1

INTRODUCTION AND CAUSAL CONCEPTS

OBJECTIVES

After reading this chapter, you should be able to:

1. Explain the history of causal thinking about disease and scientific inference from an epidemiologic perspective.

2. Explain component-cause models and how they can be used to measure disease association and the proportion of disease attributable to a causal factor.

3. Explain the basis of causal-web models.

4. Describe the counterfactual concept and its utility in understanding disease causation and the estimation of causal effects.

5. Explain how observational studies and field experiments seek to estimate causal effects and how these relate to counterfactual and component-cause models.

6. Construct a logical causal diagram based on your area of research interest as a guide for your study design and analysis.

7. Apply a set of causal criteria to your own research as an aid to interpreting published literature and planning future research.

1.1 INTRODUCTION

What is epidemiology? Historically, MacMahon and Pugh (1970) saw epidemiology as being largely concerned with disease prevention, and therefore with the "succession of events which result in the exposure of specific types of individual to specific types of environment" (*ie* causal factors, or exposures). More recently, Parodi *et al* (2006) noted that "modern epidemiology aims to improve the health of populations by integrating data from different disciplines and proposing interventions on the basis of scientific evidence". More specifically, epidemiologists seek to identify exposures, be they demographic factors, infectious or toxic agents, nutritional factors, or elements of lifestyle, and evaluate their associations with various outcomes of interest (*eg* health, disease, quality of life).

Hence this book is about **associations** which are likely to be causal in nature and which, once identified, can be manipulated to improve the health and life quality of people. Epidemiologists see these associations as part of a complex web of relationships involving organisms and all aspects of their environment (Martin, 2008). Thus, epidemiologists strive to improve their study designs and data analysis with the aim of identifying valid associations between the exposure(s) and outcome(s) and at least some components of this complex web. As it is only by studying exposure-disease associations under real-world conditions that we can begin to understand this web of causal relationships, we also come to understand that epidemiology is a field-based discipline.

As a starting place, for describing epidemiologic research methods, we will review the history of the concept(s) of multiple interrelated causes. This will provide a sense of how we have arrived at our current concepts of disease causation, and where we might need to go in the future (Ness *et al*, 2009). Because we want to identify associations which are likely to be causal (or at the very least useful for disease control (Olsen, 2003)), it also is appropriate to review the relevant areas of the philosophy of science that relate to causal inference. We will then proceed with overviews of the key components of epidemiologic studies, and discuss some current concepts of disease causation. Our objective here is to provide a foundation on which a deeper understanding of epidemiologic principles and methods can be built.

1.2 A BRIEF HISTORY OF MULTIPLE CAUSATION CONCEPTS

Parodi *et al* (2006) observed that, throughout the history of epidemiology, there has been a struggle between two major points of view about disease causation and control: one oriented toward biology and mechanisms of causation, the other toward populations and their interactions with the environment (social, biotic, and abiotic). This tension is essential for the success of epidemiology. Epidemiologic research is based on the idea that 'causes' (exposures) and 'outcomes' (health events) are part of a complex web of relationships. Consequently, epidemiologists accept that there are multiple causes for almost every outcome and that a single cause can have multiple effects. This perspective is not shared by all health researchers. In the current era of great advances in understanding the genetic components of illnesses, a significant proportion of medical research is focused on identifying genes that are associated with disease, and on the characteristics of direct causal agents and how they interact with the genetic makeup of the host of interest. However, as Diez-Roux (1998a) points out, while it is true that genetic abnormalities are important precursors of many diseases, in terms of maintaining health, the real questions relate to the extent to which our current environmental exposures (Kanarek and

Anderson, 2007) and lifestyles lead to genetic defects, as well as the extent to which these exposures and lifestyles allow specific genetic patterns to complete a sufficient cause of disease. (The concept of 'sufficient cause' is discussed in Section 1.6.1) Kaprio (2000) describes the current and future states of genetic epidemiology, whereas Weed (2006) discusses what epidemiologists need to do to improve their ability to meet their objectives.

The acceptance of the concept(s) of multiple interacting causes varied with the dominant causal paradigm of the era. However, the roots of this concept can be traced back to at least 400 BC, when Hippocrates wrote *Air, Waters and Places*, in which he outlined the environmental features that should be noted in order to understand the health of populations. Based on this aspect of his writings, it is clear that Hippocrates had a strong multicausal concept about exposure factors in the environment being important 'causes' of disease occurrence. He also discussed the importance of the inhabitant's lifestyle (now often referred to as social epidemiology) as a key determinant of health status, further expanding the 'web of causation'. Nonetheless, his concepts linking the state of the environment and lifestyle to the occurrence of disease seem to have been short-lived; between 5 AD and 1750 AD, humoral imbalances (events within the individual) became the major paradigm of disease causation (Schwabe, 1982).

Between 1750 and 1885, the multifactorial nature of disease causation regained its credence as man-made environmental degradation came to be accepted as a central cause of disease. During this era, the prevalent causal paradigm was that disease was due to the effects of 'miasmas' (*ie* bad air). However, during the same era, medical statistics became well-established in France and Britain (Parodi *et al*, 2006). In France, physician Pierre Louis disproved the efficacy of blood-letting as a treatment for pneumonia, and Simon Poisson stressed the need to study large numbers of cases. In England, William Farr used demographic and health-related data to demonstrate the association of disease with urban poverty. During the mid-1800s, William Budd and John Snow conducted studies seeking to link contaminated water as the cause of cholera (Frerichs, 2001). Using a combination of astute observations about the lack of spread of the disease among health workers, the geographical distribution of cholera, the results of a series of observational studies, and their use of natural as well as contrived (removal of the Broad Street pump handle) experiments, Snow correctly concluded that cholera was being spread by water contaminated with sewage effluent. It is noteworthy that he arrived at this conclusion almost 30 years before the organism (*Vibrio cholera*) was discovered, thus demonstrating an important principle: disease can be prevented without knowing the proximal (*ie* direct) causal agent. During this same period, German pathologist Friedrich Henle was debunking miasma and demonstrating the role of 'microscopic living beings' (Parodi *et al*, 2006). Nonetheless, German pharmacologist Max Pettenkofer developed sound principles of public health based on preventing the untoward effects of miasma.

A few years later (*ie* in the 1880s–1890s), Theobald Smith (a physician) and Frederick Kilborne (a veterinarian) determined that an insect vector (a tick: *Boophilus annulatus*) was associated with a cattle disease called 'Texas Fever', even though the direct causal agent of the disease (a parasite: *Babesia bigemina*) was not discovered until many years later (Schwabe, 1984). Their initial associations were based on the similar geographical distributions between the disease and the extent of the tick's natural range; theirs was the first demonstration of a parasite requiring development within a vector before transmission. Their work also provided the basis for disease control before knowing the actual agent of the disease. At about the same time, malaria, caused by infection with protozoans of the genus *Plasmodium and* transmitted by female *Anopheles* species mosquitoes, was being studied. According to Cox (2010), Alphonse Laveran discovered

parasites in the blood of malaria patients in 1880, while William MacCallum discovered the sexual stages in the blood in birds infected with a related haematozoan—*Haemoproteus columbae*—in 1897. The transmission cycle in culicine mosquitoes and birds infected with *Plasmodium relictum* (the epidemiological component of the natural history of malaria) was elucidated by Ronald Ross in 1897.

The multifactorial causal concept lost adherents during the late 1800s to mid-1900s, as the search for specific etiological agents (usually microbiological or toxicological, as in the work of Robert Koch) gained predominance in medical research. The 'golden era' of microbiology led to a number of successes including mass-testing, immunisation, specific treatment, and vector control (*eg* the mosquito vector of malaria was now known) as methods of disease control. Nonetheless, Rudolph Virchow, the German cellular pathologist, believed that medicine was a social science and backed the earlier population-oriented work of Max von Pettenkofer from the perspective of disease control (Kottke, 2011).

Control of many specific infectious diseases meant that by the early-to-mid 1900s, chronic, non-infectious diseases were becoming relatively more important as causes of morbidity and mortality in humans in developed countries. For example, pellagra was originally thought to result from food poisoning from an unknown agent associated with poor quality maize. However, Joseph Goldberger observed the natural disease in cotton workers in the southern US and found nothing to suggest an infectious origin. Rather, he was concerned about the potential for a nutritional etiology. A key observation was that patients in mental institutions had the disease, but not the staff, nurses, or doctors. This convinced Goldberger about the lack of person-to-person spread, and he downplayed an infectious etiology. Subsequently, Goldberger chose 3 institutions, 2 orphanages and 1 mental institution in which to conduct his early studies. Both were heavily pellagra endemic before their diets were improved and virtually no new cases of pellagra occurred after the new diets had been put in place (Rajakumar, 2000). This work and Goldberger's later demonstration of the role of niacin as the direct cause of pellagra was based on a mechanistic view of causation. Yet, his proposal to prevent the disease through the use of land reform emerged from his population perspective of disease causation and control.

By the mid 1900s, it was recognised that single agents were not likely responsible for the increase in prevalence of chronic diseases. Consequently, large-scale, population-based studies examining the potential multiple causes of these diseases were initiated. For example, the Framingham Heart Study pioneered long-term surveillance and study of causes of disease beginning in 1949. Shortly thereafter, a series of observational studies on smoking and lung cancer were published, and this spurred much discussion about causal inference from observational data (Berlivet, 2005). This included debate over specificity of effect as the studies had indicated numerous health impacts of smoking. Thus, by the early 1960s, there was once again a growing awareness of the complex web of causation.

A decade later, multiple interacting causes of diseases returned as a major paradigm of disease causation. Building on the knowledge from the microbiological revolution, the concept of the agent-host-environment causal triad appeared in an early epidemiology text (Hogue, 2008; MacMahon, 1970). In this conceptual model, a number of component causes were required to come together (either sequentially or simultaneously) in order to produce disease; later, the complex of factors that would produce disease became known as a sufficient cause, and it was assumed that most diseases had a number of sufficient causes. In addition to multiple causes, the component cause model was not constrained to have all causal factors at the same level of

organisation. A traditional veterinary example used to portray some of these concepts is yellow shanks in poultry (Martin *et al*, 1987). When poultry with the specific genetic defect (an individual-level factor) are fed corn (ration is usually a herd/flock level factor), they develop a discolouration of the skin and legs. If all poultry are fed corn, then the cause of the disease would be a genetic defect; however, if all the birds had the genetic defect, then the cause of the disease would be deemed to be the feed. In reality, both factors are required, and the disease can be prevented either by removing the genetic defect, or changing the feed, or both, depending on the specific context.

The 1970s appeared to be a period of peak interest in causation (Kaufman and Poole, 2000). Susser's text (1973) on causal thinking appeared in 1973 (unfortunately, it has never been reprinted) and, 3 years later, the concepts of necessary and sufficient causes were published by Rothman (1976), followed in 1977 by a set of criteria to help assess causation by Susser (1995). Large-scale monitoring of diseases also began in this period.

No major new causal concepts were brought forward in the 1980s. Hence it was asserted that the aforementioned web of causation had become restricted to studying individual-level, direct-causal factors which focussed on biological malfunctioning (Krieger, 1994). However, it is instructive to read the story of the discovery of folic acid deficiency as the cause of neural tube disease (Oakley, 2009), the linking of delinquent behaviour and bone lead levels (Needleman, 2009) and of the descriptive and observational studies in the early 1990s that led to the discovery of prone sleeping as a major cause of SIDs (Dwyer and Ponsonby, 2009). These were indeed triumphs of epidemiology (Ness, 2009).

In 1990, epigenesis was proposed as a formal model of multivariable causation that attempted to link, explicitly, causal structures to observed risks of disease (Koopman and Weed, 1990). While this proved to be an interesting and exciting proposal, the limitations of this approach were soon realised (Thompson, 1991) and the approach remained only a concept.

Since the mid 1990s, there has been a lot of introspective writing by epidemiologists with much concern over an excess focus on individuals as the units of study and analysis. Others decried the apparently large number of false positive and false negative findings in epidemiological research (Taubes (1995); see Willett *et al*, (1995) for rejoinder). We shall not review these debates in detail because excellent discussions on these topics are available elsewhere (Diez-Roux, 1998a; Diez-Roux, 1998b; McMichael, 1999). What is apparent is that, whenever possible, elements of the social, physical, and biological features of the defined ecosystem should be included in each study. This feature was clearly evident in a noteworthy book entitled *Why Are Some People Healthy and Others Not?: The Determinants of Health of Populations* (Evans *et al*, 1994). The impact of social and economic factors on population disease was, and is, striking. As we explain later in this text, the unit of concern can range from the individual to groups (families), villages or communities, watersheds or larger ecosystems. Today, epidemiologic research remains rooted in the concept of multiple interrelated causal factors for disease occurrence, and hence for disease prevention. This conceptual basis has been supported by substantial progress in the development of epidemiologic research methodologies and these are the subject of this book.

In the first decade of the twenty-first century, perhaps the most visible shift in the focus of veterinary and medical epidemiology was to reassert the 'One Health' or 'One Medicine' approach to world health. Historically, a number of people, including Canadian physician Dr William Osler and German pathologist Rudolf Virchow, had stressed and contributed to the

One Medicine approach. Dr Calvin Schwabe was an early (1960s) veterinary leader in this regard and his classic text *Veterinary Medicine and Human Health* (1984) is a unique and valuable resource in this regard (Kahn *et al*, 2008). The One Health initiative in medical epidemiology (www.onehealthinitiative.com) was led by two physicians (Drs Kahn and Monath) and a veterinarian (Dr Kaplan) beginning in 2006, and the movement, in both medicines, has continued to expand (Fisman and Laupland, 2010; Steele, 2008).

For example, the One Medicine/One Health movement appeared to gain momentum following the occurrence of a number of serious disease outbreaks around the world including the bovine spongiform encephalopathy (BSE) epidemic, severe acute respiratory syndrome (SARS), and H5N1 avian influenza. In November 2005, *The Veterinary Record* and the *British Medical Journal* published simultaneous issues exploring how the veterinary and medical professions could collaborate for mutual benefit. In 2006, The American Medical Association and the American Veterinary Medical Association approved resolutions supporting One Medicine or One Health approaches that bridge the 2 professions. The importance of epidemiology in supporting this movement seems obvious, and the potential benefits from collaboration between the two medical professions have been discussed by Fisman and Laupland (2010) and Sargeant (2008).

1.3 A BRIEF HISTORY OF SCIENTIFIC INFERENCE

Epidemiology relies primarily on observational studies to identify associations between exposures and outcomes. The reasons are entirely pragmatic. First, many health-related problems cannot be studied under controlled laboratory conditions. The major reasons include our limited ability to create suitable models of human 'disease' in experimental animals, ethical concerns about causing disease and suffering in humans, and the cost of studying diseases in their natural hosts (humans and other animals) under laboratory conditions. Most importantly, though, if we want to understand the complex web of relationships that affects humans, then we must study them in their 'natural' state (*ie* where and how we live on a daily basis). This requires the use of observational studies, and drawing inferences from these studies uses both inductive and deductive reasoning.

Philosophical discussion of causal inferences appears to be limited mainly to fields where observation (in which we attempt to discern the cause) rather than experimentation (in which we try to discern or demonstrate the effect) is the chief approach to research. While the latter approach is powerful, one cannot assume that the results of even the best-designed experiments are infallible. Recent discussions have included approaches to identifying and understanding causal factors in complex systems (De Vreese, 2009; Rickles, 2009; Ward, 2009a). Nonetheless, because epidemiologists rely heavily on observational studies and field experiments for the majority of our research, a brief review of the basis for scientific inference is in order. We pursue this review in the context that epidemiology is a pragmatic discipline, that our activities are tied to health promotion and disease prevention, and that the key for disease prevention is to identify causal factors that we can manipulate, regardless of the level of organisation at which they act (Susser, 1991). We will briefly present the concepts and history of inductive and deductive reasoning in this process. More complete reviews on the philosophy of causal inference are available elsewhere (Aiello and Larson, 2002; Weed, 2002; White, 2001).

Inductive reasoning is the process of making generalised inferences about (in our context) 'causation' based on repeated observations. Simply put, it is the process of drawing conclusions

about the state of nature from carefully recorded and analysed observations. Francis Bacon (1620) first presented inductive reasoning as a method of making generalisations from observations to general laws of nature. As 2 examples, John Snow's observations during the cholera outbreaks of the mid 1800s led to a correct inference about the mechanism of the spread of the disease. Edward Jenner's observations that milkmaids who developed cowpox didn't get smallpox led to his conclusion that cowpox might prevent smallpox. This, in turn, led to the development of a crude vaccine which was found to be effective when tested in humans in 1796. These were both dramatic examples of the application of inductive reasoning to important health problems. In 1843, John Stuart Mill proposed a set of canons (rules) for inductive inference. Indeed, Mill's canons might have been the origin of our concepts about the set of component causes that are necessary or sufficient to cause disease (White, 2000).

While it is easy to identify important advances in human health that have been based on inductive reasoning, proponents of deductive reasoning have been critical of the philosophical basis (or lack thereof) of inductive logic. David Hume (1740) stated that "there is no logical force to inductive reasoning". He stated further that "we cannot perceive a causal connection, only a series of events". The fact that the sun comes up every day after the rooster crows, should not result in a conclusion that the rooster crowing causes the sun to rise. He noted further that many repetitions of the 2 events might be consistent with a hypothesis about causation but do not prove it true. Bertrand Russell (1872-1970) continued the discussion of the limitations of inductive reasoning and referred to it as "the fallacy of affirming the consequence" (*eg* we might imply that if A is present, then B occurs; so if B occurs, A must have been present).

Deductive reasoning is the process of inferring that a general 'law of nature' exists and has application in a specific, or local, instance. The process starts with a hypothesis about a 'law of nature', and observations are then made in an attempt to either prove or refute that assumption. The greatest change in our thinking about causal inferences in the past century has been attributed to Karl Popper, who stated that scientific hypotheses can never be proved or evaluated as true, but additional evidence might suggest they are false. This philosophy is referred to as **refutationism**. Based on Popper's philosophy, a scientist should not collect data to try and prove a hypothesis (which Popper states is impossible, anyway), but that scientists should try to disprove their theory; this can be accomplished with only one study. Once a hypothesis has been disproven, the information gained can be used to develop a revised hypothesis, which should once again be subjected to rigorous attempts to disprove it. Popper argues that, only by disproving hypotheses can we make scientific progress. It is partially for this reason that, when conducting statistical analyses, we usually form our hypothesis in the null (*ie* that a factor is not associated with an outcome) and, if our data are inconsistent with that hypothesis, we can accept the alternative hypothesis, that the factor is associated with the outcome. Thus, the current paradigm in deductive reasoning is to conjecture and then attempt to refute that conjecture. A major benefit of using Popper's approach is that it helps narrow the scope of epidemiologic studies. It suggests that we carefully review what is already known and then formulate a few specific hypotheses that are testable with a reasonable amount of data. This contrasts with the data-mining approach, in which we often generate long, multipage questionnaires and end up with dozens if not hundreds of variables to test (most without any pre-specified hypothesis). In general, epidemiologic investigations which start with a clear hypothesis are inevitably more focused and more likely to result in valid conclusions than those based on unfocused recording and analysis of observations.

Two other important concepts that relate to scientific inference are worth noting. Thomas Bayes, a Presbyterian minister and mathematician, stated in 1764 that "all forms of inference are based on the validity of their premises" and that "no inference can be known with certainty". He noted that scientific observations do not exist in a vacuum, and that the information we have prior to making a series of observations will influence our interpretation of those observations. His work has given rise to a branch of statistics known as **Bayesian analysis**, some of which appear in Chapter 24.

More recently, Thomas Kuhn reminds us that, although one observation can disprove a hypothesis, the particular observation might have been anomalous and that the hypothesis could remain true in many situations. Thus, often the scientific community will come to a decision about the usefulness, if not the truth, of a particular theory. This is the role of **consensus** in scientific thinking. Schwabe (1993) refers to this as a paradigm shift. While hard to justify on a philosophical basis, it plays a large role in shaping our current thinking about causes of disease.

Philosophical debates on causal inference will undoubtedly continue (Robins, 2001; White, 2001). As a summary of this section, we note that "... all of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature ... the tentativeness of our knowledge does not prevent practical applications, but it should keep us sceptical and critical" (Rothman and Greenland, 2005). While keeping these historical and philosophical bases in mind, we will now proceed to an outline of the key components of epidemiologic research.

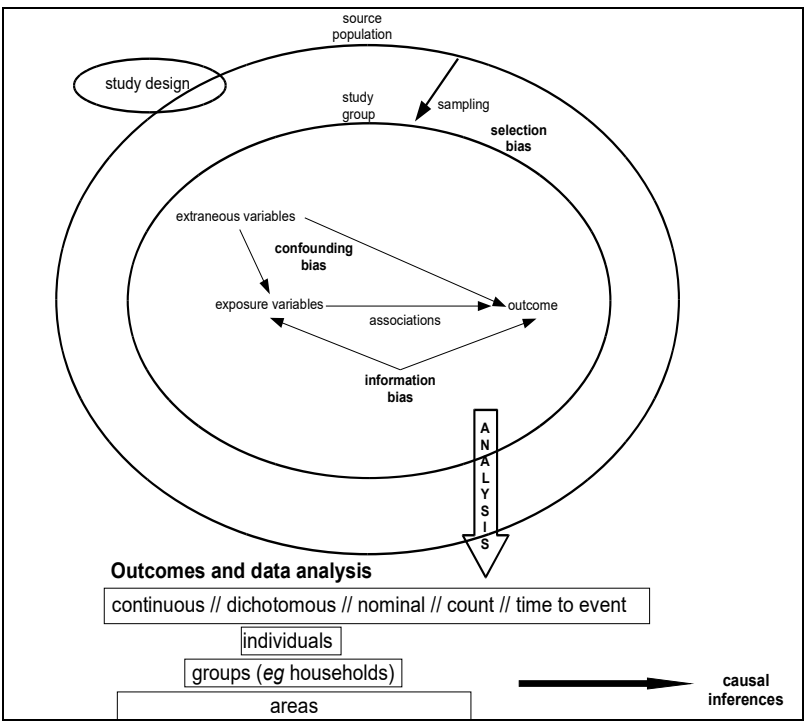


Fig. 1.1 Key components of epidemiologic research

1.4 KEY COMPONENTS OF EPIDEMIOLOGIC RESEARCH

Fig. 1.1 (previous page) summarises the key components of epidemiologic research. It might be risky to simplify a complex discipline such as epidemiology and present it in a single diagram, but we believe it is beneficial to have an overview of the process of evaluating associations between exposure and outcome as a guide to the rest of the book.

Our rationale for doing research is to identify potential causal associations between exposures and outcomes (the centre of the diagram). In many instances, the exposures are potential risk factors and the outcome is a disease of interest. However, this is not the only scenario; for example, our outcome of interest might be a measure of quality of life, and the exposures might include certain diseases (*eg* the impact of diabetes on future health). Ultimately, we aim to make **causal inferences** (bottom right of diagram) about relationships between exposure and disease in the source population as a preliminary step toward developing policy and programs to maintain health and prevent disease in the target population.

An overview of the contents of this text is shown below:

- Chapter 1 gives a brief history of epidemiology and the scientific process, and discusses some important **concepts of causation** as they relate to epidemiologic research.
- Field research starts with an overall **study design** and the main observational study types are discussed in Chapters 7–10, with **controlled trial** designs being presented in Chapter 11. In all studies, it is important to identify the **target population** and obtain our **study group** from the **source population** in a manner that does not lead to **selection bias**. **Sampling** is discussed in Chapter 2, and selection bias in Chapter 12.
- Once we have identified our study subjects, it is necessary to obtain data on exposure variables, extraneous variables, and the outcome in a manner that does not lead to **information bias** (Chapter 12). Two important tools that are used in that process are **questionnaires** (Chapter 3) and **diagnostic and screening tests** (Chapter 5).
- In order to start the process of establishing an association between exposure and outcome, we need to settle on a **measure of disease frequency** (Chapter 4) and select a **measure of association** (Chapter 6) that fits the context. In many cases, the study design will determine the measures that are appropriate.
- **Confounding bias** is a major concern in observational studies, and the identification of factors that should be controlled as confounders is featured in Chapter 13, along with a variety of techniques to prevent this bias.
- With our data in hand, we are now able to begin to model relationships with the intent of estimating causal effects of exposure (Chapter 13). Individual chapters are dedicated to the analyses appropriate for outcomes that are **continuous** (Chapter 14), **dichotomous** (Chapter 16), **nominal/ordinal** (Chapter 17), **count** (Chapter 18) and **time-to-event data** (Chapter 19). Chapter 15 presents some general guidelines on model-building techniques that are applicable to all types of model.
- In epidemiologic research, we often encounter **clustered** or **correlated data**, and these present major challenges in their analyses. Chapter 20 introduces these, while Chapters 21 and 22 focus on mixed (random effects) models for continuous and discrete outcomes. Chapter 23 focuses on the specific issue of analysing repeated measures data.
- In Chapter 24, we introduce **Bayesian analysis.** The Bayesian approach formally incorporates the degree of certainty we hold about a hypothesis before we see

additional data (prior probability). It then modifies this based on the information gained from new data to update the prior and obtain new (posterior) estimates about the certainty of that hypothesis.

- Chapters 25 and 26 present the basics of **geographical information systems** and **spatial statistics** that we use in epidemiology. These fields have developed a number of unique approaches that are useful in the study of diseases in populations.
- Chapter 27 describes **infectious disease epidemiology.** The ability of the living agent(s) to spread from subject to subject creates 'dependencies' (correlations) and other phenomenon such as herd immunity that must be accounted for in our research efforts.
- **Systematic reviews** and assessments of the literature in the form of **meta-analyses** are becoming increasingly important and are introduced in Chapter 28.
- Not all studies allow us to collect data on exposures and outcomes at the individual level, and yet there is much that we can learn by studying disease in groups (*eg* families, villages, cities). Thus, **ecologic studies** are introduced in Chapter 29.
- Finally, we complete the text with Chapter 30, which provides a 'road map' for investigators starting into the analysis of a complex epidemiologic dataset.

With this background, it is time to delve deeper into this discipline called epidemiology. And at the outset it is important to stress that epidemiology is a biological discipline that incorporates social science theory and methods, and relies heavily on quantitative (statistical) methods for its research methodology. Epidemiologists focus on the relationships (*eg* between micro-organisms, toxic agents, lifestyle, nutrition, environment, physiological factors, and health), rather than the entities themselves which are the remit of other disciplines. The integration of these facets, with a clear understanding of epidemiologic principles makes for successful epidemiologic research. To help meet this goal, this book is divided roughly equally into chapters dealing with epidemiologic principles and study design, and chapters dealing with quantitative methods.

1.5 SEEKING CAUSES

De Vreese (2009) noted: "The true subject matter of epidemiologic practice and of textbooks of epidemiology is research design and methods for disentangling causes and effects". In other words, we need to identify causes of health and disease in populations. That might seem like a simple enough task, but it is, in fact, complex. Here we want to focus on what a cause is and how we might best determine whether a factor is a cause. For practical purposes, **a cause is any factor that produces a change in the severity or frequency of the outcome**. Some prefer to separate biological causes (those operating within individuals) from population causes (those operating at or beyond the level of the individual). For example, infection with a specific microorganism could be viewed as a biological cause of disease within individuals, whereas lifestyle, nutrition, or other factors that act at the group level—or beyond (*eg* weather)—and affect whether or not an individual is exposed to the microorganism, or affect the individual's susceptibility to the effects of exposure, would be deemed as population causes. We recognise that whereas disease occurs in individuals, "epidemiology deals with groups of individuals because the methods for determining causality require it" (De Vreese, 2009). Vineis and Kriebel (2006) review concepts of causality "from Koch to Rothman". Further, as noted, it is vital that we include social as well as biological factors in our study of health and disease (Berkman, 2009; Kaplan, 2004).

In searching for causes, we stress a holistic approach to health. The term holistic might suggest that we try to identify and measure every suspected causal factor for the outcome of interest. Yet, clearly we cannot consider every factor in a single study. Rather, we place limits on the portion of the 'real world' we study and, within this, we constrain the list of factors we identify for investigation, using the current state of knowledge plus our specific objectives and hypotheses. Being pragmatists, we prefer to identify causal factors that we can manipulate in order to help prevent disease, while recognising that some non-manipulatable factors also may be crucial to our understanding of disease patterns in populations. As noted, usually extant knowledge and current belief are the bases for selecting factors for study. Because of this, having a concept of causation and a causal model in mind can help clarify the data needed, the key measures of disease frequency and the interpretation of associations between exposure and disease. We also need to differentiate between a conceptual or metaphysical view of causation which we develop at the individual level (*eg* counterfactual states) and the techniques we use to achieve our objectives at the population level (Hernan, 2004; Hernan and Robins, 2006a). We begin with an overview of 2 important models of causation.

1.6 MODELS OF CAUSATION

Given our belief in multiple causes of an effect and multiple effects of a specific cause, epidemiologists have sought to develop conceptual models of causation. Usually, however, the actual causal model is unknown and the statistical measures of association we use reflect, but do not explain, the number of ways in which the exposure might cause disease. Furthermore, although our main interest in a particular study might focus on one exposure factor, we need to take into account the effects of other causes of the outcome that are related to the exposure (this process is usually referred to as control of confounding) if we are to learn the 'truth' about the potential causal effect of our exposure of interest.

Because our inferences about causation are based, at least in the main, on the observed difference in outcome frequency, or severity, between exposed and unexposed subjects, we will continue our discussion by examining the relationship between a postulated causal model and the resultant, observed, outcome frequencies. We begin with a description of the component-cause model followed by the causal-web model of causation.

1.6.1 Component-cause model

The component-cause model is based on the concepts of necessary and sufficient causes (Rothman, 1976). In this model, a **necessary cause** is one without which the disease cannot occur (*ie* the factor will always be present if the disease occurs). In contrast, a **sufficient cause** always produces the disease (*ie* if the factor is present, the disease invariably follows). Both experience and formal research have indicated that very few exposures (factors) are sufficient in and of themselves, rather different groupings of factors combine and become a sufficient cause. Thus, a **component-cause** is one of a number of factors that, in combination, constitute a sufficient cause. The factors might be present concomitantly, or they might follow one another in a temporal chain of events. In turn, when there are a number of chains with one or more factors in common, we can conceptualise the web of causal chains (*ie* a causal-web). This concept will be explained further under the causal-web model (Section 1.6.2).

As an example of component causes, in Table 1.1 we portray the causal relationships of 4 risk factors for childhood respiratory disease (CRD) (Chibuk *et al*, 2010; Korppi *et al*, 2003). These include:

- a bacterium, *Streptococcus pneumoniae* (STREP)
- a virus, the respiratory syncytial virus (RSV)
- fluctuating damp cool/cold weather (called Stressors), and
- other bacteria such as *Mycoplasma pneumoniae* (MP).

Table 1.1 Four hypothetical sufficient causes of childhood respiratory disease (CRD)

	Sufficient causes			
Component causes	**I**	**II**	**III**	**IV**
STREP	+	+		
RSV	+		+	
Stressors		+	+	+
Other organism (eg MP)				+

In this deterministic portrayal, there are 4 sufficient causes, each one containing 2 specific components; we assume that the 4 different, 2-factor combinations each form a sufficient cause. Hence, whenever these combinations occur in the same child, respiratory disease will occur (as mentioned, one can conceive that these factors might not need to be present concomitantly, they could be sequential exposures in a given child). Some children could have more than 2 causal factors (*eg* STREP, RSV, Stressors), but the dual exposure to any 2 of the 3 factors would be sufficient to produce CRD. Note that, in our model, we have indicated that only some specific 2-factor combinations act as sufficient causes; STREP is a component of 2 of the sufficient causes, as is RSV, MP is present only in one sufficient cause. Because no factor is included in all sufficient causes, there is no necessary cause in our model of CRD. Obviously, if you have not guessed by now, you should be aware that the number of causal factors and their arrangement into sufficient causes, as presented here, are purely for the pedagogical purposes of this example.

Now, against this backdrop of causal factors, we will assume that we plan to measure only the STREP and RSV components (*ie* obtain nasal swabs for culture and/or blood samples for antibody titres) in our research. Nonetheless, we are aware that, although unmeasured, the other components (Stressors and/or MP) might operate as components of one or more of the sufficient causes. In terms of the 2 measured factors, we observe that some children with CRD will have both factors, some will have only STREP, and some only the RSV components. Because of the causal effects of the other unmeasured factors (*eg* Stressors and MP forming sufficient cause IV), there will be some children with CRD that have neither of the 2 measured factors.

The effect of risk factor prevalence on disease risk

Two benefits of thinking about causation in this manner are that it helps us understand how the prevalence and distribution of causal factors drives the observed frequency of disease and the strength of association between an exposure and the disease. In addition, it demonstrates how the prevalence of a co-factor can impact the strength of association that epidemiologists might find between the exposure factor and the outcome of interest. For example, assume that we are interested principally in the strength of association between infection with STREP and the

occurrence of CRD (the various measures of association are explained in Chapter 6). According to our model of causation in Table 1.1, STREP produces disease when present with RSV, but also without RSV when combined with Stressors. What might not be apparent, however, is that changes in the prevalence of the virus or of the Stressors (since they are components of the same sufficient cause) can change the strength of association between STREP and CRD.

To demonstrate these points, note that the 2 populations shown in Examples 1.1 and 1.2 are based on the component-cause model shown in Table 1.1. We include only 3 of the 4 causal factors—STREP, RSV, and Stressors—for simplicity. The frequency of each factor indicated above the body of the tables in Examples 1.1 and 1.2 is the same {p(Stressors)=0.4 and p(STREP)=0.6}, except that the frequency of RSV is increased from 30% in Example 1.1 to 70% in Example 1.2. In our examples, all 3 factors are distributed independently of each other; this is not likely true in the field, but it allows us to examine the effect of single factors without concerning ourselves with the biasing (*ie* confounding) effects of the other factors. Furthermore, to keep the example simple, we have not shown cases of CRD arising from the Stressors-MP pathway.

If infection with STREP is our exposure factor of interest, it would be apparent that some but not all children with STREP develop CRD, and that some children without STREP also develop

Example 1.1 Causal complement prevalence and disease risk—Part I

The number and risk of CRD cases produced by 2 measured and 1 unknown exposure factors assuming joint exposure to any 2 factors is sufficient to cause the disease are shown below. *Streptococcus pneumoniae* (STREP) is the exposure of interest (total population size is 10,000; p(Stressors)=0.4; p(STREP)=0.6; p(RSV)=0.3).

Measured factors				
Unmeasured Stressors	RSV	STREP	Population number	Number diseased
1	1	1	720	720
1	1	0	480	480
1	0	1	1680	1680
1	0	0	1120	0
0	1	1	1080	1080
0	1	0	720	0
0	0	1	2520	0
0	0	0	1680	0

Risk of disease among the STREP+	3480/6000 = 0.58
Risk of disease among the STREP-	480/4000 = 0.12
Risk difference in STREP+	0.58 - 0.12 = 0.46
Risk ratio if STREP+	0.58/0.12 = 4.83

Example 1.2 Causal complement prevalence and disease risk—Part II

The number and risk of CRD cases produced by 2 measured and one unknown exposure factors assuming joint exposure to any 2 factors is sufficient to cause the disease are shown below. *Streptococcus pneumoniae* (STREP) is the exposure of interest (total population size is 10,000; p(Stressors)=0.4; p(STREP)=0.6; p(RSV)=0.7).

	Measured factors			
Unmeasured Stressors	RSV	STREP	Population number	Number diseased
1	0	1	720	720
1	0	0	480	0
1	1	1	1680	1680
1	1	0	1120	1120
0	0	1	1080	0
0	0	0	720	0
0	1	1	2520	2520
0	1	0	1680	0

Risk of disease among the STREP+	4920/6000 = 0.82
Risk of disease among the STREP-	1120/4000 = 0.28
Risk difference in STREP+	0.82 - 0.28 = 0.54
Risk ratio if STREP+	0.82/0.28 = 2.93

CRD. Thus, STREP infection by itself is neither a necessary nor sufficient cause of CRD. Similarly for RSV, only some RSV-infected children develop CRD, while some RSV-non-infected children also develop CRD. In order to ascertain if the occurrence of CRD is associated with STREP exposure, we need to measure and contrast the risk of CRD among the exposed (STREP+) versus the non-exposed (STREP-). In Example 1.1, these frequencies are 58% and 12%, respectively, and we can express the proportions relative to one another using a statistic called the risk ratio, which is 58/12=4.83. This means that the frequency of CRD is 4.83 times higher in STREP+ children than in STREP- children. We could also measure the association between STREP and CRD using a risk difference (*RD*); in this instance, the *RD* is 0.58-0.12=0.46 or 46%. These measures are consistent with STREP being a cause of CRD, but do not prove the association to be causal.

In Example 1.2, the frequency of RSV is increased, and the risk ratio for STREP+ children becomes smaller (2.93) and the *RD* larger (0.54 or 54%). Thus, we might be tempted to think that exposure to STREP+ in some sense acts differently from a causal perspective in one example to another, yet the underlying causal relationship of STREP exposure to the occurrence of CRD has not changed. The difference in the measure of association is due to a change in the frequency of the other components of the sufficient causes, namely RSV in this example. The additional factors that can combine with the factor of interest to form sufficient causes are

called the **causal complement** to the exposure factor. Here, with sets of 2 factors being sufficient causes, the causal complements of STREP are RSV or Stressors, but not both (the latter children would have developed CRD from being stressed and having RSV, even if STREP was absent). Although some children have all 3 component causes, our theory says that once a child is infected with any two of them, CRD will occur.

In general, we note that when the prevalence of causal complements is high, measures of association between the factor of interest and the outcome that are based on risk differences will be increased (especially when the prevalence of exposure is low) (Pearce, 1989). Some, but not all, ratio or relative measures of association could have the opposite relationship with the prevalence of causal complements. In any event, although the causal mechanism remains constant, the strength of association will vary depending on the distribution of the co-factors, many of which we do not know about, or remain unmeasured, for practical reasons. As will be discussed, strength of association is one criterion of causation, but it is not a fixed measure and we need to bear the phenomenon just discussed in mind when making causal inferences. In addition to the above observations, you might verify that the impact of RSV on CRD as measured by the risk ratio would be the same (*RR*=3.2) in both Examples 1.1 and 1.2, even though its (*ie* RSV) prevalence has changed. Although this is only one example, we could state the general rule that the strength of association for a given factor depends on the frequency of the causal complements but, providing the distribution of the other causal factors is fixed, changes in the prevalence of the factor of interest do not alter its strength of association with the outcome. In the unlikely event that we could measure all the co-factors including Stressors and the other causal component factors, the picture would change considerably. For example, if the Stressors were the only other causes of CRD, it would be obvious that, in the non-stressed children, CRD occurred only when both STREP and RSV were present together. This would be clear evidence of biological synergism, a feature that is detected numerically as statistical interaction (*ie* the joint effect of the 2 factors would be different than the sum of their individual effects—in this instance, they would have no 'individual' effect, only a joint effect) (see Chapter 13 for a discussion of interaction and for more advanced reading see VanderWeele and Robins (2007c)). In weather-stressed children, all children exposed to STREP or RSV would get CRD, but there would be no evidence of interaction because 100% of singly as well as jointly exposed stressed children would develop CRD.

Because changes in the prevalence of the 'unknown' or 'unmeasured' factor(s) (*eg* MP) will alter the magnitude of effect for the measured exposure, we accept that we need to think of measures of association as 'population specific'. Only after several studies have found a similar magnitude of effect in different populations should we begin to think of the effect as in some sense a biological constant. Further, as shown in our examples, even if the cases have arisen from an assumed model that incorporates biological synergism, because of the distribution of the unknown causal factors, interaction (indicating synergism) might not be evident in the observed data (VanderWeele and Robins, 2007c). Flanders (2006) and VanderWeele and Hernan (2006) discuss the component-cause model and its relationship to the counterfactual model in a more complex multifactorial setting than we describe here.

So far, we have pursued the component-cause model as deterministic. However, in reality, because we virtually never know all of the component causes of a disease, there will be circumstances where it appears that a factor is causal and other circumstances where it appears to have no effect. The statistical models we use to identify possible component-causes average these effects across individuals. Indeed, it is possible that a factor which appears to elevate the risk of disease in a population can have no effect, or a sparing effect, on some individuals

within their population (Rothman and Greenland, 2005). Because of this, we need to stress that epidemiological measures of association are for groups rather than for individuals (this was also stressed in early writings on epidemiological methods by McMahon and Pugh in 1970). Koopman and Lynch (1999) indicated the need to broaden the scope of the sufficient cause model in individuals to include the effects of interactions between individuals and a population-based approach, particularly when studying infectious diseases. Diez-Roux (2007) extends this approach to integrate social and biological risk factors in a systems approach to disease prevention in populations. Traditionally, social risk factors are deemed to be very indirect causes of disease through their impact on more proximate biological causes. In her view, a systems approach would not investigate individual risk factors (or individuals) one at a time, but would investigate the behaviour and relationships of multiple elements in a particular population system while it is functioning. In this regard, Reiber (2009) discusses the advantages of combining epidemiology and evolutionary theory.

Although the component-cause model is somewhat simplistic, we believe it has great merit in determining which factors to include in the study of a specific disease. As noted, clues about the potential influence of a factor from studies in basic biological sciences, or from other epidemiologic studies identifying potential causal factors, are more useful than an unfocused data-mining approach in which factors are studied merely because we already have data on them, or because the data are easily available. Nonetheless, by "studying disease causation in large groups makes us ... able to answer the question of what causes diseases without knowing much about the precise biological and chemical mechanisms involved" (De Vreese, 2009).

Proportion of disease explained by risk factors

Using the concepts of necessary and sufficient causes, we also gain a better understanding of how much disease in the population is attributable to that exposure (or alternatively the proportion of disease that we could prevent by completely removing the exposure factor).

As explained in Chapter 6, this is called the population attributable fraction (AF_p). For example, if we assume that the prevalence of each of the 4 sufficient causes from Table 1.1 is as shown in Table 1.2, then, if we examine the amount of disease that can be attributed to each of the component causes, it appears that we can explain more than 100% of the disease. Of course, we really can't; it is simply because the components are involved in more than one sufficient cause and we are double-counting the role that each component cause plays as a cause of the disease.

Table 1.2 Hypothetical sufficient causes of childhood respiratory disease and relationship to population attributable fraction

	Sufficient causes				
Component causes	**I**	**II**	**III**	**IV**	**AF_P (%)**
STREP	+	+			75
RSV	+		+		60
Stressors		+	+	+	55
Other organism (eg MP)				+	10
Prevalence of sufficient cause (%)	45	30	15	10	

Another important observation is that when 2 or more factors are both essential for disease occurrence, it is difficult to attribute a specific proportion of the disease occurrence to any single causal factor. For example, in children that had all 3 factors—STREP, RSV and Stressors—it would be impossible to decide the unique importance of each factor. Our model indicates that once any 2 of the 3 were present, then CRD would occur and the presence of the third factor is of no importance causally; thus 'timing is everything'. Certainly, because the frequency of co-factors can vary from subgroup to subgroup, as with relative risk measures, one should not think of AF_p as being a 'universal' measure of importance.

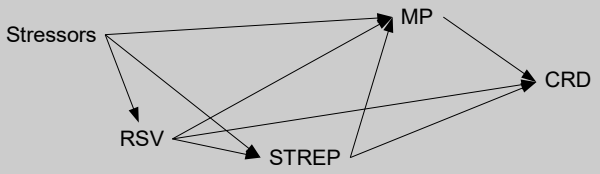
Whereas the AF_p is an extremely useful measure of importance, we also need to be aware of the **prevention paradox** (De Vreese, 2009). As an example, suppose that the AF_p for a factor (*eg* a vaccine) is 50%. This would suggest that if the prevalence of the disease in unvaccinated children was 6%, and if we fully vaccinated the population, only 3% of the subjects would develop the disease. Indeed, this is an important reduction for the population. However, 94% of the subjects we vaccinated would not have developed the disease if left unvaccinated, and half of those who would have developed the disease in the absence of vaccination developed it anyway despite being vaccinated. Thus, when proposing to implement our findings, we need to be aware of the costs and the possible side-effects of the proposed policy or program. The average person in the target population might not perceive the same benefits as we do.

1.6.2 Causal-web model

A second way of conceptualising how multiple factors can combine to cause disease is through a causal web (Example 1.3) consisting of multiple indirect and direct causes (Krieger, 1994). This concept is based on a series of interconnected causal chains or web structures; in a sense, it takes the factors portrayed in the sufficient-cause approach and links them temporally. In this model, a direct cause has no known intervening variable between that factor and the disease (diagrammatically, the exposure is adjacent to the outcome). **Direct causes** often are the **proximal causes** emphasised in therapy, such as specific microorganisms or toxins. In contrast, an **indirect cause** is one in which the effects of the exposure on the outcome are mediated through one or more intervening variables. It is important to recognise that, in terms of disease control, direct causes may be no more valuable than indirect causes. In fact, many large-scale control efforts are based on manipulating indirect rather than direct causes. Historically, this was also true: whether it was John Snow's work on cholera control through improved water supply, or preventing SIDs by not placing babies in the prone position to sleep. In both instances, disease control was possible before the actual direct causes (*Vibrio cholerae* and a yet-to-be-discovered pathophysiological event) were known, and the control programme was not focused directly on the proximal cause (as in Goldberger's suggestions to prevent pellagra).

One possible web of causation of respiratory disease (CRD) based on the 3 factors in Examples 1.1 and 1.2 might have the structure shown in Example 1.3. The causal-web model complements the component-cause model, but there is no direct equivalence between them. As we show in Chapter 13, formal causal-web diagrams are useful to guide our analyses and interpretation of data (VanderWeele *et al*, 2008), and can be related to the component cause model (VanderWeele and Robins, 2007a). Our example model indicates that Stressors make the child susceptible to STREP, RSV, and MP; that RSV increases the susceptibility to both STREP and MP, and that RSV can 'cause' CRD directly (this might be known to be true, or it might reflect the lack of knowledge about the existence of an intervening factor such as MP which is missing from the causal model). The diagram also indicates that STREP is an indirect

Example 1.3 A causal-web model of CRD based on component causes from Example 1.1



cause of CRD via MP, as well as being a direct cause of CRD. If this causal model is true, it suggests that we could reduce CRD occurrence by removing an indirect cause such as stress, even though it has no direct effect on CRD. We could also control CRD by preventing the action of the direct causes STREP, RSV, and MP (*eg* by vaccination). All things being equal, an RSV vaccine would be particularly effective, because of its direct and indirect effects on CRD. As mentioned, this model claims that Stressors do not cause CRD without STREP, RSV, or MP infection and thus suggests a number of 2- or 3-factor groupings of component causes into sufficient causes. However, it does not explicitly indicate whether some of the proximal causes can produce disease in and of themselves (*ie* it is not apparent, from our diagram whether RSV can cause CRD by itself or if it needs an additional unmeasured factor). From the previous examples, the outcome frequencies in RSV-infected and non-infected children will depend on the distribution of the other component causes and whether, in reality, it can be a sufficient cause by itself. In Section 1.8, we will discuss the relationship of the causal structure to the design of our studies and as a guide to the correct approach in our analyses and interpretation of the study data.

1.7 COUNTERFACTUAL CONCEPTS OF CAUSATION FOR A SINGLE EXPOSURE

Currently, the most widely accepted conceptual basis for determining causation in epidemiology is called the counterfactual or potential outcomes model (Greenland, 2005; Hofler, 2005a). In a sense, it reflects the way many of us would make causal inferences and can be the basis for forming clearly defined questions for future research. Both Greenland (2005) and Hernan (2004) give examples (through hypothetical interventions) of the specificity required in counterfactual questions if we are to make progress in resolving complex health problems.

The following discussion closely follows that of Hernan (2004). Suppose we are interested in whether or not a vaccine would protect against a disease (*eg* CRD), while having some concerns that the side-effects of the vaccine might be harmful. If we saw a vaccinated subject who developed the disease, we might begin to think that the exposure (or its side-effects) either failed to prevent, or even caused, the disease in that subject. If we imagined the same subject in the same period except he/she was not vaccinated, this would be referred to as the counterfactual state. Obviously, this individual does not exist, but it is what we would ideally like to observe in order to make valid causal inferences about the effect of vaccination. If the disease did not occur in this hypothetical counterfactual individual, we would surely conclude that the exposure (vaccination) had caused the observed disease in that individual. Conversely, if the disease occurred in this non-exposed counterfactual individual, we likely would conclude that the vaccination was not a cause of the disease in that subject (since the disease occurred

regardless of exposure). Vaccination did not alter the outcome in that person. We can make this thought process more formal by denoting the potential outcome in exposed subjects as D_{E+}, and the potential outcome in the same subjects if they were unexposed as D_{E-}. Our thought process concludes that there is a causal effect in that subject if $D_{E+} \neq D_{E-}$. Note that a causal exposure need not be causal in all individuals, principally because the other factors needed to complete a sufficient cause are absent. Further, in reality, we cannot determine a causal effect at the individual level because only one exposure level is observed and the data relating to what might have happened at the other exposure level (*ie* the counterfactual state) are missing.

In Table 1.3, we summarise the observed exposure, actual disease outcome and counterfactual outcomes in 20 subjects based on Hernan (2004). Here, 'i' is the subject counter, *C* a confounder, *E* the exposure, and *D* the outcome. (See Chapter 13 for a discussion of confounding.) A 1 indicates the presence, and 0 the absence of the factor or the outcome. In the 2 columns on the right side of Table 1.3 are the counterfactual outcomes for the counterfactual exposed and unexposed populations. We have set this example up such that the exposure (vaccination) causes (or would have caused) the disease in 3 individuals (subjects 7, 9, and 11) and prevents (or would have prevented) it in 3 others (subjects 1, 12, and 18). It had no effect in the remaining 14 individuals (their outcome did not change under the counterfactual state). Recall from our discussion of component causes that these effects (if real) could be due to the presence or absence of other component causes.

If we expand our thought process to the population level, we could compare the potential frequency of disease in the population if all of its members were exposed, $p(D_{E+})$, to the potential frequency of disease in that same population if none of its members was exposed $p(D_{E-})$ (recall again that because the subject might be missing a key component cause, not everyone would get the disease if exposed, and that because of other unknown sufficient cause complexes, not everyone would be disease-free if unexposed to a specific factor such as vaccination). We would infer that there is a causal effect in the population if there is a difference in counterfactual means $p(D_{E+})$-$p(D_{E-})\neq 0$. An equivalent measure of causal effect based on relative frequencies is $p(D_{E+})/p(D_{E-})\neq 1$. In our example, since both ΣD_{E+} and ΣD_{E-} equal 10, there is no causal effect in the population. We should note that in the counterfactual (or potential outcome) setting, our inference about cause is made by comparing the potential outcomes in the exact same subjects under different exposure scenarios. Although these population measures are not directly observable, unlike at the individual level, we can estimate them under specific conditions; *ie* through the use of randomisation in the perfect experiment.

In a perfect experiment in which we randomly assign subjects to receive vaccination or not, both the vaccinated and unvaccinated groups exist in the same population at the same time so we can assume that all things are equal except for the fact of vaccination (this might not be the best design for a vaccination study but bear with us for this example). Furthermore, whatever the disease frequency is in each group, we can assume that these would not change if the groups were switched (*ie* assignment to vaccination was reversed—either deliberately or by error). This creates the condition of exchangeability and is the closest estimate we have of the counterfactual outcomes. Thus, in our perfect trial we can calculate a causal effect by contrasting the disease frequency in 2 similar (exchangeable) but different groups of subjects, one of which is exposed and one of which is not. The observed probability (or risk) of the outcome in the exposed is defined as $p(D+|E+)$ and $p(D+|E-)$ in the non-exposed. In our perfect trial, we would observe $p(D+|E+)=p(D+|E-)=0.5$ and conclude that there is no causal effect of vaccination. Although the perfect trial mimics the counterfactual, the observed risks/rates in a real trial would depend on data from two different subsets (the exposed and

Table 1.3 Observed and counterfactual results of an exposure (E) and disease (D)

Subject (i)	Confounder (C)	Actual Exposure (E)	Actual Outcome (D)	Counterfactual Results $D_{E=1}$	Counterfactual Results $D_{E=0}$
1	1	0	1	0	1
2	0	0	1	1	1
3	1	1	1	1	1
4	1	1	0	0	0
5	1	1	1	1	1
6	1	1	1	1	1
7	0	0	0	1	0
8	0	1	1	1	1
9	0	0	0	1	0
10	0	1	0	0	0
11	1	1	1	1	0
12	1	0	1	0	1
13	1	1	1	1	1
14	0	1	0	0	0
15	1	1	1	1	1
16	0	0	0	0	0
17	1	0	0	0	0
18	1	1	0	0	1
19	0	1	0	0	0
20	1	1	0	0	0
Totals	12	13	10	10	10
				$p(D_{E=1}=1)=0.5$	$p(D_{E=0}=1)=0.5$

Observed p(D+|E+)=7/13=0.54

Observed p(D+|E-)=3/7=0.43

	E=1	E=0
p(D+|C+)	6/9=0.67	2/3=0.67
p(D+|C-)	1/4=0.25	1/4=0.25

unexposed subjects) of the target population. Given a lack of compliance, measurement errors, and other biases, our observed disease frequencies in a real trial reflect 'associations' and not necessarily causal effects.

Now, in the absence of a perfect trial, our practical approach to estimating the causal effect is to use an observational study and, in this instance, $p(D+|E+)=7/13$ (0.54) subjects who were actually exposed developed the disease, but $p(D+|E-)=3/7$ (0.43) actually unexposed subjects developed the disease. Thus, the association measure from our observational study does not equal the causal effect in this instance (and in general we will try and remind ourselves that association does not necessarily imply causation throughout this text).

So, what accounts for the fact that the observed risks do not equal the counterfactual risks? And, given this, how do we design studies to obtain data suitable for causal inferences? The problem is that our comparison group ($E-$) is not a good counterfactual group, in that it differs systematically from the $E+$ group in a manner that alters the risk of the outcome. In Table 1.3, we note that in the $E+$ group 9/13=0.69 of the individuals were $C+$, but in the $E-$ group, only 3/7=0.43 were $C+$. Consequently, the groups were not exchangeable. Recall that in a perfect trial with complete randomisation, the confounding factor C would have been distributed equally in the vaccinated and unvaccinated groups, and hence would not bias the disease frequency.

Thus, as shown in this example, a major issue in our use of observational studies is that the exposed and unexposed groups of subjects are rarely exchangeable. The most likely reason for this difference is the presence of other factors that are related to the exposure and the disease (in our example, the $C+$ and $C-$ subjects might have different exposure levels to infection and this impacts on disease occurrence). These factors are called confounders and the phenomenon of confounding will be explored in detail in Chapter 13. The presence or absence of this confounding factor is shown in the second column of Table 1.3. Given these data, we note that among those possessing the confounding factor (*eg* being exposed to a high risk of infection), 75% were exposed (*ie* vaccinated) whereas among those not exposed to a high risk of infection, only 50% were exposed (vaccinated).

We will delay the discussion of techniques to control confounding until Chapter 13, and at this point only introduce the fact that there are a number of ways of trying to ensure that the observed risks would equal the counterfactual risks. One way is to view the problem as trying to uncover the mechanism behind the allocation of exposure, as stated by Rubin (1991). For example, the observed data could have resulted from a controlled experiment where the researchers decided to vaccinate a higher percentage of high risk than low risk study subjects.

This approach leads to the development of propensity scores and proportional weighting of stratum-specific outcome frequencies to obtain unbiased estimates of causal parameters (Hernan and Robins, 2006a). Related to this is the use of standardised risks/rates to adjust for the distribution of exposure in the different levels of the confounder(s) (Sato and Matsuyama, 2003). Other researchers have developed methods to prevent confounding based on 'instrument variables' (Hernan and Robins, 2006b). The traditional analytic approach has involved stratifying the data based on the levels of the confounder(s) and using adjusted measures of association (risk ratios and odds ratios) initially developed by Mantel and Haenszel (1959) (see Chapter 13 for details). Here, we note that the risks of disease are the same in the exposed and unexposed subjects once the subjects are divided into those with a high risk of infection and those with a lower risk of infection. In order to make valid causal inferences, a major

underlying assumption of all of these methods is that there is no residual confounding given the control of (adjustment for) measured confounders—this produces 'exchangeability' within the strata formed by the combinations of measured confounders.

1.8 EXPERIMENTAL VERSUS OBSERVATIONAL EVIDENCE OF CAUSATION

Experimental evidence

Traditionally, the gold standard approach to identifying causal factors is to perform an experiment. In most 2-arm experiments (see Chapter 11), we randomise some people (or other units of concern) to receive the factor and some to receive nothing, a placebo, or a standard intervention (treatment). After a suitable time period, we then assess the outcome in the study subjects and proceed to assess if there are differences in the outcome between the 2 groups. As an alternative design, we might more nearly approach the counterfactual state by using a cross-over design in which subjects are randomly assigned to receive the treatment of interest, or serve as controls, in the first period of the experiment. After a suitable 'wash-out period', the subjects then receive the other level of the treatment (*ie* if they received the treatment in the first period they would receive the placebo in the second and vice-versa). This allows the subject to serve as their own control, as in the counterfactual setting. In both of these experimental designs, the exposure (now denoted as X) explicitly precedes the outcome (denoted as Y) temporally, and all other variables (known and unknown) that do not intervene between X and Y are made independent of X through the process of randomisation (this means that extraneous variables do not confound or bias the results we attribute to the exposure X). This independence of all factors from the treatment X produces exchangeability in the treatment groups; that is, the same outcome would be observed (except for sampling error) if the assignments of treatment to study subjects had been reversed (*ie* if the treated group had been assigned to be untreated). In an experiment, the formal application of randomisation provides the probabilistic basis for the validity of this assumption. Factors that are positioned temporally or causally between X and Y are not measured and are of no concern with respect to answering the causal objective of the trial. In these experimental contexts, exposure X would be a proven cause of outcome Y if the value or state of Y changed following the manipulation of X.

The measure of causation in this ideal trial is called the causal-effect coefficient, and indicates the difference in the outcome between the 'treated' and 'non-treated' groups (*ie* those with different levels of factor X). For example, if the risk of the outcome in the group receiving the treatment is denoted R_1 and the risk in the group not receiving the treatment is R_0, then we might choose to measure the effect of treatment using either an absolute measure (*ie* risk difference—*RD*) or a relative measure (*ie* risk ratio—*RR*) as shown in Chapter 6. If this difference is greater than what could be attributed to chance, then we could say that we have proved that the factor is a cause of the outcome event. A key point is that all causal-effect statements are based on contrasts of outcomes in the different treatment groups; the outcome in the treated group cannot be interpreted without knowing the outcome in the untreated group.

Observational evidence

In observational studies, we estimate the difference in values of Y between subjects that happen to have different values of X. In contrast to the experimental setting, we do not control whether a subject is, or is not, exposed. As we have seen in Table 1.3 measures of association do not necessarily reflect causation. Variables related to both X and Y, and which do not intervene between X and Y, must be controlled to prevent confounding bias and support the estimation of causal effects. The major differences between observational studies and field experiments lie in

the ability to prevent selection, misclassification and confounding bias, and dealing with the impact of unknown or unmeasured factors. Thus, by themselves, observational studies produce measures of association but cannot 'prove' causation. However, in the ideal observational study, with total control of bias, the measure of association will estimate the causal-effect coefficient. Nonetheless, in a given setting, experimental evidence is deemed to provide more solid evidence of causality than observational studies because, in reality, "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)." (Attributed to Box (1966) in Snedecor and Cochran (1980)).

Limits of experimental study evidence

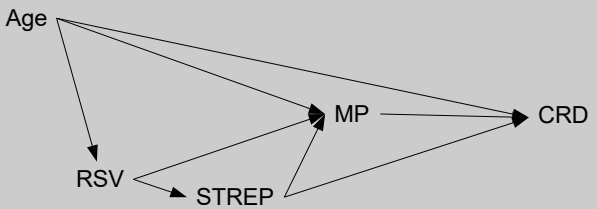
Despite their advantages, performing 'ideal' experiments is not easy even at the best of times (see Chapter 11) and, furthermore, many potential causal factors of interest to epidemiologists would be difficult to study using a controlled-trial format. For example, it would be impossible to perform the perfect experiment to answer the question of whether or not respiratory syncytial virus causes pneumonia in children. Laboratory studies are useful to demonstrate what can happen when animals, which serve as models for pneumonia in children, are exposed to a specific agent (*eg* can agent A cause outcome B), but if the circumstances are too contrived (very large dose, challenge by an unnatural route, limited range of co-factors), laboratory results might not be much help in deciding the issue of causation under normal, everyday conditions for humans.

In addition, in field trials that have an element of non-compliance, we often have to decide how to manage the non-compliance in assessing the role of the treatment on the outcome and, although any given field trial might provide more valid evidence for or against causation than any given observational study, it is not uncommon for differences in results to exist among apparently similar field trials. Hence, despite their advantages, the ability to make perfect inferences based on field trials is illusionary. In addition, in many instances, it is impossible to carry out experiments under conditions that even remotely resemble 'real-world' conditions. Rickles (2009) has discussed the particular limitations of interpreting causal effects when using experimental designs to intervene in complex systems.

1.9 CONSTRUCTING A CAUSAL DIAGRAM

Causal diagrams are helpful for displaying relationships among a number of possible causal variables (age, sex, vaccination status *etc*) that we wish to study) as well as for deducing statistical associations that might arise from a given set of underlying causal relationships. The cause-and-effect relationships and correlations are best shown in a causal diagram (also called directed acyclic graphs, or modified path models). To construct a causal diagram, we begin by imposing a plausible biological causal structure on the set of variables we plan to investigate, and translate this structure into graphical form that explains our hypothesised and known relationships among the variables. The causal-ordering assumption is usually based on known time sequence and/or plausibility considerations. For example, it might be known that one variable precedes another temporally, or current knowledge and/or common sense might suggest that it is possible for one factor to cause another (*ie* alter the risk of another factor), but not vice-versa. We explain the process in Example 1.4, in which we build on our model of pneumonia from Examples 1.1 and 1.2. Causal diagrams are discussed further in Section 13.5.1.

Example 1.4 A causal diagram of factors affecting childhood respiratory disease (CRD)



The easiest way to construct the causal diagram is to begin at the left with variables that are predetermined and progress to the right, listing the variables in their causal order. The variation of these variables (those to the extreme left such as Age in Example 1.4) is considered to be due to factors outside of the model. The remaining variables are placed in the diagram in their presumed causal order; variables to the left could 'cause' the state of variables to their right to change (so our diagram suggests that age could alter the risk of infection with various micro-organisms and hence the risk of CRD. It also suggests that Age can have a direct impact on CRD—now this seems far-fetched, but if there is an additional, but unknown, organism on the pathway between Age and CRD, we might draw the causal arrow directly between the two. (As an aside, this is no different than knowing that citrus fruits prevent scurvy; today we would place vitamin C on the pathway but that would not negate the prior former direct path between citrus fruit and scurvy.) If it is known or strongly believed that a variable does not cause a change in one or more variables to its right, then no causal arrow should be drawn between them. Once completed, if the proposed model is correct, the analyses will not only be more informative about which variables we need to include in our study, it will also provide more powerful analyses than approaches that ignore the underlying structure. The only causal models to be described here are called recursive; that is, there are no causal feedback loops (if these are believed to exist, they can be formulated as a series of causal structures).

Suppose the model is postulated to assess if infectious agents impact on the outcome, specifically pneumonia. In our model, Age is assumed to be a direct cause of CRD, RSV and MP but not STREP. (This means that the risk of RSV and MP change with the age of the child, as does the risk of CRD.) Note that STREP and MP are intervening variables between RSV and the outcome CRD. We will assume that our objective is to estimate the causal effect of RSV on CRD based on the association between these 2 variables.

The model indicates that Age can cause changes in CRD directly and also by a series of pathways involving one or more of the 3 infectious agents. It also indicates that Age is not a direct cause of STREP. In terms of understanding relationships implied by the causal diagram, the easiest way to explain them is to think of getting (perhaps walking) from an exposure variable (*eg* RSV) to a subsequent variable (*eg* CRD). As we pass through variables in the direction of the arrows, we trace out a **causal path**. The rule for tracing the causal pathways is that you can start backwards from any variable but once you start forward on the arrows you cannot back up. Paths which start backwards from a variable are **spurious causal paths** and reflect the impact of confounders. In displaying the relationships, if there are variables that we believe are correlated because of an unknown or unmeasured common cause, we use a non-headed line to indicate this, and you can travel in either direction between these variables. If 2

variables are adjacent (connected by a single direct arrow), their causal relationship is deemed to be **directly causal**. Paths which start forward from one variable and pass through intervening variables to reach the outcome are deemed to be **indirect causal paths** (*eg* RSV can cause CRD through its effect on MP, or on STREP, but not directly). The combined effects through indirect and direct paths represent the **total causal effect** of the variable.

Okay, so, how does this help us? Well, in order to estimate the causal effect, we must prevent any spurious (confounded) effects, so the variables preceding an exposure factor of interest (RSV) that have arrows pointing toward it (*ie* from Age) and through which CRD (the outcome) can be reached on a path must be controlled. In this instance, that variable is Age. The model also asserts that we do not control intervening variables, so STREP and MP are not placed in the statistical model when estimating the causal effect of RSV. If we assume that there are no other confounders that are missing from the model, our analyses will estimate the causal effect of RSV on CRD. (This also assumes the statistical model is correct, but that is another story.)

We should note that if we did control for STREP and MP in this model, we would not obtain the correct estimate of causal effect. Rather, we would only obtain the direct effect of RSV on CRD if that direct effect existed (and in our example no direct effect exists). This feature will be discussed again when regression models (*eg* Chapter 14) are described as this is a major reason why we can inadvertently break down a causal web. In the causal diagram used here, we explicitly assume there is no direct causal relationship between RSV and CRD (so this would be an inappropriate analysis for this reason also). However, RSV can impact on CRD indirectly through the agents STREP and/or MP, and controlling these variables would block these indirect pathways. Thus, only by excluding STREP and MP from our model, and controlling for Stressors, can we obtain the correct causal-effect estimate for RSV.

Greenland and Brumback (2002) discuss relations among causal diagrams, counterfactual models, component-cause models and structural equation (*ie* path) models. Howards *et al* (2007) provide a good discussion on the use of causal diagrams with linkages to appropriate regression models for estimating the associations and examples of causal diagrams based on potential causes of perinatal disease. For more advanced reading, see VanderWeele and Robins (2007a; 2007b). Hernan and Cole (2009) discuss how to describe 4 types of measurement error, as well as confounding and selection bias using causal diagrams.

1.10 CAUSAL CRITERIA

Given that researchers seek to make advances in identifying potential causes of disease using observational study techniques, a number of workers have proposed a set of causal guidelines (these seek to bring uniformity to decisions about causation (Evans, 1995; Susser, 1995). Because these depend on value judgements, we should accept that different individuals might view the same facts differently (Poole, 2001). The recent origin of these guidelines is attributed to Hill (1965) who proposed a list of criteria for making valid causal inferences (not all of which had to be fully met in every instance). These guidelines include: time sequence, strength of association, dose-response, plausibility, consistency, specificity, analogy and experimental evidence. Today, we might add evidence from meta-analysis to this list. Over the years, the first 4 of these have dominated our inference-making efforts (Weed, 2002) and recently, researchers have investigated how we use these and other criteria for making inferences (Waldmann and Hagmayer, 2001). In one study, a group of 135 epidemiologists were given a variety of realistic

but contrived examples and varying amounts of information about each scenario. At the end of the exercise, they had agreed on causal inferences in only 66% of the examples. This stresses the individuality of interpreting the same evidence. Nonetheless, since we think that reference to a set of criteria for causal inferences is a useful aid to decision-making (they provide "a road map through complicated territory" (Rothman *et al*, 2008), we will briefly comment on Hill's list of items and give our view of their role in causal inference.

Doll (2002) describes the application of Hill's guidelines of causation when deducing causation from epidemiological observations. Franco *et al* (2004) provide a good discussion of causal criteria in cancer epidemiology and examples of the collaborative impact between epidemiologists and laboratory scientists. Rothman and Greenland (2005) comment that we should "avoid the temptation to use causal criteria simply to buttress pet theories at hand, and instead .... focus on evaluating competing causal theories using crucial observations". Phillips and Goodman (2006) debate the value of causal criteria, but appear to accept their utility provided they do not degenerate into black-box algorithms which might replace 'scientific common sense'. Hofler (2005a) chooses to interpret Hill's criteria in a counterfactual setting; a setting he had elaborated upon in an earlier paper (2005b). Lipton and Odegaard (2005) suggest that causal expressions are not required for the development of policy to prevent disease and are not as defendable as just stating clearly the methods used to arrive at the statistical association(s) between an exposure and disease. Lash (2007) accepts the utility of causal criteria, but warns that in many instances researchers underestimate the magnitude of systematic errors and uncertainties in their data and fail to fully recognise 'countervailing external information'. Shapiro (2008a; 2008b; 2008c) provide a recent summary of the utility of guidelines for inferring causation. Ward (2009a; 2009b) published an extensive review of the use of causal criteria. He claims that their application does not fully satisfy either deductive or inductive reasoning, but that their application does provide a consistent basis for arriving at the best explanation for the statistical association.

At the outset, we must be clear about the context for inferring causation. As Rose (2001) stated, it is important to ask whether we are trying to identify causes of disease in individuals or causes of disease in populations. Indeed, with the expansion of molecular studies, the appropriate level at which to make causal inferences, and whether such inferences are valid across different levels of organisation remains open to debate. However, clear decisions about the appropriate level to use (think back to the objectives when choosing this) will guide the study design as well as inferences about causation. The following set of criteria for causation can be applied at any level of organisation, and the criteria are based on individual judgement, not a set of defined rules.

1.10.1 Study design and statistical issues

As will be evident after delving into study design (Chapters 7–10), some designs are less open to bias than others. For example, case-control studies are often assumed to be subject to more bias than cohort studies. However, much of this criticism is based on case-control studies using hospital or registry databases. We think it important that every study be assessed on its own merits and we need to be aware of selection, misclassification, and confounding bias in all study designs.

Most often we do not make inferences about causation unless there is a statistically significant association between the exposure and the outcome (and one that is not likely to be explained by

one or more of the previous biases). Certainly, if the differences observed in a well-designed study have P-values above 0.4, this would not provide any support for a causal relationship. However, outside of extremely large P-values, statistical significance should not play a pivotal role in assessing causal relationships. Like other researchers, we suggest an effect-estimation approach based on confidence limits as opposed to a hypothesis-testing approach. Despite this, recent research indicates that P-values continue to be used frequently to guide causal inferences: P-values of 0.04 are assumed to be consistent with causal associations and P-values of 0.06 inconsistent. At the very least, we believe this is an overemphasis of the role of assessing sampling variability *vis-a-vis* a causal association and is not a recommended practice.

1.10.2 Time sequence

While a cause must precede its effect, demonstrating this fact provides only weak support for causation. Further, the same factor could occur after disease in some individuals, and this would not disprove causation except in these specific instances. Many times it is not clear which came first; for example, did the viral infection precede or follow respiratory disease? This becomes a greater problem when we must use surrogate measures of exposure (*eg* antibody titre to indicate recent exposure). Nonetheless, for inferring causation we would like to be able to demonstrate that an exposure preceded the effect, or at least develop a rational argument for believing that it did—sometimes these arguments are based largely on plausibility (*ie* which time sequence is more plausible) rather than on demonstrable facts.

1.10.3 Strength of association

This is usually measured by ratio measures such as risk ratio or odds ratio, but could also be measured by risk or rate differences. The belief in larger (stronger) associations being causal appears to relate to the likelihood that unknown or residual confounding might have produced this effect. However, because the strength of the association also depends on the distribution of other components of a sufficient cause, an association should not be discounted merely because it is weak. Also, when studying diseases with very high frequency, risk ratio measures of association will tend to be weaker than with less common diseases. White (2004) studied the influence of relative prevalence of the agents when making causal inferences about the roles of 2 potential causal factors. It appeared that the agent with the higher prevalence (and in his studies the larger the etiologic fraction (see Section 6.3.1)) was deemed to be more important causally. Hence, some could posit that we should base our judgement more on etiologic fractions than on risk ratios. In further work (White, 2005) it was shown that whereas people do put a lot of weight on what we would call etiologic fractions, they often modify their judgements based on the impact of the second cause when the first is not present, and the judgements appeared to differ when the apparent effect was sparing instead of harmful.

1.10.4 Dose-response relationship

If we had a continuous, or ordinal, exposure variable and the risk of disease increased directly with the level of exposure, then this evidence supports causation as it tends to reduce the likelihood of confounding and is consistent with biological expectations. However, in some instances, there might be a cutpoint of exposure such that nothing happens until a threshold exposure is reached and there is no further increase in frequency at higher levels of exposure.

These circumstances require considerable knowledge about the causal structures for valid inferences. Because certain physiological factors can function to stimulate production of hormones or enzymes at low doses and yet act to reduce production of these at higher levels, one should not be too dogmatic in demanding monotonic relationships.

1.10.5 Coherence or plausibility

The essence of this criterion is that if an association is biologically sensible, it is more likely causal than one that isn't. However, be careful with this line of reasoning. A number of fundamentally important causal inferences have proved to be valid although initially they were dismissed because they did not fit with the current paradigm of disease causation. For example, John Snow's initial suggestion that cholera was 'caused' by bad water was initially met with great scepticism because the prevailing belief was that miasma was the cause of ill-health.

Coherence requires that the observed association is explicable in terms of what we know about disease mechanisms. However, our knowledge is a dynamic state and ranges all the way from the observed association being assessed as 'reasonable' (without any biological supporting evidence) to requiring that 'all the facts be known' (a virtually nonexistent state currently). Postulating a biological mechanism to explain an association after the fact is deemed to be insufficient for causal inferences unless there is some additional evidence supporting the existence of that mechanism.

1.10.6 Consistency

If the same association is found in different studies by different workers, this gives support to causality. Lack of consistency doesn't mean that we should ignore the results of the first study on a topic, but we should temper our interpretation of the results until they are repeated. This would prevent a lot of false positive scares in both human and veterinary medicine. The same approach might be applied to the results of field trials and, because there is less concern over confounding, we might not need to be as strict. Research has indicated that once 12 studies have reached the same essential conclusion, further studies reaching the same conclusion are given little additional weight in making causal inferences (Holman *et al*, 2001).

Meta-analysis is used to combine results from a number of studies on a specific exposure factor in a rigorous, well-defined manner (Weed, 2000) and consequently helps with the evaluation of consistency. Evidence for or against a hypothesis can be obtained as opposed to dichotomising study results into those that support a hypothesis and those that do not. In addition, explanation of the methods used in meta-analysis tends to provide a clearer picture of the reviewer's criteria for causation than many qualitative reviews (see Chapter 28).

1.10.7 Specificity of association

Based on rigid criteria for causation such as Henle-Koch's postulates (Hill, 1965), it used to be thought that, if a factor was associated with only one disease, it was more likely causal than a factor that was associated with numerous disease outcomes. We no longer believe this and specificity, or the lack thereof, has no valid role in assessing causation—the numerous effects of smoking (heart, lungs, infant birth weight, infant intelligence) and the numerous causes for each of these outcomes should be proof enough on this point.

1.10.8 Analogy

This is not a very important criterion for assessing causation, although there are examples of its being used to good purpose. This approach tends to be used to infer relationships in cases of human diseases based on experimental results in other species. Today, many of us have inventive minds and explanations can be developed for almost any observation, so this criterion is not particularly useful to help differentiate between causal and non-causal associations.

1.10.9 Experimental evidence

This criterion perhaps relates partly to biological plausibility and partly to the additional control that is exerted in well-designed experiments. We tend to place more importance on experimental evidence if the same target species is used and the routes of challenge, or nature of the treatment, are in line with what one might expect under field conditions. Experimental evidence from other species in more contrived settings is given less weight in our assessment of causation. Indeed, the experimental approach is just another way to test the hypothesis, so this is not really a distinct criterion for causation in its own right.

Swaen and van Amelsvoort (2009) developed a process for formalising the application of these causal criteria, for assessing the extent to which each criterion was true, and the application of a formal weighting of the criteria using discriminant analysis to estimate the probability that the observed associations between an exposure and an outcome were causal. However, details of this procedure are beyond the scope of this text.

REFERENCES

Aiello AE, Larson EL. Causal inference: the case of hygiene and health. Am J Inf Control. 2002;30(8):503-11.

Berkman LF. Social epidemiology: social determinants of health in the United States: are we losing ground? Ann Rev Public Health. 2009;30:27-41.

Berlivet L. "Association or causation?" The debate on the scientific status of risk factor epidemiology, 1947-c. 1965. Clio medica (Amsterdam, Netherlands). 2005;75:39-74.

Chibuk TK, Robinson JL, Hartfield DS. Pediatric complicated pneumonia and pneumococcal serotype replacement: trends in hospitalized children pre and post introduction of routine vaccination with Pneumococcal Conjugate Vaccine (PCV7). Eur J Pediatrics. 2010;169(9):1123-8.

Cox FE. History of the discovery of the malaria parasites and their vectors. Parasites & vectors. 2010;3(1):5.

De Vreese L. Epidemiology and causation. Medicine, health care, and philosophy. 2009;12(3):345-53.

Diez-Roux AV. On genes, individuals, society, and epidemiology. Am J Epidemiol. 1998a;148(11):1027-32.

Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health. 1998b;88(2):216-22.

Diez Roux AV. Integrating social and biologic factors in health research: a systems view. Annals Epidemiol. 2007;17(7):569-74.

Doll R. Proof of causality: deduction from epidemiological observation. Persp Biol and Med. 2002;45(4):499-515.

Dwyer T, Ponsonby AL. Sudden infant death syndrome and prone sleeping position. Annals Epidemiol. 2009;19(4):245-9.

Evans AS. Causation and disease: a chronological journey. The Thomas Parran Lecture. 1978. Am J Epidemiol. 1995;142(11):1126-35; discussion 5.

Evans RG, Barer ML, Marmor TR. Why Are Some People Healthy and Others Not? The Determinants of Health of Populations. New York: Aldine de Gruyter; 1994.

Fisman DN, Laupland KB. The 'One Health' paradigm: Time for infectious diseases clinicians to take note? Can J Inf Dis & Med Micro. 2010;21(3):111-4.

Flanders WD. On the relationship of sufficient component cause models with potential outcome (counterfactual) models. Europ J Epidemiol. 2006;21(12):847-53.

Franco EL, Correa P, Santella RM, Wu X, Goodman SN, Petersen GM. Role and limitations of epidemiology in establishing a causal association. Seminars in Cancer Biol. 2004;14(6):413-26.

Frerichs RR. Readers respond to "Cholera in Paris". Am J Public Health. 2001;91(8):1170.

Greenland S, Brumback B. An overview of relations among causal modelling methods. Int J Epidemiol. 2002;31(5):1030-7.

Greenland S. Epidemiologic measures and policy formulation: lessons from potential outcomes. Emerging Themes Epidemiol. 2005;2:5.

Hernan MA. A definition of causal effect for epidemiological research. J Epidemiology and Commun Health. 2004;58(4):265-71.

Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol and Comm Health. 2006a;60(7):578-86.

Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiol. 2006b;17(4):360-72.

Hernan MA, Cole SR. Invited Commentary: Causal diagrams and measurement bias. Am J Epidemiol. 2009;170(8):959-62; discussion 63-4.

Hill AB. The Environment and Disease: Association Or Causation? Proc R Soc Med. 1965;58:295-300.

Hofler M. Causal inference based on counterfactuals. BMC Med Res Meth. 2005a;5:28.

Hofler M. The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes Epidemiol. 2005b;2:11.

Hogue CJ. The triangular future of epidemiology. Annals Epidemiol. 2008;18(11):862-4.

Holman CD, Arnold-Reed DE, de Klerk NH, English DR, Donovan RJ. Epidemiologists'

characteristics had little influence on causal inference. Epidemiol. 2001;12(6):752-3.

Howards PP, Schisterman EF, Heagerty PJ. Potential confounding by exposure history and prior outcomes: an example from perinatal epidemiology. Epidemiol. 2007;18(5):544-51.

Kahn LH, Kaplan B, Monath TP, Steele JH. Teaching "one medicine, one health". Am J Med. 2008;121(3):169-70.

Kanarek MS, Anderson HA. Environmental epidemiology practitioners: looking to the future October 11, 2006. Annals Epidemiol. 2007;17(11):911-3.

Kaplan GA. What's wrong with social epidemiology, and how can we make it better? Epidemiol Rev. 2004;26:124-35.

Kaprio J. Science, medicine, and the future. Genetic epidemiology. BMJ (Clin Res). 2000;320(7244):1257-9.

Kaufman JS, Poole C. Looking back on "causal thinking in the health sciences". Ann Rev Public Health. 2000;21:101-19.

Koopman JS, Weed DL. Epigenesis theory: a mathematical model relating causal concepts of pathogenesis in individuals to disease patterns in populations. Am J Epidemiol. 1990;132(2):366-90.

Koopman JS, Lynch JW. Individual causal models and population system models in epidemiology. Am J Public Health. 1999;89(8):1170-4.

Korppi M, Heiskanen-Kosma T, Kleemola M. Mycoplasma pneumoniae causes over 50% of community-acquired pneumonia in school-aged children. Scand J Inf Dis. 2003;35(4):294.

Kottke TE. The need for social epidemiology: now as much as ever. Annals Epidemiol. 2011;21(12):875-6.

Krieger N. Epidemiology and the web of causation: has anyone seen the spider? Social science & medicine. 1994;39(7):887-903.

Lash TL. Heuristic thinking and inference from observational epidemiology. Epidemiol. 2007;18(1):67-72.

Lipton R, Odegaard T. Causal thinking and causal language in epidemiology: it's in the details. Epidemiol Persp & Innov. 2005;2:8.

MacMahon B PTF. Epidemiology: Principles and Methods. Boston: Little Brown; 1970.

Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Nat Cancer Res. 1959;22(4):719-48.

Martin SW, Meek AH, Willeberg P. Veterinary Epidemiology: Principles and Methods. Ames Io: Iowa State Press; 1987.

Martin W. Linking causal concepts, study design, analysis and inference in support of one epidemiology for population health. Prev Vet Med. 2008;86(3-4):270-88.

McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. Am J Epidemiol. 1999;149(10):887-97.

Needleman H. Low level lead exposure: history and discovery. Annals Epidemiol. 2009;19(4):235-8.

Ness RB. Introduction. Triumphs in epidemiology. Annals Epidemiol. 2009;19(4):225.

Ness RB, Andrews EB, Gaudino JA, Jr., Newman AB, Soskolne CL, Sturmer T, et al. The future of epidemiology. Academic Medicine. 2009;84(11):1631-7.

Oakley GP, Jr. The scientific basis for eliminating folic acid-preventable spina bifida: a modern miracle from epidemiology. Annals Epidemiol. 2009;19(4):226-30.

Olsen J. What characterises a useful concept of causation in epidemiology? J Epidemiol and Comm Health. 2003;57(2):86-8.

Parodi A, Neasham D, Vineis P. Environment, population, and biology: a short history of modern epidemiology. Persp Biol and Med. 2006;49(3):357-68.

Pearce N. Analytical implications of epidemiological concepts of interaction. Int J Epidemiol. 1989;18(4):976-80.

Phillips CV, Goodman KJ. Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. Emerging Themes Epidemiol. 2006;3:5.

Poole C. Causal values. Epidemiol. 2001;12(2):139-41.

Rajakumar K. Pellagra in the United States: a historical perspective. Southern Med J. 2000;93(3):272-7.

Reiber C. Evolution for epidemiologists. Annals Epidemiol. 2009;19(4):276-9.

Rickles D. Causality in complex interventions. Medicine, health care, and philosophy. 2009;12(1):77-90.

Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiol. 2001;12(3):313-20.

Rose G. Sick individuals and sick populations. Int J Epidemiol. 2001;30(3):427-32; discussion 33-4.

Rothman KJ. Causes. Am J Epidemiol. 1976;104(6):587-92.

Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health. 2005;95 Suppl 1:S144-50.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd Ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics. 1991;47(4):1213-34.

Sargeant JM. The influence of veterinary epidemiology on public health: past, present and future. Prev Vet Med. 2008;86(3-4):250-9.

Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiol. 2003;14(6):680-6.

Schwabe CW. The current epidemiolgical revolution in veterinary medicine. Part I. Prev Vet Med. 1982;1:5-15.

Schwabe CW. Veterinary Medicine and Human Health. Anonymous B, editor: Williams and Wilkins, Baltimore; 1984.

Schwabe CW. The current epidemiological revolution in veterinary medicine. Part II. Prev Vet Med. 1993;18:3-16.

Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. Part 3: principles of causality in epidemiological research: statistical stability, dose- and duration-response effects, internal and external consistency, analogy and biological plausibility. J Family Planning & Repro Health Care. 2008a;34(4):261-4.

Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy Part 2. Principles of causality in epidemiological research: confounding, effect modification and strength of association. J Family Planning & Repro Health Care. 2008b;34(3):185-90.

Shapiro S. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. Part 1. Principles of causality in epidemiological research: time order, specification of the study base and specificity. J Family Planning & Repro Health Care. 2008c;34(2):83-7.

Snedecor GW, Cochran WG. Statistical Methods; page 356. Ames: Iowa State University Press; 1980.

Steele JH. Veterinary public health: past success, new opportunities. Prev Vet med. 2008;86(3-4):224-43.

Susser M. Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology. Anonymous B, editor: Oxford University Press, Toronto; 1973.

Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol. 1991;133(7):635-48.

Susser M. Judgment and causal inference: criteria in epidemiologic studies. 1977. Am J Epidemiol. 1995;141(8):701-15; discussion 699-700.

Swaen G, van Amelsvoort L. A weight of evidence approach to causal inference. J Clin Epidemiol. 2009;62(3):270-7.

Taubes G. Epidemiology faces its limits. Science. 1995;269(5221):164-9.

Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol. 1991;44(3):221-32.

VanderWeele TJ, Hernan MA. From counterfactuals to sufficient component causes and vice versa. Europ J Epidemiol. 2006;21(12):855-8.

VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. Am J Epidemiol. 2007a;166(9):1096-104.

VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiol. 2007b;18(5):561-8.

VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-

cause framework. Epidemiol. 2007c;18(3):329-39.

VanderWeele TJ, Hernan MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias. Epidemiology. 2008;19(5):720-8.

Vineis P, Kriebel D. Causal models in epidemiology: past inheritance and genetic future. Environ Health. 2006;5:21.

Waldmann MR, Hagmayer Y. Estimating causal strength: the role of structural knowledge and processing effort. Cognition. 2001;82(1):27-58.

Ward A. Causal criteria and the problem of complex causation. Medicine, health care, and philosophy. 2009a;12(3):333-43.

Ward AC. The role of causal criteria in causal inferences: Bradford Hill's "aspects of association". Epidemiol Persp & Innov. 2009b;6:2.

Weed DL. Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. Int J Epidemiol. 2000;29(3):387-90.

Weed DL. Environmental epidemiology: basics and proof of cause-effect. Toxicology. 2002;181-182:399-403.

Weed DL. Commentary: rethinking epidemiology. Int J Epidemiol. 2006;35(3):583-6; discussion 93-6.

White P. Judgement of two causal candidates from contingency information: effects of relative prevalence of the two causes. The Quarterly J Exp Psychology. 2004;57(6):961-91.

White PA. Causal attribution and Mill's methods of experimental inquiry: past, present and prospect. Brit J Social Psychology. 2000;39 (Pt 3)(Pt 3):429-47.

White PA. Causal judgments about relations between multilevel variables. J Exp Psychology. 2001;27(2):499-513.

White PA. Judgement of two causal candidates from contingency information: II. Effects of information about one cause on judgements of the other cause. The Quarterly J Exp Psychology. 2005;58(6):999-1021.

Willett W, Greenland S, MacMahon B, Trichopoulos D, Rothman K, Thomas D, et al. The discipline of epidemiology. Science. 1995;269(5229):1325-6.