construction of models (in the same way that an architect may create a model building). The DAG is a model. The equation describing the DAG is also a model. The DAG needs specification of exposure (A) and outcome (B) and, ideally, on the ancestors and descendants (see Table 5.4) of A and B.

A relationship between A and B can be expressed as A – B, so A and B are connected by a line that we call an edge. The variables are called nodes. Does A cause B or does B cause A (the problem of reverse causation)? As it is designed to help causal analysis, the DAG method does not permit two-headed (bidirectional) arrows. The word acyclic means there are no cycles in the graph. When complex cyclical relationships are to be studied, other types of causal graphs need to be used (theory is available). Of course, we could create two DAGs, one for A → B and the other for A ← B. We could plan and run the analyses separately. Let us assume that our interest is in A causes B, then we can express that as

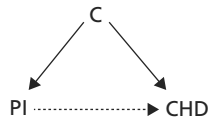
$$A \rightarrow B$$

If A → B, it could either increase B (positive association) or decrease B (negative association). The DAG can be used to show this, for example, by having one colour for positive relationships and another for negative ones or plus/minus signs. Clearly, the researchers need a high level of knowledge to set out these kinds of decisions in advance.

Let us assume that our understanding of the relationship between physical inactivity and CHD is not based on our reading of fabricated, seriously biased, or chance research results (i.e. that it is a reasonable proposition).

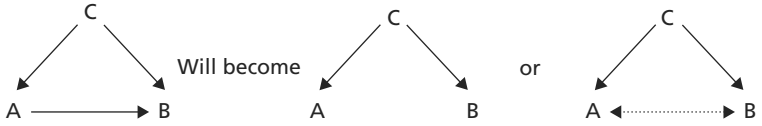
What else do we need to consider in this DAG to make it informative? First, we must consider the possibility that a third variable (C) is creating the relationship A → B, that is, there is confounding (known as a backdoor path in DAG terms). The confounding factor could be age, sex, socioeconomic status, or similar variables.

This can be expressed as

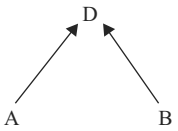


Here the dashed line denotes the possibility of a confounded relationship. As we already know if we adjust for C, and C is a confounding factor creating this relationship, then the association between A and B will greatly reduce or even disappear. The dashed line will disappear. This is described as closing the backdoor path.

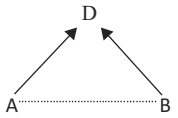
If we control for a confounding variable and the association between A and B disappears, we can redraw the DAG either without an arrow between A and B, or with a dotted line between A and B to signify that relationship is confounded.



When both our variables of interest are associated with a third variable, presumed causally in that direction, this variable is called a collider. If there is a collider D, then we have



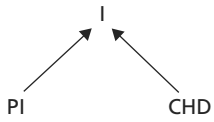
If we control for collider D, we induce a spurious association between A and B, which is denoted as



The dotted line has no arrows, which distinguishes it from confounding.

It is possible that both physical inactivity and CHD cause an effect that is important in the relationship, for example, social isolation (I)

So



I (social isolation) is called a collider. The collider blocks the path from P1 to CHD but adjusting for it opens it up, the opposite of adjusting for confounding (closing the back door path). Here a biased association may be created between PI and CHD because of adjustment for social isolation. This is not at all an unfamiliar concept (even though it is still an unfamiliar term) in epidemiology,

but the use of DAGs has made it clearer. We have already discussed this in Chapter 4 with the example of Berkson's bias. In this example, let us assume that, in fact, there is no association between PI and CHD as in the above diagram (no line connecting the two). If we condition on the collider I, we would hold it constant. One of several ways of doing that is to study only socially isolated people, that is, stratify. In this group, there will be an association between PI and CHD, if these two variables are among the causes of social isolation, which seems plausible. The term *collider stratification bias* includes a range of biases that have varying names in epidemiology (e.g. Berkson's bias).

In analysis, our aim is to help separate non-causal and causal associations so we can see we must not control for colliders (or their descendants or ancestors) and mediators (or their ancestors or descendants), but we should control for confounders. If the study is free of error/bias, and all confounding factors are included, then the DAG reflects a causal structure and the resulting analysis reflects a causal relationship between A and B. Of course, this ideal state is not actually achieved.

We may ask how physical inactivity causes CHD. One possibility, among others, is that it does so through obesity (O).

$$\text{PI} \rightarrow \text{O} \rightarrow \text{CHD}$$

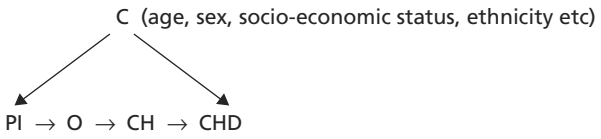
Obesity might, for example, work through raised cholesterol (CH).

$$\text{PI} \rightarrow \text{O} \rightarrow \text{CH} \rightarrow \text{CHD}$$

Obesity and cholesterol would, therefore, be postulated to be intermediate variables (synonym, mediators). If, however, we adjust statistically for the intermediate variable O and the relationship between PI and CHD remains completely unaltered then it is not, actually, acting as an intermediate variable. If the association weakens or disappears, then the case for O being a mediator is strengthened. However, this interpretation requires that the outcome, here CHD, does not cause the intermediate (i.e. obesity), which common sense tells us is plausible but we make the assumption this in not true. (In our DAG we may either use different colours for the lines/arrows for different kinds of variables, or use different line widths or use broken lines.)

Therefore, our revised causal proposition might be that after adjusting for potential confounding factors, we think physical inactivity leads to CHD through obesity and cholesterol.

So, this DAG will be

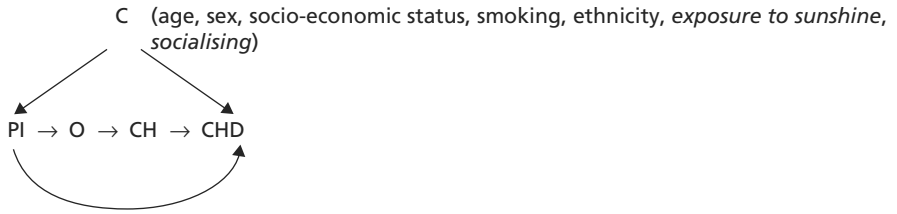


We can see the DAG has helped to think through and clearly express, in a way that everyone can understand, complex relationships between confounders (C), the exposure/risk factor (PI), intermediate variables (O, CH) and the outcome (CHD). Our knowledge and data sets both tend to be incomplete, so this is a provisional model. The model should also include uncertainties. Uncertainties can be added, for example, as U for unmeasured variables and error terms.

If we know of variables that are potentially important in the relationship but have not been measured, they should be added to the DAG so we can immediately see the limitations of the study. So, for example, it may not be physical inactivity that is important but the fact that physically inactive persons may spend less time outdoors, thereby being less exposed to sunshine or to the social benefits of meeting neighbours and people on the street. These could be unmeasured confounders or mediators. Our study may well have no data, either because these variables were

forgotten, omitted deliberately, or most probably it was not feasible to collect the data, possibly for reasons of time and cost. The DAG helps by being explicit on the limitations of the causal analysis.

Again, these unmeasured variables could be picked out using a different colour or different kind of line. Here, these are shown as unmeasured confounders in *light grey*. These unmeasured factors are treated here as confounders, but they may actually be on the causal path. Investigators need knowledge and judgement to decide where the variables belong.



Of course, physical inactivity may have effects on CHD either directly (without mediators as shown by the direct arrow) or through a completely different pathway than obesity (e.g. through endothelial dysfunction). The effect of the direct and indirect paths can be quantified. This extra, endothelial pathway, could be added (but this is beyond our purposes here). There may be confounding in the relationships between the intermediate variables and both exposure and outcome, which needs to be considered. Ideally, the DAG (and statistical model) should include confounders not only between PI and CHD, but also between other variables on the path.

The DAG is also telling us an important story about the variables that are not shown. The DAG is telling us that the investigators do not think these are important in this particular causal path. For example, diet is not mentioned. If that is not an oversight, then we can deduce the investigators think that diet has no relevance here. If so, excluding diet is the correct decision. Variables that do not connect to causal and confounding variables on the DAG should be excluded, even if they connect to the outcome through some other causal path, for example, smoking cigarettes and CHD. Ideally, investigators would explain why variables like this were not included. DAGs do not include statistical interaction, although this area is being developed.

A variable may be a moderator, mediator, and a confounder variable (e.g. socio-economic position). The elements of such a variable that are thought to be moderators or mediators need to be separated from those thought to be confounders, and should be labelled differently, for example, SEP1, (income), SEP2 (education), SEP3 (housing tenure), etc.

Software is available for drawing DAGs, for example, DAGitty (www.dagitty.net). This software uses colour coding with confounders in red, mediators in green, and no effect in grey. Currently, a variable can only have one status in this software. The DAG software shows which variables are, and which are not, essential for measuring the association between A and B.

Exemplar 5.1 illustrates a study that shows how the process works.

Exemplar 5.1: Factors associated with prediabetes (Bardenheier *et al.* 2013)

Bardenheier *et al.* used the cross-sectional United States National Health and Nutrition Examination Surveys (NHANES) 2001–2006 to test a hypothetical causal model for prediabetes in people aged 50 years or more. NHANES had data on 2230 eligible people without diabetes. The study was motivated by observing that while many risk factors for prediabetes had been identified, they had not been examined simultaneously as a coherent system or model. They identified a statistical method called structural equation modelling (see glossary) as a way of testing a hypothetical model. Their hypothetical models before and after analysis can

be examined in their original coloured format at this URL: http://care.diabetesjournals.org/content/36/9/2655.full.pdf

The variables that were measured were shown in rectangles, and those that were not measured but reflected in or derived from observed variables were shown in ovals and are known as latent variables. The arrows indicated the postulated direction of association.

Based on the investigators' knowledge, itself reflecting prior research, 10 major variables were identified as either direct or indirect predictors on the causal path to prediabetes. Arrows showed 27 paths from these variables to the outcome. Some of the variables were composite ones based on several measured variables, for example, socio-economic position (a latent variable; SEP), while others were single (and therefore measured); for example, high blood pressure.

The model was assessed using structural equation modelling with factor and path analysis. Factor analysis can group interrelated variables (factors are the lowest common denominators for a number). Path analysis assesses the direct and indirect effects of the factors identified. The factor analysis approach reduced the number of individual variables, which has statistical advantages, as well as easing examination of the model.

In the model, there were variables that had no number (i.e. age, race/ethnicity, and sex). These were identified as potential confounding variables. They were not entered as potential direct effects, while family history was. The authors justified this as follows: 'Because age, sex, and race/ethnicity are strong, non-modifiable confounders related to most of the other factors in the model, their direct effects, while included, are not shown in the graphic of the final model. Although family history is non-modifiable, it is specific to diabetes risk and therefore is examined as a factor of interest.'

The analysis used advanced statistics that is both beyond the scope of the book and unnecessary to understand the key points.

The structural equation model indicated that in the SEP factor (group of variables) the number of family members did not contribute. In the poor diet factor, saturated fats and processed meats did not contribute. The latent constructs were shown to be correlated with each other. To create a model fitting the data better, total cholesterol and BMI were removed. Then the direct effect of diet on high-density lipoprotein (HDL) cholesterol was dropped. Age, sex, and ethnicity were shown to have direct effects on most factors.

The model was reconstructed with the 10 postulated directly causal variables.

Concluding remarks

The outcome of prediabetes is not a complex one compared to many chronic diseases or syndromes such as CHD or asthma. The paper shows, however, how complex the putative causal relationships are. Yet, the authors have also taken a pragmatic decision to treat three major variables (age, sex, ethnicity) as confounding variables, rather than as direct effects or intermediate variables. This simplifies the thinking greatly. The great merit of this paper is the immense effort the authors have expended in creating a credible, theory-based causal diagram. They have then used data to assess the model. The data suggested a slightly different model that provided the best fit between the postulated model and the calculated model. The authors have, after this mammoth effort, made no claims to having produced a definitive model. Rather, recognizing the limitations of their data, in particular the cross-sectional design, they have urged further examination of the model, but with cohort data. They have also showed in what respect this model aligns with the published literature. This paper exemplifies how causal epidemiology might be done.

In my view, all papers working with associations aiming to contribute to the causal basis of a subject should provide the causal diagram based on current evidence (in the introduction) and the refined version following their research (in the discussion), as Bardenheier and colleagues did.

Source: data from Bardenheier *et al.* (2013) A novel use of structural equation models to examine factors associated with prediabetes among adults aged 50 years and older: National Health and Nutrition Examination Survey 2001–2006. *Diabetes Care*, Volume 36, Issue 9, pp. 2655–62.

DAGs and related diagrams help to design studies and to analyse data. By pointing to data which are needed, they can make research more efficient. Using DAGs is also changing basic concepts in epidemiology, an example being the realization that adjusting for a collider can cause bias, that is, a spurious association. If a variable is both a collider and a confounder then adjustment for it removes confounding, but can induce bias. Obviously, stratifying a sample into separate groups is a selection bias, but so is selecting a sample (except perfectly randomly), or non-response, missing data, subgroup analysis, or adjusting for a variable in regression analysis. So, selection bias (and hence colliders) are unavoidable in the practice of epidemiology.

The strength of the DAG approach is that it openly presents the assumptions guiding the analysis and interpretation of the data. The DAG, unlike the statistical analysis, is a causal model, albeit a postulated one.

DAGs can be very complicated so variables can be grouped, for example, as demographic variables, or as behaviour-related ones to help simplify the diagram.

The output is not a causal truth, but an opportunity for the investigators to assess whether the data fit the model. If there is a close match with most of the variances explained, then it would be reasonable to infer that the DAG is reflecting a causal structure, at least as portrayed in these data. The ultimate goal is to strip out bias and error, so what is left must be causal. This process is called identification.

The field of causal diagrams is enriching epidemiology and we are seeing these diagrams increasingly, though mostly in specialist journals. They have not, however, 'mathematized causality' (a claim by Judea Pearl) in epidemiology, which is still based on judgement as discussed in section 5.6. DAGs have the potential to contribute substantially to the formation of such judgements.

An even more complex approach is the logic diagram, which tries to produce a model that shows the full complexity of reality including non-causal, non-linear relationships. One of the best known of these diagrams is that produced by the Foresight Report on obesity (see http://www.noo.org.uk/NOO_about_obesity/causes). (Government Office for Science 2007).

5.6 Guidelines (sometimes erroneously called criteria) for epidemiological reasoning on cause and effect

5.6.1 Comparison of epidemiological and other guidelines for causal reasoning

Turning epidemiological data into an understanding of cause and effect is challenging and perhaps the most difficult aspect of the subject. Unfortunately, there is a widespread tendency to reach easy, but often premature or wrong conclusions. The commonest problem is either to declare, or interpret, an association as causal when it quite possibly a result of confounding or reverse causality. This problem may be becoming more common, partly because of media involvement. It is not easy to present the nuances of data interpretation in a news sound bite or even press release.

Conclusions with strong caveats are not newsworthy, and, indeed, may not attract the attention of the top journals either.

To convince colleagues and the public, epidemiologists need an explicit mode of reasoning. Scientists, like all other human beings, rely on intuition in evaluating evidence and making judgements. Einstein intuitively understood the theory of relativity years before he published it and before there was empirical evidence to support his findings. The theorems of the mathematical genius Srinivasan Ramanujan were intuitive and many have yet to be resolved, although they are generally accepted as correct based on precedent. These are only two of many examples. The lesson for epidemiology is that subjective judgements on cause and effect should not be dismissed but tested empirically. Epidemiologists place much emphasis on the evaluation of empirical data, and have devised (and adopted from other disciplines) so-called criteria for causality. *Criteria* is an inappropriate word, as it encourages a checklist approach; *guidelines* is better.

The use of such guidelines for reaching causal judgements in epidemiology is controversial. They are not, and must not be used as, a checklist or algorithm for causality. There is no causality score. Rothman and Greenland (1998) provide a vigorous critique of the limitations of causal guidelines as stated so clearly by Bradford Hill (1965). A set of guidelines has been so closely linked to this exposition that they are commonly known as the Bradford Hill criteria, although he called them considerations. Similar principles had already been developed, indeed published, long before their inclusion in the 1964 United States Surgeon General's report on smoking and health, so their close association to one person is not appropriate. These kinds of guidelines are only sporadically used and it is evident that new scholarship and research is needed to update them. Some work has been published examining the meaning of such guidelines in the context of genetic epidemiology. The guideline-free approach—simply relying on measures of the association alone, has not been sufficient. The existing principles are valuable and I have distilled them here, together with some additional thoughts. Clearly, such guidelines should be seen as a framework for thought about the totality of evidence including from non-epidemiological studies.

Epidemiological causal reasoning comes under frequent attack, particularly from people and organizations that do not agree with particular research findings. It is a cliché, but one with barbs, that epidemiological results on cause and effect which are making headlines one week will be replaced by results reporting the opposite the next. One such result making headlines as I write is that dairy fats are good for cardiovascular health, overturning some 30–40 years of epidemiologically based views to the contrary. All sciences refine, re-adjust, and sometimes reverse their conclusions but unlike epidemiology, their debates do not usually make headline news internationally. Exemplar 5.2 considers this.

Exemplar 5.2: Headline: 'Toasties get you laid, fat prevents dementia and I'm a sex god' (*The Sunday Times*, 12 April 2015, p. 19)

Epidemiologically based health stories are common, often making headlines in newspapers, television, and social media. This article in *The Sunday Times*, one of the United Kingdom's important and serious newspapers, exemplified the mixture of amusement, bemusement, and humorous derision that journalists, including specialist health journalists, subject epidemiology to. The journalist Rod Liddle writes a funny article, but one with serious points to make.

A survey associating eating cheese toast (toasties) with a good sex life is not taken seriously by Liddle, but it does show that people like to collect, publish, and discuss these kinds of statistics. Seventy-three per cent (73%) of those eating toasted cheese sandwiches reported enjoying sex

at least once a month, but only 63 per cent of those preferring other snacks did so. This report came from a dating website. This lighthearted discourse is then followed by a serious one.

Liddle then moves on to a study published in a prestigious journal of two million people showing that being overweight was associated with lower risk of dementia. To quote: 'This is an awful thing for our state sponsored health fascists to contemplate.' The result of our efforts to curb obesity, he says, may well lead to a nation of agreeably thin people who have no idea of what they are doing. He then lambasts conflicting advice: 'One day eggs are bad, the next day they are good. One day Lurpak* is an agent of Satan, the next you are advised to spread it thickly and maybe put some bacon on it.' He says you end up trusting nothing as it is all disproved next week.

Concluding remarks

If there were not so much truth in this amusing and lighthearted derision, it would be funny. However, it rings so true that we epidemiologists need to take it seriously. The reporting of lighthearted statistics and serious epidemiology is unfortunate because the general reading public might not appreciate the difference.

(*a brand of butter).

The more serious criticisms are that epidemiologists' reasoning lacks a theoretical basis and it falls short of the more rigorous thinking in the experimental sciences. These criticisms are unhelpful and unjustified. Causal thinking in epidemiology draws upon the theories and principles of other disciplines including philosophy, the laboratory sciences, and the social sciences and is theoretically grounded, though this may not be obvious. Epidemiology is predominantly an observational and not experimental science, as are demography, astronomy, geology, evolutionary biology, palaeontology, and archaeology. Epidemiology is far more complex than most sciences, and experimentation in epidemiology is strictly limited by ethical constraints on human research. Epidemiology has, moreover, contributed new ways of thinking about causality when experiment is not possible. Epidemiological guidelines are, furthermore, designed for thinking about the causes of disease in populations and not in individuals. When applied to the individual, as in the courtroom, they are wanting but that is a criticism of those who misapply them rather than of the discipline.

Table 5.6 summarizes some of the cause and effect thinking in microbiology, health economics, philosophy, and epidemiology. There are commonalities of reasoning. The approach to establishing causality in the experimental medical sciences is illustrated by the Henle–Koch postulates, as discussed in detail by Susser (1977) (Table 5.6, column 1). These postulates also have limitations. First, consider the postulate the organism must be present in every case. This is impossible to show for many bacterial diseases including tuberculosis. (In clinical practice, a trial of antituberculosis therapy is sometimes required when the patient has a clinical picture of tuberculosis but the organism cannot be grown in the laboratory.) Second, the organism must be grown in pure culture. Viral organisms are particularly hard to grow, and so are some bacteria such as the mycobacterium causing leprosy. Third, when inoculated into a susceptible animal (or human) the specific disease should occur. Animal models are sometimes not available, and even when they are the induced disease may be different from the human version. Fourth, the organism must be recovered from the animal (or human), but this is often not achieved.

The Henle–Koch postulates are a counsel of perfection and too stringent. Evans (1978) points out that even when they were developed, it was recognized that they were not to be applied rigidly, and that Koch believed that the cholera bacillus caused cholera even though the postulates were not achieved. According to Evans, leprosy, typhoid fever, syphilis, malaria, mycoplasma pneumonia, and *Chlamydia trachomatis* infection are among the microbial diseases which have causes

Table 5.6 A comparison of four modes of thinking about causality

Microbiology: Henle–Koch's postulates	Philosophy: Mill's canons	Economics	Epidemiology: some related modes of reasoning for causality[1]
The microorganism causing the disease can be demonstrated in every case of the disease	Method of concomitant variation: the phenomenon which varies when another phenomenon varies in a specific way is either a cause, an effect, or connected through some fact of causation	The future cannot predict the present	The cause precedes the effect (temporality)
The organism can be isolated and grown in pure culture	Method of agreement: if there is only one circumstance in common in instances of the phenomenon, then the common circumstance is the cause of effect	The effect (y) can be predicted more accurately by using values of the cause (x) than by not using them	The disease is commoner in those exposed to the cause (strength)
Animals (or humans) exposed to the cultured organism develop the disease	Method of difference: if there is only one difference in the circumstances when a phenomenon occurs compared with when it does not occur, that difference is part of the cause or effect	Instantaneous causation does not exist, since there is a time difference between independent actions. If A, itself, causes B, and A did not exist, B would not have occurred	The amount of exposure relates to the amount of disease (dose–response)
The organism can be grown from the experimentally exposed animal (or human)	The method of residues: remove from the phenomenon any part known to be the effect of known antecedents (causes), and the remainder is the effect of the remaining antecedents	One cause can have many effects and one effect many causes The putative cause A may have an effect by itself or be a part of the cause	The causes are linked to diseases in specific and relevant ways (specificity) Altering the amount of exposure to the cause leads to change in the disease pattern (experiment or natural experiment) Different types of studies reach similar conclusions (consistency)

[1]The guidelines for causality have been reduced to six by the author for simplicity. Biological plausibility is discussed in the text and is, strictly, not an epidemiological concept.

Source: data from Susser M. *Causal thinking in the health sciences*, Second Edition. New York: Oxford University Press, Copyright © 1977, pp. 70–71; Charemza WW and Deadman DF. *New directions in econometric practice: general to specific modelling, cointegration, and vector autoregression* Second Edition, Cheltenham: Elgar, Copyright © Elgar; Hicks J. *Causality in economics*. Blackwell, Oxford, Copyright © 1979 John Wiley and Sons.

that do not meet the criteria. Furthermore, with new technologies such as antibody tests and DNA sequencing available, the postulates are being superseded. Epidemiologists need to be aware of such criteria, both as a standard to incorporate into their own work, and so they can discuss causality in the context of infectious disease epidemiology.

Some philosophers' ideas were considered in section 5.1. John Stuart Mill (1806–1873) was a British philosopher and economist who succinctly offered a practical interpretation of causal thinking in philosophy, the nub of which is now known as Mill's canons (Table 5.6, column 2). Susser (1977) has discussed these in the epidemiological context. The principles are of importance to epidemiology and are essentially incorporated into its own guidelines. The method of concomitant variation corresponds to current ideas on correlation and association (see section 8.15); the method of agreement to the search for a factor in common (e.g. in an outbreak of Legionnaires' disease, all those sick may have been to a particular air-conditioned hotel); the method of difference is at the core of epidemiological thinking (e.g. why do some people get heart disease and others of the same age and sex do not?); and the method of residues echoes modern ideas of experiments of preventive action, to establish what proportion of disease can be prevented, or where this is not possible, calculations of attributable risk (see Chapter 8). (Readers should note that the order in which the canons are presented in Table 5.3 does not correspond to Mills's numbering of his canons, e.g. the method of concomitant variation is the fifth in his list.)

Economics also evaluates associations in similar ways (Table 5.6, column 3). Even more than epidemiology, health economics relies on observation and modelling, with the scope for experiment being extremely limited. According to Charemza and Deadman (1997), the operational meaning of causality in economics is more on the lines of 'to predict' than 'to produce' (an effect). A scan of the third and fourth columns shows the similarity in concept, if not detail, between economics and epidemiology.

The nub of epidemiological reasoning (Table 5.3, column 4) is that the cause:

- must precede the effect
- should raise the incidence of the disease in a population
- should have a greater effect in greater quantity
- be associated with specific and relevant effects
- should show consistent effects across a number of studies.

These epidemiological ideas are similar to Mill's canons and to thinking in health economics.

Evidence from experimentation, natural or by design, on humans or animals, may show that manipulating exposures changes the disease. Experimentation may also elucidate the mechanisms by which this happens. The cause–effect relationship should make biological sense. These latter ideas, now integral to epidemiology, are those of the other biological sciences. The epidemiological guidelines for causality are not an idiosyncratic epidemiological invention. Their validity, as a collective, needs to be assessed empirically.

In the modern era an amalgam of epidemiological and basic science guidelines are adopted as the standard for causal thinking, as shown in the example in Box 5.13 and in the ensuing examples. Before reading on, try the exercise in Box 5.13.

In Kaposi's sarcoma (Box 5.13), the first and second items of evidence match the ideas underpinning the Henle–Koch postulates. The third and fourth match epidemiological concepts (strength of association) and the data could be converted to a measure of strength, such as relative risk (see Chapter 8). The fifth item is a mixture of microbiology (distribution in tissues) and epidemiology (transmission). The sixth item is, again, epidemiology, as is the seventh (temporality).

The principle is this: causation is established by judgement based on evidence from all disciplines. Failure to meet some guidelines (with the exception that the cause must precede the effect, which is not easy to establish conclusively) does not dismiss causality and achievement

Box 5.13 An exercise on causal thinking in medical science and causal guidelines

Can you see the links between the evidence listed below and the causal guidelines in Table 5.3? Which of these pieces of evidence match the guidelines for causality?

Aetiology of Kaposi's sarcoma: Evidence cited for a herpesvirus as the cause;

1. Viral sequences (DNA) can be detected in sarcoma tissues in most cases.
2. Such sequences are rarely detected in other tissues.
3. Virus is detected in blood cells in 50 per cent of cases but not in controls.
4. HIV positive patients who had the virus in blood cells had a greater risk of developing sarcoma than comparable patients without the virus.
5. The virus is probably sexually transmitted and is found in semen and other genital tissues of healthy adults.
6. Antibody levels in blood correlate with presence of sarcoma.
7. Antibody levels rise before Kaposi's sarcoma appears.

Conclusion: Kaposi's sarcoma is caused by a herpesvirus (Beiser 1997, p. 581).

Source: data from Berisera C. Recent advances: HIV infection-II. *British Medical Journa*l, Volume 314, Issue 7080, pp. 579–582, Copyright © 1997 BMJ Publishing Group Ltd.

of some guidelines does not ensure it. The currently popular 'hierarchy of evidence' places the systematic review and the human experiment (trial) at the apex of the evidence pyramid. This is understandable for studying the effectiveness of interventions on humans. In doing this in causal epidemiology, one side effect is to dismiss too readily other kinds of evidence. This matters for causal thinking, especially where the question is difficult to resolve using epidemiology alone (e.g. alcohol and heart disease). Just as we would expect laboratory scientists to examine the evidence from human population studies in reaching their conclusions, it is important that population scientists consider the findings of laboratory work.

Epidemiology often establishes cause in populations unequivocally, but this information only applies to individuals in a probabilistic way, which does not prove cause and effect at the individual level. Try the exercise in Box 5.14 before reading on.

The answer is that we do not know the cause at the individual level. If the person is a non-smoker, the cancer may have arisen from passive exposure to tobacco but perhaps is more likely to be due to other factors. If the person is a smoker, the cause is most likely smoking, but may result from other factors such as exposure to radiation or asbestos. There is no way, at present, to distinguish a lung cancer (or heart attack) resulting from smoking from a lung cancer (or heart attack) arising from another cause.

Box 5.14 Causes of lung cancer in populations and in individuals

If 90 per cent of all lung cancer in a population is due to smoking, and assuming that is correct, what is the likelihood that in an individual with lung cancer the cause was smoking?

A drug or public health intervention may be effective in a population but harmful to an individual. For example, exercise may be good generally but lead to collapse and death in some individuals. Some people are harmed by alcohol while others benefit, and the net effect on the health of the population as a whole is unclear. In contrast to alcohol, the net health effect of tobacco consumption is overwhelmingly negative.

These kinds of counterintuitive observations on causality lie at the heart of disputes between population scientists and those whose work is based on individuals. Immunization may or may not harm individuals as is sometimes claimed (e.g. MMR and autism, and whooping cough vaccine and neurological disorders), but the benefits far outweigh these harms at population level epidemiological studies show. To prove harm in these particular circumstances, that is, to the individual, is beyond the scope of epidemiology and requires other kinds of sciences. Equally, to prove there is no harm to individuals is also impossible for epidemiology. We can, however, show there is no sizeable harm to the population. The problem is that sciences based on individuals are also unable to predict well at the individual level. As a result, epidemiology is used in a setting and context for which it is not designed, in the absence of better alternatives. A great deal of the criticism of epidemiology arises from failing to separate what it can and cannot do.

Epidemiological data are, therefore, difficult (possibly impossible) to apply in legal cases about individuals. To quote Evans discussing the issue in the United States of America:

> Legal requirements are concerned with the risk in the *individual*, the plaintiff, and whether the preponderance of evidence supports the conclusion that *that* exposure 'more likely than not' resulted in *that* illness or injury in *that* person.
>
> (1978, p. 194)

Evans contests that a higher order of proof and specificity is required in legal proof than in epidemiological proof, concluding that epidemiological evidence is often inapplicable in this context. Epidemiology is a science based on studies of groups and cannot be directly applicable to individuals, and this is an inherent limitation. Equally, a factor demonstrated to cause a disease in an individual, by a science of individuals, say toxicology or pathology, may not be demonstrable as harmful in the population, possibly because harmful effects are balanced by beneficial ones. This is an inherent limitation of a science of individuals. The problem lies not with epidemiology itself, but with those who apply epidemiology in these circumstances. The law also extrapolates from population data to the individual. The standard of proof in epidemiology is not of a lower order than in law, but it is of a different order and for a different purpose. The problem is that so often the best we can offer the individual is average risk derived from the study of groups similar to that individual. That is a limitation of medical sciences collectively. We now consider how epidemiological guidelines for causality help to analyse the causal basis of associations observed at the population level.

5.6.2 Application of guidelines to associations

The association (or link or relationship) between disease and postulated causal factors lies at the core of epidemiological thinking. Mostly, such associations are found by observing that disease varies with time, place, or characteristics of persons in observational data. An association rarely reflects a causal relationship, but it may. The preceding chapters on variation and error showed how to separate the probably not causal association from the possibly causal one. Having accounted for chance, error and bias, and confounding, logically, what remains is a causal relationship. The problem is that accounting for these factors is not a foolproof process. That is why we need to go to the next step of synthesizing the evidence to strengthen our conclusions. Once we learn how to conduct perfect studies and data analysis, this step will become unnecessary. For the foreseeable

future, however, it is essential. We have a further problem, however, and that is that there is no foolproof, or even consensual, process of synthesizing the evidence. The approach described here is a distillation of now-standard concepts developed over the last 50–70 years. These and other approaches need more work.

Table 5.7 begins the questioning and reasoning process often used in epidemiology to make the difficult judgement on whether an association may be causal. These six guidelines are a distillation of, and echo, the 10 Alfred Evans postulates in *A Dictionary of Epidemiology*, the reasoning in the United States Surgeon General's report on smoking and health, and the nine Bradford Hill considerations. The causal challenge is illustrated by the pyramid of associations and causes in Figure 5.12. The pyramid displays the axiom 'association is not causation' with my qualification, ' . . . but rarely it may be'. Try the exercise in Box 5.15 before reading on.

We shall now look at these guidelines in more detail. A visual summary is in Figure 5.13.

i. Temporality

Did the supposed cause precede the effect? In Figure 5.13, the first panel shows the relative risk (RR) of disease rising over time following exposure. In this image, a lag period is shown with the risk increasing over time and then stabilizing. We would not expect the effect to be immediate. Some causes (e.g. carcinogens) take many decades to have an effect. If the effect is simultaneous with or precedes the proposed cause, the association is definitely not causal in the direction postulated. There may be reverse causality, that is, the outcome actually affects the supposed cause; for example, early undetected cancer leads to low weight, not low weight leads to cancer. If there is no clear answer the judgement will be tentative, irrespective of other data, no matter how convincing these are. If the effect follows the action of a proposed cause, the association may be a causal one and the analysis can proceed. This matter of timing is referred to as temporality. Demonstrating that this guideline is satisfied does not usually establish causality. Why might this be so? Reflect on this question before reading on.

The supposed cause under study may not be the only exposure that changed. There may have been many changes, and the temporal relation under study may be coincidental. If a perfect experiment (trial) only changed one exposure and the outcome is altered, this evidence alone would, usually, be accepted as causal. Before reading on, do the exercise in Box 5.16.

Thunder follows lightning, at least as perceived by humans, but is not caused by it. Both are generated simultaneously by an electrical discharge in clouds. The later arrival of the thunder is simply a result of the slower speed of sound than of light. Without an understanding of the nature of thunder and lightning, erroneous conclusions about cause and effect are likely. Empirical observation seduces us to err. Generating alternative explanations is an essential discipline in epidemiology. The epidemiological imagination needs to be cultivated for this. Our alternative explanations can be put to the test. It would be hard to test the lightning and thunder association. Earlier, when discussing Hume (section 5.1.1) we considered the association between flicking a switch and a light going on. If the act of flicking similar switches in other settings turns on a light, we are likely to accept a cause and effect relation on empirical grounds. The empirical observation has no explanatory power for exceptions, for example, when the light does not go on because of a break in the wiring, or when it goes on even without the switch being flicked, by water penetration. When there is a deeper understanding of the nature and action of electrical circuits, the association may be agreed as causal, especially if it explains exceptions. Just because B follows A does not, of itself, prove a causal relation. Deeper understanding, opening and understanding the black box, is essential.

Questions underlying the guidelines for causality and implications of answers for interpretation of associations in populations

derlying guideline	Label for guideline	Evidence		
		Unsure	**No**	**Yes**
pposed cause precede or other effect)?	Temporality	Judgement premature	Not causal	Causal relation possible
re to the supposed cause dence of disease?	Strength of association	Judgement premature	Not causal in the population context but does not rule out causal effects in individuals	Causal relation in populations possible
exposure to the use lead to varying disease?	Dose–response	Not critical	Causal relation still possible if there is a threshold effect	Strengthens case for a causal judgement
tion between supposed sease(s) limited in range	Specificity	Not critical	Not critical but extra caution—a sign of artefact	Strengthens causal claim
tion between supposed utcome consistent across dies and across populations ps?	Consistency	Defer decision, and await further research unless an immediate judgement is essential	Judgement will require explanation for inconsistent results	Strengthens causal claim
ulating the level of exposure sed cause change disease	Experimental confirmation	Not always possible, so not critical	Caution needed for a causal claim	Strong confirmation of a causal relation
at the supposed cause ect on disease understood?	Biological plausibility	Not critical	Not critical but great caution needed for causal claim	Causal judgement strengthened

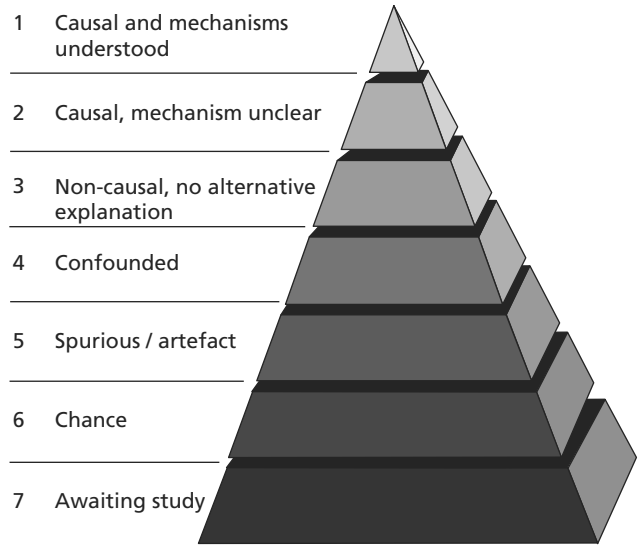


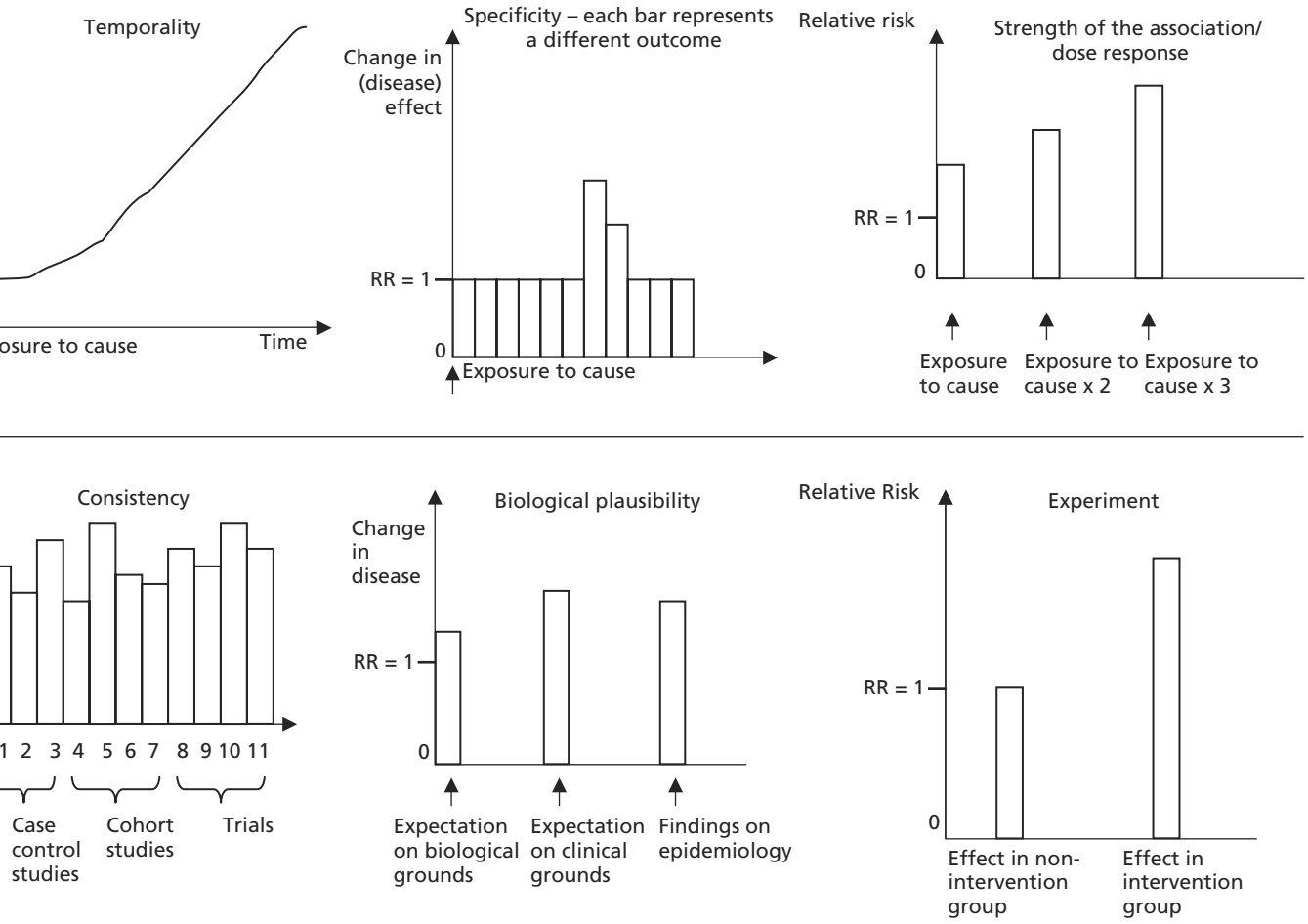
Fig. 5.12 Pyramid of associations and causes.

Box 5.15 Linking associations and causality in risk factors/outcome relationships

Can you think of an exposure/risk factor variable with an association that fits with each category in the pyramid? Once you have attempted this, have a look at Table 5.8.

Table 5.8 Examples of associations fitting each level of the pyramid of associations

Level	Example
1. Causal, understood	Common in infectious and toxic disease but still quite rare in chronic diseases with examples being tobacco, asbestos, radiation, and lung cancer
2. Causal, unclear	Posture when sleeping and sudden infant death syndrome
3. Non-causal, no alternative explanation	Suicide and smoking
4. Confounded	Alcohol and lung cancer
5 Artefact	Numerous
6. Chance	Numerous
7. Awaiting study	By definition, no examples but most associations are yet to be discovered



elines for causality: graphic representation.

Box 5.16 The deduction of cause and effect from the linkage of events

Reflect on whether the linkage of two events provides convincing evidence on cause and effect. For example, thunder follows lightning. Does lightning cause thunder?

If you observe this once or a thousand times, does it make a difference? What other explanations might there be?

ii. Strength and dose–response

Does exposure to the cause change disease incidence, and if so, what is the shape of this relationship? If not, or we are unsure, there is no epidemiological basis for a conclusion on cause and effect. The failure to demonstrate this does not, however, disprove a causal role. Reflect on Box 5.17 before reading on.

The cause or its effect may be so rare that there are insufficient cases available to reach a conclusion. Alternatively, or additionally, the effect may be very small. Epidemiology is not good at demonstrating causal links when the rise in disease incidence is low, for example, less than 10 per cent excess. Alternatively, there may be some people in the population in whom the cause is operative while in others it has an opposite effect, leading to the view that there is no association. It might be reasonable to say that the cause studied was operative in causing disease in individuals but not at the population level. The cause may only be operative in the presence of a cofactor, that is, as part of a package of sufficient causes. The cofactor may be absent in the time, place, or population you studied. The same study somewhere else may have reached a different conclusion. The cause may be operative on everyone. If oxygen is the cause of, say, pancreatic cancer we cannot show this epidemiologically (or possibly in any way).

The most usual way of assessing strength and dose–response is the relative risk (see Chapter 8). Other ways of measurement include the correlation coefficient, the regression coefficient, absolute rates, odds ratios, and other measures of associations and effects. (Some of these and other measures are considered in Chapters 7 and 8.) The greater the relative risk, the greater the strength of the association.

Does the disease incidence vary with the level of exposure? If yes, this is a dose dependent effect and the case for a causal relation is advanced, but if not, the effects may be independent of the amount of exposure. Allergy is one example where trivial doses of substances such as peanuts can cause life-threatening reactions. For most exposures, the relationship with disease is not linear, but the principle that more exposure leads to more disease tends to hold. For high blood pressure there is a threshold above and below which adverse effects arise. Above the threshold, the dose–response concept applies. Below the threshold, the effects are unclear but some minimal blood pressure is needed for life. For weight and alcohol consumption, there is an apparent adverse effect at both low values and high values (called a J-shaped distribution). Dose–response can be considered as a development of the concept of the strength of the association; that is, does

Box 5.17 Epidemiology fails to uncover a cause

Can you think of circumstances when exposure to a causal factor does not change, or cannot be demonstrated to change disease incidence, yet the exposure is a cause?

Box 5.18 Strength of association in a perfect, counterfactual study

Imagine that we had the perfect study. In an imaginary world, we can even have the counterfactual data, that is, what the risk would be in the same population if that population did not have the risk factor. Would these data alone—that tell us the strength of the association—suffice for a cause and effect judgement? What about in the real world?

the strength of the association vary with the level of exposure? The second panel in Figure 5.13 shows the risk of the disease increasing in people exposed to the cause (the unexposed group has an RR of 1 by definition), with the risk increasing with greater exposure. It is generally accepted that the size of the relative risk matters in causal interpretation. In a well-designed study, if the relative risk is 5 or more we veer towards causal explanations, whereas if it is 1.5 we worry more about non-causal effects. There is no single metric to describe the strength of the association but by custom and practice, the following is a rough guide: weak associations imply RR <2; moderate associations imply RRs of 2 to 3.9; strong associations imply RRs of 4 or more. Before reading on, try the exercise in Box 5.18.

Assuming temporality was established, from a population perspective this evidence from the perfect study in Box 5.18 would be causal. The closest we come to this, in reality, is the well-designed experiment, in epidemiology the trial (Chapter 9). If a trial were free of all errors and biases, the strength of the association would, together with temporality, reflect causality. As yet, all epidemiological trials are imperfect and on selected populations. If trials were perfect, we would need only one on any topic and we would not need any systematic reviews or meta-analyses. In our imperfect but real epidemiological world, the other guidelines are, therefore, of importance.

iii. Specificity

Is the association of the supposed cause limited to relevant diseases and are diseases associated with a limited number of supposed causes? If so, we would say the association is specific. This idea is called specificity. Imagine a factor which was linked to all health effects. Why would that be? It is hard to explain (except through some error). Unless the links to a broad range of diseases can be explained, the case for causality is weakened. Non-specificity is characteristic of spurious associations (e.g. underestimating the size of the population denominator; Chapter 7). Some factors do have broad effects, for example, poverty and cigarette smoking. However, even these are not associated with more of every health problem. In the United Kingdom, poverty is associated with less malignant melanoma, a skin cancer that includes within its sufficient causes, sunburn and excess exposure to sunlight. This observation makes sense because these causes are often a result of holidays in hot climates. Those in poverty are less able to afford to expose themselves to these causes than richer people. Panel 3 in Figure 5.13 shows that following exposure to a possible cause, the RR was increased for two adjacent disease outcomes of 11 studied. This is indicating reasonable specificity. While specificity is not a critically important guideline, epidemiologists should take advantage of the reasoning power it offers.

iv. Consistency

Is the evidence within and between studies consistent? It is wise to be tentative if it is not. Unless the inconsistency can be explained, the case for causality is weakened. Consistency is linked to generalizability of findings. Experience tells us that causal effects tend to be widely applicable,

while spurious associations are often local. The systematic review and meta-analysis are ways of assessing consistency in a rigorous way and are considered briefly in Chapter 9.

Panel 4 in Figure 5.13 shows 11 studies using three kinds of design. While the relative risk varies, in every case it is above the reference value of 1. This is compelling evidence. It is possible, however, that all 11 studies are making the same kind of error, although given a few are trials it is hard to imagine a non-causal explanation. Trials are experiments, the next guideline to be considered.

v. Experiment

Does changing the exposure to the supposed cause change disease incidence? If yes, this is experimental confirmation. Sometimes there have been natural experiments, with changes over time in exposure to risk factors. For example, a spill of a pollutant into a water supply, the closure of a factory, the availability of a new product, redundancy in a factory, economic collapse of a society, or a change of policy (e.g. putting fluoride into a water supply). Perhaps the greatest of these experiments is the mixing of the genes in Mendelian randomization (see Chapter 3 section 3.3.1 and Chapter 9). These natural experiments can be vitally important. Often there is no such evidence, and some form of deliberate experimentation will be necessary.

The problem is that human experiments or trials are sometimes impossible on ethical grounds and are always difficult and expensive to organize. Ethically, the individual involved must have the potential to benefit and yet there must be uncertainty on the question posed. For risk factors, as opposed to protective factors, there may be no such benefit. Then the experimental approach requires a valid *in vitro* or animal model. Causal understanding can be greatly advanced by laboratory and experimental observations. Such data must be integrated with epidemiological observations, to ensure that the theoretically predicted effects do occur in free-living human populations. Experimental methods are introduced in section 9.7 on trials. The ethics of epidemiology are particularly important to these studies (section 10.10). Panel 5 in Figure 5.13 shows an experiment with an intervention. The non-intervention group by definition is assigned a relative risk of one. We see the risk is increased in the intervention group. There is an effect, but it is an adverse one. If the trial were perfect, this would be causal evidence at least in the type of population studied.

vi. Biological plausibility

Is there a biological mechanism by which the supposed cause can induce the effect? This is the guideline of biological plausibility. If there is plausibility, the case for a causal effect will be easier to advance. For truly novel advances, however, the biological plausibility may not be apparent. For example, it is biologically plausible that laying an infant on its back to sleep may lead to it inhaling its own vomit. This biologically plausible theory, which informed parenting behaviour for decades, has been overturned by the biologically implausible observation that laying a child on its back halves the risk of cot death compared to the side or front. The mechanisms are still being worked out. That said, biological plausibility remains important, particularly in confirming causality. The analogy is with the light switch; when there is understanding of the electrical circuit, the causal basis of flicking the light switch is confirmed. An understanding of electrical discharges in clouds explains the association between thunder and lightning (see i. Temporality). Ultimately, biological processes govern all diseases and adverse health outcomes, without exception. This applies to social and physiological processes alike—so the ill effects of economic deprivation on health must, ultimately, occur through biology. Understanding these processes is important. Clinical and other scientists are not easily persuaded by epidemiological evidence that does not fit into biological understanding.

Demonstrating biological plausibility is not part of epidemiological methods. This does not, however, mean epidemiologists can forget about it. Epidemiologists need to understand the biology of the diseases they study, explain their hypotheses in biological terms, and propose and promote (sometimes even lead) biological research to test hypotheses. The precedent and inspiration for such work is abundant, as we saw in Chapters 1 and 2 with syndrome X (pellagra) and the work of Joseph Goldberger (and his colleague Edwin Syderstricker).

Panel 6 in Figure 5.13 shows a situation where on biological and clinical grounds we might expect an increased risk of disease. If epidemiology demonstrates this prior expectation, that is powerful. It is less powerful, but still useful, when in retrospect, biological and clinical ideas are invoked to explain an epidemiological finding. An example of the latter is my four-stage model synthesizing the published evidence to try and explain a well-established epidemiological observation that South Asians have about four times the risk of type 2 diabetes mellitus compared to White Europeans in the same context (Bhopal, 2013). The model can be tested by further epidemiological studies.

5.6.3 Judging the causal basis of the association

The investigator can now proceed to a conclusion, but the interpretation ought to be tentative as judgements on cause and effect are not necessarily universal. An association which meets many or even all of the causal guidelines may, at least theoretically, be non-causal. George Davey Smith, Andrew N. Phillips, and James D. Neaton (1992) have shown, for example, that the association between cigarette smoking and suicide meets many (but not all) of the guidelines for causality including temporality, strength, and dose–response. Yet, they argue, the association is not causal.

The guidelines are particularly valuable in exposing the lack of, or contradictory nature of, evidence for causality, for indicating the need for further research and for avoiding premature conclusions. This said, sometimes firm judgements are possible, and at other times are forced upon us, even in the face of limited evidence. A judgement may be essential when policy is to be made. Using a causal framework makes the judgement explicit. Table 5.9 indicates how the questions implicit in causal guidelines can be applied to weigh up evidence.

Three examples of the case for causality (illustrating the need for a systematic mode of analysis) are shown in Table 5.9: diethylstilboestrol as a cause of adenocarcinoma of the vagina (Herbst *et al*. 1971); smoking as a cause of lung cancer (Doll and Bradford Hill 1956); and residential proximity to a coking works as a cause of ill health (Bhopal *et al*. 1994). Before reading on, reflect on the exercise in Box 5.19. Readers are invited to read the original studies (listed in References).

At the time that the key studies referred to in Table 5.6 were published, the authors claimed that the smoking–lung cancer association was causal (true, but many remained unconvinced), that diethylstilboestrol had caused adenocarcinoma of the vagina (this was accepted), and that residential proximity to a coking works had caused respiratory morbidity, but not mortality. The latter case was not, however, accepted as solid, though it was the best that was achievable.

5.6.4 Interpretation of data, paradigms, study design, and causal guidelines

Causal knowledge is born in the investigator's imagination and understanding. Scientific data do not, in themselves, offer knowledge. Indeed, the same data can be interpreted in quite different ways, depending on the investigator's way of thought. For example, data that one scientist, Samuel Morton, interpreted as showing clear differences by race in cranial capacity and hence brain size and ultimately intelligence, was interpreted by another, Stephen Gould (1984), as showing no noteworthy differences. This arose because of differences in the way they saw the world—including the research world. This way of seeing the world is often referred to as the paradigm.

examples of applying the guidelines for causality

	Smoking and lung cancer	**Diethylstilboestrol and adenocarcinoma of the vagina**	**Living near a coking works and ill health**
d cause precede t) (temporality)	Yes, clearly so	Yes, maternal exposure to diethylstilboestrol preceded the disease in the offspring	Yes, the coking works was functioning before most people in the study were born
es exposure to the cidence of disease?	Greatly and as much as 20 to 30-fold in smokers of 20 or more cigarettes per day	Greatly, as estimated from the first case–control study	The excess of disease is modest, varying for each specific cause but is rarely more than 30–50% greater than expected
osure lead to varying sponse)	Yes, there is a clear relationship and more smoking causes more disease	No clear evidence from the study	The evidence is suggestive that the closer the residence to the coking works, the greater the effect on health
ad to a rise in a ases? (specificity)	No, numerous diseases show an association with smoking	Only one outcome was studied	Yes, the association is restricted mainly to some respiratory diseases
consistent across nd between	Yes, the association is demonstrable in men and women, and across social groups internationally	Decisions had to be taken on the one study	There are no directly comparable studies, but it fits with understanding of the role of industrial air pollution
e cause exerts its understood? ility)	Only partly. The tar in cigarettes contains important carcinogens	At the time of the discovery, no	Generally, yes, specifically no. Coking works produce complex mixtures of emissions. Most knowledge is on single components of air pollution, not mixtures
g the level of ause change e? firmation)	Yes. Reducing consumption of cigarettes reduces risk. Persuading people to smoke more would be unethical. Tobacco is carcinogenic to animals	At the time this was unknown	Not known. An experiment is not possible, but the plant closed during the research, producing a natural experiment. Closure of the coking plant was not linked to changes in consultation with a general practitioner, but on days when pollution levels were high, the consultation rates were also high

Box 5.19 Reaching a judgement on cause and effect

Reflect on the evidence in Table 5.9 and deliver a verdict on whether the associations between smoking and lung cancer, diethylstilboestrol and adenocarcinoma of the vagina, and living close to a coking works and ill health are causal.

The paradigm within which epidemiologists work will determine the nature of the causal links they see and emphasize. There is a strong case for researchers to make their guiding research philosophy and paradigm explicit (see also section 10.5).

Causal thinking and study design (Chapter 9) are distinct, though interlinked, issues. No epidemiological design, in practice, confirms causality and no design is incapable of adding important evidence. In all studies, there are limitations and pitfalls. There are differences among the various study designs in both the type of pitfalls and their likelihood (see Chapter 9). While a single observation may spark off causal understanding, it would be wise to exercise great caution until further observations confirm or refute the idea. Exceptionally, however, there may be no time to delay.

Table 5.10 indicates the potential contributions of various study designs to the epidemiological guidelines for causality. Note that with the exception of consistency, to which all designs contribute, and biological plausibility, to which no epidemiological designs contribute directly, all epidemiological studies contribute to some but not all guidelines. This must not be confused with the hierarchy of evidence that has emerged in relation to the effectiveness and cost-effectiveness of interventions, or even of measuring the burden of disease. Each purpose requires its own hierarchy. We have already discussed that the experiment is the pre-eminent causal method, but it can only be applied rarely. Otherwise, there is no hierarchy of causality.

Table 5.10 Potential contributions of study design to causal guidelines

Guideline	Case series	Cross-sectional	Case-control	Cohort	Trial
Temporality	Sometimes	Sometimes	Sometimes	Often	Usually
Strength or dose–response	Sometimes	Sometimes	Often	Always	Always
Specificity	Sometimes for exposure	Sometimes	Sometimes for exposure	Sometimes for outcome	Sometimes for outcome
Consistency	Yes	Yes	Yes	Yes	Yes
Experimental confirmation	Sometimes, in the case of natural experiment	Sometimes, in case of repeated studies, following an intervention	Seldom (but this design is not advised for assessing interventions)	Sometimes, following natural changes	Always
Biological Plausibility	Not directly	Not directly	Not directly	Not directly	Not directly

5.7 Epidemiological theory illustrated by this chapter

Several theories underpin epidemiological causal thinking. First, there is the theory that diseases arise from a complex interaction of genetic, social, and environmental factors. This is not the kind of theory that is common in other medical sciences, but it is also fundamental in social sciences applied to health. Second, there is a theory that causes of disease in populations may not necessarily be demonstrable as causes of disease in individuals and vice versa. This theory is very rarely discussed. It is not easy for a science that is applied so much at the individual level to expose such limitations. The third (and pragmatic) epidemiological theory of causation is that reliable cause and effect judgements are achievable through hypothesis generation and testing, with data interpreted using a logical framework of analysis, which draws on multidisciplinary perspectives. This is a theoretical perspective that differs from purely quantitative disciplines. At present, at least, epidemiology does not work with a theory, derived from a mix of mathematics, statistics and computational science, that causality can be inferred from the mathematics, as articulated by Judea Pearl. The ideas underpinning this mathematical theory of causality, however, are being incorporated into epidemiology quite rapidly.

5.8 Conclusion

The most important aim of epidemiology is to generate cause and effect theories, to break the links between disease and its causes, and to improve health. The misapplication of such theory may have serious repercussions including deaths on a mass scale, while its proper application can transform the control of disease.

It is difficult to achieve trustworthy causal knowledge because of the complexity of diseases, the long timescales over which many human diseases develop, and ethical restraints on human experimentation. Nonetheless, there is an imperative to act, even when our knowledge is incomplete, for lives depend on our science. In the words of Bradford Hill (1965):

> All scientific work is incomplete—whether it be observational or experimental. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time.
>
> Bradford Hill (1965, p. 300)

A rigorous analysis of all the scientific data available is essential, though to quote Bradford Hill again, 'this does not imply crossing every "t", and swords with every critic, before we act'.

Epidemiology engages with health policy-makers and planners who are the users of much of the work. Rothman and colleagues (1998) have helped to open up the prickly question, posed by Lanes, of whether epidemiologists (as scientists) ought to be engaged in choosing between theories of causation, or whether they should simply present the evidence and the theory options to policy-makers and leave the choices to them. Mostly, there are alternative interpretations of data, and it is our theoretical perspective that governs our preference. To take an example, small amounts of alcohol have been shown in many (but not all) kinds of study to be associated with cardiovascular health benefits. Reaching a causal judgement and drawing a policy conclusion is tricky. It may well be that in doing so, the epidemiologist adopts the role of advocate. To refuse to draw a conclusion, however, potentially leaves us paralysed for it is improbable that lay decision-makers would make policy in these circumstances. Generally, epidemiologists tend to draw policy conclusions, and often possibly do so too early. Readers need to ponder on this question and form their own views. The debate continues in the scientific journals. Whatever viewpoint prevails, epidemiology has a responsibility to understand the theories of causation used by other disciplines, and to educate others about the mode of thought in epidemiology.

Notions of epidemiology about causality, for example, that a cause is something which raises the incidence of disease, are not particularly helpful in persuading sceptics who can retort that association is not causation. Demonstrating causation to the sceptics' satisfaction is complex and requires detailed understanding by both parties of causal reasoning in epidemiology. Developing effective actions, a difficult challenge usually achieved in cross-disciplinary partnerships, is also demanding in epidemiological knowledge.

Epidemiology provides a broad perspective on the causes of disease, which contrasts with the narrower one of the physical and most biological sciences. This complementary perspective is a great strength. The causal models reinforce this perspective and provide a framework to organize ideas.

The prevailing attitude in epidemiology, that all judgements of cause and effect are tentative, is both pragmatic and in line with modern thinking about the nature of scientific advances. The increasing understanding that the data do not hold an unequivocal answer, and that the answers derived are dependent on human judgement (symbolized in Fig. 5.14) though apparently common sense, are harder to accept because they give space for subjectivity whereas science prefers objectivity.

Epidemiologists should be alert for the play of chance, error, and bias, reverse causality and confounding; should create causal diagrams and related analysis plans; and should apply guidelines for causality as an aid to thinking and not as a checklist. Only rarely will causal mechanisms be understood, as symbolized by the peak of the pyramid of associations in Figure 5.14. Finally, epidemiology should seek corroboration from other scientific disciplines in terms of both data and scientific frameworks for cause and effect.

Seeking causal associations is like panning for gold, for it usually yields nothing but grit and mud (error and bias). Often we find gold flakes and specks (risk factors, relating to the causal pathway in ways that remain obscure). Sometimes we get a nugget (causal factor). Rarely a gold mine is discovered (a causal theory). Like panning for gold, a great deal of hard work is required to find a gold mine. When it is discovered, panning alone will not be enough, and much sophisticated equipment and skills will be needed to take full advantage of the discovery. In epidemiology, this implies working with other laboratory and population-based scientific disciplines (including social sciences) to gain understanding of the mechanisms by which the cause operates.

In the second edition of this book, I said that epidemiology needs an international council to assess evidence to establish the causal credentials of the multiplicity of associations being generated. Associations would be categorized by this council according to the pyramid of associations and causes, to help guide both future research and public health action, and to plan and deliver

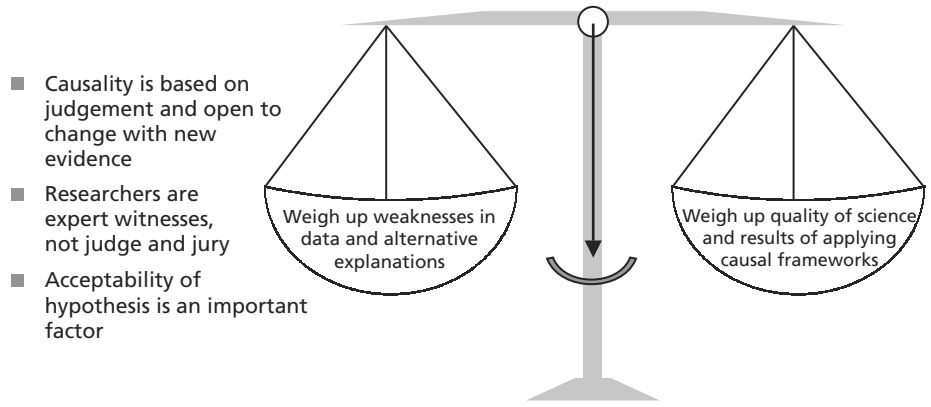


Fig. 5.14 Cause and effect: judgement.

the necessary health interventions. This proposal has been developed and discussed under the interim title of World Council for Epidemiology and Causality (Bhopal 2009) and further discussion is underway. Workshop-based and other discussions have pointed to the difficulties of the task. However, there are precedents. First, the World Council on Epidemiology and Causation (WCEC) would focus on associations in the same way that the Cochrane Collaboration and the UK organization National Institute for Clinical and Public Health Excellence (NICE) do for effectiveness of interventions. Second, this would be an extension of the monographs of the World Health Organization's (WHO) cancer agency, International Agency for Research on Cancer, dealing with the causal effects of carcinogens. However it is done, the task is essential in epidemiology. The vision now is for an independent body working with many partner organizations to spur on concepts, methods in causality, and structured, comprehensive analysis of specific associations, to reach speedier and more robust judgements on whether these might represent cause and effect.

Summary

Cause and effect understanding is the highest form of scientific knowledge, for it permits prediction and generalization, one of the main purposes of science. Understanding of cause and effect has also been a preoccupation of philosophy. A comparison of epidemiological and other forms of causal thinking shows similarity, reflecting the debt which epidemiologists owe to other, older disciplines. Epidemiology is increasingly influencing thought in other sciences.

An association between disease and the postulated causal factors lies at the core of the science of epidemiology. Causal knowledge can be greatly advanced by experimental observations on what happens to disease incidence when the causal factors are manipulated.

In epidemiology, demonstrating causality experimentally is difficult because of the long and complex natural history of many human diseases and because of ethical restraints. Epidemiologists should: hold the attitude that all judgements of cause and effect are tentative; understand that causal thinking demands a judgement; be alert for the play of chance, error, and bias; always consider reverse causality and confounding; utilize the power of causal models that broaden causal perspectives; apply guidelines for causality as an aid to thinking and not as a checklist; and look for corroboration of causality from other scientific frameworks for assessment of cause and effect.

The many guidelines for evaluating the causal basis of associations have here been distilled to six: temporality; strength, and dose–response; specificity; consistency; experimental confirmation; and biological plausibility. Causal models and tools introduced as aids to causal thinking include the line, triangle, wheel, and web of causation; the component cause model and causal diagrams including the directed acyclic graph.

The ultimate aim of epidemiology is to use cause and effect knowledge to break links between disease and its causes, and to improve health. The application of erroneous knowledge has serious repercussions.

Sample questions

Many of the questions at the end of Chapter 9 are highly relevant to this chapter, and you may need to read that chapter before doing some of the questions here.

Question 1 The phrase 'association is not causation' is often used in epidemiology. List five non-causal factors that can lead to association.

Answer Associations that are non-causal can be generated by chance, errors, bias in the selection of study populations, bias in the quality of information, failing to compare like-with-like (confounding), fraud, etc.

Question 2 Which of the guidelines (also erroneously known as criteria) for causal reasoning can clinical trials make a contribution to? Which do they not contribute to?

Answer Trials can contribute to the following causal guidelines:

- Temporality—intervention comes first, outcome is observed
- Strength of the association (dose–response may be possible if varying levels of the intervention are studied)
- Consistency—if they support other studies
- Specificity of the disease outcome, but not exposure (usually only one exposure is changed)
- Experimental confirmation—trials are experiments
- Trials do not contribute, at least directly, to biological plausibility unless specific effort is made to collect data on the biological mechanisms within the trial

Question 3 Which of the guidelines (also erroneously known as criteria) for causal reasoning can cohort studies make a contribution to? Which do they not contribute to?

Answer Cohort studies can contribute to temporality by showing that the exposure precedes disease; to strength of the association by measuring the relative risk; to dose–response by measuring the association as the exposure increases; to consistency either by comparison with other kinds of studies or with other cohort studies; and to specificity, that is, the range of outcomes that exposures lead to. They cannot contribute directly to biological plausibility, although they can test a biologically derived hypothesis. They cannot give experimental confirmation though (a) they can be used to study natural experiments and (b) to find people to do trials (experiments) on.

Question 4 What is the value of the concept of 'consistency' in helping you assess an association?

Answer 'Consistency' of an association is linked to generalizability of findings. The consistent association is similar in different populations and in studies using different kinds of study design. This helps to evaluate an association. For example, causal effects tend to be widely applicable, while spurious associations are often local. Unless the lack of consistency in findings can be explained, the case for an association is weakened.

Question 5 The phrase 'risk factor' is often used in epidemiology. Explain the meaning of this phrase, and discuss how it differs from causal factor.

Answer Risk in epidemiology usually refers to the likelihood (probability) of dying or developing a disease, or its precursors. In epidemiology, our prime interest is in the interaction between the probability of disease, or risk, and those environmental, individual, and social characteristics which influence the risk. Where there is an association with an increased probability of disease in those with such characteristics, the characteristics are called risk factors.

The phrase 'risk factor' does not necessarily imply the characteristic has a causal effect (association is not causation). No causal relationship is presumed, though there is great interest in assessing whether one exists. When a causal relationship is agreed between disease and the risk factor, the phrase causal factor, or simply cause, is used.

References

Bardenheier, B.H., Bullard, K.M., Caspersen, C.J., Cheng, Y.J., Gregg, E.W., and Geiss, L.S. (2013) A novel use of structural equation models to examine factors associated with prediabetes among adults aged 50 years and older: National Health and Nutrition Examination Survey 2001–2006. *Diabetes Care*, **36**, 2655–62.

Beiser, C. (1997) Recent advances: HIV infection–II. *British Medical Journal*, **314**, 579.

Bhopal, R.S., Phillimore, P., Moffatt, S., and Foy, C. (1994) Is living near a coking works harmful to health? A study of industrial air pollution. *Journal of Epidemiology and Community Health*, **48**, 237–47.

Bhopal, R. (2009) Seven mistakes and potential solutions in epidemiology, including a call for a World Council of Epidemiology and Causality. *Emerging Themes in Epidemiololgy*, **6**, 6.

Bhopal, R.S. (2013) A four-stage model explaining the higher risk of Type 2 diabetes mellitus in South Asians compared with European populations. *Diabetic Medicine*, **30**, 35–42.

Bradford Hill, A. (1965) The environment and disease: association or causation? *Occupational Medicine*, **58**, 295–300.

Chadwick, J. and Mann, W.N. (1950) *The Medical Works of Hippocrates*. Oxford, UK: Blackwell Scientific.

Charemza, W.W. and Deadman, D.F. (1997) *New Directions in Econometric Practice: General to Specific Modelling, Cointegration, and Vector Autoregression*, 2nd edn. Cheltenham, UK: Elgar.

Cottingham, J. (1996) *Western Philosophy—An Anthology*. Oxford, UK: Blackwell.

Davey Smith, G. Phillips, A.N. and Neaton J.D. (1992) Confounding in epidemiological studies: why 'independent' effects may not be all they seem. *British Medical Journal*, **305**, 757–9.

Doll, R. and Bradford Hill, A. (1956) Lung cancer and other causes of death in relation to smoking. *British Medical Journal*, **2**, 1071–81.

Evans, A. (1978) Causation and disease: a chronological journey. *American Journal of Epidemiology*, **108**, 249–58.

Gould, S.J. (1984) *The Mismeasure of Man*. London, UK: Pelican.

Government Office for Science. (2007) *Foresight. Tackling Obesities: Future Choices—Project Report*, 2nd edn.

Hammond, E.C., Selikoff, I.J., and Seidman, H. (1979) Asbestos exposure, cigarette smoking and death rates. *Annals of the New York Academy of Science*, **330**, 473–90.

Herbst, A., Ulfelder, H., and Poskanzer, D. (1971) Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumour appearance in young women. *New England Journal of Medicine*, **284**, 878–81. (Reprinted in Buck *et al*. 1988, pp. 446–50.)

Krieger, N. (1994) Epidemiology and the web of causation: has anyone seen the spider? *Social Science and Medicine*, **39**, 887–903.

Liddle, R. (2015) Toasties get you laid, fat prevents dementia and I'm a sex god. *The Sunday Times*, 12 April 2015, p. 19.

Mausner, J.S. and Kramer, S. (1985) *Epidemiology*, 2nd edn. Philadelphia, PA: W.B. Saunders.

Pearl, J. (2009) *Causality*, 2nd edn. Cambridge, UK: Cambridge University Press.

Porta, M. (2014) *A Dictionary of Epidemiology*, 6th edn. New York, NY: Oxford University Press.

Porter, R. (1997) *The Greatest Benefit to Mankind: A Medical history of Humanity from Antiquity to the present*. Harper Collins, London.

Rothman, K.J. and Greenland, S. (1998) *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven.

Rothman, K.J., Adami, H., and Trichopoulos, D. (1998) Should the mission of epidemiology include the eradication of poverty? *Lancet*, **352**, 810–13.

Semmelweis, I. (1983) *The Etiology, Concept and Prophylaxis of Childbed Fever*, translated by **Codell Carter, K.** University of Wisconsin, Madison. (Excerpted and reprinted in Buck *et al*. 1988, pp. 46–59.)

Skrabanek, P. (1994) The emptiness of the black box. *Epidemiology*, **5**, 553–5.