PHW250B Week 2 Reader

Topic 1: Estimating Causal Effects

Topic 2: Directed acyclic graphs

Modern Epidemiology. 3rd Chapter 2

CHAPTER 2

Causation and Causal Inference

Kenneth J. Rothman, Sander Greenland,
Charles Poole, and Timothy L. Lash

CAUSALITY

A rudimentary understanding of cause and effect seems to be acquired by most people on their own much earlier than it could have been taught to them by someone else. Even before they can speak, many youngsters understand the relation between crying and the appearance of a parent or other adult, and the relation between that appearance and getting held, or fed. A little later, they will develop theories about what happens when a glass containing milk is dropped or turned over, and what happens when a switch on the wall is pushed from one of its resting positions to another. While theories such as these are being formulated, a more general causal theory is also being formed. The more general theory posits that some events or states of nature are causes of specific effects. Without a general theory of causation, there would be no skeleton on which to hang the substance of the many specific causal theories that one needs to survive.

Nonetheless, the concepts of causation that are established early in life are too primitive to serve well as the basis for scientific theories. This shortcoming may be especially true in the health and social sciences, in which typical causes are neither necessary nor sufficient to bring about effects of interest. Hence, as has long been recognized in epidemiology, there is a need to develop a more refined conceptual model that can serve as a starting point in discussions of causation. In particular, such a model should address problems of multifactorial causation, confounding, interdependence of effects, direct and indirect effects, levels of causation, and systems or webs of causation (MacMahon and Pugh, 1967; Susser, 1973). This chapter describes one starting point, the sufficient-component cause model (or sufficient-cause model), which has proven useful in elucidating certain concepts in individual mechanisms of causation. Chapter 4 introduces the widely used potential-outcome or counterfactual model of causation, which is useful for relating individual-level to population-level causation, whereas Chapter 12 introduces graphical causal models (causal diagrams), which are especially useful for modeling causal systems.

Except where specified otherwise (in particular, in Chapter 27, on infectious disease), throughout the book we will assume that disease refers to a nonrecurrent event, such as death or first occurrence of a disease, and that the outcome of each individual or unit of study (e.g., a group of persons) is not affected by the exposures and outcomes of other individuals or units. Although this assumption will greatly simplify our discussion and is reasonable in many applications, it does not apply to contagious phenomena, such as transmissible behaviors and diseases. Nonetheless, all the definitions and most of the points we make (especially regarding validity) apply more generally. It is also essential to understand simpler situations before tackling the complexities created by causal interdependence of individuals or units.

A MODEL OF SUFFICIENT CAUSE AND COMPONENT CAUSES

To begin, we need to define *cause*. One definition of the cause of a specific disease occurrence is an antecedent event, condition, or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are fixed. In other words, a cause of a disease occurrence is an event, condition, or characteristic that preceded the disease onset and that, had the event, condition, or characteristic been different in a specified way, the disease either would not have occurred at all or would not have occurred until some later time. Under this definition, if someone walking along an icy path falls and breaks a hip, there may be a long list of causes. These causes might include the weather on the day of the incident, the fact that the path was not cleared for pedestrians, the choice of footgear for the victim, the lack of a handrail, and so forth. The constellation of causes required for this particular person to break her hip at this particular time can be depicted with the sufficient cause diagrammed in Figure 2–1. By *sufficient cause* we mean a complete causal mechanism, a minimal set of conditions and events that are sufficient for the outcome to occur. The circle in the figure comprises five segments, each of which represents a causal component that must be present or have occured in order for the person to break her hip at that instant. The first component, labeled A, represents poor weather. The second component, labeled B, represents an uncleared path for pedestrians. The third component, labeled C, represents a poor choice of footgear. The fourth component, labeled D, represents the lack of a handrail. The final component, labeled U, represents all of the other unspecified events, conditions, and characteristics that must be present or have occured at the instance of the fall that led to a broken hip. For etiologic effects such as the causation of disease, many and possibly all of the components of a sufficient cause may be unknown (Rothman, 1976a). We usually include one component cause, labeled U, to represent the set of unknown factors.

All of the component causes in the sufficient cause are required and must be present or have occured at the instance of the fall for the person to break a hip. None is superfluous, which means that blocking the contribution of any component cause prevents the sufficient cause from acting. For many people, early causal thinking persists in attempts to find single causes as explanations for observed phenomena. But experience and reasoning show that the causal mechanism for any effect must consist of a constellation of components that act in concert (Mill, 1862; Mackie, 1965). In disease etiology, a sufficient cause is a set of conditions sufficient to ensure that the outcome will occur. Therefore, completing a sufficient cause is tantamount to the onset of disease. Onset here may refer to the onset of the earliest stage of the disease process or to any transition from one well-defined and readily characterized stage to the next, such as the onset of signs or symptoms.

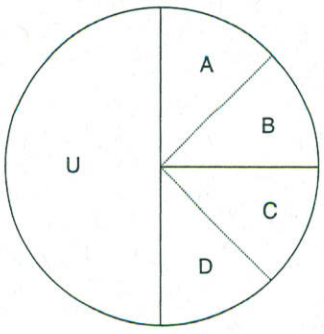


FIGURE 2–1 • Depiction of the constellation of component causes that constitute a sufficient cause for hip fracture for a particular person at a particular time. In the diagram, A represents poor weather, B represents an uncleared path for pedestrians, C represents a poor choice of footgear, D represents the lack of a handrail, and U represents all of the other unspecified events, conditions, and characteristics that must be present or must have occured at the instance of the fall that led to a broken hip.

Consider again the role of the handrail in causing hip fracture. The absence of such a handrail may play a causal role in some sufficient causes but not in others, depending on circumstances such as the weather, the level of inebriation of the pedestrian, and countless other factors. Our definition links the lack of a handrail with this one broken hip and does not imply that the lack of this handrail by itself was sufficient for that hip fracture to occur. With this definition of cause, no specific event, condition, or characteristic is sufficient by itself to produce disease. The definition does not describe a complete causal mechanism, but only a component of it. To say that the absence of a handrail is a component cause of a broken hip does not, however, imply that every person walking down the path will break a hip. Nor does it imply that if a handrail is installed with properties sufficient to prevent that broken hip, that no one will break a hip on that same path. There may be other sufficient causes by which a person could suffer a hip fracture. Each such sufficient cause would be depicted by its own diagram similar to Figure 2–1. The first of these sufficient causes to be completed by simultaneous accumulation of all of its component causes will be the one that depicts the mechanism by which the hip fracture occurs for a particular person. If no sufficient cause is completed while a person passes along the path, then no hip fracture will occur over the course of that walk.

As noted above, a characteristic of the naive concept of causation is the assumption of a one-to-one correspondence between the observed cause and effect. Under this view, each cause is seen as "necessary" and "sufficient" in itself to produce the effect, particularly when the cause is an observable action or event that takes place near in time to the effect. Thus, the flick of a switch appears to be the singular cause that makes an electric light go on. There are less evident causes, however, that also operate to produce the effect: a working bulb in the light fixture, intact wiring from the switch to the bulb, and voltage to produce a current when the circuit is closed. To achieve the effect of turning on the light, each of these components is as important as moving the switch, because changing any of these components of the causal constellation will prevent the effect. The term *necessary cause* is therefore reserved for a particular type of component cause under the sufficient-cause model. If any of the component causes appears in every sufficient cause, then that component cause is called a "necessary" component cause. For the disease to occur, any and all necessary component causes must be present or must have occurred. For example, one could label a component cause with the requirement that one must have a hip to suffer a hip fracture. Every sufficient cause that leads to hip fracture must have that component cause present, because in order to fracture a hip, one must have a hip to fracture.

The concept of complementary component causes will be useful in applications to epidemiology that follow. For each component cause in a sufficient cause, the set of the other component causes in that sufficient cause comprises the complementary component causes. For example, in Figure 2–1, component cause A (poor weather) has as its complementary component causes the components labeled B, C, D, and U. Component cause B (an uncleared path for pedestrians) has as its complementary component causes the components labeled A, C, D, and U.

THE NEED FOR A SPECIFIC REFERENCE CONDITION

Component causes must be defined with respect to a clearly specified alternative or reference condition (often called a *referent*). Consider again the lack of a handrail along the path. To say that this condition is a component cause of the broken hip, we have to specify an alternative condition against which to contrast the cause. The mere presence of a handrail would not suffice. After all, the hip fracture might still have occurred in the presence of a handrail, if the handrail was too short or if it was old and made of rotten wood. We might need to specify the presence of a handrail sufficiently tall and sturdy to break the fall for the absence of that handrail to be a component cause of the broken hip.

To see the necessity of specifying the alternative event, condition, or characteristic as well as the causal one, consider an example of a man who took high doses of ibuprofen for several years and developed a gastric ulcer. Did the man's use of ibuprofen cause his ulcer? One might at first assume that the natural contrast would be with what would have happened had he taken nothing instead of ibuprofen. Given a strong reason to take the ibuprofen, however, that alternative may not make sense. If the specified alternative to taking ibuprofen is to take acetaminophen, a different drug that might have been indicated for his problem, and if he would not have developed the ulcer had he used acetaminophen, then we can say that using ibuprofen caused the ulcer. But ibuprofen did not cause

his ulcer if the specified alternative is taking aspirin and, had he taken aspirin, he still would have developed the ulcer. The need to specify the alternative to a preventive is illustrated by a newspaper headline that read: "Rare Meat Cuts Colon Cancer Risk." Was this a story of an epidemiologic study comparing the colon cancer rate of a group of people who ate rare red meat with the rate in a group of vegetarians? No, the study compared persons who ate rare red meat with persons who ate highly cooked red meat. The same exposure, regular consumption of rare red meat, might have a preventive effect when contrasted against highly cooked red meat and a causative effect or no effect in contrast to a vegetarian diet. An event, condition, or characteristic is not a cause by itself as an intrinsic property it possesses in isolation, but as part of a causal contrast with an alternative event, condition, or characteristic (Lewis, 1973; Rubin, 1974; Greenland et al., 1999a; Maldonado and Greenland, 2002; see Chapter 4).

APPLICATION OF THE SUFFICIENT-CAUSE MODEL TO EPIDEMIOLOGY

The preceding introduction to concepts of sufficient causes and component causes provides the lexicon for application of the model to epidemiology. For example, tobacco smoking is a cause of lung cancer, but by itself it is not a sufficient cause, as demonstrated by the fact that most smokers do not get lung cancer. First, the term *smoking* is too imprecise to be useful beyond casual description. One must specify the type of smoke (e.g., cigarette, cigar, pipe, or environmental), whether it is filtered or unfiltered, the manner and frequency of inhalation, the age at initiation of smoking, and the duration of smoking. And, however smoking is defined, its alternative needs to be defined as well. Is it smoking nothing at all, smoking less, smoking something else? Equally important, even if smoking and its alternative are both defined explicitly, smoking will not cause cancer in everyone. So who is susceptible to this smoking effect? Or, to put it in other terms, what are the other components of the causal constellation that act with smoking to produce lung cancer in this contrast?

Figure 2–2 provides a schematic diagram of three sufficient causes that could be completed during the follow-up of an individual. The three conditions or events—A, B, and E—have been defined as binary variables, so they can only take on values of 0 or 1. With the coding of A used in the figure, its reference level, $A = 0$, is sometimes causative, but its index level, $A = 1$, is never causative. This situation arises because two sufficient causes contain a component cause labeled "$A = 0$," but no sufficient cause contains a component cause labeled "$A = 1$." An example of a condition or event of this sort might be $A = 1$ for taking a daily multivitamin supplement and $A = 0$ for taking no vitamin supplement. With the coding of B and E used in the example depicted by Figure 2–2, their index levels, $B = 1$ and $E = 1$, are sometimes causative, but their reference levels, $B = 0$ and $E = 0$, are never causative. For each variable, the index and reference levels may represent only two alternative states or events out of many possibilities. Thus, the coding of B might be $B = 1$ for smoking 20 cigarettes per day for 40 years and $B = 0$ for smoking 20 cigarettes per day for 20 years, followed by 20 years of not smoking. E might be coded $E = 1$ for living in an urban neighborhood with low average income and high income inequality, and $E = 0$ for living in an urban neighborhood with high average income and low income inequality.

$A = 0$, $B = 1$, and $E = 1$ are individual component causes of the sufficient causes in Figure 2–2. U_1, U_2, and U_3 represent sets of component causes. U_1, for example, is the set of all components other than $A = 0$ and $B = 1$ required to complete the first sufficient cause in Figure 2–2. If we decided not to specify $B = 1$, then $B = 1$ would become part of the set of components that are causally complementary to $A = 0$; in other words, $B = 1$ would then be absorbed into U_1.

Each of the three sufficient causes represented in Figure 2–2 is minimally sufficient to produce the disease in the individual. That is, only one of these mechanisms needs to be completed for

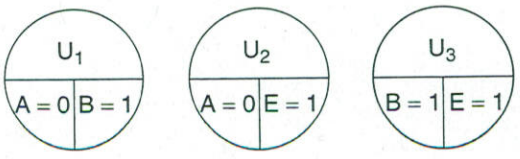


FIGURE 2–2 • Three classes of sufficient causes of a disease (sufficient causes I, II, and III from left to right).

disease to occur (sufficiency), and there is no superfluous component cause in any mechanism (minimality)—each component is a required part of that specific causal mechanism. A specific component cause may play a role in one, several, or all of the causal mechanisms. As noted earlier, a component cause that appears in all sufficient causes is called a *necessary* cause of the outcome. As an example, infection with HIV is a component of every sufficient cause of acquired immune deficiency syndrome (AIDS) and hence is a necessary cause of AIDS. It has been suggested that such causes be called "universally necessary," in recognition that every component of a sufficient cause is necessary for that sufficient cause (mechanism) to operate (Poole 2001a).

Figure 2–2 does not depict aspects of the causal process such as sequence or timing of action of the component causes, dose, or other complexities. These can be specified in the description of the contrast of index and reference conditions that defines each component cause. Thus, if the outcome is lung cancer and the factor B represents cigarette smoking, it might be defined more explicitly as smoking at least 20 cigarettes a day of unfiltered cigarettes for at least 40 years beginning at age 20 years or earlier (B = 1), or smoking 20 cigarettes a day of unfiltered cigarettes, beginning at age 20 years or earlier, and then smoking no cigarettes for the next 20 years (B = 0).

In specifying a component cause, the two sides of the causal contrast of which it is composed should be defined with an eye to realistic choices or options. If prescribing a placebo is not a realistic therapeutic option, a causal contrast between a new treatment and a placebo in a clinical trial may be questioned for its dubious relevance to medical practice. In a similar fashion, before saying that oral contraceptives increase the risk of death over 10 years (e.g., through myocardial infarction or stroke), we must consider the alternative to taking oral contraceptives. If it involves getting pregnant, then the risk of death attendant to childbirth might be greater than the risk from oral contraceptives, making oral contraceptives a preventive rather than a cause. If the alternative is an equally effective contraceptive without serious side effects, then oral contraceptives may be described as a cause of death.

To understand prevention in the sufficient-component cause framework, we posit that the alternative condition (in which a component cause is absent) prevents the outcome relative to the presence of the component cause. Thus, a preventive effect of a factor is represented by specifying its causative alternative as a component cause. An example is the presence of A = 0 as a component cause in the first two sufficient causes shown in Figure 2–2. Another example would be to define a variable, F (not depicted in Fig. 2–2), as "vaccination (F = 1) or no vaccination (F = 0)". Prevention of the disease by getting vaccinated (F = 1) would be expressed in the sufficient-component cause model as causation of the disease by not getting vaccinated (F = 0). This depiction is unproblematic because, once both sides of a causal contrast have been specified, causation and prevention are merely two sides of the same coin.

Sheps (1958) once asked, "Shall we count the living or the dead?" Death is an event, but survival is not. Hence, to use the sufficient-component cause model, we must count the dead. This model restriction can have substantive implications. For instance, some measures and formulas approximate others only when the outcome is rare. When survival is rare, death is common. In that case, use of the sufficient-component cause model to inform the analysis will prevent us from taking advantage of the rare-outcome approximations.

Similarly, etiologies of adverse health outcomes that are conditions or states, but not events, must be depicted under the sufficient-cause model by reversing the coding of the outcome. Consider spina bifida, which is the failure of the neural tube to close fully during gestation. There is no point in time at which spina bifida may be said to have occurred. It would be awkward to define the "incidence time" of spina bifida as the gestational age at which complete neural tube closure ordinarily occurs. The sufficient-component cause model would be better suited in this case to defining the event of complete closure (no spina bifida) as the outcome and to view conditions, events, and characteristics that prevent this beneficial event as the causes of the adverse condition of spina bifida.

PROBABILITY, RISK, AND CAUSES

In everyday language, "risk" is often used as a synonym for probability. It is also commonly used as a synonym for "hazard," as in, "Living near a nuclear power plant is a risk you should avoid." Unfortunately, in epidemiologic parlance, even in the scholarly literature, "risk" is frequently used for many distinct concepts: rate, rate ratio, risk ratio, incidence odds, prevalence, etc. The more

specific, and therefore more useful, definition of *risk* is "probability of an event during a specified period of time."

The term *probability* has multiple meanings. One is that it is the relative frequency of an event. Another is that probability is the tendency, or propensity, of an entity to produce an event. A third meaning is that probability measures someone's degree of certainty that an event will occur. When one says "the probability of death in vehicular accidents when traveling >120 km/h is high," one means that the proportion of accidents that end with deaths is higher when they involve vehicles traveling >120 km/h than when they involve vehicles traveling at lower speeds (frequency usage), that high-speed accidents have a greater tendency than lower-speed accidents to result in deaths (propensity usage), or that the speaker is more certain that a death will occur in a high-speed accident than in a lower-speed accident (certainty usage).

The frequency usage of "probability" and "risk," unlike the propensity and certainty usages, admits no meaning to the notion of "risk" for an individual beyond the relative frequency of 100% if the event occurs and 0% if it does not. This restriction of individual risks to 0 or 1 can only be relaxed to allow values in between by reinterpreting such statements as the frequency with which the outcome would be seen upon random sampling from a very large population of individuals deemed to be "like" the individual in some way (e.g., of the same age, sex, and smoking history). If one accepts this interpretation, whether any actual sampling has been conducted or not, the notion of individual risk is replaced by the notion of the frequency of the event in question in the large population from which the individual was sampled. With this view of risk, a risk will change according to how we group individuals together to evaluate frequencies. Subjective judgment will inevitably enter into the picture in deciding which characteristics to use for grouping. For instance, should tomato consumption be taken into account in defining the class of men who are "like" a given man for purposes of determining his risk of a diagnosis of prostate cancer between his 60th and 70th birthdays? If so, which study or meta-analysis should be used to factor in this piece of information?

Unless we have found a set of conditions and events in which the disease does not occur at all, it is always a reasonable working hypothesis that, no matter how much is known about the etiology of a disease, some causal components remain unknown. We may be inclined to assign an equal risk to all individuals whose status for some components is known and identical. We may say, for example, that men who are heavy cigarette smokers have approximately a 10% lifetime risk of developing lung cancer. Some interpret this statement to mean that all men would be subject to a 10% probability of lung cancer if they were to become heavy smokers, as if the occurrence of lung cancer, aside from smoking, were purely a matter of chance. This view is untenable. A probability may be 10% conditional on one piece of information and higher or lower than 10% if we condition on other relevant information as well. For instance, men who are heavy cigarette smokers and who worked for many years in occupations with historically high levels of exposure to airborne asbestos fibers would be said to have a lifetime lung cancer risk appreciably higher than 10%.

Regardless of whether we interpret probability as relative frequency or degree of certainty, the assignment of equal risks merely reflects the particular grouping. In our ignorance, the best we can do in assessing risk is to classify people according to measured risk indicators and then assign the average risk observed within a class to persons within the class. As knowledge or specification of additional risk indicators expands, the risk estimates assigned to people will depart from average according to the presence or absence of other factors that predict the outcome.

STRENGTH OF EFFECTS

The causal model exemplified by Figure 2–2 can facilitate an understanding of some key concepts such as *strength of effect* and *interaction*. As an illustration of strength of effect, Table 2–1 displays the frequency of the eight possible patterns for exposure to A, B, and E in two hypothetical populations. Now the pie charts in Figure 2–2 depict classes of mechanisms. The first one, for instance, represents all sufficient causes that, no matter what other component causes they may contain, have in common the fact that they contain $A = 0$ and $B = 1$. The constituents of U_1 may, and ordinarily would, differ from individual to individual. For simplification, we shall suppose, rather unrealistically, that U_1, U_2, and U_3 are always present or have always occured for everyone and Figure 2–2 represents all the sufficient causes.

TABLE 2–1

Exposure Frequencies and Individual Risks in Two Hypothetical Populations According to the Possible Combinations of the Three Specified Component Causes in Fig. 2–2

Exposures					Frequency of Exposure Pattern	
A	B	E	Sufficient Cause Completed	Risk	Population 1	Population 2
1	1	1	III	1	900	100
1	1	0	None	0	900	100
1	0	1	None	0	100	900
1	0	0	None	0	100	900
0	1	1	I, II, or III	1	100	900
0	1	0	I	1	100	900
0	0	1	II	1	900	100
0	0	0	none	0	900	100

Under these assumptions, the response of each individual to the exposure pattern in a given row can be found in the response column. The response here is the risk of developing a disease over a specified time period that is the same for all individuals. For simplification, a deterministic model of risk is employed, such that individual risks can equal only the value 0 or 1, and no values in between. A stochastic model of individual risk would relax this restriction and allow individual risks to lie between 0 and 1.

The proportion getting disease, or incidence proportion, in any subpopulation in Table 2–1 can be found by summing the number of persons at each exposure pattern with an individual risk of 1 and dividing this total by the subpopulation size. For example, if exposure A is not considered (e.g., if it were not measured), the pattern of incidence proportions in population 1 would be those in Table 2–2.

As an example of how the proportions in Table 2–2 were calculated, let us review how the incidence proportion among persons in population 1 with $B = 1$ and $E = 0$ was calculated: There were 900 persons with $A = 1$, $B = 1$, and $E = 0$, none of whom became cases because there are no sufficient causes that can culminate in the occurrence of the disease over the study period in persons with this combination of exposure conditions. (There are two sufficient causes that contain $B = 1$ as a component cause, but one of them contains the component cause $A = 0$ and the other contains the component cause $E = 1$. The presence of $A = 1$ or $E = 0$ blocks these etiologic mechanisms.) There were 100 persons with $A = 0$, $B = 1$, and $E = 0$, all of whom became cases because they all had U_1, the set of causal complements for the class of sufficient causes containing $A = 0$ and

TABLE 2–2

Incidence Proportions (IP) for Combinations of Component Causes B and E in Hypothetical Population 1, Assuming That Component Cause A Is Unmeasured

	B = 1, E = 1	B = 1, E = 0	B = 0, E = 1	B = 0, E = 0
Cases	1,000	100	900	0
Total	1,000	1,000	1,000	1,000
IP	1.00	0.10	0.90	0.00

TABLE 2–3

Incidence Proportions (IP) for Combinations of Component Causes B and E in Hypothetical Population 2, Assuming That Component Cause A Is Unmeasured

	B = 1, E = 1	B = 1, E = 0	B = 0, E = 1	B = 0, E = 0
Cases	1,000	900	100	0
Total	1,000	1,000	1,000	1,000
IP	1.00	0.90	0.10	0.00

B = 1. Thus, among all 1,000 persons with B = 1 and E = 0, there were 100 cases, for an incidence proportion of 0.10.

If we were to measure strength of effect by the difference of the incidence proportions, it is evident from Table 2–2 that for population 1, E = 1 has a much stronger effect than B = 1, because E = 1 increases the incidence proportion by 0.9 (in both levels of B), whereas B = 1 increases the incidence proportion by only 0.1 (in both levels of E). Table 2–3 shows the analogous results for population 2. Although the members of this population have exactly the same causal mechanisms operating within them as do the members of population 1, the relative strengths of causative factors E = 1 and B = 1 are reversed, again using the incidence proportion difference as the measure of strength. B = 1 now has a much stronger effect on the incidence proportion than E = 1, despite the fact that A, B, and E have no association with one another in either population, and their index levels (A = 1, B = 1 and E = 1) and reference levels (A = 0, B = 0, and E = 0) are each present or have occured in exactly half of each population.

The overall difference of incidence proportions contrasting E = 1 with E = 0 is (1,900/2,000) − (100/2,000) = 0.9 in population 1 and (1,100/2,000) − (900/2,000) = 0.1 in population 2. The key difference between populations 1 and 2 is the difference in the prevalence of the conditions under which E = 1 acts to increase risk: that is, the presence of A = 0 or B = 1, but not both. (When A = 0 *and* B = 1, E = 1 completes all three sufficient causes in Figure 2–2; it thus does not increase anyone's risk, although it may well shorten the time to the outcome.) The prevalence of the condition, "A = 0 or B = 1 but not both" is 1,800/2,000 = 90% in both levels of E in population 1. In population 2, this prevalence is only 200/2,000 = 10% in both levels of E. This difference in the prevalence of the conditions sufficient for E = 1 to increase risk explains the difference in the strength of the effect of E = 1 as measured by the difference in incidence proportions.

As noted above, the set of all other component causes in all sufficient causes in which a causal factor participates is called the *causal complement* of the factor. Thus, A = 0, B = 1, U_2, and U_3 make up the causal complement of E = 1 in the above example. This example shows that the strength of a factor's effect on the occurrence of a disease in a population, measured as the absolute difference in incidence proportions, depends on the prevalence of its causal complement. This dependence has nothing to do with the etiologic mechanism of the component's action, because the component is an equal partner in each mechanism in which it appears. Nevertheless, a factor will appear to have a strong effect, as measured by the difference of proportions getting disease, if its causal complement is common. Conversely, a factor with a rare causal complement will appear to have a weak effect.

If strength of effect is measured by the ratio of proportions getting disease, as opposed to the difference, then strength depends on more than a factor's causal complement. In particular, it depends additionally on how common or rare the components are of sufficient causes in which the specified causal factor does *not* play a role. In this example, given the ubiquity of U_1, the effect of E = 1 measured in ratio terms depends on the prevalence of E = 1's causal complement and on the prevalence of the conjunction of A = 0 and B = 1. If many people have both A = 0 and B = 1, the "baseline" incidence proportion (i.e., the proportion of not-E or "unexposed" persons getting disease) will be high and the proportion getting disease due to E will be comparatively low. If few

people have both A = 0 and B = 1, the baseline incidence proportion will be low and the proportion getting disease due to E = 1 will be comparatively high. Thus, strength of effect measured by the incidence proportion ratio depends on more conditions than does strength of effect measured by the incidence proportion difference.

Regardless of how strength of a causal factor's effect is measured, the public health significance of that effect does not imply a corresponding degree of etiologic significance. Each component cause in a given sufficient cause has the same etiologic significance. Given a specific causal mechanism, any of the component causes can have strong or weak effects using either the difference or ratio measure. The actual identities of the components of a sufficient cause are part of the mechanics of causation, whereas the strength of a factor's effect depends on the time-specific distribution of its causal complement (if strength is measured in absolute terms) plus the distribution of the components of all sufficient causes in which the factor does not play a role (if strength is measured in relative terms). Over a span of time, the strength of the effect of a given factor on disease occurrence may change because the prevalence of its causal complement in various mechanisms may also change, even if the causal mechanisms in which the factor and its cofactors act remain unchanged.

INTERACTION AMONG CAUSES

Two component causes acting in the same sufficient cause may be defined as *interacting causally* to produce disease. This definition leaves open many possible mechanisms for the interaction, including those in which two components interact in a direct physical fashion (e.g., two drugs that react to form a toxic by-product) and those in which one component (the *initiator* of the pair) alters a substrate so that the other component (the *promoter* of the pair) can act. Nonetheless, it excludes any situation in which one component E is merely a cause of another component F, with no effect of E on disease except through the component F it causes.

Acting in the same sufficient cause is not the same as one component cause acting to produce a second component cause, and then the second component going on to produce the disease (Robins and Greenland 1992, Kaufman et al., 2004). As an example of the distinction, if cigarette smoking (vs. never smoking) is a component cause of atherosclerosis, and atherosclerosis (vs. no atherosclerosis) causes myocardial infarction, both smoking and atherosclerosis would be component causes (cofactors) in certain sufficient causes of myocardial infarction. They would not necessarily appear in the same sufficient cause. Rather, for a sufficient cause involving atherosclerosis as a component cause, there would be another sufficient cause in which the atherosclerosis component cause was replaced by all the component causes that brought about the atherosclerosis, including smoking. Thus, a sequential causal relation between smoking and atherosclerosis would not be enough for them to interact synergistically in the etiology of myocardial infarction, in the sufficient-cause sense. Instead, the causal sequence means that smoking can act indirectly, through atherosclerosis, to bring about myocardial infarction.

Now suppose that, perhaps in addition to the above mechanism, smoking reduces clotting time and thus causes thrombi that block the coronary arteries if they are narrowed by atherosclerosis. This mechanism would be represented by a sufficient cause containing both smoking and atherosclerosis as components and thus would constitute a synergistic interaction between smoking and atherosclerosis in causing myocardial infarction. The presence of this sufficient cause would not, however, tell us whether smoking also contributed to the myocardial infarction by causing the atherosclerosis. Thus, the basic sufficient-cause model does not alert us to indirect effects (effects of some component causes mediated by other component causes in the model). Chapters 4 and 12 introduce potential-outcome and graphical models better suited to displaying indirect effects and more general sequential mechanisms, whereas Chapter 5 discusses in detail interaction as defined in the potential-outcome framework and its relation to interaction as defined in the sufficient-cause model.

PROPORTION OF DISEASE DUE TO SPECIFIC CAUSES

In Figure 2–2, assuming that the three sufficient causes in the diagram are the only ones operating, what fraction of disease is caused by E = 1? E = 1 is a component cause of disease in two of the sufficient-cause mechanisms, II and III, so all disease arising through either of these two mechanisms is attributable to E = 1. Note that in persons with the exposure pattern A = 0, B = 1, E = 1, all three

sufficient causes would be completed. The first of the three mechanisms to be completed would be the one that actually produces a given case. If the first one completed is mechanism II or III, the case would be causally attributable to E = 1. If mechanism I is the first one to be completed, however, E = 1 would not be part of the sufficient cause producing that case. Without knowing the completion times of the three mechanisms, among persons with the exposure pattern A = 0, B = 1, E = 1 we cannot tell how many of the 100 cases in population 1 or the 900 cases in population 2 are etiologically attributable to E = 1.

Each of the cases that is etiologically attributable to E = 1 can also be attributed to the other component causes in the causal mechanisms in which E = 1 acts. Each component cause interacts with its complementary factors to produce disease, so each case of disease can be attributed to every component cause in the completed sufficient cause. Note, though, that the attributable fractions added across component causes of the same disease do not sum to 1, although there is a mistaken tendency to think that they do. To illustrate the mistake in this tendency, note that a necessary component cause appears in every completed sufficient cause of disease, and so by itself has an attributable fraction of 1, without counting the attributable fractions for other component causes. Because every case of disease can be attributed to every component cause in its causal mechanism, attributable fractions for different component causes will generally sum to more than 1, and there is no upper limit for this sum.

A recent debate regarding the proportion of risk factors for coronary heart disease attributable to particular component causes illustrates the type of errors in inference that can arise when the sum is thought to be restricted to 1. The debate centers around whether the proportion of coronary heart disease attributable to high blood cholesterol, high blood pressure, and cigarette smoking equals 75% or "only 50%" (Magnus and Beaglehole, 2001). If the former, then some have argued that the search for additional causes would be of limited utility (Beaglehole and Magnus, 2002), because only 25% of cases "remain to be explained." By assuming that the proportion explained by yet unknown component causes cannot exceed 25%, those who support this contention fail to recognize that cases caused by a sufficient cause that contains any subset of the three named causes might also contain unknown component causes. Cases stemming from sufficient causes with this overlapping set of component causes could be prevented by interventions targeting the three named causes, or by interventions targeting the yet unknown causes when they become known. The latter interventions could reduce the disease burden by much more than 25%.

As another example, in a cohort of cigarette smokers exposed to arsenic by working in a smelter, an estimated 75% of the lung cancer rate was attributable to their work environment and an estimated 65% was attributable to their smoking (Pinto et al., 1978; Hertz-Picciotto et al., 1992). There is no problem with such figures, which merely reflect the multifactorial etiology of disease. So, too, with coronary heart disease; if 75% of that disease is attributable to high blood cholesterol, high blood pressure, and cigarette smoking, 100% of it can still be attributable to other causes, known, suspected, and yet to be discovered. Some of these causes will participate in the same causal mechanisms as high blood cholesterol, high blood pressure, and cigarette smoking. Beaglehole and Magnus were correct in thinking that if the three specified component causes combine to explain 75% of cardiovascular disease (CVD) and we somehow eliminated them, there would be only 25% of CVD cases remaining. But until that 75% is eliminated, any newly discovered component could cause up to 100% of the CVD we currently have.

The notion that interventions targeting high blood cholesterol, high blood pressure, and cigarette smoking could eliminate 75% of coronary heart disease is unrealistic given currently available intervention strategies. Although progress can be made to reduce the effect of these risk factors, it is unlikely that any of them could be completely eradicated from any large population in the near term. Estimates of the public health effect of eliminating diseases themselves as causes of death (Murray et al., 2002) are even further removed from reality, because they fail to account for all the effects of interventions required to achieve the disease elimination, including unanticipated side effects (Greenland, 2002a, 2005a).

The debate about coronary heart disease attribution to component causes is reminiscent of an earlier debate regarding causes of cancer. In their widely cited work, *The Causes of Cancer,* Doll and Peto (1981, Table 20) created a table giving their estimates of the fraction of all cancers caused by various agents. The fractions summed to nearly 100%. Although the authors acknowledged that any case could be caused by more than one agent (which means that, given enough agents, the attributable

fractions would sum to far more than 100%), they referred to this situation as a "difficulty" and an "anomaly" that they chose to ignore. Subsequently, one of the authors acknowledged that the attributable fraction could sum to greater than 100% (Peto, 1985). It is neither a difficulty nor an anomaly nor something we can safely ignore, but simply a consequence of the fact that no event has a single agent as the cause. The fraction of disease that can be attributed to known causes will grow without bound as more causes are discovered. Only the fraction of disease attributable to a single component cause cannot exceed 100%.

In a similar vein, much publicity attended the pronouncement in 1960 that as much as 90% of cancer is environmentally caused (Higginson, 1960). Here, "environment" was thought of as representing all nongenetic component causes, and thus included not only the physical environment, but also the social environment and all individual human behavior that is not genetically determined. Hence, environmental component causes must be present to some extent in every sufficient cause of a disease. Thus, Higginson's estimate of 90% was an underestimate.

One can also show that 100% of any disease is inherited, even when environmental factors are component causes. MacMahon (1968) cited the example given by Hogben (1933) of yellow shanks, a trait occurring in certain genetic strains of fowl fed on yellow corn. Both a particular set of genes and a yellow-corn diet are necessary to produce yellow shanks. A farmer with several strains of fowl who feeds them all only yellow corn would consider yellow shanks to be a genetic condition, because only one strain would get yellow shanks, despite all strains getting the same diet. A different farmer who owned only the strain liable to get yellow shanks but who fed some of the birds yellow corn and others white corn would consider yellow shanks to be an environmentally determined condition because it depends on diet. In humans, the mental retardation caused by phenylketonuria is considered by many to be purely genetic. This retardation can, however, be successfully prevented by dietary intervention, which demonstrates the presence of an environmental cause. In reality, yellow shanks, phenylketonuria, and other diseases and conditions are determined by an interaction of genes and environment. It makes no sense to allocate a portion of the causation to either genes or environment separately when both may act together in sufficient causes.

Nonetheless, many researchers have compared disease occurrence in identical and nonidentical twins to estimate the fraction of disease that is inherited. These twin-study and other heritability indices assess only the relative role of environmental and genetic causes of disease in a particular setting. For example, some genetic causes may be necessary components of every causal mechanism. If everyone in a population has an identical set of the genes that cause disease, however, their effect is not included in heritability indices, despite the fact that the genes are causes of the disease. The two farmers in the preceding example would offer very different values for the heritability of yellow shanks, despite the fact that the condition is always 100% dependent on having certain genes.

Every case of every disease has some environmental and some genetic component causes, and therefore every case can be attributed both to genes and to environment. No paradox exists as long as it is understood that the fractions of disease attributable to genes and to environment overlap with one another. Thus, debates over what proportion of all occurrences of a disease are genetic and what proportion are environmental, inasmuch as these debates assume that the shares must add up to 100%, are fallacious and distracting from more worthwhile pursuits.

On an even more general level, the question of whether a given disease does or does not have a "multifactorial etiology" can be answered once and for all in the affirmative. All diseases have multifactorial etiologies. It is therefore completely unremarkable for a given disease to have such an etiology, and no time or money should be spent on research trying to answer the question of whether a particular disease does or does not have a multifactorial etiology. They all do. The job of etiologic research is to identify components of those etiologies.

INDUCTION PERIOD

Pie-chart diagrams of sufficient causes and their components such as those in Figure 2–2 are not well suited to provide a model for conceptualizing the *induction period,* which may be defined as the period of time from causal action until disease initiation. There is no way to tell from a pie-chart diagram of a sufficient cause which components affect each other, which components must come before or after others, for which components the temporal order is irrelevant, etc. The crucial

information on temporal ordering must come in a separate description of the interrelations among the components of a sufficient cause.

If, in sufficient cause I, the sequence of action of the specified component causes must be A = 0, B = 1 and we are studying the effect of A = 0, which (let us assume) acts at a narrowly defined point in time, we do not observe the occurrence of disease immediately after A = 0 occurs. Disease occurs only after the sequence is completed, so there will be a delay while B = 1 occurs (along with components of the set U_1 that are not present or that have not occured when A = 0 occurs). When B = 1 acts, if it is the last of all the component causes (including those in the set of unspecified conditions and events represented by U_1), disease occurs. The interval between the action of B = 1 and the disease occurrence is the induction time for the effect of B = 1 in sufficient cause I.

In the example given earlier of an equilibrium disorder leading to a later fall and hip injury, the induction time between the start of the equilibrium disorder and the later hip injury might be long, if the equilibrium disorder is caused by an old head injury, or short, if the disorder is caused by inebriation. In the latter case, it could even be instantaneous, if we define it as blood alcohol greater than a certain level. In general, conditions and events can be component causes and two or more component causes can have the same induction time, including zero.

Defining an induction period of interest is tantamount to specifying the characteristics of the component causes of interest. A clear example of a lengthy induction time is the cause–effect relation between exposure of a female fetus to diethylstilbestrol (DES) and the subsequent development of adenocarcinoma of the vagina. The cancer is usually diagnosed between ages 15 and 30 years. Because the causal exposure to DES occurs early in pregnancy, there is an induction time of about 15 to 30 years for the carcinogenic action of DES. During this time, other causes presumably are operating; some evidence suggests that hormonal action during adolescence may be part of the mechanism (Rothman, 1981).

It is incorrect to characterize a disease itself as having a lengthy or brief induction period. The induction time can be conceptualized only in relation to a specific component cause operating in a specific sufficient cause. Thus, we say that the induction time relating DES to clear-cell carcinoma of the vagina is 15 to 30 years, but we should not say that 15 to 30 years is the induction time for clear-cell carcinoma in general. Because each component cause in any causal mechanism can act at a time different from the other component causes, each can have its own induction time. For the component causes that act last, the induction time equals zero. If another component cause of clear-cell carcinoma of the vagina that acts during adolescence were identified, it would have a much shorter induction time for its carcinogenic action than DES. Thus, induction time characterizes a specific cause–effect pair rather than just the effect.

In carcinogenesis, the terms *initiator* and *promotor* have been used to refer to some of the component causes of cancer that act early and late, respectively, in the causal mechanism. Cancer itself has often been characterized as a disease process with a long induction time. This characterization is a misconception, however, because any late-acting component in the causal process, such as a promotor, will have a short induction time. Indeed, by definition, the induction time will always be zero for at least one component cause, the last to act. The mistaken view that diseases, as opposed to cause–disease relationships, have long or short induction periods can have important implications for research. For instance, the view of adult cancers as "diseases of long latency" may induce some researchers to ignore evidence of etiologic effects occurring relatively late in the processes that culminate in clinically diagnosed cancers. At the other extreme, the routine disregard for exposures occurring in the first decade or two in studies of occupational carcinogenesis, as a major example, may well have inhibited the discovery of occupational causes with very long induction periods.

Disease, once initiated, will not necessarily be apparent. The time interval between irreversible disease occurrence and detection has been termed the *latent period* (Rothman, 1981), although others have used this term interchangeably with induction period. Still others use *latent period* to mean the total time between causal action and disease detection. We use *induction period* to describe the time from causal action to irreversible disease occurrence and *latent period* to mean the time from disease occurrence to disease detection. The latent period can sometimes be reduced by improved methods of disease detection. The induction period, on the other hand, cannot be reduced by early detection of disease, because disease occurrence marks the end of the induction period. Earlier detection of disease, however, may reduce the apparent induction period (the time between causal action and disease detection), because the time when disease is detected, as a practical matter, is

usually used to mark the time of disease occurrence. Thus, diseases such as slow-growing cancers may appear to have long induction periods with respect to many causes because they have long latent periods. The latent period, unlike the induction period, is a characteristic of the disease and the detection effort applied to the person with the disease.

Although it is not possible to reduce the induction period proper by earlier detection of disease, it may be possible to observe intermediate stages of a causal mechanism. The increased interest in biomarkers such as DNA adducts is an example of attempting to focus on causes more proximal to the disease occurrence or on effects more proximal to cause occurrence. Such biomarkers may nonetheless reflect the effects of earlier-acting agents on the person.

Some agents may have a causal action by shortening the induction time of other agents. Suppose that exposure to factor $X = 1$ leads to epilepsy after an interval of 10 years, on average. It may be that exposure to a drug, $Z = 1$, would shorten this interval to 2 years. Is $Z = 1$ acting as a catalyst, or as a cause, of epilepsy? The answer is both: A catalyst is a cause. Without $Z = 1$, the occurrence of epilepsy comes 8 years later than it comes with $Z = 1$, so we can say that $Z = 1$ causes the onset of the early epilepsy. It is not sufficient to argue that the epilepsy would have occurred anyway. First, it would not have occurred at that time, and the time of occurrence is part of our definition of an event. Second, epilepsy will occur later only if the individual survives an additional 8 years, which is not certain. Not only does agent $Z = 1$ determine when the epilepsy occurs, it can also determine whether it occurs. Thus, we should call any agent that acts as a catalyst of a causal mechanism, speeding up an induction period for other agents, a cause in its own right. Similarly, any agent that postpones the onset of an event, drawing out the induction period for another agent, is a preventive. It should not be too surprising to equate postponement to prevention: We routinely use such an equation when we employ the euphemism that we "prevent" death, which actually can only be postponed. What we prevent is death at a given time, in favor of death at a later time.

SCOPE OF THE MODEL

The main utility of this model of sufficient causes and their components lies in its ability to provide a general but practical conceptual framework for causal problems. The attempt to make the proportion of disease attributable to various component causes add to 100% is an example of a fallacy that is exposed by the model (although MacMahon and others were able to invoke yellow shanks and phenylketonuria to expose that fallacy long before the sufficient-component cause model was formally described [MacMahon and Pugh, 1967, 1970]). The model makes it clear that, because of interactions, there is no upper limit to the sum of these proportions. As we shall see in Chapter 5, the epidemiologic evaluation of interactions themselves can be clarified, to some extent, with the help of the model.

Although the model appears to deal qualitatively with the action of component causes, it can be extended to account for dose dependence by postulating a set of sufficient causes, each of which contains as a component a different dose of the agent in question. Small doses might require a larger or rarer set of complementary causes to complete a sufficient cause than that required by large doses (Rothman, 1976a), in which case it is particularly important to specify both sides of the causal contrast. In this way, the model can account for the phenomenon of a shorter induction period accompanying larger doses of exposure, because a smaller set of complementary components would be needed to complete the sufficient cause.

Those who believe that chance must play a role in any complex mechanism might object to the intricacy of this seemingly deterministic model. A probabilistic (stochastic) model could be invoked to describe a dose–response relation, for example, without the need for a multitude of different causal mechanisms. The model would simply relate the dose of the exposure to the probability of the effect occurring. For those who believe that virtually all events contain some element of chance, deterministic causal models may seem to misrepresent the indeterminism of the real world. However, the deterministic model presented here can accommodate "chance"; one way might be to view chance, or at least some part of the variability that we call "chance," as the result of deterministic events that are beyond the current limits of knowledge or observability.

For example, the outcome of a flip of a coin is usually considered a chance event. In classical mechanics, however, the outcome can in theory be determined completely by the application of physical laws and a sufficient description of the starting conditions. To put it in terms more familiar

to epidemiologists, consider the explanation for why an individual gets lung cancer. One hundred years ago, when little was known about the etiology of lung cancer; a scientist might have said that it was a matter of chance. Nowadays, we might say that the risk depends on how much the individual smokes, how much asbestos and radon the individual has been exposed to, and so on. Nonetheless, recognizing this dependence moves the line of ignorance; it does not eliminate it. One can still ask what determines whether an individual who has smoked a specific amount and has a specified amount of exposure to all the other known risk factors will get lung cancer. Some will get lung cancer and some will not, and if all known risk factors are already taken into account, what is left we might still describe as chance. True, we can explain much more of the variability in lung cancer occurrence nowadays than we formerly could by taking into account factors known to cause it, but at the limits of our knowledge, we still ascribe the remaining variability to what we call chance. In this view, chance is seen as a catchall term for our ignorance about causal explanations.

We have so far ignored more subtle considerations of sources of unpredictability in events, such as chaotic behavior (in which even the slightest uncertainty about initial conditions leads to vast uncertainty about outcomes) and quantum-mechanical uncertainty. In each of these situations, a random (stochastic) model component may be essential for any useful modeling effort. Such components can also be introduced in the above conceptual model by treating unmeasured component causes in the model as random events, so that the causal model based on components of sufficient causes can have random elements. An example is treatment assignment in randomized clinical trials (Poole 2001a).

OTHER MODELS OF CAUSATION

The sufficient-component cause model is only one of several models of causation that may be useful for gaining insight about epidemiologic concepts (Greenland and Brumback, 2002; Greenland, 2004a). It portrays qualitative causal mechanisms within members of a population, so its fundamental unit of analysis is the causal mechanism rather than a person. Many different sets of mechanisms can lead to the same pattern of disease within a population, so the sufficient-component cause model involves specification of details that are beyond the scope of epidemiologic data. Also, it does not incorporate elements reflecting population distributions of factors or causal sequences, which are crucial to understanding confounding and other biases.

Other models of causation, such as potential-outcome (counterfactual) models and graphical models, provide direct representations of epidemiologic concepts such as confounding and other biases, and can be applied at mechanistic, individual, or population levels of analysis. Potential-outcome models (Chapters 4 and 5) specify in detail what would happen to individuals or populations under alternative possible patterns of interventions or exposures, and also bring to the fore problems in operationally defining causes (Greenland, 2002a, 2005a; Hernán, 2005). Graphical models (Chapter 12) display broad qualitative assumptions about causal directions and independencies. Both types of model have close relationships to the structural-equations models that are popular in the social sciences (Pearl, 2000; Greenland and Brumback, 2002), and both can be subsumed under a general theory of longitudinal causality (Robins, 1997).

PHILOSOPHY OF SCIENTIFIC INFERENCE

Causal inference may be viewed as a special case of the more general process of scientific reasoning. The literature on this topic is too vast for us to review thoroughly, but we will provide a brief overview of certain points relevant to epidemiology, at the risk of some oversimplification.

INDUCTIVISM

Modern science began to emerge around the 16th and 17th centuries, when the knowledge demands of emerging technologies (such as artillery and transoceanic navigation) stimulated inquiry into the origins of knowledge. An early codification of the scientific method was Francis Bacon's *Novum Organum,* which, in 1620, presented an inductivist view of science. In this philosophy, scientific reasoning is said to depend on making generalizations, or inductions, from observations to general laws of nature; the observations are said to induce the formulation of a natural law in the mind of

the scientist. Thus, an inductivist would have said that Jenner's observation of lack of smallpox among milkmaids induced in Jenner's mind the theory that cowpox (common among milkmaids) conferred immunity to smallpox. Inductivist philosophy reached a pinnacle of sorts in the canons of John Stuart Mill (1862), which evolved into inferential criteria that are still in use today.

Inductivist philosophy was a great step forward from the medieval scholasticism that preceded it, for at least it demanded that a scientist make careful observations of people and nature rather than appeal to faith, ancient texts, or authorities. Nonetheless, in the 18th century the Scottish philosopher David Hume described a disturbing deficiency in inductivism. An inductive argument carried no logical force; instead, such an argument represented nothing more than an *assumption* that certain events would in the future follow the same pattern as they had in the past. Thus, to argue that cowpox caused immunity to smallpox because no one got smallpox after having cowpox corresponded to an unjustified assumption that the pattern observed to date (no smallpox after cowpox) would continue into the future. Hume pointed out that, even for the most reasonable-sounding of such assumptions, there was no logical necessity behind the inductive argument.

Of central concern to Hume (1739) was the issue of causal inference and failure of induction to provide a foundation for it:

> Thus not only our reason fails us in the discovery of the ultimate connexion of causes and effects, but even after experience has inform'd us of their constant conjunction, 'tis impossible for us to satisfy ourselves by our reason, why we shou'd extend that experience beyond those particular instances, which have fallen under our observation. We suppose, but are never able to prove, that there must be a resemblance betwixt those objects, of which we have had experience, and those which lie beyond the reach of our discovery.

In other words, no number of repetitions of a particular sequence of events, such as the appearance of a light after flipping a switch, can prove a causal connection between the action of the switch and the turning on of the light. No matter how many times the light comes on after the switch has been pressed, the possibility of coincidental occurrence cannot be ruled out. Hume pointed out that observers cannot perceive causal connections, but only a series of events. Bertrand Russell (1945) illustrated this point with the example of two accurate clocks that perpetually chime on the hour, with one keeping time slightly ahead of the other. Although one invariably chimes before the other, there is no direct causal connection from one to the other. Thus, assigning a causal interpretation to the pattern of events cannot be a logical extension of our observations alone, because the events might be occurring together only because of a shared earlier cause, or because of some systematic error in the observations.

Causal inference based on mere association of events constitutes a logical fallacy known as *post hoc ergo propter hoc* (Latin for "after this therefore on account of this"). This fallacy is exemplified by the inference that the crowing of a rooster is necessary for the sun to rise because sunrise is always preceded by the crowing.

The *post hoc* fallacy is a special case of a more general logical fallacy known as the *fallacy of affirming the consequent*. This fallacy of confirmation takes the following general form: "We know that if H is true, B must be true; and we know that B is true; therefore H must be true." This fallacy is used routinely by scientists in interpreting data. It is used, for example, when one argues as follows: "If sewer service causes heart disease, then heart disease rates should be highest where sewer service is available; heart disease rates are indeed highest where sewer service is available; therefore, sewer service causes heart disease." Here, H is the hypothesis "sewer service causes heart disease" and B is the observation "heart disease rates are highest where sewer service is available." The argument is logically unsound, as demonstrated by the fact that we can imagine many ways in which the premises could be true but the conclusion false; for example, economic development could lead to both sewer service and elevated heart disease rates, without any effect of sewer service on heart disease. In this case, however, we also know that one of the premises is not true—specifically, the premise, "If H is true, B must be true." This particular form of the fallacy exemplifies the problem of *confounding*, which we will discuss in detail in later chapters.

Bertrand Russell (1945) satirized the fallacy this way:

> 'If p, then q; now q is true; therefore p is true.' E.g., 'If pigs have wings, then some winged animals are good to eat; now some winged animals are good to eat; therefore pigs have wings.' This form of inference is called 'scientific method.'

REFUTATIONISM

Russell was not alone in his lament of the illogicality of scientific reasoning as ordinarily practiced. Many philosophers and scientists from Hume's time forward attempted to set out a firm logical basis for scientific reasoning.

In the 1920s, most notable among these was the school of logical positivists, who sought a logic for science that could lead inevitably to correct scientific conclusions, in much the way rigorous logic can lead inevitably to correct conclusions in mathematics. Other philosophers and scientists, however, had started to suspect that scientific hypotheses can never be proven or established as true in any logical sense. For example, a number of philosophers noted that scientific statements can only be found to be consistent with observation, but cannot be proven or disproven in any "airtight" logical or mathematical sense (Duhem, 1906, transl. 1954; Popper 1934, transl. 1959; Quine, 1951). This fact is sometimes called the problem of *nonidentification* or *underdetermination* of theories by observations (Curd and Cover, 1998). In particular, available observations are always consistent with several hypotheses that themselves are mutually inconsistent, which explains why (as Hume noted) scientific theories cannot be logically proven. In particular, consistency between a hypothesis and observations is no proof of the hypothesis, because we can always invent alternative hypotheses that are just as consistent with the observations.

In contrast, a valid observation that is inconsistent with a hypothesis implies that the hypothesis as stated is false and so refutes the hypothesis. If you wring the rooster's neck before it crows and the sun still rises, you have disproved that the rooster's crowing is a necessary cause of sunrise. Or consider a hypothetical research program to learn the boiling point of water (Magee, 1985). A scientist who boils water in an open flask and repeatedly measures the boiling point at 100°C will never, no matter how many confirmatory repetitions are involved, prove that 100°C is always the boiling point. On the other hand, merely one attempt to boil the water in a closed flask or at high altitude will refute the proposition that water always boils at 100°C.

According to Popper, science advances by a process of elimination that he called "conjecture and refutation." Scientists form hypotheses based on intuition, conjecture, and previous experience. Good scientists use deductive logic to infer predictions from the hypothesis and then compare observations with the predictions. Hypotheses whose predictions agree with observations are confirmed (Popper used the term "corroborated") only in the sense that they can continue to be used as explanations of natural phenomena. At any time, however, they may be refuted by further observations and might be replaced by other hypotheses that are more consistent with the observations. This view of scientific inference is sometimes called *refutationism* or *falsificationism*. Refutationists consider induction to be a psychologic crutch: Repeated observations did not in fact induce the formulation of a natural law, but only the belief that such a law has been found. For a refutationist, only the psychologic comfort provided by induction explains why it still has advocates.

One way to rescue the concept of induction from the stigma of pure delusion is to resurrect it as a psychologic phenomenon, as Hume and Popper claimed it was, but one that plays a legitimate role in hypothesis formation. The philosophy of conjecture and refutation places no constraints on the origin of conjectures. Even delusions are permitted as hypotheses, and therefore inductively inspired hypotheses, however psychologic, are valid starting points for scientific evaluation. This concession does not admit a logical role for induction in confirming scientific hypotheses, but it allows the process of induction to play a part, along with imagination, in the scientific cycle of conjecture and refutation.

The philosophy of conjecture and refutation has profound implications for the methodology of science. The popular concept of a scientist doggedly assembling evidence to support a favorite thesis is objectionable from the standpoint of refutationist philosophy because it encourages scientists to consider their own pet theories as their intellectual property, to be confirmed, proven, and, when all the evidence is in, cast in stone and defended as natural law. Such attitudes hinder critical evaluation, interchange, and progress. The approach of conjecture and refutation, in contrast, encourages scientists to consider multiple hypotheses and to seek crucial tests that decide between competing hypotheses by falsifying one of them. Because falsification of one or more theories is the goal, there is incentive to depersonalize the theories. Criticism leveled at a theory need not be seen as criticism of the person who proposed it. It has been suggested that the reason why certain fields of science advance rapidly while others languish is that the rapidly advancing fields are propelled by scientists

who are busy constructing and testing competing hypotheses; the other fields, in contrast, "are sick by comparison, because they have forgotten the necessity for alternative hypotheses and disproof" (Platt, 1964).

The refutationist model of science has a number of valuable lessons for research conduct, especially of the need to seek alternative explanations for observations, rather than focus on the chimera of seeking scientific "proof" for some favored theory. Nonetheless, it is vulnerable to criticisms that observations (or some would say their interpretations) are themselves laden with theory (sometimes called the *Duhem-Quine thesis*; Curd and Cover, 1998). Thus, observations can never provide the sort of definitive refutations that are the hallmark of popular accounts of refutationism. For example, there may be uncontrolled and even unimagined biases that have made our refutational observations invalid; to claim refutation is to assume as true the unprovable theory that no such bias exists. In other words, not only are theories underdetermined by observations, so are refutations, which are themselves theory-laden. The net result is that logical certainty about either the truth or falsity of an internally consistent theory is impossible (Quine, 1951).

CONSENSUS AND NATURALISM

Some 20th-century philosophers of science, most notably Thomas Kuhn (1962), emphasized the role of the scientific community in judging the validity of scientific theories. These critics of the conjecture-and-refutation model suggested that the refutation of a theory involves making a choice. Every observation is itself dependent on theories. For example, observing the moons of Jupiter through a telescope seems to us like a direct observation, but only because the theory of optics on which the telescope is based is so well accepted. When confronted with a refuting observation, a scientist faces the choice of rejecting either the validity of the theory being tested or the validity of the refuting observation, which itself must be premised on scientific theories that are not certain (Haack, 2003). Observations that are falsifying instances of theories may at times be treated as "anomalies," tolerated without falsifying the theory in the hope that the anomalies may eventually be explained. An epidemiologic example is the observation that shallow-inhaling smokers had higher lung cancer rates than deep-inhaling smokers. This anomaly was eventually explained when it was noted that lung tissue higher in the lung is more susceptible to smoking-associated lung tumors, and shallowly inhaled smoke tars tend to be deposited higher in the lung (Wald, 1985).

In other instances, anomalies may lead eventually to the overthrow of current scientific doctrine, just as Newtonian mechanics was displaced (remaining only as a first-order approximation) by relativity theory. Kuhn asserted that in every branch of science the prevailing scientific viewpoint, which he termed "normal science," occasionally undergoes major shifts that amount to scientific revolutions. These revolutions signal a decision of the scientific community to discard the scientific infrastructure rather than to falsify a new hypothesis that cannot be easily grafted onto it. Kuhn and others have argued that the consensus of the scientific community determines what is considered accepted and what is considered refuted.

Kuhn's critics characterized this description of science as one of an irrational process, "a matter for mob psychology" (Lakatos, 1970). Those who believe in a rational structure for science consider Kuhn's vision to be a regrettably real description of much of what passes for scientific activity, but not prescriptive for any good science: Although many modern philosophers reject rigid demarcations and formulations for science such as refutationism, they nonetheless maintain that science is founded on reason, albeit possibly informal common sense (Haack, 2003). Others go beyond Kuhn and maintain that attempts to impose a singular rational structure or methodology on science hobbles the imagination and is a prescription for the same sort of authoritarian repression of ideas that scientists have had to face throughout history (Feyerabend, 1975 and 1993).

The philosophic debate about Kuhn's description of science hinges on whether Kuhn meant to describe only what has happened historically in science or instead what ought to happen, an issue about which Kuhn (1970) has not been completely clear:

> Are Kuhn's [my] remarks about scientific development. . . to be read as descriptions or prescriptions? The answer, of course, is that they should be read in both ways at once. If I have a theory of how and why science works, it must necessarily have implications for the way in which scientists should behave if their enterprise is to flourish.

The idea that science is a sociologic process, whether considered descriptive or normative, is an interesting thesis, as is the idea that from observing how scientists work we can learn about how scientists ought to work. The latter idea has led to the development of *naturalistic* philosophy of science, or "science studies," which examines scientific developments for clues about what sort of methods scientists need and develop for successful discovery and invention (Callebaut, 1993; Giere, 1999).

Regardless of philosophical developments, we suspect that most epidemiologists (and most scientists) will continue to function as if the following classical view is correct: The ultimate goal of scientific inference is to capture some objective truths about the material world in which we live, and any theory of inference should ideally be evaluated by how well it leads us to these truths. This ideal is impossible to operationalize, however, for if we ever find any ultimate truths, we will have no way of knowing that for certain. Thus, those holding the view that scientific truth is not arbitrary nevertheless concede that our knowledge of these truths will always be tentative. For refutationists, this tentativeness has an asymmetric quality, but that asymmetry is less marked for others. We may believe that we know a theory is false because it consistently fails the tests we put it through, but our tests could be faulty, given that they involve imperfect reasoning and sense perception. Neither can we know that a theory is true, even if it passes every test we can devise, for it may fail a test that is as yet undevised.

Few, if any, would disagree that a theory of inference should be evaluated at least in part by how well it leads us to detect errors in our hypotheses and observations. There are, however, many other inferential activities besides evaluation of hypotheses, such as prediction or forecasting of events, and subsequent attempts to control events (which of course requires causal information). Statisticians rather than philosophers have more often confronted these problems in practice, so it should not be surprising that the major philosophies concerned with these problems emerged from statistics rather than philosophy.

BAYESIANISM

There is another philosophy of inference that, like most, holds an objective view of scientific truth and a view of knowledge as tentative or uncertain, but that focuses on evaluation of knowledge rather than truth. Like refutationism, the modern form of this philosophy evolved from the writings of 18th-century thinkers. The focal arguments first appeared in a pivotal essay by the Reverend Thomas Bayes (1764), and hence the philosophy is usually referred to as Bayesianism (Howson and Urbach, 1993), and it was the renowned French mathematician and scientist Pierre Simon de Laplace who first gave it an applied statistical format. Nonetheless, it did not reach a complete expression until after World War I, most notably in the writings of Ramsey (1931) and DeFinetti (1937); and, like refutationism, it did not begin to appear in epidemiology until the 1970s (e.g., Cornfield, 1976).

The central problem addressed by Bayesianism is the following: In classical logic, a deductive argument can provide no information about the truth or falsity of a scientific hypothesis unless you can be 100% certain about the truth of the premises of the argument. Consider the logical argument called *modus tollens:* "If H implies B, and B is false, then H must be false." This argument is logically valid, but the conclusion follows only on the assumptions that the premises "H implies B" and "B is false" are true statements. If these premises are statements about the physical world, we cannot possibly know them to be correct with 100% certainty, because all observations are subject to error. Furthermore, the claim that "H implies B" will often depend on its own chain of deductions, each with its own premises of which we cannot be certain.

For example, if H is "Television viewing causes homicides" and B is "Homicide rates are highest where televisions are most common," the first premise used in *modus tollens* to test the hypothesis that television viewing causes homicides will be: "If television viewing causes homicides, homicide rates are highest where televisions are most common." The validity of this premise is doubtful—after all, even if television does cause homicides, homicide rates may be low where televisions are common because of socioeconomic advantages in those areas.

Continuing to reason in this fashion, we could arrive at a more pessimistic state than even Hume imagined. Not only is induction without logical foundation, *deduction* has limited scientific utility because we cannot ensure the truth of all the premises, even if a logical argument is valid.

The Bayesian answer to this problem is partial in that it makes a severe demand on the scientist and puts a severe limitation on the results. It says roughly this: If you can assign a degree of certainty, or personal probability, to the premises of your valid argument, you may use any and all the rules of probability theory to derive a certainty for the conclusion, and this certainty will be a logically valid consequence of your original certainties. An inescapable fact is that your concluding certainty, or *posterior probability,* may depend heavily on what you used as initial certainties, or *prior probabilities*. If those initial certainties are not the same as those of a colleague, that colleague may very well assign a certainty to the conclusion different from the one you derived. With the accumulation of consistent evidence, however, the data can usually force even extremely disparate priors to converge into similar posterior probabilities.

Because the posterior probabilities emanating from a Bayesian inference depend on the person supplying the initial certainties and so may vary across individuals, the inferences are said to be subjective. This subjectivity of Bayesian inference is often mistaken for a subjective treatment of truth. Not only is such a view of Bayesianism incorrect, it is diametrically opposed to Bayesian philosophy. The Bayesian approach represents a constructive attempt to deal with the dilemma that scientific laws and facts should not be treated as known with certainty, whereas classic deductive logic yields conclusions only when some law, fact, or connection is asserted with 100% certainty.

A common criticism of Bayesian philosophy is that it diverts attention away from the classic goals of science, such as the discovery of how the world works, toward psychologic states of mind called "certainties," "subjective probabilities," or "degrees of belief" (Popper, 1959). This criticism, however, fails to recognize the importance of a scientist's state of mind in determining what theories to test and what tests to apply, the consequent influence of those states on the store of data available for inference, and the influence of the data on the states of mind.

Another reply to this criticism is that scientists already use data to influence their degrees of belief, and they are not shy about expressing those degrees of certainty. The problem is that the conventional process is informal, intuitive, and ineffable, and therefore not subject to critical scrutiny; at its worst, it often amounts to nothing more than the experts announcing that they have seen the evidence and here is how certain they are. How they reached this certainty is left unclear, or, put another way, is not "transparent." The problem is that no one, even an expert, is very good at informally and intuitively formulating certainties that predict facts and future events well (Kahneman et al., 1982; Gilovich, 1993; Piattelli-Palmarini, 1994; Gilovich et al., 2002). One reason for this problem is that biases and prior prejudices can easily creep into expert judgments. Bayesian methods force experts to "put their cards on the table" and specify explicitly the strength of their prior beliefs and why they have such beliefs, defend those specifications against arguments and evidence, and update their degrees of certainty with new evidence in ways that do not violate probability logic.

In any research context, there will be an unlimited number of hypotheses that could explain an observed phenomenon. Some argue that progress is best aided by severely testing (empirically challenging) those explanations that seem most probable in light of past research, so that shortcomings of currently "received" theories can be most rapidly discovered. Indeed, much research in certain fields takes this form, as when theoretical predictions of particle mass are put to ever more precise tests in physics experiments. This process does not involve mere improved repetition of past studies. Rather, it involves tests of previously untested but important predictions of the theory. Moreover, there is an imperative to make the basis for prior beliefs criticizable and defensible. That prior probabilities can differ among persons does not mean that all such beliefs are based on the same information, nor that all are equally tenable.

Probabilities of auxiliary hypotheses are also important in study design and interpretation. Failure of a theory to pass a test can lead to rejection of the theory more rapidly when the auxiliary hypotheses on which the test depends possess high probability. This observation provides a rationale for preferring "nested" case-control studies (in which controls are selected from a roster of the source population for the cases) to "hospital-based" case-control studies (in which the controls are "selected" by the occurrence or diagnosis of one or more diseases other than the case-defining disease), because the former have fewer mechanisms for biased subject selection and hence are given a higher probability of unbiased subject selection.

Even if one disputes the above arguments, most epidemiologists desire some way of expressing the varying degrees of certainty about possible values of an effect measure in light of available data. Such expressions must inevitably be derived in the face of considerable uncertainty about

methodologic details and various events that led to the available data and can be extremely sensitive to the reasoning used in its derivation. For example, as we shall discuss at greater length in Chapter 19, conventional confidence intervals quantify only random error under often questionable assumptions and so should not be interpreted as measures of total uncertainty, particularly for nonexperimental studies. As noted earlier, most people, including scientists, reason poorly in the face of uncertainty. At the very least, subjective Bayesian philosophy provides a methodology for sound reasoning under uncertainty and, in particular, provides many warnings against being overly certain about one's conclusions (Greenland 1998a, 1988b, 2006a; see also Chapters 18 and 19).

Such warnings are echoed in refutationist philosophy. As Peter Medawar (1979) put it, "I cannot give any scientist of any age better advice than this: the intensity of the conviction that a hypothesis is true has no bearing on whether it is true or not." We would add two points. First, the intensity of conviction that a hypothesis is false has no bearing on whether it is false or not. Second, Bayesian methods do not mistake beliefs for evidence. They use evidence to modify beliefs, which scientists routinely do in any event, but often in implicit, intuitive, and incoherent ways.

IMPOSSIBILITY OF SCIENTIFIC PROOF

Vigorous debate is a characteristic of modern scientific philosophy, no less in epidemiology than in other areas (Rothman, 1988). Can divergent philosophies of science be reconciled? Haack (2003) suggested that the scientific enterprise is akin to solving a vast, collective crossword puzzle. In areas in which the evidence is tightly interlocking, there is more reason to place confidence in the answers, but in areas with scant information, the theories may be little better than informed guesses. Of the scientific method, Haack (2003) said that "there is less to the 'scientific method' than meets the eye. Is scientific inquiry categorically different from other kinds? No. Scientific inquiry is continuous with everyday empirical inquiry—only more so."

Perhaps the most important common thread that emerges from the debated philosophies is that proof is impossible in empirical science. This simple fact is especially important to observational epidemiologists, who often face the criticism that proof is impossible in epidemiology, with the implication that it is possible in other scientific disciplines. Such criticism may stem from a view that experiments are the definitive source of scientific knowledge. That view is mistaken on at least two counts. First, the nonexperimental nature of a science does not preclude impressive scientific discoveries; the myriad examples include plate tectonics, the evolution of species, planets orbiting other stars, and the effects of cigarette smoking on human health. Even when they are possible, experiments (including randomized trials) do not provide anything approaching proof and in fact may be controversial, contradictory, or nonreproducible. If randomized clinical trials provided proof, we would never need to do more than one of them on a given hypothesis. Neither physical nor experimental science is immune to such problems, as demonstrated by episodes such as the experimental "discovery" (later refuted) of cold fusion (Taubes, 1993).

Some experimental scientists hold that epidemiologic relations are only suggestive and believe that detailed laboratory study of mechanisms within single individuals can reveal cause–effect relations with certainty. This view overlooks the fact that *all* relations are suggestive in exactly the manner discussed by Hume. Even the most careful and detailed mechanistic dissection of individual events cannot provide more than associations, albeit at a finer level. Laboratory studies often involve a degree of observer control that cannot be approached in epidemiology; it is only this control, not the level of observation, that can strengthen the inferences from laboratory studies. And again, such control is no guarantee against error. In addition, neither scientists nor decision makers are often highly persuaded when only mechanistic evidence from the laboratory is available.

All of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature, even when the work itself is carried out without mistakes. The tentativeness of our knowledge does not prevent practical applications, but it should keep us skeptical and critical, not only of everyone else's work, but of our own as well. Sometimes etiologic hypotheses enjoy an extremely high, universally or almost universally shared, degree of certainty. The hypothesis that cigarette smoking causes lung cancer is one of the best-known examples. These hypotheses rise above "tentative" acceptance and are the closest we can come to "proof." But even

these hypotheses are not "proved" with the degree of absolute certainty that accompanies the proof of a mathematical theorem.

CAUSAL INFERENCE IN EPIDEMIOLOGY

Etiologic knowledge about epidemiologic hypotheses is often scant, making the hypotheses themselves at times little more than vague statements of causal association between exposure and disease, such as "smoking causes cardiovascular disease." These vague hypotheses have only vague consequences that can be difficult to test. To cope with this vagueness, epidemiologists usually focus on testing the negation of the causal hypothesis, that is, the null hypothesis that the exposure does *not* have a causal relation to disease. Then, any observed association can potentially refute the hypothesis, subject to the assumption (auxiliary hypothesis) that biases and chance fluctuations are not solely responsible for the observation.

TESTS OF COMPETING EPIDEMIOLOGIC THEORIES

If the causal mechanism is stated specifically enough, epidemiologic observations can provide crucial tests of competing, non-null causal hypotheses. For example, when toxic-shock syndrome was first studied, there were two competing hypotheses about the causal agent. Under one hypothesis, it was a chemical in the tampon, so that women using tampons were exposed to the agent directly from the tampon. Under the other hypothesis, the tampon acted as a culture medium for staphylococci that produced a toxin. Both hypotheses explained the relation of toxic-shock occurrence to tampon use. The two hypotheses, however, led to opposite predictions about the relation between the frequency of changing tampons and the rate of toxic shock. Under the hypothesis of a chemical agent, more frequent changing of the tampon would lead to more exposure to the agent and possible absorption of a greater overall dose. This hypothesis predicted that women who changed tampons more frequently would have a higher rate than women who changed tampons infrequently. The culture-medium hypothesis predicts that women who change tampons frequently would have a lower rate than those who change tampons less frequently, because a short duration of use for each tampon would prevent the staphylococci from multiplying enough to produce a damaging dose of toxin. Thus, epidemiologic research, by showing that infrequent changing of tampons was associated with a higher rate of toxic shock, refuted the chemical theory in the form presented. There was, however, a third hypothesis that a chemical in some tampons (e.g., oxygen content) improved their performance as culture media. This chemical-promotor hypothesis made the same prediction about the association with frequency of changing tampons as the microbial toxin hypothesis (Lanes and Rothman, 1990).

Another example of a theory that can be easily tested by epidemiologic data relates to the observation that women who took replacement estrogen therapy had a considerably elevated rate of endometrial cancer. Horwitz and Feinstein (1978) conjectured a competing theory to explain the association: They proposed that women taking estrogen experienced symptoms such as bleeding that induced them to consult a physician. The resulting diagnostic workup led to the detection of endometrial cancer at an earlier stage in these women, as compared with women who were not taking estrogens. Horwitz and Feinstein argued that the association arose from this detection bias, claiming that without the bleeding-induced workup, many of these cancers would not have been detected at all. Many epidemiologic observations were used to evaluate these competing hypotheses. The detection-bias theory predicted that women who had used estrogens for only a short time would have the greatest elevation in their rate, as the symptoms related to estrogen use that led to the medical consultation tended to appear soon after use began. Because the association of recent estrogen use and endometrial cancer was the same in both long- and short-term estrogen users, the detection-bias theory was refuted as an explanation for all but a small fraction of endometrial cancer cases occurring after estrogen use. Refutation of the detection-bias theory also depended on many other observations. Especially important was the theory's implication that there must be a huge reservoir of undetected endometrial cancer in the typical population of women to account for the much greater rate observed in estrogen users, an implication that was not borne out by further observations (Hutchison and Rothman, 1978).

The endometrial cancer example illustrates a critical point in understanding the process of causal inference in epidemiologic studies: Many of the hypotheses being evaluated in the interpretation of epidemiologic studies are auxiliary hypotheses in the sense that they are independent of the presence, absence, or direction of any causal connection between the study exposure and the disease. For example, explanations of how specific types of bias could have distorted an association between exposure and disease are the usual alternatives to the primary study hypothesis. Much of the interpretation of epidemiologic studies amounts to the testing of such auxiliary explanations for observed associations.

CAUSAL CRITERIA

In practice, how do epidemiologists separate causal from noncausal explanations? Despite philosophic criticisms of inductive inference, inductively oriented considerations are often used as criteria for making such inferences (Weed and Gorelic, 1996). If a set of necessary and sufficient causal criteria could be used to distinguish causal from noncausal relations in epidemiologic studies, the job of the scientist would be eased considerably. With such criteria, all the concerns about the logic or lack thereof in causal inference could be subsumed: It would only be necessary to consult the checklist of criteria to see if a relation were causal. We know from the philosophy reviewed earlier that a set of sufficient criteria does not exist. Nevertheless, lists of causal criteria have become popular, possibly because they seem to provide a road map through complicated territory, and perhaps because they suggest hypotheses to be evaluated in a given problem.

A commonly used set of criteria was based on a list of considerations or "viewpoints" proposed by Sir Austin Bradford Hill (1965). Hill's list was an expansion of a list offered previously in the landmark U.S. Surgeon General's report *Smoking and Health* (1964), which in turn was anticipated by the inductive canons of John Stuart Mill (1862) and the rules given by Hume (1739). Subsequently, others, especially Susser, have further developed causal considerations (Kaufman and Poole, 2000).

Hill suggested that the following considerations in attempting to distinguish causal from noncausal associations that were already "perfectly clear-cut and beyond what we would care to attribute to the play of chance": (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) biologic gradient, (6) plausibility, (7) coherence, (8) experimental evidence, and (9) analogy. Hill emphasized that causal inferences cannot be based on a set of rules, condemned emphasis on statistical significance testing, and recognized the importance of many other factors in decision making (Phillips and Goodman, 2004). Nonetheless, the misguided but popular view that his considerations should be used as criteria for causal inference makes it necessary to examine them in detail.

Strength

Hill argued that strong associations are particularly compelling because, for weaker associations, it is "easier" to imagine what today we would call an unmeasured confounder that might be responsible for the association. Several years earlier, Cornfield et al. (1959) drew similar conclusions. They concentrated on a single hypothetical confounder that, by itself, would explain entirely an observed association. They expressed a strong preference for ratio measures of strength, as opposed to difference measures, and focused on how the observed estimate of a risk ratio provides a minimum for the association that a completely explanatory confounder must have with the exposure (rather than a minimum for the confounder–disease association). Of special importance, Cornfield et al. acknowledged that having only a weak association does not rule out a causal connection (Rothman and Poole, 1988). Today, some associations, such as those between smoking and cardiovascular disease or between environmental tobacco smoke and lung cancer, are accepted by most as causal even though the associations are considered weak.

Counterexamples of strong but noncausal associations are also not hard to find; any study with strong confounding illustrates the phenomenon. For example, consider the strong relation between Down syndrome and birth rank, which is confounded by the relation between Down syndrome and maternal age. Of course, once the confounding factor is identified, the association is diminished by controlling for the factor.

These examples remind us that a strong association is neither necessary nor sufficient for causality, and that weakness is neither necessary nor sufficient for absence of causality. A strong association

bears only on hypotheses that the association is entirely or partially due to unmeasured confounders or other source of modest bias.

Consistency

To most observers, consistency refers to the repeated observation of an association in different populations under different circumstances. Lack of consistency, however, does not rule out a causal association, because some effects are produced by their causes only under unusual circumstances. More precisely, the effect of a causal agent cannot occur unless the complementary component causes act or have already acted to complete a sufficient cause. These conditions will not always be met. Thus, transfusions can cause infection with the human immunodeficiency virus, but they do not always do so: The virus must also be present. Tampon use can cause toxic-shock syndrome, but only rarely, when certain other, perhaps unknown, conditions are met. Consistency is apparent only after all the relevant details of a causal mechanism are understood, which is to say very seldom. Furthermore, even studies of exactly the same phenomena can be expected to yield different results simply because they differ in their methods and random errors. Consistency serves only to rule out hypotheses that the association is attributable to some factor that varies across studies.

One mistake in implementing the consistency criterion is so common that it deserves special mention. It is sometimes claimed that a literature or set of results is inconsistent simply because some results are "statistically significant" and some are not. This sort of evaluation is completely fallacious even if one accepts the use of significance testing methods. The results (effect estimates) from a set of studies could all be identical even if many were significant and many were not, the difference in significance arising solely because of differences in the standard errors or sizes of the studies. Conversely, the results could be significantly in conflict even if all were all were nonsignificant individually, simply because in aggregate an effect could be apparent in some subgroups but not others (see Chapter 33). The fallacy of judging consistency by comparing P-values or statistical significance is not eliminated by "standardizing" estimates (i.e., dividing them by the standard deviation of the outcome, multiplying them by the standard deviation of the exposure, or both); in fact it is worsened, as such standardization can create differences where none exists, or mask true differences (Greenland et al., 1986, 1991; see Chapters 21 and 33).

Specificity

The criterion of specificity has two variants. One is that a cause leads to a single effect, not multiple effects. The other is that an effect has one cause, not multiple causes. Hill mentioned both of them. The former criterion, specificity of effects, was used as an argument in favor of a causal interpretation of the association between smoking and lung cancer and, in an act of circular reasoning, in favor of ratio comparisons and not differences as the appropriate measures of strength. When ratio measures were examined, the association of smoking to diseases looked "quantitatively specific" to lung cancer. When difference measures were examined, the association appeared to be nonspecific, with several diseases (other cancers, coronary heart disease, etc.) being at least as strongly associated with smoking as lung cancer was. Today we know that smoking affects the risk of many diseases and that the difference comparisons were accurately portraying this lack of specificity. Unfortunately, however, the historical episode of the debate over smoking and health is often cited today as justification for the specificity criterion and for using ratio comparisons to measure strength of association. The proper lessons to learn from that episode should be just the opposite.

Weiss (2002) argued that specificity can be used to distinguish some causal hypotheses from noncausal hypotheses, when the causal hypothesis predicts a relation with one outcome but no relation with another outcome. His argument is persuasive when, in addition to the causal hypothesis, one has an alternative noncausal hypothesis that predicts a nonspecific association. Weiss offered the example of screening sigmoidoscopy, which was associated in case-control studies with a 50% to 70% reduction in mortality from distal tumors of the rectum and tumors of the distal colon, within the reach of the sigmoidoscope, but no reduction in mortality from tumors elsewhere in the colon. If the effect of screening sigmoidoscopy were not specific to the distal colon tumors, it would lend support not to all noncausal theories to explain the association, as Weiss suggested, but only to those noncausal theories that would have predicted a nonspecific association. Thus, specificity can

come into play when it can be logically deduced from the causal hypothesis in question and when nonspecificity can be logically deduced from one or more noncausal hypotheses.

Temporality

Temporality refers to the necessity that the cause precede the effect in time. This criterion is inarguable, insofar as any claimed observation of causation must involve the putative cause C preceding the putative effect D. It does *not,* however, follow that a reverse time order is evidence against the hypothesis that C can cause D. Rather, observations in which C followed D merely show that C could not have caused D in these instances; they provide no evidence for or against the hypothesis that C can cause D in those instances in which it precedes D. Only if it is found that C cannot precede D can we dispense with the causal hypothesis that C *could* cause D.

Biologic Gradient

Biologic gradient refers to the presence of a dose–response or exposure–response curve with an expected shape. Although Hill referred to a "linear" gradient, without specifying the scale, a linear gradient on one scale, such as the risk, can be distinctly nonlinear on another scale, such as the log risk, the odds, or the log odds. We might relax the expectation from linear to strictly monotonic (steadily increasing or decreasing) or even further merely to monotonic (a gradient that never changes direction). For example, more smoking means more carcinogen exposure and more tissue damage, hence more opportunity for carcinogenesis. Some causal associations, however, show a rapid increase in response (an approximate threshold effect) rather than a strictly monotonic trend. An example is the association between DES and adenocarcinoma of the vagina. A possible explanation is that the doses of DES that were administered were all sufficiently great to produce the maximum effect from DES. Under this hypothesis, for all those exposed to DES, the development of disease would depend entirely on other component causes.

The somewhat controversial topic of alcohol consumption and mortality is another example. Death rates are higher among nondrinkers than among moderate drinkers, but they ascend to the highest levels for heavy drinkers. There is considerable debate about which parts of the J-shaped dose–response curve are causally related to alcohol consumption and which parts are noncausal artifacts stemming from confounding or other biases. Some studies appear to find only an increasing relation between alcohol consumption and mortality, possibly because the categories of alcohol consumption are too broad to distinguish different rates among moderate drinkers and nondrinkers, or possibly because they have less confounding at the lower end of the consumption scale.

Associations that do show a monotonic trend in disease frequency with increasing levels of exposure are not necessarily causal. Confounding can result in a monotonic relation between a noncausal risk factor and disease if the confounding factor itself demonstrates a biologic gradient in its relation with disease. The relation between birth rank and Down syndrome mentioned earlier shows a strong biologic gradient that merely reflects the progressive relation between maternal age and occurrence of Down syndrome.

These issues imply that the existence of a monotonic association is neither necessary nor sufficient for a causal relation. A nonmonotonic relation only refutes those causal hypotheses specific enough to predict a monotonic dose–response curve.

Plausibility

Plausibility refers to the scientific plausibility of an association. More than any other criterion, this one shows how narrowly systems of causal criteria are focused on epidemiology. The starting point is an epidemiologic association. In asking whether it is causal or not, one of the considerations we take into account is its plausibility. From a less parochial perspective, the entire enterprise of causal inference would be viewed as the act of determining how plausible a causal *hypothesis* is. One of the considerations we would take into account would be epidemiologic associations, if they are available. Often they are not, but causal inference must be done nevertheless, with inputs from toxicology, pharmacology, basic biology, and other sciences.

Just as epidemiology is not essential for causal inference, plausibility can change with the times. Sartwell (1960) emphasized this point, citing remarks of Cheever in 1861, who had been commenting on the etiology of typhus before its mode of transmission (via body lice) was known:

> It could be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship to ascribe the typhus, which he there contracted, to the vermin with which bodies of the sick might be infested. An adequate cause, one reasonable in itself, must correct the coincidences of simple experience.

What was to Cheever an implausible explanation turned out to be the correct explanation, because it was indeed the vermin that caused the typhus infection. Such is the problem with plausibility: It is too often based not on logic or data, but only on prior beliefs. This is not to say that biologic knowledge should be discounted when a new hypothesis is being evaluated, but only to point out the difficulty in applying that knowledge.

The Bayesian approach to inference attempts to deal with this problem by requiring that one quantify, on a probability (0 to 1) scale, the certainty that one has in prior beliefs, as well as in new hypotheses. This quantification displays the dogmatism or open-mindedness of the analyst in a public fashion, with certainty values near 1 or 0 betraying a strong commitment of the analyst for or against a hypothesis. It can also provide a means of testing those quantified beliefs against new evidence (Howson and Urbach, 1993). Nevertheless, no approach can transform plausibility into an objective causal criterion.

Coherence

Taken from the U.S. Surgeon General's *Smoking and Health* (1964), the term *coherence* implies that a cause-and-effect interpretation for an association does not conflict with what is known of the natural history and biology of the disease. The examples Hill gave for coherence, such as the histopathologic effect of smoking on bronchial epithelium (in reference to the association between smoking and lung cancer) or the difference in lung cancer incidence by sex, could reasonably be considered examples of plausibility, as well as coherence; the distinction appears to be a fine one. Hill emphasized that the absence of coherent information, as distinguished, apparently, from the presence of conflicting information, should not be taken as evidence against an association being considered causal. On the other hand, the presence of conflicting information may indeed refute a hypothesis, but one must always remember that the conflicting information may be mistaken or misinterpreted. An example mentioned earlier is the "inhalation anomaly" in smoking and lung cancer, the fact that the excess of lung cancers seen among smokers seemed to be concentrated at sites in the upper airways of the lung. Several observers interpreted this anomaly as evidence that cigarettes were not responsible for the excess. Other observations, however, suggested that cigarette-borne carcinogens were deposited preferentially where the excess was observed, and so the anomaly was in fact consistent with a causal role for cigarettes (Wald, 1985).

Experimental Evidence

To different observers, experimental evidence can refer to clinical trials, to laboratory experiments with rodents or other nonhuman organisms, or to both. Evidence from human experiments, however, is seldom available for epidemiologic research questions, and animal evidence relates to different species and usually to levels of exposure very different from those that humans experience. Uncertainty in extrapolations from animals to humans often dominates the uncertainty of quantitative risk assessments (Freedman and Zeisel, 1988; Crouch et al., 1997).

To Hill, however, experimental evidence meant something else: the "experimental, or semi-experimental evidence" obtained from reducing or eliminating a putatively harmful exposure and seeing if the frequency of disease subsequently declines. He called this the strongest possible evidence of causality that can be obtained. It can be faulty, however, as the "semi-experimental" approach is nothing more than a "before-and-after" time trend analysis, which can be confounded or otherwise biased by a host of concomitant secular changes. Moreover, even if the removal of exposure does causally reduce the frequency of disease, it might not be for the etiologic reason hypothesized. The draining of a swamp near a city, for instance, would predictably and causally reduce the rate of yellow fever or malaria in that city the following summer. But it would be a mistake to call this observation the strongest possible evidence of a causal role of miasmas (Poole, 1999).

Analogy

Whatever insight might be derived from analogy is handicapped by the inventive imagination of scientists who can find analogies everywhere. At best, analogy provides a source of more elaborate hypotheses about the associations under study; absence of such analogies reflects only lack of imagination or experience, not falsity of the hypothesis.

We might find naive Hill's examples in which reasoning by analogy from the thalidomide and rubella tragedies made it more likely to him that other medicines and infections might cause other birth defects. But such reasoning is common; we suspect most people find it more credible that smoking might cause, say, stomach cancer, because of its associations, some widely accepted as causal, with cancers in other internal and gastrointestinal organs. Here we see how the analogy criterion can be at odds with either of the two specificity criteria. The more apt the analogy, the less specific are the effects of a cause or the less specific the causes of an effect.

Summary

As is evident, the standards of epidemiologic evidence offered by Hill are saddled with reservations and exceptions. Hill himself was ambivalent about their utility. He did not use the word *criteria* in the speech. He called them "viewpoints" or "perspectives." On the one hand, he asked, "In what circumstances can we pass from this observed *association* to a verdict of *causation*?" (emphasis in original). Yet, despite speaking of verdicts on causation, he disagreed that any "hard-and-fast rules of evidence" existed by which to judge causation: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non" (Hill, 1965).

Actually, as noted above, the fourth viewpoint, temporality, is *a sine qua non* for causal explanations of observed associations. Nonetheless, it does not bear on the hypothesis that an exposure is capable of causing a disease in situations as yet unobserved (whether in the past or the future). For suppose every exposed case of disease ever reported had received the exposure after developing the disease. This reversed temporal relation would imply that exposure had not caused disease among these reported cases, and thus would refute the hypothesis that it had. Nonetheless, it would *not* refute the hypothesis that the exposure is *capable* of causing the disease, or that it had caused the disease in unobserved cases. It would mean only that we have no worthwhile epidemiologic evidence relevant to that hypothesis, for we had not yet seen what became of those exposed before disease occurred relative to those unexposed. Furthermore, what appears to be a causal sequence could represent reverse causation if preclinical symptoms of the disease lead to exposure, and then overt disease follows, as when patients in pain take analgesics, which may be the result of disease that is later diagnosed, rather than a cause.

Other than temporality, there is no necessary or sufficient criterion for determining whether an observed association is causal. Only when a causal hypothesis is elaborated to the extent that one can predict from it a particular form of consistency, specificity, biologic gradient, and so forth, can "causal criteria" come into play in evaluating causal hypotheses, and even then they do not come into play in evaluating the general hypothesis per se, but only some specific causal hypotheses, leaving others untested.

This conclusion accords with the views of Hume and many others that causal inferences cannot attain the certainty of logical deductions. Although some scientists continue to develop causal considerations as aids to inference (Susser, 1991), others argue that it is detrimental to cloud the inferential process by considering checklist criteria (Lanes and Poole, 1984). An intermediate, refutationist approach seeks to transform proposed criteria into deductive tests of causal hypotheses (Maclure, 1985; Weed, 1986). Such an approach helps avoid the temptation to use causal criteria simply to buttress pet theories at hand, and instead allows epidemiologists to focus on evaluating competing causal theories using crucial observations. Although this refutationist approach to causal inference may seem at odds with the common implementation of Hill's viewpoints, it actually seeks to answer the fundamental question posed by Hill, and the ultimate purpose of the viewpoints he promulgated:

> What [the nine viewpoints] can do, with greater or less strength, is to help us to make up our minds on the fundamental question—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect? (Hill, 1965)

The crucial phrase "equally or more likely than cause and effect" suggests to us a subjective assessment of the certainty, or probability of the causal hypothesis at issue relative to another hypothesis. Although Hill wrote at a time when expressing uncertainty as a probability was unpopular in statistics, it appears from his statement that, for him, causal inference is a subjective matter of degree of personal belief, certainty, or conviction. In any event, this view is precisely that of subjective Bayesian statistics (Chapter 18).

It is unsurprising that case studies (e.g., Weed and Gorelick, 1996) and surveys of epidemiologists (Holman et al., 2001) show, contrary to the rhetoric that often attends invocations of causal criteria, that epidemiologists have *not* agreed on a set of causal criteria or on how to apply them. In one study in which epidemiologists were asked to employ causal criteria to fictional summaries of epidemiologic literatures, the agreement was only slightly greater than would have been expected by chance (Holman et al., 2001). The typical use of causal criteria is to make a case for a position for or against causality that has been arrived at by other, unstated means. Authors pick and choose among the criteria they deploy, and define and weight them in *ad hoc* ways that depend only on the exigencies of the discussion at hand. In this sense, causal criteria appear to function less like standards or principles and more like values (Poole, 2001b), which vary across individual scientists and even vary within the work of a single scientist, depending on the context and time. Thus universal and objective causal criteria, if they exist, have yet to be identified.

A definition of causal effect for epidemiological research

M A Hernán

J Epidemiol Community Health 2004;**58**:265–271. doi: 10.1136/jech.2002.006361

Estimating the causal effect of some exposure on some outcome is the goal of many epidemiological studies. This article reviews a formal definition of causal effect for such studies. For simplicity, the main description is restricted to dichotomous variables and assumes that no random error attributable to sampling variability exists. The appendix provides a discussion of sampling variability and a generalisation of this causal theory. The difference between association and causation is described—the redundant expression ''causal effect'' is used throughout the article to avoid confusion with a common use of ''effect'' meaning simply statistical association—and shows why, in theory, randomisation allows the estimation of causal effects without further assumptions. The article concludes with a discussion on the limitations of randomised studies. These limitations are the reason why methods for causal inference from observational data are needed.

INDIVIDUAL CAUSAL EFFECTS

Zeus is a patient waiting for a heart transplant. On 1 January, he received a new heart. Five days later, he died. Imagine that we can somehow know, perhaps by divine revelation, that had Zeus not received a heart transplant on 1 January (all others things in his life being unchanged) then he would have been alive five days later. Most people equipped with this information would agree that the transplant caused Zeus' death. The intervention had a causal effect on Zeus' five day survival.

Another patient, Hera, received a heart transplant on 1 January. Five days later she was alive. Again, imagine we can somehow know that had Hera not received the heart on 1 January (all other things being equal) then she would still have been alive five days later. The transplant did not have a causal effect on Hera's five day survival.

These two vignettes illustrate how human reasoning for causal inference works: we compare (often only mentally) the outcome when action A is present with the outcome when action A is absent, all other things being equal. If the two outcomes differ, we say that the action A has a causal effect, causative or preventive, on the outcome. Otherwise, we say that the action A has no causal effect on the outcome. In epidemiology, A is commonly referred to as exposure or treatment.

The next step is to make this causal intuition of ours amenable to mathematical and statistical analysis by introducing some notation. Consider a dichotomous exposure variable A (1: exposed, 0: unexposed) and a dichotomous outcome variable Y (1: death, 0: survival). Table 1 shows the data from a heart transplant observational study with 20 participants. Let $Y_{a=1}$ be the outcome variable that would have been observed under the exposure value $a = 1$, and $Y_{a=0}$ the outcome variable that would have been observed under the exposure value $a = 0$. (Lowercase a represents a particular value of the variable A.) As shown in table 2, Zeus has $Y_{a=1} = 1$ and $Y_{a=0} = 0$ because he died when exposed but would have survived if unexposed.

We are now ready to provide a formal definition of causal effect for each person: exposure has a causal effect if $Y_{a=0} \neq Y_{a=1}$. Table 2 is all we need to decide that the exposure has an effect on Zeus' outcome because $Y_{a=0} \neq Y_{a=1}$, but not on Hera's outcome because $Y_{a=0} = Y_{a=1}$. When the exposure has no causal effect for any subject—that is, $Y_{a=0} = Y_{a=1}$ for all subjects—we say that the *sharp causal null hypothesis* is true.

The variables $Y_{a=1}$ and $Y_{a=0}$ are known as potential outcomes because one of them describes the subject's outcome value that would have been observed under a potential exposure value that the subject did not actually experience. For example, $Y_{a=0}$ is a potential outcome for exposed Zeus, and $Y_{a=1}$ is a potential outcome for unexposed Hera. Because these outcomes would have been observed in situations that did not actually happen (that is, in counter to the fact situations), they are also known as *counterfactual outcomes*. For each subject, one of the counterfactual outcomes is actually factual—the one that corresponds to the exposure level or treatment regimen that the subject actually received. For example, if $A = 1$ for Zeus, then $Y_{a=1} = Y_{a=A} = Y$ for him.

The fundamental problem of causal inference should now be clear. Individual causal effects are defined as a contrast of the values of counterfactual outcomes, but only one of those values is observed. Table 3 shows the observed data and each subject's observed counterfactual outcome: the one corresponding to the exposure value actually experienced by the subject. All other counterfactual outcomes are missing. The unhappy conclusion is that, in general, individual causal effects cannot be identified because of missing data.

POPULATION CAUSAL EFFECT

We define the probability $\Pr[Y_a = 1]$ as the proportion of subjects that would have developed

Correspondence to:
Dr M Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA; miguel_hernan@post.harvard.edu

Accepted for publication
29 August 2003

Table 1 Data from a study with dichotomous exposure *A* and outcome *Y*

ID	*A*	*Y*
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Circe	0	0
Ares	1	1
Athene	1	1
Eros	1	1
Aphrodite	1	1
Prometheus	1	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

Table 2 Counterfactual outcomes of subjects in a study with dichotomous exposure *A* and outcome *Y*

ID	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Circe	0	1
Ares	1	1
Athene	1	1
Eros	0	1
Aphrodite	0	1
Prometheus	0	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

the outcome Y had all subjects in the population of interest received exposure value a. We also refer to $\Pr[Y_a = 1]$ as the risk of Y_a. The exposure has a causal effect in the population if $\Pr[Y_{a=1} = 1] \neq \Pr[Y_{a=0} = 1]$.

Suppose that our population is comprised by the subjects in table 2. Then $\Pr[Y_{a=1} = 1] = 10/20 = 0.5$, and $\Pr[Y_{a=0} = 1] = 10/20 = 0.5$. That is, 50% of the patients would have died had everybody received a heart transplant, and 50% would have died had nobody received a heart transplant. The exposure has no effect on the outcome at the population level. When the exposure has no causal effect in the population, we say that the *causal null hypothesis* is true.

Unlike individual causal effects, population causal effects *can* sometimes be computed—or, more rigorously, consistently estimated (see appendix)—as discussed below. Hereafter we refer to the ''population causal effect'' simply as ''causal effect''. Some equivalent definitions of causal effect are

(a) $\Pr[Y_{a=1} = 1] - \Pr[Y_{a=0} = 1] \neq 0$

(b) $\Pr[Y_{a=1} = 1] / \Pr[Y_{a=0} = 1] \neq 1$

(c) $(\Pr[Y_{a=1} = 1] / \Pr[Y_{a=1} = 0]) / (\Pr[Y_{a=0} = 1] / \Pr[Y_{a=0} = 0]) \neq 1$

where the left hand side of inequalities (a), (b), and (c) is the causal risk difference, risk ratio, and odds ratio, respectively. The causal risk difference, risk ratio, and odds ratio (and other causal parameters) can also be used to quantify the strength of the causal effect when it exists. They measure the same causal effect in different scales, and we refer to them as *effect measures*.

ASSOCIATION AND CAUSATION

To characterise association, we first define the probability $\Pr[Y = 1 | A = a]$ as the proportion of subjects that developed the outcome Y among those subjects in the population of interest that happened to receive exposure value a. We also refer to $\Pr[Y = 1 | A = a]$ as the risk of Y given $A = a$. Exposure and outcome are associated if $\Pr[Y = 1 | A = 1] \neq \Pr[Y = 1 | A = 0]$. In our population of

Table 3 Data and observed counterfactual outcomes from a study with dichotomous exposure *A* and outcome *Y*

ID	*A*	*Y*	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Circe	0	0	0	?
Ares	1	1	?	1
Athene	1	1	?	1
Eros	1	1	?	1
Aphrodite	1	1	?	1
Prometheus	1	1	?	1
Selene	1	1	?	1
Hermes	1	0	?	0
Eos	1	0	?	0
Helios	1	0	?	0

table 1, exposure and outcome are associated because $\Pr[Y=1|A=1]=7/13$, and $\Pr[Y=1|A=0]=3/7$. Some equivalent definitions of association are

(a) $\Pr[Y=1|A=1]-\Pr[Y=1|A=0]\neq 0$

(b) $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]\neq 1$

(c) $(\Pr[Y=1|A=1]/\Pr[Y=0|A=1])/(\Pr[Y=1|A=0]/\Pr[Y=0|A=0])\neq 1$

where the left hand side of the inequalities (a), (b), and (c) is the associational risk difference, risk ratio, and odds ratio, respectively. The associational risk difference, risk ratio, and odds ratio (and other association parameters) can also be used to quantify the strength of the association when it exists. They measure the same association in different scales, and we refer to them as *association measures*.

When A and Y are not associated, we say that A does not predict Y, or vice versa. Lack of association is represented by $Y \coprod A$ (or, equivalently, $A \coprod Y$), which is read as Y and A are independent.

Note that the risk $\Pr[Y=1|A=a]$ is computed using the subset of subjects of the population that meet the condition "having actually received exposure a" (that is, it is a conditional probability), whereas the risk $\Pr[Y_a=1]$ is computed using *all* subjects of the population had they received the counterfactual exposure a (that is, it is an unconditional or marginal probability). Therefore, association is defined by a different risk in two disjoint subsets of the population determined by the subjects' actual exposure value, whereas causation is defined by a different risk in the same subset (for example, the entire population) under two potential exposure values (fig 1). This radically different definition accounts for the well known adage "association is not causation." When an association measure differs from the corresponding effect measure, we say that there is *bias* or *confounding*.

COMPUTATION OF CAUSAL EFFECTS VIA RANDOMISATION

Unlike association measures, effect measures cannot be directly computed because of missing data (see table 3). However, effect measures can be computed—or, more rigorously, consistently estimated (see appendix)—in randomised experiments.

Suppose we have a (near-infinite) population and that we flip a coin for each subject in such population. We assign the subject to group 1 if the coin turns tails, and to group 2 if it turns heads. Next we administer the treatment or exposure of interest ($A=1$) to subjects in group 1 and placebo ($A=0$) to those in group 2. Five days later, at the end of the study, we compute the mortality risks in each group, $\Pr[Y=1|A=1]$ and $\Pr[Y=1|A=0]$. For now, let us assume that this randomised experiment is ideal in all other respects (no loss to follow up, full compliance with assigned treatment, blind assignment).

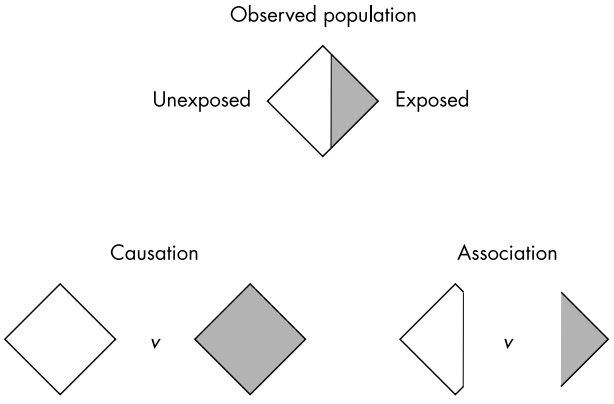


Figure 1 Causation is defined by a different risk in the entire population under two potential exposure values; association is defined by a different risk in the subsets of the population determined by the subjects' actual exposure value.

We will show that, in such a study, the observed risk $\Pr[Y=1|A=a]$ is equal to the counterfactual risk $\Pr[Y_a=1]$, and therefore the associational risk ratio equals the causal risk ratio.

First note that, when subjects are randomly assigned to groups 1 and 2, the proportion of deaths among the exposed, $\Pr[Y=1|A=1]$, will be the same whether subjects in group 1 receive the exposure and subjects in group 2 receive placebo, or vice versa. Because group membership is randomised, both groups are "comparable": which particular group got the exposure is irrelevant for the value of $\Pr[Y=1|A=1]$. (The same reasoning applies to $\Pr[Y=1|A=0]$.) Formally, we say that both groups are exchangeable.

Exchangeability means that the risk of death in group 1 would have been the same as the risk of death in group 2 had subjects in group 1 received the exposure given to those in group 2. That is, the risk under the potential exposure value a among the exposed, $\Pr[Y_a=1|A=1]$, equals the risk under the potential exposure value a among the unexposed, $\Pr[Y_a=1|A=0]$, for $a=0$ and $a=1$. An obvious consequence of these (conditional) risks being equal in all subsets defined by exposure status in the population is that they must be equal to the (marginal) risk under exposure value a in the whole population: $\Pr[Y_a=1|A=1]=\Pr[Y_a=1|A=0]=\Pr[Y_a=1]$. In other words, under exchangeability, the actual exposure does not predict the counterfactual outcome; they are independent, or $Y_a \coprod A$ for all values a. Randomisation produces exchangeability.

We are only one step short of showing that the observed risk $\Pr[Y=1|A=a]$ equals the counterfactual risk $\Pr[Y_a=1]$ in ideal randomised experiments. By definition, the value of the counterfactual outcome Y_a for subjects who actually received exposure value a is their observed outcome value Y. Then, among those who actually received exposure value a, the risk under the potential exposure value a is trivially equal to the observed risk. That is, $\Pr[Y_a=1|A=a]=\Pr[Y=1|A=a]$.

Let us now combine the results from the two previous paragraphs. Under exchangeability, $Y_a \coprod A$ for all a, the conditional risk among those exposed to a is equal to the marginal risk had the whole population been exposed to a: $\Pr[Y_a=1|A=1]=\Pr[Y_a=1|A=0]=\Pr[Y_a=1]$. And by definition of counterfactual outcome $\Pr[Y_a=1|A=a]=\Pr[Y=1|A=a]$. Therefore, the observed risk $\Pr[Y=1|A=a]$ equals the counterfactual risk $\Pr[Y_a=1]$. In ideal randomised experiments, association *is* causation. On the other hand, in non-randomised (for example, observational) studies association is not necessarily causation because of potential lack of exchangeability of exposed and unexposed subjects. For example, in our heart transplant study, the risk of death under no treatment is different for the exposed and the unexposed: $\Pr[Y_{a=0}=1|A=1]=7/13\neq\Pr[Y_{a=0}=1|A=0]=3/7$. We say that the exposed had a worse prognosis, and therefore a greater risk of death, than the unexposed, or that $Y_a \coprod A$ does not hold for $a=0$.

INTERVENTIONS AND CAUSAL QUESTIONS

We have so far assumed that the counterfactual outcomes Y_a exist and are well defined. However, that is not always the case.

Suppose women ($S=1$) have a greater risk of certain disease Y than men ($S=0$)—that is, $\Pr[Y=1|S=1]>\Pr[Y=1|S=0]$. Does sex S has a causal effect on the risk

of Y—that is, $\Pr[Y_{s=1}=1] > \Pr[Y_{s=0}=1]$? This question is quite vague because it is unclear what we mean by the risk of Y had everybody been a woman (or a man). Do we mean the risk of Y had everybody ''carried a pair of X chromosomes'', ''been brought up as a woman'', ''had female genitalia'', or ''had high levels of oestrogens between adolescence and menopausal age''? Each of these definitions of the exposure ''female sex'' would lead to a different causal effect.

To give an unambiguous meaning to a causal question, we need to be able to describe the interventions that would allow us to compute the causal effect in an ideal randomised experiment. For example, ''administer 30 μg/day of ethinyl estradiol from age 14 to age 45'' compared with ''administer placebo.'' That some interventions sound technically unfeasible or plainly crazy simply indicates that the formulation of certain causal questions (for example, the effect of sex, high serum LDL-cholesterol, or high HIV viral load on the risk of certain disease) is not always straightforward. A counterfactual approach to causal inference highlights the imprecision of ambiguous causal questions, and the need for a common understanding of the interventions involved.

LIMITATIONS OF RANDOMISED EXPERIMENTS

We now review some common methodological problems that may lead to bias in randomised experiments. To fix ideas, suppose we are interested in the causal effect of a heart transplant on one year survival. We start with a (near-infinite) population of potential recipients of a transplant, randomly allocate each subject in the population to either transplant ($A=1$) or medical treatment ($A=0$), and ascertain how many subjects die within the next year ($Y=1$) in each group. We then try to measure the effect of heart transplant on survival by computing the associational risk ratio $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]$, which is theoretically equal to the causal risk ratio $\Pr[Y_{a=1}=1]/\Pr[Y_{a=0}=1]$. Consider the following problems:

- *Loss to follow up*. Subjects may be lost to follow up or drop out of the study before their outcome is ascertained. When this happens, the risk $\Pr[Y=1|A=a]$ cannot be computed because the value of Y is not available for some people. Instead we can compute $\Pr[Y=1|A=a, C=0]$ where C indicates whether the subject was lost (1: yes, 0: no). This restriction to subjects with $C=0$ is problematic because subjects that were lost ($C=1$) may not be exchangeable with subjects who remained through the end of the study ($C=0$). For example, if subjects who did not receive a transplant ($A=0$) and who had a more severe disease decide to leave the study, then the risk $\Pr[Y=1|A=0, C=0]$ among those remaining in the study would be lower than the risk $\Pr[Y=1|A=0]$ among those originally assigned to medical treatment. Our association measure $\Pr[Y=1|A=1, C=0]/\Pr[Y=1|A=0, C=0]$ would not generally equal the effect measure $\Pr[Y_{a=1}=1]/\Pr[Y_{a=0}=1]$.
- *Non-compliance*. Subjects may not adhere to the assigned treatment. Let A be the exposure to which subjects were randomly assigned, and B the exposure they actually received. Suppose some subjects that had been assigned to medical treatment ($A=0$) obtained a heart transplant outside of the study ($B=1$). In an ''intention to treat'' analysis, we compute $\Pr[Y=1|A=a]$, which equals $\Pr[Y_a=1]$. However, we are not interested in the causal effect of assignment A, a misclassified version of the true exposure B, but on the causal effect of B itself. The alternative ''as treated'' approach—using $\Pr[Y=1|B=b]$ for causal inference—is problematic. For example, if the most severely ill subjects in the $A=0$ group seek a heart transplant ($B=1$) outside of the study, then the group $B=1$ would include a higher proportion of severely ill subjects than the group $B=0$. The groups $B=1$ and $B=0$ would not be exchangeable—that is, $\Pr[Y=1|B=b] \neq \Pr[Y_b=1]$. In the presence of non-compliance, an intention to treat analysis guarantees exchangeability of the groups defined by a misclassified exposure (the original assignment), whereas an as treated analysis guarantees a correct classification of exposure but not exchangeability of the groups defined by this exposure. However, the intention to treat analysis is often preferred because, unlike the as treated analysis, it provides an unbiased association measure if the sharp causal null hypothesis holds for the exposure B.
- *Unblinding*. When the study subjects are aware of the treatment they receive (as in our heart transplant study), they may change their behaviour accordingly. For example, those who received a transplant may change their diet to keep their new heart healthy. The equality $\Pr[Y=1|A=a] = \Pr[Y_a=1]$ still holds, but now the causal effect of A combines the effects of the transplant and the dietary change. To avoid this problem, knowledge of the level of exposure assigned to each group is withheld from subjects and their doctors (they are ''blinded''), when possible. The goal is to ensure that the whole effect, if any, of the exposure assignment A is solely attributable to the exposure received B (the heart transplant in our example). When this goal is achieved, we say that the *exclusion restriction* holds—that is, $Y_{a=0,b} = Y_{a=1,b}$ for all subjects and all values b and, specifically, for the value B observed for each subject. In non-blinded studies, or when blinding does not work (for example, the well known side effects of a treatment make apparent who is taking it), the exclusion restriction cannot be guaranteed, and therefore the intention to treat analysis may not yield an unbiased association measure even under the sharp causal null hypothesis for exposure B.

In summary, the fact that exchangeability $Y_a \coprod A$ holds in a well designed randomised experiment does not guarantee an unbiased estimate of the causal effect because: *i*) Y may not be measured for all subjects (loss to follow up), *ii*) A may be a misclassified version of the true exposure (non-compliance), and *iii*) A may be a combination of the exposure of interest plus other actions (unblinding). Causal inference from randomised studies in the presence of these problems requires similar assumptions and analytical methods as causal inference from observational studies.

Leaving aside these methodological problems, randomised experiments may be unfeasible because of ethical, logistic, or financial reasons. For example, it is questionable that an ethical committee would have approved our heart transplant study. Hearts are in short supply and society favours assigning them to subjects who are more likely to benefit from the transplant, rather than assigning them randomly among potential recipients. Randomised experiments of harmful exposures (for example, cigarette smoking) are generally unacceptable too. Frequently, the only option is conducting observational studies in which exchangeability is not guaranteed.

BIBLIOGRAPHICAL NOTES

Hume[1] hinted a counterfactual theory of causation, but the application of counterfactual theory to the estimation of causal effects via randomised experiments was first formally proposed by Neyman.[2] Rubin[3 4] extended Neyman's theory to the estimation of the effects of fixed exposures in randomised *and* observational studies. Fixed exposures are exposures that either are applied at one point in time only or never change over time. Examples of fixed exposures in epidemiology

are a surgical intervention, a traffic accident, a one dose immunisation, or a medical treatment that is continuously administered during a given period regardless of its efficacy or side effects. Rubin's counterfactual model has been discussed by Holland and others.[5]

Robins[6 7] proposed a more general counterfactual model that permits the estimation of total and direct effects of fixed *and* time varying exposures in longitudinal studies, whether randomised or observational. Examples of time varying exposures in epidemiology are a medical treatment, diet, cigarette smoking, or an occupational exposure. For simplicity of presentation, our article was restricted to the effects of fixed exposures. The use of the symbol $\amalg$ to denote independence was introduced by Dawid.[8]

ACKNOWLEDGEMENTS

The author is deeply indebted to James Robins for his contributions to earlier versions of this manuscript.

Funding: NIH grant KO8-AI-49392

Conflicts of interest: none declared.

APPENDIX

A1 SAMPLING VARIABILITY

Our descriptions of causal effect and exchangeability have relied on the idea that we somehow collected information from all the subjects in the population of interest. This simplification has been useful to focus our attention on the conceptual aspects of causal inference, by keeping them separate from aspects related to random statistical variability. We now extend our definitions to more realistic settings in which random variability exists.

Many real world studies are based on samples of the population of interest. The first consequence of working with samples is that, even if the counterfactual outcomes of all subjects in the study were known, one cannot obtain the exact proportion of subjects in the population who had the outcome under exposure value a—that is, the probability $\Pr[Y_{a=0}=1]$ cannot be directly computed. One can only estimate this probability. Consider the subjects in table 2. We have previously viewed them as forming a 20 person population. Let us now view them as a random sample of a much larger population. In this sample, the proportion of subjects who would have died if unexposed is $\hat{\Pr}[Y_{a=0}=1] = 10/20 = 0.5$, which does not have to be exactly equal to the proportion of subjects who would have died if the entire population had been unexposed, $\Pr[Y_{a=0}=1]$. We use the sample proportion $\hat{\Pr}[Y_a=1]$ to estimate the population probability $\Pr[Y_a=1]$. (The ''hat'' over Pr indicates that $\hat{\Pr}[Y_a=1]$ is an estimator.) We say that $\hat{\Pr}[Y_a=1]$ is a *consistent estimator* of $\Pr[Y_a=1]$ because the larger the number of subjects in the sample, the smaller the difference between $\hat{\Pr}[Y_a=1]$ and $\Pr[Y_a=1]$ is expected to be. In the long run (that is, if the estimator is applied to infinite samples of the population), the mean difference is expected to become zero.

There is a causal effect of A on Y in such population if $\Pr[Y_{a=1}=1] \neq \Pr[Y_{a=0}=1]$. This definition, however, cannot be directly applied because the population probabilities $\Pr[Y_a=1]$ cannot be computed, but only consistently estimated by the sample proportions $\hat{\Pr}[Y_a=1]$. Therefore, one cannot conclude with certainty that there is (or there is not) a causal effect. Rather, standard statistical procedures are needed to test the causal null hypothesis $\Pr[Y_{a=1}=1] = \Pr[Y_{a=0}=1]$ by comparing $\hat{\Pr}[Y_{a=1}=1]$ and $\hat{\Pr}[Y_{a=1}=1]$, and to compute confidence intervals for the effect measures. The availability of data from only a sample of subjects in the population, even if the values of all their counterfactual outcomes were known, is the first reason why statistics is necessary in causal inference.

The previous discussion assumes that one can have access to the values of both counterfactual outcomes for each subject in the sample (as in table 2), whereas in real world studies one can only access the value of one counterfactual outcome for each subject (as in table 3). Therefore, whether one is working with the whole population or with a sample, neither the probability $\Pr[Y_a=1]$ or its consistent estimator $\hat{\Pr}[Y_a=1]$ can be directly computed for any value a. Instead, one can compute the sample proportion of subjects that develop the outcome among the exposed, $\hat{\Pr}[Y=1|A=1] = 7/13$, and among the unexposed, $\hat{\Pr}[Y=1|A=0] = 3/7$. There are two major conceptualisations of this problem:

(1) The population of interest is near infinite and we hypothesise that all subjects in the population are randomly assigned to either $A=1$ or $A=0$. Exchangeability of the exposed and unexposed would hold in the population—that is, $\Pr[Y_a=1] = \Pr[Y=1|A=a]$. Now we can see our sample as a random sample from this population where exposure is randomly assigned. The problem boils down to standard statistical inference with the sample proportion $\hat{\Pr}[Y=1|A=a]$ being a consistent estimator of the population probability $\Pr[Y=1|A=a]$. This is the simplest conceptualisation.

(2) Only the subjects in our sample, not all subjects in the entire population, are randomly assigned to either $A=1$ or $A=0$. Because of the presence of random sampling variability, we do not expect that exchangeability will exactly hold in our sample. For example, suppose that 100 subjects are randomly assigned to either heart transplant ($A=1$) or medical treatment ($A=0$). Each subject can be classified as good or bad prognosis at the time of randomisation. We say that the groups $A=0$ and $A=1$ are exchangeable if they include exactly the same proportion of subjects with bad prognosis. By chance, it is possible that 17 of the 50 subjects assigned to $A=1$ and 13 of the 50 subjects assigned to $A=0$ had bad prognosis. The two groups are not exactly exchangeable. However, if we could draw many additional 100 person samples from the population and repeat the randomised experiment in each of these samples (or, equivalently, if we could increase the size of our original sample), then the imbalances between the groups $A=1$ and $A=0$ would be increasingly attenuated. Under this conceptualisation, the sample proportion $\hat{\Pr}[Y=1|A=a]$ is a consistent estimator of $\hat{\Pr}[Y_a=1]$, and $\hat{\Pr}[Y_a=1]$ is a consistent estimator of the population proportion $\Pr[Y_a=1]$ if our sample is a random sample of the population of interest. This is the most realistic conceptualisation.

Under either conceptualisation, standard statistical procedures are needed to test the causal null hypothesis $\Pr[Y_{a=1}=1] = \Pr[Y_{a=0}=1]$ by comparing $\hat{\Pr}[Y=1|A=1]$ and $\hat{\Pr}[Y=1|A=0]$, and to compute confidence intervals for the estimated association measures, which are consistent estimators of the effect measures. The availability of the value of only one counterfactual outcome for each subject, regardless of whether all subjects in the population of interest are or are not included the study (and regardless of which conceptualisation is used), is the second reason why statistics is necessary in causal inference.

A2 GENERALISATIONS

A2.1 Definition of causal effect

We defined causal effect of the exposure on the outcome, $\Pr[Y_{a=1}=1] \neq \Pr[Y_{a=0}=1]$, as a difference between the counterfactual risk of the outcome had everybody in the

population of interest been exposed and the counterfactual risk of the outcome had everybody in the population been unexposed. In some cases, however, investigators may be more interested in the causal effect of the exposure in a subset of the population of interest (rather than the effect in the entire population). This causal effect is defined as a contrast of counterfactual risks in that subset of the population of interest.

A common choice is the subset of the population comprised by the subjects that were actually exposed. Thus, we can define the *causal effect in the exposed* as $\Pr[Y_{a=1}=1|A=1] \neq \Pr[Y_{a=0}=1|A=1]$ or, by definition of counterfactual outcome, $\Pr[Y=1|A=1] \neq \Pr[Y_{a=0}=1|A=1]$. That is, there is a causal effect in the exposed if the risk of the outcome among the exposed subjects in the population of interest does not equal the counterfactual risk of the outcome had the exposed subjects in the population been unexposed. The causal risk difference in the exposed is $\Pr[Y=1|A=1]-\Pr[Y_{a=0}=1|A=1]$, the causal risk ratio in the exposed is $\Pr[Y=1|A=1]/\Pr[Y_{a=0}=1|A=1]$, and the causal odds ratio in the exposed is $(\Pr[Y=1|A=1]/\Pr[Y=0|A=1])/(\Pr[Y_{a=0}=1|A=1]/\Pr[Y_{a=0}=0|A=1])$.

The causal effect in the entire population can be computed under the condition that the exposed and the unexposed are exchangeable—that is, $Y_a \coprod A$ for $a=0$ and $a=1$. On the other hand, the causal effect in the exposed can be computed under the weaker condition that the exposed and the unexposed are exchangeable had they been unexposed—that is, $Y_a \coprod A$ for $a=0$ only. Under this weaker exchangeability condition, the risk of the outcome under no exposure is equal for the exposed and the unexposed: $\Pr[Y_{a=0}=1|A=1] = \Pr[Y_{a=0}=1|A=0]$. By definition of a counterfactual outcome $\Pr[Y_{a=0}=1|A=0] = \Pr[Y=1|A=0]$. Therefore, when the exposed and unexposed are exchangeable under $a=0$, $\Pr[Y_{a=0}=1|A=1] = \Pr[Y_{a=0}=1|A=0] = \Pr[Y=1|A=0]$. We decided to restrict our discussion to the causal effect in the entire population and not to the causal effect in the exposed because the latter cannot be directly generalised to time varying exposures.

A2.2 Non-dichotomous outcome and exposure

The definition of causal effect can be generalised to non-dichotomous exposure A and outcome Y. Let $\mathrm{E}[Y_a]$ be the mean counterfactual outcome had all subjects in the population received exposure level a. For discrete outcomes, the expected value $\mathrm{E}[Y_a]$ is defined as the weighted sum $\sum_y y p_{Y_a}(y)$ over all possible values y of the random variable Y_a, where $p_{Y_a}(\cdot)$ is the probability mass function of Y_a—that is, $p_{Y_a}(y) = \Pr[Y_a = y]$. For continuous outcomes, the expected value $\mathrm{E}[Y_a]$ is defined as the integral $\int y\, f_{Y_a}(y)dy$ over all possible values y of the random variable Y_a, where $f_{Y_a}(\cdot)$ is the probability density function of Y_a. A common representation of the expected value for discrete and continuous outcomes is $\mathrm{E}[Y_a] = \int y\, dF_{Y_a}(y)$, where $F_{Y_a}(\cdot)$ is the cumulative density function (cdf) of the random variable Y_a.

We say that there is a population *average causal effect* if $\mathrm{E}[Y_a] \neq \mathrm{E}[Y_{a'}]$ for any two values a and a'. In ideal randomised experiments, the expected value $\mathrm{E}[Y_a]$ can be consistently estimated by the average of Y among subjects with $A=a$. For dichotomous outcomes, $\mathrm{E}[Y_a] = \Pr[Y_a = 1]$.

The average causal effect is defined by the contrast of $\mathrm{E}[Y_a]$ and $\mathrm{E}[Y_{a'}]$. When we talk of ''the causal effect of heart transplant (A)'' we mean the contrast between ''receiving a heart transplant ($a=1$)'' and ''not receiving a heart transplant ($a=0$).'' In this case, we may not need to be explicit about the particular contrast because there are only two possible actions, and therefore only one possible contrast. But for non-dichotomous exposure variables A, the particular contrast of interest needs to be specified. For example, ''the causal effect of aspirin'' is meaningless unless we specify that the contrast of interest is, say, ''taking 150 mg of aspirin daily for five years'' compared with ''not taking aspirin''. Note that this causal effect is well defined even if counterfactual outcomes under interventions other than those involved in the causal contrast of interest are not well defined or even do not exist (for example, ''taking 1 kg of aspirin daily for five years'').

The average causal effect, defined as a contrast of means of counterfactual outcomes, is the most commonly used causal effect. However, the causal effect may also be defined by a contrast of, say, medians, variances, or cdfs of counterfactual outcomes. In general, the causal effect can be defined as a contrast of any functional of the distributions of counterfactual outcomes under different exposure values. The causal null hypothesis refers to the particular contrast of functionals (means, medians, variances, cdfs, ...) used to define the causal effect.

A2.3 Non-deterministic counterfactual outcomes

We have defined the counterfactual outcome Y_a as the subject's outcome had he experienced exposure value a. For example, in our first vignette, Zeus would have died if treated and would have survived if untreated. This definition of counterfactual outcome is deterministic because each subject has a fixed value for each counterfactual outcome, for example, $Y_{a=1}=1$ and $Y_{a=0}=0$ for Zeus. However, we could imagine a world in which Zeus has certain probability of dying, say $Q_{Y_{a=1}}(1)=0.9$, if treated and certain probability of dying, say $Q_{Y_{a=0}}(1)=0.1$, if untreated. This is a non-deterministic or stochastic definition of counterfactual outcome because the probabilities $Q_{Y_a}(\cdot)$ are not zero or one. In general, the probabilities $Q_{Y_a}(\cdot)$ vary across subjects (that is, they are random) because not all subjects are equally susceptible to develop the outcome. For discrete outcomes, the expected value $\mathrm{E}[Y_a]$ is then defined as the weighted sum $\sum_y y\, p_{Y_a}(y)$ over all possible values y of the random variable Y_a, where the probability mass function $p_{Y_a}(\cdot) = \mathrm{E}[Q_{Y_a}(\cdot)]$.

More generally, a non-deterministic definition of counterfactual outcome does not attach some particular value of the random variable Y_a to each subject, but rather a statistical distribution $\Theta_{Y_a}(\cdot)$ of Y_a. The deterministic definition of counterfactual outcome implies that the cdf $\Theta_{Y_a}(y)$ can only take values 0 or 1 for all y. The use of random distributions of Y_a (that is, distributions that may vary across subjects) to allow for non-deterministic counterfactual outcomes does not imply any modification in the definition of average causal effect or the methods used to estimate it. To show this, first note that $\mathrm{E}[Y_a] = \mathrm{E}[E[Y_a|\Theta_{Y_a}(\cdot)]]$. Therefore, $\mathrm{E}[Y_a] = E\left[\int y\, d\Theta_{Y_a}(y)\right] = \int y\, dE[\Theta_{Y_a}(y)] = \int y\, dF_{Y_a}(y)$ because $F_{Y_a}(\cdot) = E[\Theta_{Y_a,i}(\cdot)]$. The non-deterministic definition of causal effect is a generalisation of the deterministic definition in which $\Theta_{Y_a}(\cdot)$ is a general cdf that may take values between 0 and 1.

The choice of deterministic compared with non-deterministic counterfactual outcomes has no consequences for the definition of the average causal effect and the point estimation of effect measures based on averages of counterfactual outcomes. However, this choice has implications for the computation of confidence intervals for the effect measures.[9]

A3 NO INTERACTION BETWEEN SUBJECTS

An implicit assumption in our definition of individual causal effect is that a subject's counterfactual outcome under exposure value a does not depend on other subjects' exposure value. This assumption was labelled ''no interaction between

units'' by Cox,[10] and ''stable-unit-treatment-value assumption (SUTVA)'' by Rubin.[11] If this assumption does not hold (for example, in studies dealing with contagious diseases or educational programmes), then individual causal effects cannot be identified by using the hypothetical data in table 2. Most methods for causal inference assume that SUTVA holds.

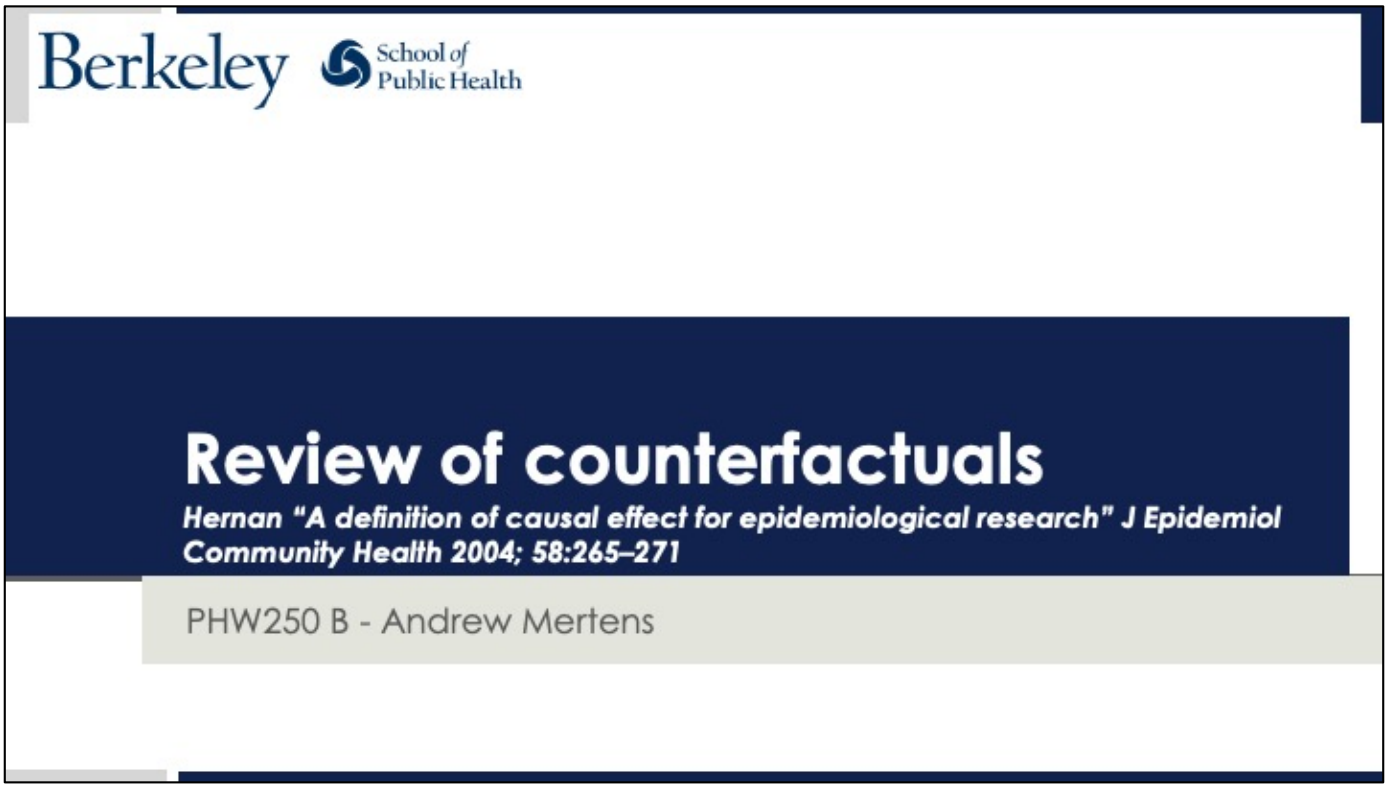
A4 POSSIBLE WORLDS

Some philosophers of science define causal effects using the concept of ''possible worlds.'' The actual world is the way things actually are. A possible world is a way things might be. Imagine a possible world a where everybody receives exposure value a, and a possible world a' where everybody received exposure value a'. The mean of the outcome is $E[Y_a]$ in the first possible world and $E[Y_{a'}]$ in the second one. There is a causal effect if $E[Y_a] \neq E[Y_{a'}]$ and the worlds a and a' are the two worlds closest to the actual world where all subjects receive exposure value a and a', respectively.

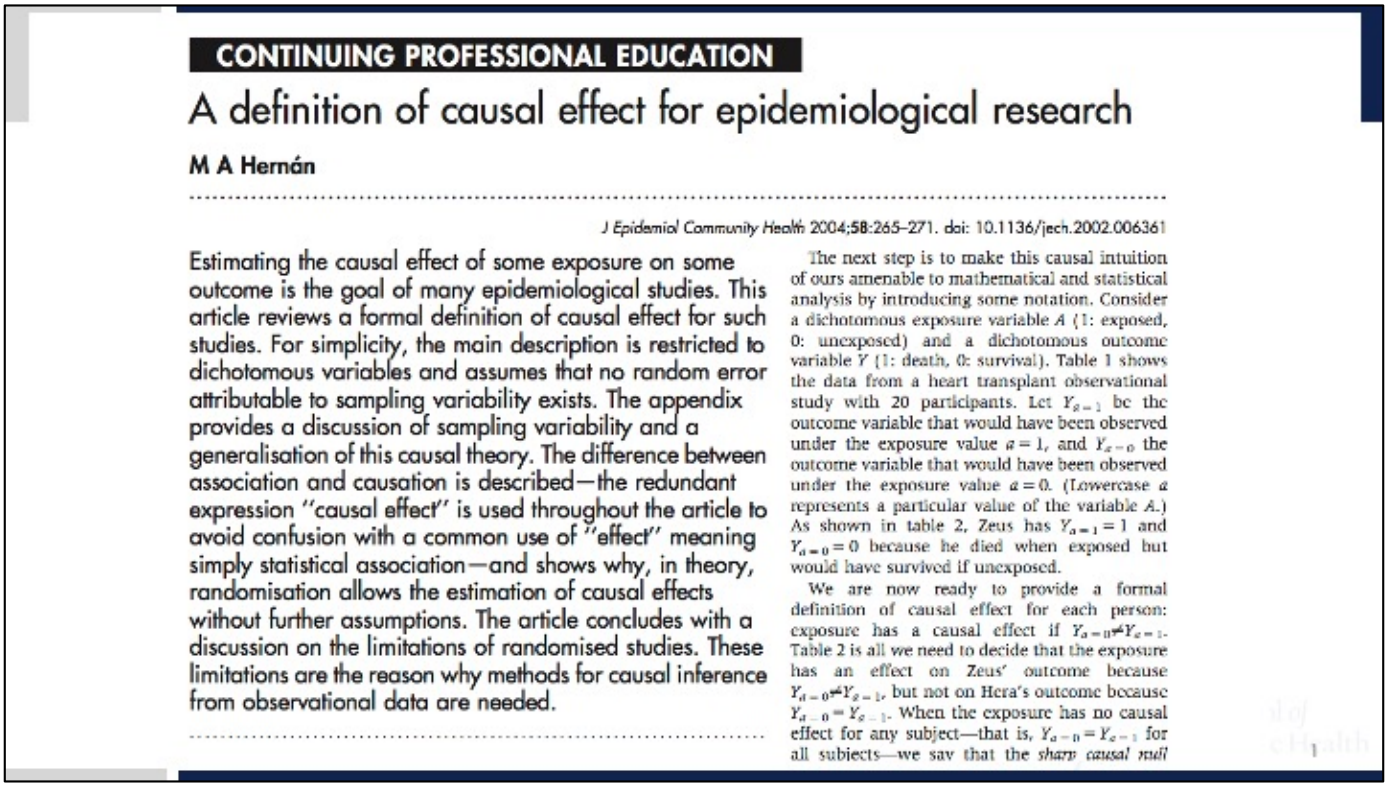
We introduced the counterfactual Y_a as the outcome of a certain subject under a well specified intervention that exposed her to a. Some philosophers prefer to think of the counterfactual Y_a as the outcome of the subject in the possible world that is closest to our world and where she was exposed to a. Both definitions are equivalent when the only difference between the closest possible world involved and the actual world is that the intervention of interest took place. The possible worlds' formulation of counterfactuals replaces the difficult problem of specifying the intervention of interest by the equally difficult problem of describing the closest possible world that is minimally different from the actual world. The two main counterfactual theories based on possible worlds, which differ only in details, have been proposed by Stalnaker[12] and Lewis.[13]

REFERENCES

1 **Hume D**. *An enquiry concerning human understanding*. [Reprinted and edited 1993]. Indianapolis/Cambridge: Hacket, 1748.
2 **Neyman J**. On the application of probability theory to agricultural experiments: essay on principles, section 9. *Translated in Statistical Science* 1923, 1990;**5**:465–80.
3 **Rubin DB**. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974;**56**:688–701.
4 **Rubin DB**. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 1978;**6**:34–58.
5 **Holland PW**. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 1986;**81**:945–61.
6 **Robins JM**. A new approach to causal Inference in mortality studies with sustained exposure periods_application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986;**7**:1393–512 (errata appeared in *Computers and Mathematics with Applications* 1987;**14**:917–21).
7 **Robins JM**. Addendum to ''A new approach to causal inference in mortality studies with sustained exposure periods_application to control of the healthy worker survivor effect''. *Computers and Mathematics with Applications* 1987;**14**:923–45. (errata appeared in *Computers and Mathematics with Applications* 1987;**18**:477).
8 **Dawid AP**. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society B* 1979;**41**:1–31.
9 **Robins JM**. Confidence intervals for causal parameters. *Stat Med* 1988;**7**:773–85.
10 **Cox DR**. *Planning of experiments*. New York: Wiley, 1958.
11 **Rubin DB**. Discussion of ''Randomized analysis of experimental data: the Fisher randomization test'' by Basu D. *Journal of the American Statistical Association* 1980;**75**:591–3.
12 **Stalnaker RC**. A theory of conditionals. In: Rescher N, ed. *Studies in logical theory*. Oxford: Blackwell, 1968. [Reprinted in Jackson F, ed. *Conditionals*. Oxford: Oxford University Press, 1991.]
13 **Lewis D**. *Counterfactuals*. Oxford: Blackwell, 1973.



In this video, we're going to review counterfactuals at the individual level and at the population level.



CONTINUING PROFESSIONAL EDUCATION

A definition of causal effect for epidemiological research

M A Hernán

J Epidemiol Community Health 2004;58:265–271. doi: 10.1136/jech.2002.006361

Estimating the causal effect of some exposure on some outcome is the goal of many epidemiological studies. This article reviews a formal definition of causal effect for such studies. For simplicity, the main description is restricted to dichotomous variables and assumes that no random error attributable to sampling variability exists. The appendix provides a discussion of sampling variability and a generalisation of this causal theory. The difference between association and causation is described—the redundant expression "causal effect" is used throughout the article to avoid confusion with a common use of "effect" meaning simply statistical association—and shows why, in theory, randomisation allows the estimation of causal effects without further assumptions. The article concludes with a discussion on the limitations of randomised studies. These limitations are the reason why methods for causal inference from observational data are needed.

The next step is to make this causal intuition of ours amenable to mathematical and statistical analysis by introducing some notation. Consider a dichotomous exposure variable A (1: exposed, 0: unexposed) and a dichotomous outcome variable Y (1: death, 0: survival). Table 1 shows the data from a heart transplant observational study with 20 participants. Let $Y_{a=1}$ be the outcome variable that would have been observed under the exposure value $a = 1$, and $Y_{a=0}$ the outcome variable that would have been observed under the exposure value $a = 0$. (Lowercase a represents a particular value of the variable A.) As shown in table 2, Zeus has $Y_{a=1} = 1$ and $Y_{a=0} = 0$ because he died when exposed but would have survived if unexposed.

We are now ready to provide a formal definition of causal effect for each person: exposure has a causal effect if $Y_{a=0} \neq Y_{a=1}$. Table 2 is all we need to decide that the exposure has an effect on Zeus' outcome because $Y_{a=0} \neq Y_{a=1}$, but not on Hera's outcome because $Y_{a=0} = Y_{a=1}$. When the exposure has no causal effect for any subject—that is, $Y_{a=0} = Y_{a=1}$ for all subjects—we say that the *sharp causal null*

And I'm going to be drawing heavily on the Hernan paper called "A definition of causal effect for epidemiologic research." And hopefully, you've read it before watching this video.

Individual causal effects

- Zeus is a patient waiting for a heart transplant.
- On 1 January, he received a new heart.
- Five days later, he died.

- Imagine that we can somehow know, perhaps by divine revelation, that had Zeus not received a heart transplant on 1 January (all others things in his life being unchanged) then he would have been alive five days later.

- We conclude that the intervention had a causal effect on Zeus' five day survival.

All right. So to start off with an example about Zeus, the Greek God to illustrate an individual causal effect. So let's say that Zeus is a patient waiting for a heart transplant. On January 1, he receives a new heart. Five days later, unfortunately, he dies. And that's information that we observed in real life.

Now let's imagine that we can somehow go back in time or know, perhaps by divine revelation, that if Zeus did not receive a heart transplant on January 1, and everything else in his life stayed the same, he would have been alive five days later. So we have two different scenarios-- one in which Zeus got the heart transplant, and died and the other which he didn't, and he survived.

Most of us would conclude from these two different scenarios that the intervention, which was the heart transplant had a causal effect on Zeus's five-day survival-- that period of time between January 1 and 6

Individual causal effects

- Table 1: Data from a heart transplant observational study with 20 participants.
- Exposure variable *A*
  - *A*=1 if exposed,
  - *A*=0 if unexposed
- Outcome variable *Y*
  - *Y*=1 if death
  - *Y*=0 if survival

Table 1 Data from a study with dichotomous exposure *A* and outcome *Y*

ID	A	Y
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Circe	0	0
Ares	1	1
Athene	1	1
Eros	1	1
Aphrodite	1	1
Prometheus	1	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

So we can formalize this with a bit of mutation here. And so over here on the right, is table one from the paper, and it's data from a study with a dichotomous exposure, A and outcome, Y.

So a dichotomous exposure is one that has two levels, in this case, 0 and 1. So we say A is equal to 1 if the person got a heart transplant, and it's equal to 0 if they did not. And we can also see right here that we can call A equal 1 if the person is exposed, or A equals 0 if the person is unexposed.

And then for the outcome variable Y, this is the variable to indicate if the person died or survives. So Y is 1 if the person died, and Y is 0 if they survived.

Individual causal effects

Table 2 Counterfactual outcomes of subjects in a study with dichotomous exposure *A* and outcome *Y*

ID	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Circe	0	1
Ares	1	1
Athene	1	1
Eros	0	1
Aphrodite	0	1
Prometheus	0	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

- Table 2: What would the outcome have been under each exposure?
- $Y_{a=1}$: potential outcome observed under the exposure value a = 1
- $Y_{a=0}$: potential outcome observed under the exposure value a = 0
- Exposure has a causal effect if $Y_{a=0} \neq Y_{a=1}$.
- Did the exposure have a causal effect on Zeus?
  - Yes because $Y_{a=0} \neq Y_{a=1}$
- What about Hera?

OK. So we can take this a step further now. So here's table two from the paper. And it lists counter-factual outcomes of subjects in a study with dichotomous exposure A in outcome Y.

And let's ask the question, what would the outcome have ban under each exposure? So we can define some people we're going to call a potential outcome. We use this subscript notation-- so here we see Y sub A equals 1 is the potential outcome observed under the exposure value A is 1.

So in our example, this is the potential outcome if a person had a heart transplant. The outcome could be died or survived. And then the second one here is the potential outcome if the person did not have a heart transplant. And these are recorded in the columns of table two. So let's look at our friend Zeus here. For Zeus, the potential outcome is he did not have the heart transplant was 0. Because, we remember from our previous slide that he survived. And the potential outcome if he does have a heart transplant is 1 because he died if he had the heart transplant.

And then, for a given exposure or intervention, we can calculate or assess if there's a causal effect by comparing the potential outcomes under A is 0 and A is 1. So in the
case of Zeus, did the heart transplant have a causal effect on him?

And the answer is yes because we concluded that before, but also using our notation, we can write it as his potential outcome under A is 0 was not the same as the potential outcome under A is 1.

Now take a minute for yourself, and take a look at Hera right above Zeus, and assess whether or not there was a causal effect of heart transplant on Hera. So Hera had a potential outcome of 0 under both cases. So whether or not her exposure was 1, meaning she got the heart transplant, and whether or not it was 0, she did not get the heart transplant, in both cases, she survived. And so there was no causal effect of heart transplant on Hera.

Now, this may all seem quite silly to you because, in real life, we wouldn't have both columns. We would just know, under one column, what the answer was. So either Zeus did or didn't get a heart transplant. There's no way we can roll the clock back and know if, on the exact same day of Zeus's life, what the outcome would have been if he hadn't gotten the heart transplant.

Individual causal effects

Table 2 Counterfactual outcomes of subjects in a study with dichotomous exposure *A* and outcome *Y*

ID	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Circe	0	1
Ares	1	1
Athene	1	1
Eros	0	1
Aphrodite	0	1
Prometheus	0	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

- Table 2: What would the outcome have been under each exposure?
- $Y_{a=1}$: potential outcome observed under the exposure value a = 1
- $Y_{a=0}$: potential outcome observed under the exposure value a = 0
- Exposure has a causal effect if $Y_{a=0} \neq Y_{a=1}$.
- Did the exposure have a causal effect on Zeus?
  - Yes because $Y_{a=0} \neq Y_{a=1}$
- What about Hera?

Key point: "Individual causal effects are defined as a contrast of the values of counterfactual outcomes, but only one of those values is observed.[...] All other counterfactual outcomes are missing. The unhappy conclusion is that, in general, **individual causal effects cannot be identified [i.e., estimated] because of missing data."**

5

Hernan has sort of summarize this nicely in the sentence here at the bottom. "Individual causal effects are defined as a contrast of the values of counterfactual outcomes, but only one of those values is observed."

OK. So first thing's first. Counterfactual outcomes are in other way of saying potential outcomes. And as we just discussed, we only observe one of these columns. All of the other counterfactual outcomes are missing. So unhappily, we conclude that, in general, individual causal effects cannot be identified, which is another way of saying estimated validly because we have missing data.

Population causal effects

- What is the population causal effect?
- $P(Y_{a=1})$: the proportion of subjects that would have developed the outcome Y had all subjects in the population of interest received exposure value a.
- The exposure has a causal effect in the population if $P(Y_{a=1}=1) \neq P(Y_{a=0}=1)$.
- Let's calculate this in Table 2:
- $P(Y_{a=1}=1)$ =
- $P(Y_{a=0}=1)$ =
- Causal risk difference =

Table 2 Counterfactual outcomes of subjects in a study with dichotomous exposure A and outcome Y

ID	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Circe	0	1
Ares	1	1
Athene	1	1
Eros	0	1
Aphrodite	0	1
Prometheus	0	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

OK. So we've been talking about individual causal effects. But what about population- level causal effects? We've got this full table here of potential outcomes for different gods and goddesses. How can we calculate the causal effect for all of them? Well, here's some new notation. So I just want to quickly point out that the paper has a slightly different way of recording this. Here, I'm saying P parentheses and then the potential outcome. And the paper uses PR square bracket. It means the exact same thing. So hopefully, that's not too confusing. This is slightly simpler notation-- same meaning.

And what it means is the proportion of subjects that would have developed the outcome Y had all subjects in the population of interest received the exposure, A. So it's the proportion of people let's, say, in column Y A equals 1 that passed away if everyone had received a heart transplant. And we can define something similar for this column.

And so then we can say the exposure had a population-level causal effect it's the probability or the proportion of subjects in

these two columns is different. So let's take a moment to calculate this for table two. So starting off with the right-hand column, what we need to do is basically take the percentage or the mean of this column.

So if we count this up, there are 10 Y equals 1 in this column and 20 total individuals in this column. So 10 over 20 is 0.5. And if we do that for the second coming here, we get the same number, 10 over 20. And then we can take 0.5 minus 0.5, which is 0, and that is our causal risk difference. So in other words, there is no difference at the population level in the proportion of subjects who died if everyone had a heart transplant or if everyone did not have a heart transplant. There's no causal effect.

Population causal effects

- We can't estimate **individual** causal effects.
- But what about **population** causal effects?
- Short answer: sometimes, and usually under multiple assumptions
- Examples of causal effects:
- $P(Y_{a=1}=1) - P(Y_{a=0}=1) \neq 0$
- $P(Y_{a=1}=1)/P(Y_{a=0}=1) \neq 1$
- $[P(Y_{a=1}=1)/P(Y_{a=1}=0)]/[P(Y_{a=0}=1)/P(Y_{a=0}=0)] \neq 1$

Table 2 Counterfactual outcomes of subjects in a study with dichotomous exposure *A* and outcome *Y*

ID	$Y_{a=0}$	$Y_{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Circe	0	1
Ares	1	1
Athene	1	1
Eros	0	1
Aphrodite	0	1
Prometheus	0	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

So we said we can't estimate individual causal effects. But what about population causal effects? The short answer is sometimes, and usually under multiple assumptions. So just to back up for a second, as we discussed, we only have one of these columns in real life. And so how can we make something out of nothing?

Well, there are lots of ways that we can design our study and collect data to help fill in the gaps. That's beyond the scope of this class. But it's just important to keep in mind that this can be possible, and it's important to articulate what assumptions are required when we want to estimate causal effects at the population level.

Here are some different kinds of causal effects. So this first one here is what we actually calculate it in the last slide. It's a causal risk difference. Take a moment for yourself, and see if you recognize the second two types of causal effects here. So hopefully, you've seen that this is very similar to the causal risk difference. This is a causal risk ratio or causal relative risk. It's the same except, instead of subtracting these two, we divide them. And here, at the bottom, we have a causal odds ratio.

Summary of key points

Individual causal effects can't be estimated with real data because we will never know the counterfactual outcome for a single person.

Population causal effects can sometimes be estimated with real data under certain assumptions.

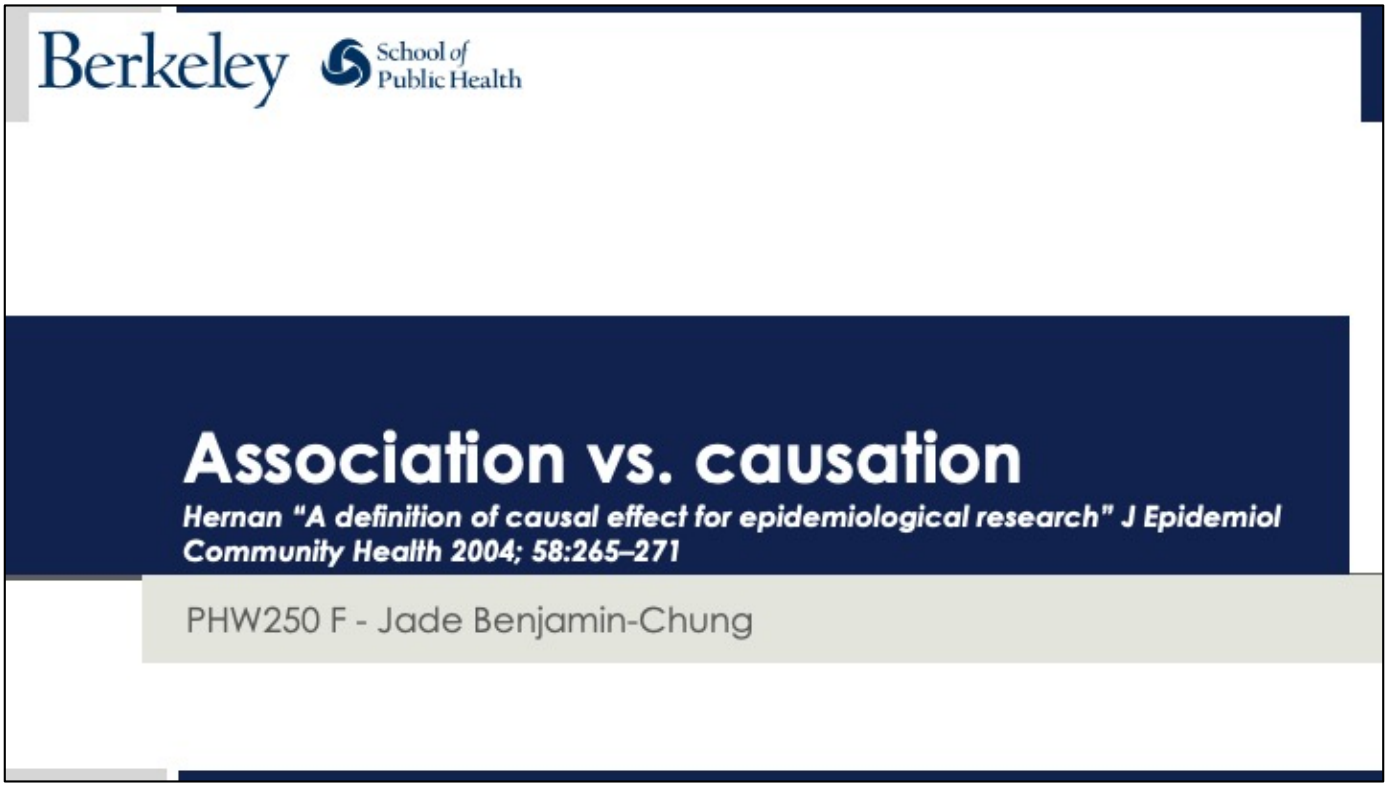
Usually when we say "causal effects" we mean "population causal effects".

Berkeley School of Public Health

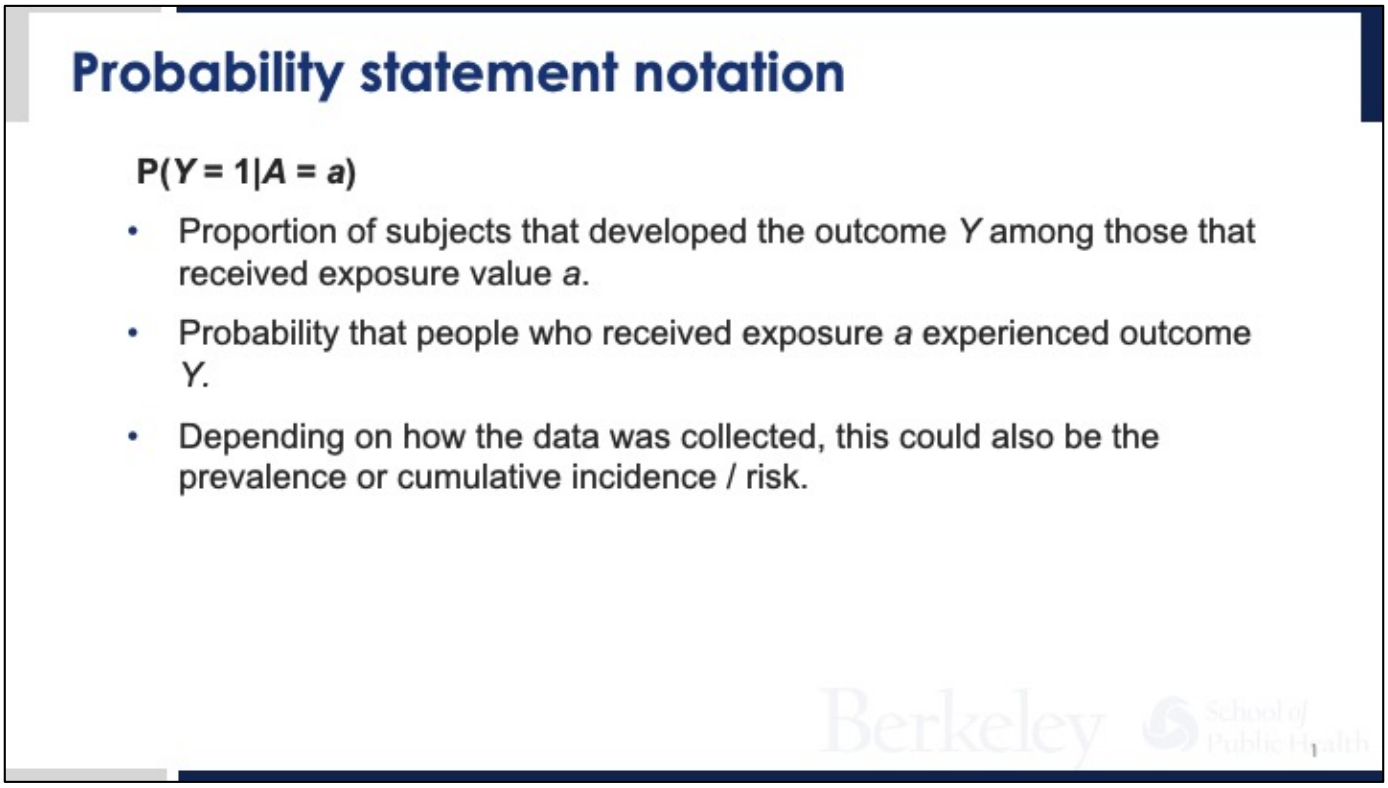
8

So to summarize, individual causal effects can't be estimated with real data, because we never know the counterfactual outcome for a single person. But at the population level, we can sometimes estimate causal effects under certain assumptions. Just one more thing to note is that, moving forward, when you see causal effects, it's going to mean population causal effects in almost every context. So please keep that in mind as we move on.

Lecture 2.1.2: Association vs. Causation



ANDREW MERTENS: Let's talk about the difference between association and causation. This video is going to draw upon the Hernan article "A definition of causal effects for epidemiologic research," just like in the last video.



Before we dive into this topic, it's worth it to pause and review probability statement notation, because it's critical to understanding the difference between association and causation. Here, we're saying that the probability that Y, our outcome, equals 1 is conditional or given that-- that's what the little vertical bar means-- that A, our exposure or our treatment, equals a.

Note that we're using italics here for Y and A, and that indicates that these are variables of interest, whereas p is not italicized, because it is a probability. And then it's worth mentioning that when we have a capital A, we're treating the exposure or the treatment as a random variable, which could take on different values. So a capital letter indicates something, generally, that could be assigned different values. So you could say that A indicates the intervention or exposure, but that intervention may be either the treatment or the control arm.

And then you have little a, which is the specific value. So little a will be assigned to a number like 0 or 1 in these slides. We can interpret this as the proportion of subjects that develops the outcome Y among those that received the exposure value a. Or we can say that it is the probability that people who receive exposure a experience the outcome Y. Depending on how the data was collected, this probability or proportion statement could also be interpreted as a prevalence or a cumulative incidence or risk measure.

So it would be prevalence if you conducted a cross-sectional survey. And if you conducted a cohort survey, where you were able to identify new cases over time, then Y equals 1 to indicate a new case. Then you could calculate the cumulative incidence measure.

Probability statement notation

$P(Y = 1|A = a)$

- Proportion of subjects that developed the outcome *Y* among those that received exposure value *a*.
- Probability that people who received exposure *a* experienced outcome *Y*.

	Disease	No disease	Total
Exposed	5	5	10
Unexposed	1	9	10
Total	6	14	20

So let's just take a look at a quick example in a simple 2 by 2 table to review how this notation works. We'll say that the exposed is indicated as A equals 1. And so if we want to get the risk or the prevalence for the exposed, we will write this as the probability that Y equals 1 conditional on A equaling 1.

And so A equals 1 is this first row here. And the number of people with the disease, Y equals 1, is 5. And the total number of people with A equals 1 is 10, so our probability is 5 over 10, or 0.5. And then we can do the same thing for the unexposed, where our A equals 0. The probability that a person is diseased, Y equals 1, among the unexposed-- there's one of these people. And there's a total of 10 people who are unexposed, and that probability is equal to 0.1. So that's just quickly to review how these probability statements work.

Using probability statement notation to report associations

- Exposure and outcome are associated if:

$$P(Y = 1|A = 1) \neq P(Y = 1|A = 0)$$

Table 1 Data from a study with dichotomous exposure *A* and outcome *Y*

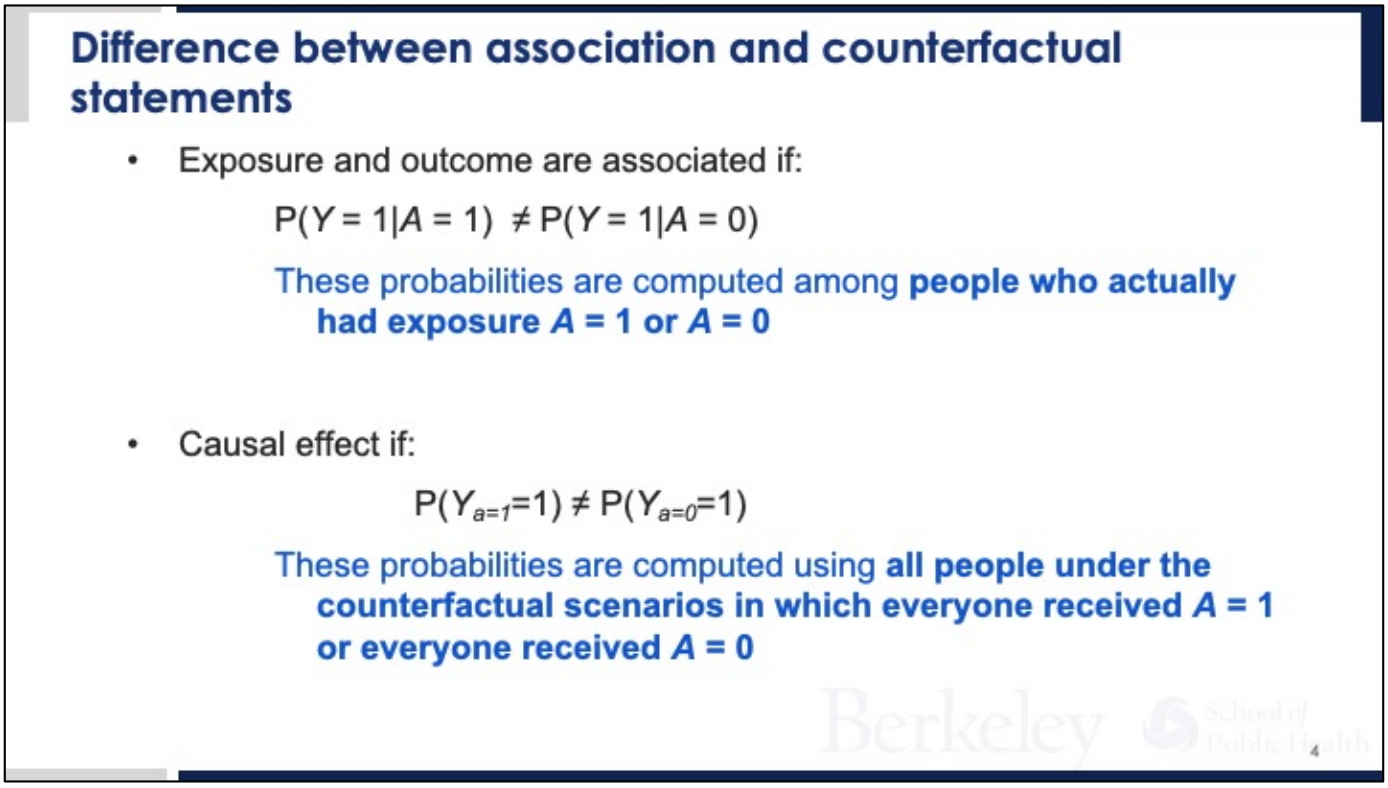
ID	A	Y
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Circe	0	0
Ares	1	1
Athene	1	1
Eros	1	1
Aphrodite	1	1
Prometheus	1	1
Selene	1	1
Hermes	1	0
Eos	1	0
Helios	1	0

3

So we can use probability statements to make reports about associations. In this case, this would be the association between A and Y. And so exposure and outcome, or A and Y, are associated if the probability of Y conditional on A equals 1 is different from the probability of Y conditional on A equals 0. So here's the table we saw in the last video. Why don't you pause the video and try to calculate each of these quantities here yourself? And then I'll go over it.

OK, so the probability that Y equals 1, given equals 1, is equal to 7 out of 13. So how did I get that? Well, if we look at the column for A and we count up the number of 1's, there's 13 of them. And then just looking at the rows where A equals 1, we see that seven of these rows where Y is also equal to 1. And then we use the same approach for the probability that Y is 1 conditional on A equaling 0.

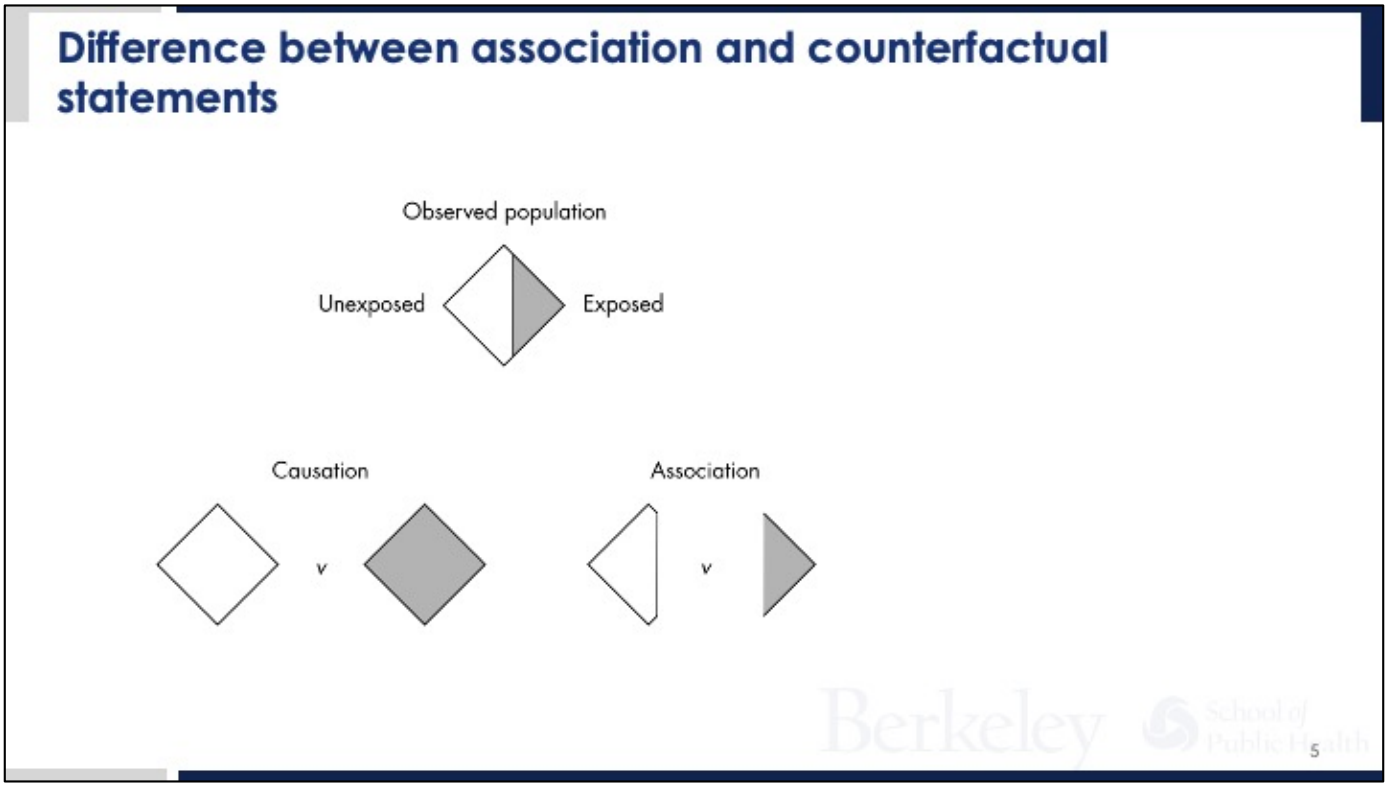
We look at the first column and we see that seven rows were a 0. And then, just within those rows, there are three people who have the disease, Y equals 1. And so that's 3 out of 7. So these two quantities are not the same-- 7 out of 13 and 3 out of 7. And that means that A and Y are associated. Another way of saying that is that knowing the value of A provides you with information about the possible value of Y.



So let's contrast that with the statements we made in the previous video. We're going to contrast association statements with counterfactual or causal statements. The exposure and the outcome are associated if this probability of Y equals 1 conditional on A equals 1 is not the same as the probability conditional on A equals 0. These are probabilities that are calculated among people who actually had exposure, A equals 1 or A equals 0. That's in our observed data from the real world.

The causal effect uses these counterfactual statements. So the probability that the potential outcome under exposure is 1 is not the same as the probability that the counterfactual under no exposure is 1. So these statements are computed among all people under each counterfactual scenario, who receives A equals 1 and A equals 0. So everyone getting A equals 1 or everyone is getting A equals 0. So that is the key difference between an associational statement and a counterfactual statement.

And as we have discussed, obviously, it's impossible to know what the potential outcome would have been for everyone under two different scenarios at the exact same point in time. So this is really theoretical. This isn't something that we can estimate with real data, but we can approximate it. And we'll come back to that.

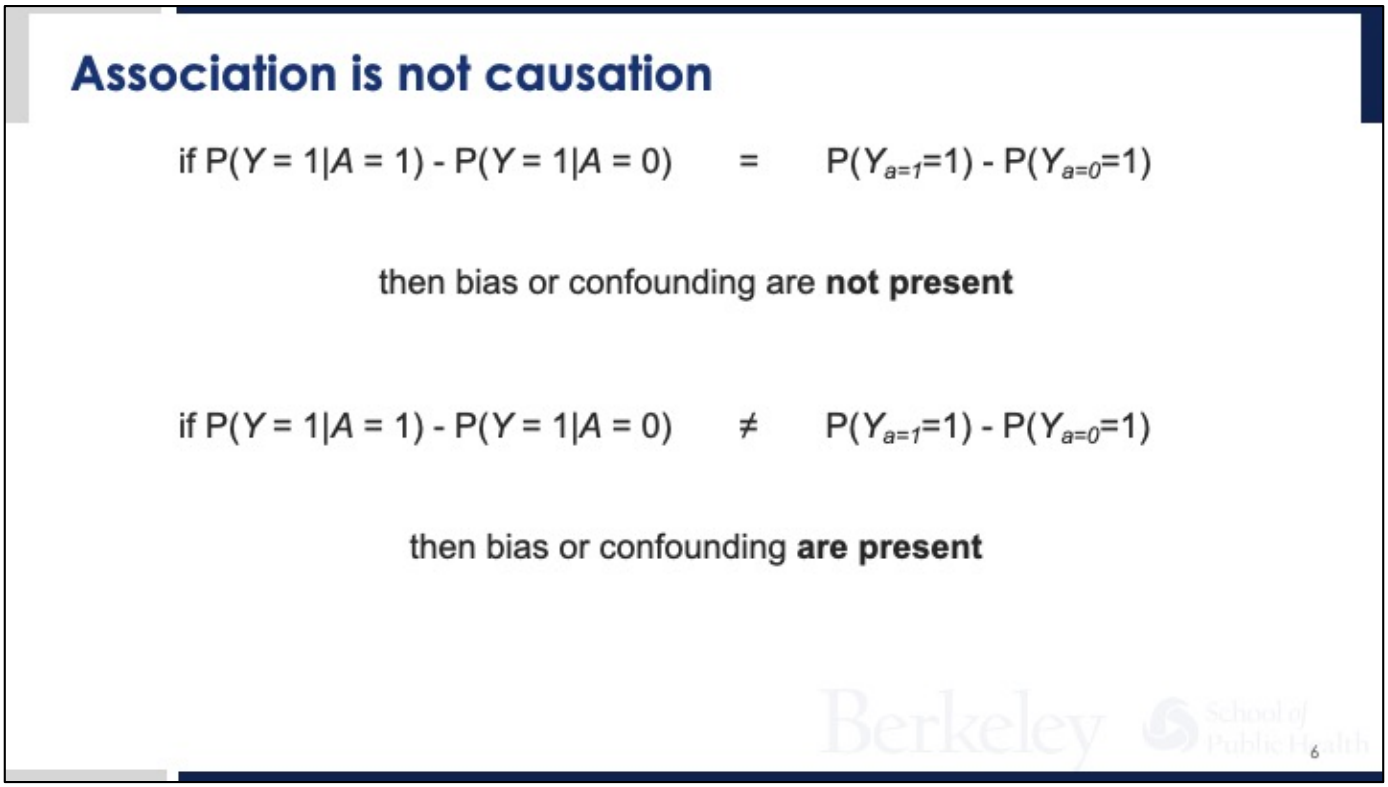


Here's another way of looking at this without notation. I'm going to fill in the notation, though, to help link the notation to this visual depiction here. Up at the top, we have our observed population, or our observed data. This diamond is split into white and gray, and we can say that the gray section is exposed, and the white is the unexposed segment of the population. So this is like real life, right? Some people, for example, smoke, and some don't smoke.

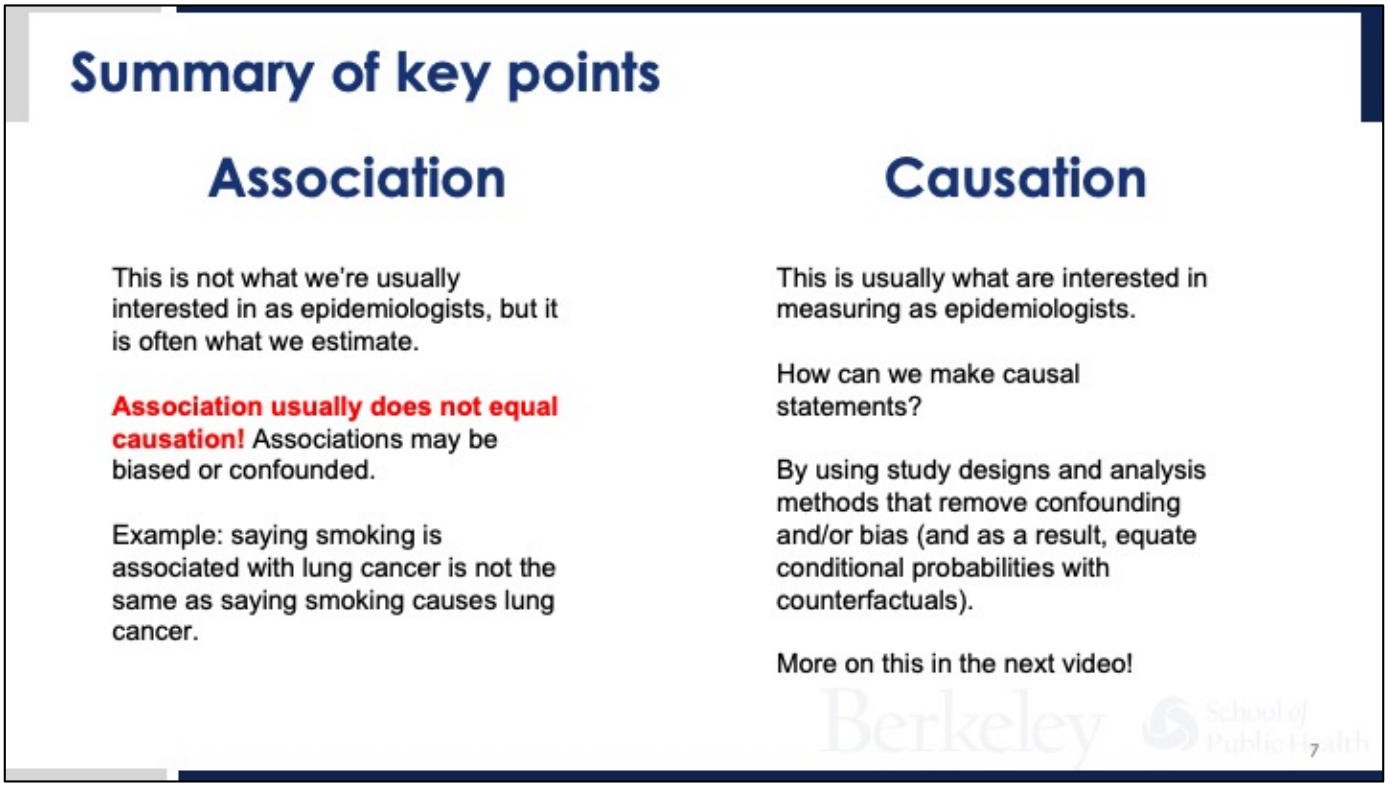
And as I was just mentioning, when we're making causal statements, we think about if the whole population smoked or if the whole population didn't smoke at the same point in time. A causal statement is comparing the mean outcome, or the prevalence, or the risk in the population under these two scenarios. The association statement is looking at the outcome under these proportions of the population-- so the portion of the population who smokes compared to the portion of the population who doesn't smoke.

And as we know, people who are in these two different portions of the population, whether they're smokers, non-smokers, or some other kind of unexposed or exposed groupings, are often systematically different from each other. There's other factors about them that split them into these two groups besides just the exposure-- in this example, smoking on its own. And so when we make a statement using this kind of population subset, the statement that we make could be a result of bias or confounding mixed together with the true effect.

So that's why association and causation are not the same thing. So what we're looking at is the probability that Y equals 1 conditional on A equals 1. So A equals 1 here is representing this gray portion right here of the diamond, and A equals 0 is representing this white portion of the diamond. And then for causal statements, we can write this as the probability of the potential outcome in this scenario in which everyone got A equals 1 and then, for the white diamond, the probability of the potential outcome under the scenario under which nobody got the exposure, A equals 0.



OK, so just to reiterate, because this is a really key point, association is not causation. Here, I show how that is depicted mathematically. When our measure of association is equal to our causal effect, then neither bias or confounding are present. When either is present, then these two qualities are not equal, and that, unfortunately, is almost always the case.



So to summarize, when it comes to associations, they're usually not what we're interested in as epidemiologists.

But unfortunately, because we don't have information on counterfactual scenarios, it's usually what we have to estimate. Association does not equal causation, and that's because usually there is some amount of bias or confounding in our data that prevent us from making causal statements with it. So saying smoking is associated with lung cancer is a very different thing from saying smoking causes lung cancer. Causation is what we're actually interested in. And how do we make causal statements?

Well, we can use study designs and analytic methods that we'll learn throughout this course to remove confounding or bias. And when that happens, we see those equations from the previous slide become equal to each other-- the conditional probabilities for association and the counterfactual statements. And we're going to delve into this more in the next video.

ANDREW MERTENS: This video focuses on exchangeability, which is a key assumption required to make causal inferences. And again, we're going to be following along with the Hernan article called "A Definition of Causal Effects for Epidemiological Research."

Example trial

- **Randomization**: Flip a coin
  - Heads = assign to group 1
  - Tails = assign to group 2
- **Treatment**
  - Give treatment (A = 1) to subjects in group 1
  - Give placebo (A = 0) to subjects in group 2

- **Follow-up:** Five days later, we compute the mortality risks in each group, $P(Y = 1|A = 1)$ and $P(Y = 1|A = 0)$.

- Assume it is a **perfect randomized trial** (no loss to follow up, full compliance with assigned treatment, blind assignment).

So to set the stage, let's use a really simple example of a trial where we flip a coin. And if it's heads, we assign that person to group one. And if it's tails, we assign that person to group two. We have a treatment that we give to subjects in group one, and a placebo that we give to subjects in group two. We follow these people for five days, and then we compute mortality risk within each group.

Here's the notation. So let's assume for the purpose of this example that it's a perfect randomized trial, so no loss to follow up. We have information on each person for all five days. There's also full compliance with everyone's assigned treatments, so no one is switching between group one and group two without permission. And the assignment is blinded.

Exchangeability

- Because of randomization, $P(Y = 1|A = 1)$ is the same when:
  - Group 1 gets treatment, Group 2 gets placebo
  - Group 2 gets treatment, Group 1 gets placebo
- The same is true for $P(Y = 1|A = 0)$
- Which group gets the treatment is irrelevant for the measure of disease among the exposed AND the measure of disease among the unexposed.
- This is called **exchangeability**. In other words, you could **exchange** one group for the other and the only difference is their treatment assignment.

Person #	Group (*A*)	Death (*Y*)
1	1	0
2	1	1
3	0	1
4	0	1

So because we use randomization, the probability that y equals 1, conditional on A equaling 1, is the same when group one gets the treatment and group two gets the placebo as when group two gets the treatment and group one gets the placebo. Let's delve into that idea bit more, and then we'll come back to this slide.

Exchangeability & observed risk

id #	Group	Treat-ment (*A*)	Death (*Y*)
1	1	1	0
2	1	1	1
3	2	0	1
4	2	0	1

- $P(Y = 1|A = 1) =$
- $P(Y = 1|A = 0) =$

id #	Group	Treat-ment (*A*)	Death (*Y*)
1	1	0	1
2	1	0	1
3	2	1	1
4	2	1	0

- $P(Y = 1|A = 1) =$
- $P(Y = 1|A = 0) =$

3

So here are two different tables showing the results from two different repetitions of the same trial. Let me walk you through it. Let's first look at this table here on the left. Each row is an ID number for a person in the study. Just to keep things super simple, we're going to pretend that there are only four people in this trial. We'd never do that in a real trial because it's a very small sample size, but it allows for easy calculation.

And so the first two people got a heads when they flipped the coin, and the second two people got a tails when they flipped the coin, and that's denoted with yellow or green cells. And so in this first table, if you flip the head, you received the treatment that's indicated in the third column with an A equals 1. And therefore those who flipped the tail received A equals 0, or the placebo arm of the trial. And then looking in the fourth column, you see an indicator for the outcome Death where Y equals 1, and the subject died.

So then we can calculate the probability that Y equals 1 conditional on A equals 1, or the probability of death within the treatment arm. And that's done by looking at the first two rows here, where A equals 1. And we can see that in the top two rows. One person died out of two, so one out of two is a mortality rate of 50%. And then that same probability, but for A equals 0 in the bottom two rows, where two people died out of a total of two in the placebo arm, so 100% probability of death.

Now, this table on the right-hand side shows the exact same experiment, except if we had given the treatment to those who flipped tails and the placebo to those who

flipped heads. And once again, we can calculate the probability that Y equals 1 given A equals 1 by looking at the bottom two rows now. And we can see once again that one out of two individuals died, for a probability of 50%. And then we can look at the top two rows and see that two died out of two, or a probability of 100%.

So even though we flipped who received the treatment and who received the placebo based on their coin flip results, we see a same probability of death within the treatment arm. And this is called exchangeability. This simple toy example shows that regardless of which group receives the treatment or the placebo, the results stay the same.

Exchangeability

- Because of randomization, $P(Y = 1|A = 1)$ is the same when:
  - Group 1 gets treatment, Group 2 gets placebo
  - Group 2 gets treatment, Group 1 gets placebo
- The same is true for $P(Y = 1|A = 0)$
- Which group gets the treatment is irrelevant for the measure of disease among the exposed AND the measure of disease among the unexposed.
- This is called **exchangeability**. In other words, you could **exchange** one group for the other and the only difference is their treatment assignment.

Person #	Group (*A*)	Death (*Y*)
1	1	0
2	1	1
3	0	1
4	0	1

So that's just another way of saying that whichever group receives the treatment versus the placebo or control is irrelevant for the measure of disease of interest among the exposed and among the unexposed. In other words, we could exchange one group to the other, and the only difference between them would be their treatment assignment.

Summary of key points

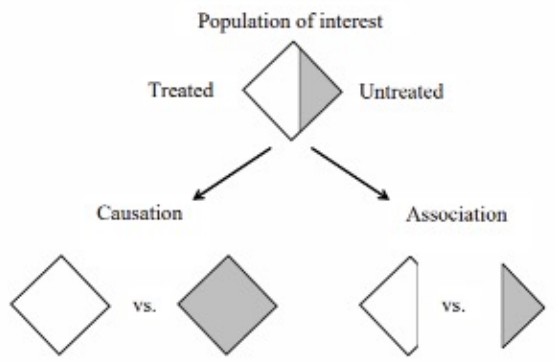


Image source: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2017/03/hernanrobins_v1.10.32.pdf

- Formally, **exchangeability** occurs when the counterfactual outcome Y_a is independent of the treatment or exposure A for all values a.
- In a perfect randomized trial,
  - the observed risk is equal to the counterfactual risk,

    $$P(Y = 1|A = a) = P(Y_a = 1)$$
  - we can estimate a causal effect from our data, and
  - the associational measure of association equals the causal measure of association.

We've gone through an example with fake hypothetical data, but when we think about this more formally, it's really about making counterfactual statements. So we can say exchangeability occurs when the counterfactual outcome Y sub a is independent of the treatment or exposure A for all values of little a. So that's A equals 1 or A equals 0. Let's flesh out what that means. In a perfect randomized trial, our observed risk will be equal to the counterfactual risk. So here on the left is our observed risk, and here on the right is our counterfactual risk.

In the last slide, we looked at probability statements that summarize the observed data. The Hernan article goes through some additional steps to basically show that exchangeability is actually formally defined in terms of counterfactual statements as opposed to associational statements from observed data. So I recommend you take a look at the article if you're interested in those details. The key thing to know is really just that the formal definition is that exchangeability occurs when the counterfactual outcome Y sub a is independent of the treatment or exposure A for all values of a.

So what's this definition getting at? Well, the concept of exchangeability is getting at the fact that when we have perfect randomization, on average, the people in each group are basically the same as each other, or as good as exchangeable. And because the only difference between them is their treatment assignment, it's almost as if they're serving as this full diamond here, where everyone in the population received the treatment compared to when no one in the population received the

treatment, or everyone gets the control. The only difference between these diamonds is the treatment assignment itself.

So in a perfect RCT, the observed risk or randomized trial, the observed risk is equal to the counterfactual risk. So the probability statement for this portion here-- the white portion of this association diamond-- is P of Y equals 1, given that A equals 1. And this one here is P of Y equals 1 given that A equals 0. And then we can write out the counterfactual statements over here-- P of Y sub 0 equals 1, and P of Y sub 1 equals 1.

So we can estimate the causal effects from our data in a perfect randomized controlled trial because these quantities are true. These pieces of the overall diamond are approximating these two diamonds from the observed data. So that means that in this case, the associational measures of association equal the causal measures of association. It's really hard to do a perfect randomized controlled trial, but a lot of times, we can get very close to it, which means that we are very close to being able to make causal inferences about our data.

Modern Epidemiology. 3rd Chapter 12

Causal Diagrams

M. Maria Glymour and Sander Greenland

INTRODUCTION

Diagrams of causal pathways have long been used to visually summarize hypothetical relations among variables of interest. Modern causal diagrams, or causal graphs, were more recently developed from a merger of graphical probability theory with path diagrams. The resulting theory provides a powerful yet intuitive device for deducing the statistical associations implied by causal relations. Conversely, given a set of observed statistical relations, a researcher armed with causal graph theory can systematically characterize all causal structures compatible with the observations. The theory also provides a visual representation of key concepts in the more general theory of longitudinal causality of Robins (1997); see Chapter 21 for further discussion and references on the latter topic.

The graphical rules linking causal relations to statistical associations are grounded in mathematics. Hence, one way to think of causal diagrams is that they allow nonmathematicians to draw logically sound conclusions about certain types of statistical relations. Learning the rules for reading statistical associations from causal diagrams may take a little time and practice. Once these rules are mastered, though, they facilitate many tasks, such as understanding confounding and selection

bias, choosing covariates for adjustment and for regression analyses, understanding analyses of direct effects and instrumental-variable analyses, and assessing "natural experiments." In particular, diagrams help researchers recognize and avoid common mistakes in causal analysis.

This chapter begins with the basic definitions and assumptions used in causal graph theory. It then describes construction of causal diagrams and the graphical separation rules linking the causal assumptions encoded in a diagram to the statistical relations implied by the diagram. The chapter concludes by presenting some examples of applications. Some readers may prefer to begin with the examples and refer back to the definitions and rules for causal diagrams as needed. The section on Graphical Models, however, is essential to understanding the examples. Full technical details of causal diagrams and their relation to causal inference can be found in Pearl (2000) and Spirtes et al. (2001), while Greenland and Pearl (2008) provide a short technical review. Less technical articles geared toward health scientists include Greenland et al. (1999a), Robins (2001), Greenland and Brumback (2002), Hernán et al. (2002), Jewell (2004), and Glymour (2006b).

PRELIMINARIES FOR CAUSAL GRAPHS

Consider two variables X and Y for which we wish to represent a causal connection from X to Y, often phrased as "X causes Y" or "X affects Y." Causal diagrams may be constructed with almost any definition of cause and effect in mind. Nonetheless, as emphasized in Chapter 4, it is crucial to distinguish causation from mere association. For this purpose we use the potential-outcome (counterfactual) concept of causation. We say that X affects Y in a population of units (which may be people, families, neighborhoods, etc.) if and only if there is at least one unit for which changing (intervening on) X will change Y (Chapter 4).

STATISTICAL INDEPENDENCE

Association of X and Y corresponds to statistical dependence of Y and X, whereby the distribution of Y differs across population strata defined by levels of X. When the distribution of Y does not differ across strata of X, we say that X and Y are statistically independent, or unassociated. If X and Y are unassociated (independent), knowing the value of X gives us no information about the value of Y. Association refers to differences in Y *between* units with different X values. Such between-unit differences do not necessarily imply that changing the value of X for any single unit will result in a change in Y (which is causation).

It is helpful to rephrase the above ideas more formally. Let $\Pr(Y = y)$ be the expected proportion of people in the population who have y for the value of Y; this expected proportion is more often called the probability that $Y = y$. If we examine the proportion who have $Y = y$ within levels or strata of a second variable X, we say that we are examining the probability of Y *given* or *conditional on* X. We use a vertical line "|" to denote "given" or "conditional on." For example, $\Pr(Y = y|X = x)$ denotes the proportion with $Y = y$ in the subpopulation with $X = x$. Independence of X and Y then corresponds to saying that for any pair of values x and y for X and Y,

$$\Pr(Y = y|X = x) = \Pr(Y = y) \qquad [12–1]$$

which means that the distribution of Y values does not differ across different subpopulations defined by the X values. In other words, the equation says that the distribution of Y given (or conditional on) a particular value of X always equals the total population (marginal or unconditional) distribution of Y. As stated earlier, if X and Y are independent, knowing the value of X and nothing more about a unit provides no information about the Y value of the unit.

Equation 12–1 involves no variable other than X and Y, and is the definition of *marginal* independence of X and Y. When we examine the relations between two variables within levels of a third variable—for example, the relation between income and mortality within levels of education—we say that we are examining the conditional relation. We examine conditional relationships in many contexts in epidemiology. We may intentionally condition on a variable(s) through features of study design such as restriction or matching, or analytic decisions, such as stratification or regression modeling. Conditioning may arise inadvertently as well, for example due to refusal to participate or

loss to follow-up. These events essentially force conditioning on variables that determine participation and ascertainment. Informally, it is sometimes said that conditioning on a variable is "holding the variable constant," but this phrase is misleading because it suggests we are actively intervening on the value of the variable, when all we are doing is separating the data into groups based on observed values of the variable and estimating the effects within these groups (and then, in some cases, averaging these estimates over the groups, see Chapter 15).

To say that X and Y are independent given Z means that for any values x, y, z for X, Y, and Z,

$$\Pr(Y = y|X = x, Z = z) = \Pr(Y = y|Z = z) \qquad [12\text{–}2]$$

which says that, within any stratum of Z, the distribution of Y does not vary with X. In other words, within any stratum defined in terms of Z alone, we should see no association between X and Y. If X and Y are independent given Z, then once one knows the Z value of a unit, finding out the value of X provides no further information about the value of Y.

CAUSATION AND ASSOCIATION

As explained in Chapter 4, causation and association are qualitatively different concepts. Causal relations are directed; associations are undirected (symmetric). Sample associations are directly observable, but causation is not. Nonetheless, our intuition tells us that associations are the result of causal forces. Most obviously, if X causes Y, this will generally result in an association between X and Y. The catch, of course, is that even if we observe X and Y without error, many other forces (such as confounding and selection) may also affect the distribution of Y and thus induce an association between X and Y that is not due to X causing Y. Furthermore, unlike causation, association is symmetric in time (nondirectional), e.g., an association of X and Y could reflect Y causing X rather than X causing Y.

A study of causation must describe plausible explanations for observed associations in terms of causal structures, assess the logical and statistical compatibility of these structures with the observations, and (in some cases) develop probabilities for those structures. Causal graphs provide schematic diagrams of causal structures, and the independencies predicted by a graph provide a means to assess the compatibility of each causal structure with the observations.

More specifically, when we see an association of X and Y, we will seek sound explanations for this observation. For example, logically, if X always precedes Y, we know that Y cannot be causing X. Given that X precedes Y, obvious explanations for the association are that X causes Y, that X and Y share a common cause (confounding), or some combination of the two (which can also lead to no association even though X affects Y). Collider bias is a third type of explanation that seems much less intuitive but is easily illustrated with graphs. We will first discuss focus on collider bias because it arises frequently in epidemiology.

COLLIDER BIAS

As described in Chapter 9, a potentially large source of bias in assessing the effect of X on Y arises when selection into the population under study or into the study sample itself is affected by both X and Y. Such selection is a source of bias even if X and Y are independent before selection. This phenomenon was first described by Joseph Berkson in 1938 (published in Berkson [1946]). *Berksonian bias* is an example of the more general phenomenon called *collider bias,* in which the association of two variables X and Y changes upon conditioning on a third variable Z if Z is affected by both X and Y. The effects of X and Y are said to "collide" somewhere along the way to producing Z.

As an example, suppose that X and Y are marginally independent and $Z = Y - X$, so Z is completely determined by X and Y. Then X and Y will exhibit perfect dependence given Z: If $Z = z$, then $Y = X + z$. As a more concrete example, body mass index (BMI) is defined as (weight in kg)/(height in meters)^2 and so is strongly affected by both height and weight. Height and weight are associated in any natural population, but not perfectly: We could not exactly tell a person's weight from his or her height. Suppose, however, we learn that the person has BMI = 25 kg/m^2;

then, upon being told (say) that the person is 2 m tall, we can compute his weight exactly, as $\text{BMI}(\text{height}^2) = 25(4) = 100$ kg.

Collider bias occurs even when the causal dependency of the collider Z on X and Y is not perfect, and when there are several intermediates between X and the collider or between Y and the collider. It can also be induced when X and Z (or Y and Z) are associated due to a common cause rather than because X influences Z.

Collider bias can result from sample selection, stratification, or covariate adjustment if X and Y affect selection or the stratifying covariates. It can be just as severe as confounding, as shown in the classic example in which X, Y, and Z were exogenous estrogen use, endometrial cancer, and uterine bleeding (Chapter 9). As discussed later, it can also can induce confounding.

SUMMARY

Four distinct causal structures can contribute to an association between X and Y: (a) X may cause Y; (b) Y may cause X; (c) X and Y may share a common cause that we have failed to condition on (confounding); or (d) we have conditioned or selected on a variable affected by X and Y, factors influenced by such a variable, or a variable that shares causes with X and Y (collider bias). Of course, the observed association may also have been affected by purely random events. As described in Part III of this book, conventional statistics focus on accounting for the resulting random variation. The remainder of this chapter focuses on the representation of causal structures via graphical models, and on the insights that these representations provide. Throughout, we focus on the causal structures underlying our observations, ignoring random influences.

GRAPHICAL MODELS

TERMINOLOGY

Causal diagrams visually encode an investigator's assumptions about causal relations among the exposure, outcomes, and covariates. We say that a variable X affects a variable Y *directly* (relative to the other variables in the diagram) if there is an arrow from X to Y. We say that X affects Y *indirectly* if there is a head-to-tail sequence of arrows (or "one-way street") from X to Y; such a sequence is called a *directed path* or *causal path*. Any variable along a causal path from X to Y is called an *intermediate variable* between X and Y. X may affect Y both directly and indirectly. In Figure 12–1, X affects Y directly and Z indirectly. The absence of a directed path between two variables represents the assumption that neither affects the other; in Figure 12–1, U and X do not affect each other.

Children of a variable X are variables that are affected directly by X (have an arrow pointing to them from X); conversely, *parents* of X are variables that directly affect X (have an arrow pointing from them to X). More generally, the *descendants* of a variable X are variables affected, either directly or indirectly, by X; conversely, the *ancestors* of X are all the variables that affect X directly or indirectly. In Figure 12–1, Y has parents U and X, and a child Z; X has one child (Y) and two descendants (Y and Z); and Z has a parent Y and three ancestors, Y, U, and X.

It is not necessary to include all causes of variables in the diagram. If two or more variables in a graph share a cause, however, then this cause must also be shown in the graph as an ancestor of those variables, or else the graph is not considered a causal graph. A variable with no parents in a causal graph is said to be *exogenous* in the graph; otherwise it is *endogenous*. Thus, all

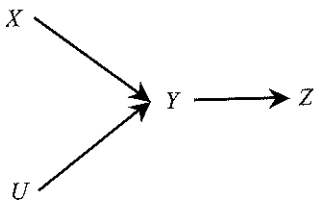


FIGURE 12–1 ● A causal diagram with no confounding.

exogenous variables in the graph are assumed to share no cause with other variables in the graph. If unknown common causes of two variables may exist, a casual graph must show them; they may be represented as unspecified variables with arrows to the variables they are thought to influence. In a slight modification of these rules, some authors (e.g., Pearl, 2000) use a two-headed arrow between two variables as a shorthand to indicate that there is at least one unknown exogenous common cause of the two variables (e.g., $X \leftrightarrow Z$ means that there is at least one unknown exogenous variable U such that $X \leftarrow U \rightarrow Z$). We assume in the remainder of this chapter that unknown common causes are represented explicitly in causal diagrams, so there is no need for two-headed arrows.

All the graphs we will consider are *acyclic*, which means that they contain no feedback loops; this means that no variable is an ancestor or descendant of itself, so if X causes Y, Y cannot also cause X at the same moment. If a prior value of Y affects X, and then X affects a subsequent value of Y, these must each be shown as separate variables (e.g., $Y_0 \rightarrow X_1 \rightarrow Y_2$) (for discussions of extensions to causal structures including feedback, see Spirtes [1995], Pearl and Dechter [1996], and Lauritzen and Richardson [2002]). In most causal graphs the only connectors between variables are one-headed arrows ($\rightarrow$), although some graphs use an undirected dashed line (- - -) to indicate associations induced by collider bias. Connectors, whether arrows or dashed lines, are also known as *edges,* and variables are often called *nodes* or *vertices* of the graph. Two variables joined by a connector are said to be *adjacent* or *neighbors*. If the only connectors in the graph are one-headed arrows, the graph is called *directed.* A directed acyclic graph or DAG is thus a graph with only arrows between variables and with no feedback loops. The remainder of our discussion applies to DAGs and graphs that result from conditioning on variables in DAGs.

A *path* between X and Y is any noncrossing and nonrepeating sequence traced out along connectors (also called edges) starting with X and ending with Y, *regardless of the direction of arrowheads.* A variable along the path from X to Y is said to *intercept* the path. Directed paths are the special case in which all the connectors in the path flow head to tail. Any other path is an *undirected path.* In Figure 12–1, $U \rightarrow Y \leftarrow X$ is an undirected path from U to X, and Y intercepts the path.

When tracing out a path, a variable on the path where two arrowheads meet is called a *collider* on that path. In Figure 12–1, Y is a collider on the path $U \rightarrow Y \leftarrow X$ from U to X. Thus, a collider on a path is a direct effect (child) of both the variable just before it and the variable just after it on the path. A directed path cannot contain a collider. If a variable on a path has neighbors on both sides but is not a collider, then the variable must be either an intermediate ($X \rightarrow Y \rightarrow Z$ or $X \leftarrow Y \leftarrow Z$) or a cause ($X \leftarrow Y \rightarrow Z$) of its immediate neighbors on the path.

Being a collider is specific to a path. In the same DAG, a variable may be a collider on one path but an intermediate on another path; e.g., in Figure 12–1, Y is an intermediate rather than a collider on the path $X \rightarrow Y \rightarrow Z$. Nonetheless, a variable with two or more parents (direct causes) is called a collider in the graph, to indicate that it is a collider on at least one path. As we will see, paths with colliders can turn out to be sources of confounding and selection bias.

RULES LINKING ABSENCE OF OPEN PATHS TO STATISTICAL INDEPENDENCIES

Given a causal diagram, we can apply the *d-separation criteria* (or directed-graph separation rules) to deduce independencies implied by the diagram. We first focus on rules for determining whether two variables are d-separated unconditionally, and then examine how conditioning on variables may d-separate or d-connect other variables in the graph. We emphasize that the deduced relations apply only "in expectation," meaning that they apply to the *expected* data distribution if the causal structure represented by the graph is correct. They do not describe the associations that may arise as a result of purely random events, such as those produced by randomization or random sampling.

Unconditional d-Separation

A path is said to be *open* or *unblocked* or *active* unconditionally if there is no collider on the path. Otherwise, if there is a collider on the path, it is said to be *closed* or *blocked* or *inactive,* and we say that the collider blocks the path. By definition a directed path has no collider, so every directed path is open, although not every open path is directed. Two variables X and Y are said to be *d-separated* if there is no open path between them; otherwise they are *d-connected.* In Figure 12–2, the only path from X to Y is open at Z_1 and Z_2 but closed at W, and hence it is closed overall; thus X and Y

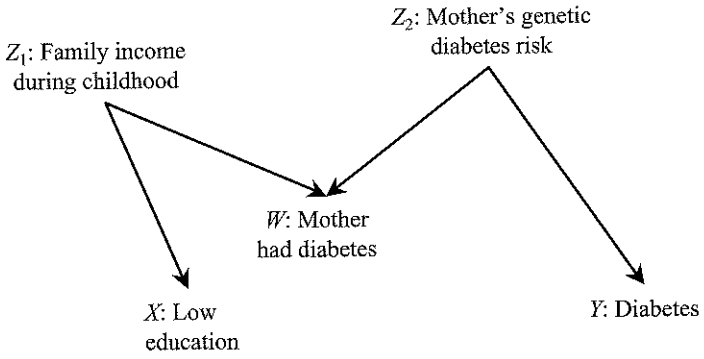


FIGURE 12–2 • A DAG under which traditional confounder-identification rules fail (an "*M* diagram").

are d-separated. When using these terms we will usually drop the "d-" prefix and just say that they are separated or connected as appropriate.

If X and Y are separated in a causal graph, then the causal assumptions encoded by the graph imply that X and Y will be unassociated. Thus, if every path from X to Y is closed, the graph predicts that X and Y will be marginally independent; i.e., for any values x and y of X and Y, $\Pr(Y = y|X = x) = \Pr(Y = y)$. More generally and informally we can say this: In a causal graph, the only sources of marginal association between variables are the open paths between them. Consider Table 12–1, which lists the causal assumptions represented by the diagram of Figure 12–1, and the associations implied by those causal assumptions. For example, the causal diagram implies that U and X are marginally independent because the only path between them passes through a collider, Y. This idea is formalized later when we define compatibility.

Conditional d-Separation

We also need the concept of graphical conditioning. Consider first conditioning on a noncollider Z on a path. Because it is a noncollider, Z must either be an intermediate between its neighbors on the path ($X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$) or a cause of its neighbors ($X \leftarrow Z \rightarrow Y$). In these cases the path is open at Z, but conditioning on Z closes the path and removes Z as a source of association between X and Y. These phenomena reflect the first criterion for blocking paths by conditioning on covariates:

- Conditioning on a noncollider Z on a path blocks the path at Z.

In contrast, conditioning on a collider requires reverse reasoning. If two variables X and Y are marginally independent, we expect them to become associated upon conditioning (stratifying) on a shared effect W. In particular, suppose we are tracing a path from X to Y and reach a segment on the path with a collider, $X \rightarrow W \leftarrow Y$. The path is blocked at W, so no association between X and Y passes through W. Nonetheless, conditioning on W or any descendant of W opens the path at W. In other words, we expect conditioning on W or any descendant to create an X–Y association via W. We thus come to the second criterion for blocking paths by conditioning on covariates:

- Conditioning on a collider W on a path, or any descendant of W, or any combination of W or its descendants, opens the path at W.

Combining these criteria, we see that conditioning on a variable reverses its status on a path: Conditioning closes noncolliders (which are open unconditionally) but opens colliders (which are closed unconditionally).

We say that a set of variables $\boldsymbol{S}$ *blocks a path* from X to Y if, after conditioning on $\boldsymbol{S}$, the path is closed (regardless of whether it was closed or open to begin with). Conversely, we say that a set of variables $\boldsymbol{S}$ unblocks a path if, after conditioning on $\boldsymbol{S}$, the path is open (regardless of whether it was closed or open to begin with). The criteria for a set of variables to block or unblock a path are summarized in Table 12–2.

TABLE 12–1

Assumptions Represented in the Directed Acyclic Graph in Figure 12–1, and Statistical Implications of These Assumptions

Causal Assumptions Represented in Figure 12–1	Independencies Implied by Figure 12–1	Marginal Associations Expected under Figure 12–1 (Assuming Faithfulness)	Conditional Associations Expected under Figure 12–1 (Assuming Faithfulness)
• X and U are each direct causes of Y (direct with respect to other variables in the diagram). • Y is a direct cause of Z. • X is not a direct cause of Z, but X is an indirect cause of Z via Y. • X is not a cause of U and U is not a cause of X. • U is not a direct cause of Z, but U is an indirect cause of Z via Y. • No two variables in the diagram (X, U, Y, or Z) share a prior cause not shown in the diagram, e.g., no variable causes both X and Y, or both X and U.	• X and U are independent (the only path between them is blocked by the collider Y). • X and Z are independent conditional on Y (conditioning on Y blocks the path between X and Z). • U and Z are independent conditional on Y.	• X and Y are associated. • U and Y are associated. • Y and Z are associated. • X and Z are associated. • U and Z are associated.	• X and U are associated conditional on Y (conditioning on a collider unblocks the path). • X and U are associated conditional on Z (Z is a descendant of the collider Y).

If $\boldsymbol{S}$ blocks every path from X to Y, we say that X and Y are *d-separated by* $\boldsymbol{S}$, or that $\boldsymbol{S}$ *separates* X and Y. This definition of d-separation includes situations in which there was no open path before conditioning on $\boldsymbol{S}$. For example, a set $\boldsymbol{S}$ may be sufficient to separate X and Y even if $\boldsymbol{S}$ includes no variables: if there is no open path between X and Y to begin with, the empty set separates them.

d-Separation and Statistical Independence

We have now specified the d-separation criteria and explained how to apply them to determine whether two variables in a graph are d-separated or d-connected, either marginally or conditionally. These concepts provide a link between the causal structure depicted in a DAG and the statistical associations we expect in data generated from that causal structure. The following two rules specify the relation between d-separation and statistical independence; these rules underlie the applications we will present.

Rule 1 (compatibility). Suppose that two variables X and Y in a causal graph are separated by a set of variables $\boldsymbol{S}$. Then if the graph is correct, X and Y will be unassociated given $\boldsymbol{S}$. In other

TABLE 12–2

Criteria for Determining Whether a Path is Blocked or Unblocked Conditional on a Set of Variables *S*

The Path from *X* to *Y* is Blocked Conditional on *S* if Either:	The Path from *X* to *Y* is Unblocked Conditional on *S* if Both:
A noncollider *Z* on the path is in ***S*** (because the path will be blocked by ***S*** at *Z*)	***S*** contains no noncollider on the path (so conditioning on ***S*** blocks no noncollider)
OR	AND
There is a collider *W* on the path that is not in ***S*** and has no descendant in ***S*** (because *W* still blocks the path after conditioning on ***S***).	Every collider on the path is either in ***S*** or has a descendant in ***S*** (because conditioning on ***S*** opens every collider).

words, if $\boldsymbol{S}$ separates X from Y, we will have $\Pr(Y = y|X = x, \boldsymbol{S} = \boldsymbol{S}) = \Pr(Y = y|\boldsymbol{S} = \boldsymbol{S})$ for every possible value x, y, $\boldsymbol{S}$ of X, Y, $\boldsymbol{S}$.

Rule 2 (weak faithfulness). Suppose that $\boldsymbol{S}$ does not separate X and Y. Then, if the graph is correct, X and Y may be associated given $\boldsymbol{S}$. In other words, if X and Y are connected given $\boldsymbol{S}$, then without further information we should not assume that X and Y are independent given $\boldsymbol{S}$.

As an illustration, consider again Figure 12–1. U and X are unassociated. Because Y is a collider, however, we expect U and X to become associated after conditioning on Y or Z or both (that is, $\boldsymbol{S}$ unblocks the path whether $\boldsymbol{S} = \{Y\}$, $\boldsymbol{S} = \{Z\}$, or $\boldsymbol{S} = \{Y, Z\}$). In contrast, X and Z are marginally associated, but become independent after conditioning on Y or $\boldsymbol{S} = \{U, Y\}$.

ASSUMPTIONS AND INTUITIONS UNDERLYING THE RULES

Although informal diagrams of causal paths go back at least to the 1920s, the mathematical theory of graphs (including DAGs) developed separately and did not at first involve causal inference. By the 1980s, however, graphs were being used to represent the structure of joint probability distributions, with d-separation being used to encode "stable" conditional independence relations (Pearl, 1988). One feature of this use of graphs is that a given distribution will have more than one graph that encodes these relations. In other words, graphical representations of probability distributions are not unique. For example, in probabilistic (associational) terms, $A \rightarrow B$ and $B \rightarrow A$ have the same implication, that A and B are dependent. By the 1990s, however, several research groups had adapted these probability graphs to causal inference by letting the arrows represent cause–effect relations, as they had in path diagrams. Many graphical representations that are probabilistically equivalent are not causally equivalent. For example, if A precedes B temporally, then $B \rightarrow A$ can be ruled out as a representation for the relation of A and B.

The compatibility and faithfulness rules define what we mean when we say that a causal model for a set of variables is consistent with a probability model for the distribution of those variables. In practice, the rules are used to identify causal graphs consistent with the observed probability distributions of the graphed variables, and, conversely, to identify distributions that are consistent with a given causal graph. When the arrows in probability graphs represent causal processes, the compatibility rule above (rule 1) is equivalent to the *causal Markov assumption* (CMA), which formalizes the idea that (apart from chance) all unconditional associations arise from ancestral causal relations. Causal explanations of an association between two variables invoke some combination of shared common causes, collider bias, and one of the variables affecting the other. These relations form the basis for Rule 1.

Specifically, the CMA states that for any variable X, conditional upon its direct causes (parents), X is independent of all other variables that it does not affect (its nondescendants). This condition asserts that if we can hold constant the direct causes of X, then X will be independent of any other variable that is not itself affected by X. Thus, assuming X precedes Y temporally, in a DAG without

conditioning there are only two sources of association between X and Y: Effects of X on Y (directed paths from X to Y), or common causes (shared ancestors) of X and Y, which introduce confounding. We will make use of this fact when we discuss control of bias.

The d-separation rule (Rule 1) and equivalent conditions such as the CMA codify common intuitions about how probabilistic relations (associations) arise from causal relations. We rely implicitly on these conditions in drawing causal inferences and predicting everyday events—ranging from assessments of whether a drug in a randomized trial was effective to predictions about whether flipping a switch on the wall will suffuse a room with light. In any sequence of events, holding constant both intermediate events and confounding events (common causes) will interrupt the causal cascades that produce associations. In both our intuition and in causal graph theory, this act of "holding constant" renders the downstream events independent of the upstream events. Conditioning on a set that d-separates upstream from downstream events corresponds to this act. This correspondence is the rationale for deducing the conditional independencies (features of a probability distribution) implied by a given causal graph from the d-separation rule.

The intuition behind Rule 2 is this: If, after conditioning on S, there is an open path between two variables, then there must be some causal relation linking the variables, and so they ought to be associated given S, apart from certain exceptions or special cases. An example of an exception occurs when associations transmitted along different open paths perfectly cancel each other, resulting in no association overall. Other exceptions can also occur. Rule 2 says only that we should not count on such special cases to occur, so that, in general, when we see an open path between two variables, we expect them to be associated, or at least we are not surprised if they are associated.

Some authors go beyond Rule 2 and assume that an open path between two variables means that they *must* be associated. This stronger assumption is called *faithfulness* or *stability* and says that if S does not d-separate X and Y, then X and Y will be associated given S. Faithfulness is thus the logical converse of compatibility (Rule 1). Compatibility says that if two variables are d-separated, then they must be independent; faithfulness says that if two variables are independent, then they must be d-separated. When both compatibility and faithfulness hold, we have *perfect compatibility,* which says that X and Y are independent given S if *and only if* S d-separates X and Y; faithfulness adds the "only if" part. For any given pattern of associations, the assumption of perfect compatibility rules out a number of possible causal structures (Spirtes et al., 2001). Therefore, when it is credible, perfect compatibility can help identify causal structures underlying observed data.

Nonetheless, because there are real examples of near-cancellation (e.g., when confounding obscures a real effect in a study) and other exceptions, faithfulness is controversial as a routine assumption, as are algorithms for inferring causal structure from observational data; see Robins (1997, section 11), Korb and Wallace (1997), Freedman and Humphreys (1999), Glymour et al. (1999), Robins and Wasserman (1999), and Robins et al. (2003). Because of this controversy, we discuss only uses of graphical models that do not rely on the assumption of faithfulness. Instead, we use Rule 2, which weakens the faithfulness condition by saying that the presence of open paths alerts us to the possibility of association, and so we should allow for that possibility.

The rules and assumptions just discussed should be clearly distinguished from the content-specific causal assumptions encoded in a diagram, which relate to the substantive question at hand. These rules serve only to link the assumed causal structure (which is ideally based on sound and complete contextual information) to the associations that we observe. In this fashion, they allow testing of those assumptions and estimation of the effects implied by the graph.

GRAPHICAL REPRESENTATION OF BIAS AND ITS CONTROL

A major use of causal graphs is to identify sources of bias in studies and proposed analyses, including biases resulting from confounding, selection, or overadjustment. Given a causal graph, we can use the definitions and rules we have provided to determine whether a set of measured variables $\boldsymbol{S}$ is sufficient to allow us to identify (validly estimate) the causal effect of X on Y.

Suppose that X precedes Y temporally and that the objective of a study is to estimate a measure of the effect of X on Y. We will call an undirected open path between X and Y a *biasing path* for the effect because such paths do not represent effects of X on Y, yet can contribute to the association of X and Y. The association of X and Y is *unconditionally unbiased* or *marginally unbiased* for the effect of X on Y if the only open paths from X to Y are the directed paths.

SUFFICIENT AND MINIMALLY SUFFICIENT CONDITIONING SETS

When there are biasing paths between X and Y, it may be possible to close these paths by conditioning on other variables. Consider a set of variables $\boldsymbol{S}$. The association of X and Y is *unbiased given* $\boldsymbol{S}$ if, after conditioning on $\boldsymbol{S}$, the open paths between X and Y are exactly (only and all) the directed paths from X to Y. In such a case we say that $\boldsymbol{S}$ is *sufficient* to control bias in the association of X and Y. Because control of colliders can open biasing paths, it is possible for a set $\boldsymbol{S}$ to be sufficient, and yet a larger set containing $\boldsymbol{S}$ and such colliders may be insufficient.

A sufficient set $\boldsymbol{S}$ is *minimally sufficient* to identify the effect of X on Y if no proper subset of $\boldsymbol{S}$ is sufficient (i.e., if removing any set of variables from $\boldsymbol{S}$ leaves an insufficient set). In practice, there may be several distinct sufficient sets and even several distinct minimally sufficient sets for bias control. Investigators may sometimes wish to adjust for more variables than are included in what appears as a minimally sufficient set in a graph (e.g., to allow for uncertainty about possible confounding paths). Identifying minimally sufficient sets can be valuable nonetheless, because adjusting for more variables than necessary risks introducing biases and reducing precision, and measuring extra variables is often difficult or expensive.

For example, the set of all parents of X is always sufficient to eliminate bias when estimating the effects of X in an unconditional DAG. Nonetheless, the set of parents of X may be far from *minimally* sufficient. Whenever X and Y share no ancestor and there is no conditioning or measurement error, the only open paths from X to Y are directed paths. In this case, there is no bias and hence no need for conditioning to prevent bias in estimating the effect of X on Y, no matter how many parents of X exist.

CHOOSING CONDITIONING SETS TO IDENTIFY CAUSAL EFFECTS

There are several reasons to avoid (where possible) including descendants of X in a set $\boldsymbol{S}$ of conditioning variables. First, conditioning on descendants of X that are intermediates will block directed (causal) paths that are part of the effect of interest, and thus create bias. Second, conditioning on descendants of X can unblock or create paths that are not part of the effect of X on Y and thus introduce another source of bias. For example, biasing paths can be created when one conditions on a descendant Z of both X and Y. The resulting bias is the Berksonian bias described earlier. Third, even when inclusion of a particular descendant of X induces no bias, it may still reduce precision in effect estimation.

Undirected paths from X to Y are termed *back-door* (relative to X) if they start with an arrow pointing into X (i.e., it leaves X from a "back door"). In Figure 12–2, the one path from X to Y is back-door because it starts with the back-step $X \leftarrow Z_1$. Before conditioning, all biasing paths in a DAG are open back-door paths, and all open back-door paths are biasing paths. Thus, to identify the causal effect of X on Y all the back-door paths between the two variables must be blocked. A set $\boldsymbol{S}$ satisfies the *back-door criterion* for identifying the effect of X on Y if $\boldsymbol{S}$ contains no descendant of X and there is no open back-door path from X to Y after conditioning on S. If $\boldsymbol{S}$ satisfies the back-door criterion, then conditioning on $\boldsymbol{S}$ alone is sufficient to control bias in the DAG, and we say that the effect of X on Y is *identified* or *estimable* given $\boldsymbol{S}$ alone. We emphasize again, however, that further conditioning may introduce bias: Conditioning on a collider may create new biasing paths, and conditioning on an intermediate will block paths that are part of the effect under study.

CONFOUNDING AND SELECTION BIAS

The terms *confounding* and *selection bias* have varying and overlapping usage in different disciplines. The traditional epidemiologic concepts of confounding and selection bias both correspond to biasing paths between X and Y. The distinction between the two concepts is not consistent across the literature, however, and many phenomena can be reasonably described as both confounding and selection bias. We emphasize that the d-separation criteria are sufficient to identify structural sources of bias, and thus there is no need to categorize each biasing path as a confounding or selection-bias path. Nonetheless, the discussion below may help illustrate the correspondence between conventional epidemiologic terms and sources of bias in causal diagrams.

Traditionally, confounding is thought of as a source of bias arising from causes of Y that are associated with but not affected by X (Chapter 9). Thus we say that a biasing path from X to Y is

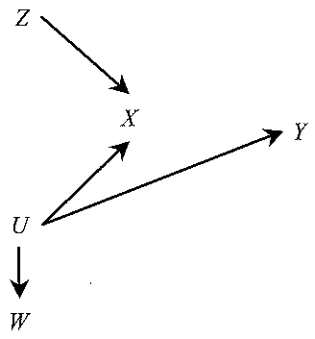


FIGURE 12–3 ● A causal diagram with confounding of the X–Y association by U but not by Z.

a *confounding path* if it ends with an arrow into Y. Bias arising from a common cause of X and Y (and thus present in the unconditional graph, e.g., U in Figure 12–3) is sometimes called "classical confounding" (Greenland, 2003a) to distinguish it from confounding that arises from conditioning on a collider. Variables that intercept confounding paths between X and Y are *confounders*.

Often, only indirect measures of the variables that intercept a confounding path are available (e.g., W in Figure 12–3). In this case, adjusting for such surrogates or markers of proper confounders may help remediate bias (Greenland and Pearl, 2008). Such surrogates are often referred to informally as confounders. Caution is needed whenever adjusting for a surrogate in an effort to block a confounding path. To the extent that the surrogate is imperfectly related to the actual confounder, the path will remain partially open. Furthermore, if variables other than the actual confounder itself influence the surrogate, conditioning on the surrogate may open new paths and introduce collider bias. More generally, adjusting for an imperfect surrogate may increase bias under certain circumstances. Related issues will be discussed in the section on residual confounding.

If a confounding path is present, we say that the dependence of Y on X is *confounded*, and if no confounding path is present we say that the dependence is *unconfounded*. Note that an unconfounded dependency may still be biased because of biasing paths that are not confounding paths (e.g., if Berksonian bias is present). Thus, $\boldsymbol{S}$ may be sufficient for confounding control (in that it blocks all confounding paths), and yet may be insufficient to control other bias (such as Berksonian bias, which is often uncontrollable).

If W is a variable representing selection into the study sample (e.g., due to intentional selection, self-selection, or survival), all analyses are conditioned on W. Selection bias is thus sometimes defined as the collider bias that arises from conditioning on selection W. For example, in Figure 12–4, we would say that, before conditioning on W, the relation between X and Y is confounded by the path $X - Z_1 - W - Y$. Conditioning on W alone opens the confounding path $X - Z_1 - W - Z_2 - Y$; the bias that results is a collider bias because the bias arises from conditioning on W, a common

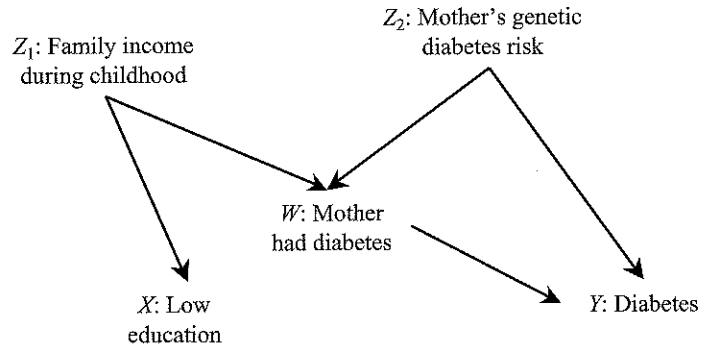


FIGURE 12–4 ● A diagram under which control of W alone might increase bias even though W is a confounder.

effect of causes of X and Y. But it can also be called confounding, because the bias arises from a path that ends with an arrow into Y.

Econometricians and others frequently use "selection bias" to refer to any form of confounding. The motivation for this terminology is that some causes of Y also influence "selection for treatment," that is, selection of the level of X one receives, rather than selection into the study sample. This terminology is especially common in discussions of confounding that arises from self-selection, e.g., choosing to take hormone-replacement therapy. Other writers call any bias created by conditioning a "selection bias," thus using the term "selection bias" for what we have called collider bias (Hernán et al., 2004); they then limit their use of "confounding" to what we have defined as "classical confounding" (confounding from a common cause of X and Y).

Regardless of terminology, it is helpful to identify the potential sources of bias to guide both design and analysis decisions. Our examples show how bias can arise in estimating the effect of X on Y if selection is influenced either by X or by factors that influence X, and is also influenced by Y or factors that influence Y. Thus, to control the resulting bias, one will need good data on either the factors that influence both selection and X or the factors that influence both selection and Y. We will illustrate these concepts in several later examples, and provide further structure to describe biases due to measurement error, missing data, and model-form misspecification.

SOME APPLICATIONS

Causal diagrams help us answer causal queries under various assumed causal structures, or causal models. Consider Figure 12–3. If we are interested in estimating the effect of X on Y, it is evident that, under the model shown in the figure, our analysis should condition on U: There is a confounding path from X to Y, and U is the only variable on the path. On the other hand, suppose that we are interested in estimating the effect of Z on Y. Under the diagram in Figure 12–3, we need not condition on U, because the relation of Z to Y is unconfounded (as is the relation of X to Z), that is, there is no confounding path from Z to Y. Because Figure 12–3 is a DAG, we can rephrase these conditions by saying that there is an open back-door path from X to Y, but not from Z to Y.

We now turn to examples in which causal diagrams can be used to clarify methodologic issues. In some cases the diagrams simply provide a convenient way to express well-understood concepts. In other examples they illuminate points of confusion regarding the biases introduced by proposed analyses or study designs. In all these cases, the findings can be shown mathematically or seen by various informal arguments. The advantage of diagrams is that they provide flexible visual explanations of the problems, and the explanations correspond to logical relations under the definitions and rules given earlier.

WHY CONVENTIONAL RULES FOR CONFOUNDING ARE NOT ALWAYS RELIABLE

In both intuition and application, the graphical and conventional criteria for confounding overlap substantially. For example, in Chapter 9, confounding was informally described as a distortion in the estimated exposure effect that results from differences in risk between the exposed and unexposed that are not due to exposure. Similarly, Hennekens and Buring (1987, p. 35) say that confounding occurs when "an observed association. . . . is in fact due to a mixing of effects between the exposure, the disease, and a third factor. . . ."

Variations on the following specific criteria for identifying confounders are frequently suggested, although, as noted in Chapter 9, these criteria do not define a confounder:

1. A confounder must be associated with the exposure under study in the source population.
2. A confounder must be a "risk factor" for the outcome (i.e., it must predict who will develop disease), though it need not actually cause the outcome.
3. The confounding factor must not be affected by the exposure or the outcome.

These traditional criteria usually agree with graphical criteria; that is, one would choose the same set of covariates for adjustment using either set of criteria. For example, in Figure 12–3, both the graphical and intuitive criteria indicate that one should condition on U to derive an unbiased estimate of the effect of X on Y. Under the graphical criteria, U satisfies the back-door criterion

for identifying the effect of X on Y: U is not an effect of X, and the only path between X and Y that contains an arrow into X can be blocked by conditioning on U. It fulfills the three traditional criteria because U and X will be associated, U will also predict Y, and U is not affected by X or Y.

Nonetheless, there are cases in which the criteria disagree, and when they diverge, it is the conventional criteria (1–3) that fail. Suppose that we are interested in whether educational attainment affects risk of type II diabetes. Figure 12–2 then depicts a situation under the causal null hypothesis in which education (X) has no effect on subject's diabetes (Y). Suppose that we have measured maternal diabetes status (W), but we do not have measures of family income during childhood (Z_1) or whether the mother had any genes that would increase risk of diabetes (Z_2). Should we adjust for W, maternal diabetes?

Figure 12–2 reflects the assumption that family income during childhood affects both educational attainment and maternal diabetes. The reasoning is that if a subject was poor as a child, his or her mother was poor as an adult, and this poverty also increased the mother's risk of developing diabetes (Robbins et al., 2005). Maternal diabetes will thus be associated with the subject's education, because under these assumptions they share a cause, family income. In Figure 12–2, this association is due entirely to confounding of the X–W (education-maternal diabetes) association. Figure 12–2 also reflects the assumption that a maternal genetic factor affects risk of both maternal diabetes and the subject's diabetes. Maternal diabetes will thus be associated with the subject's diabetes, because under these assumptions they share a cause, the genetic factor. In Figure 12–2, this association is purely confounding of the W–Y (maternal diabetes-subject's diabetes) association.

In Figure 12–2, maternal diabetes W is not affected by the subject's education level X or diabetes status Y. Thus, the mother's diabetes meets the three traditional criteria for a confounder, so these criteria could lead one to adjust for mother's diabetic status. Note, however, that both the associations on which the latter decision is based (traditional criteria 1 and 2) arise from confounding.

Turning to the graphical criteria, note first that there is only one undirected path between low education X and diabetes Y, and mother's diabetes W is a collider on that path. Thus this path is blocked at W and transmits no association between X and Y — that is, it introduces no bias. This structure means that we get an unbiased estimate if we do *not* adjust for the mother's diabetes. Because maternal diabetes is a collider, however, adjusting for it opens this undirected path, thus introducing a potential spurious association between low education and diabetes. The path opened by conditioning on W could be blocked by conditioning on either Z_1 or Z_2, but there is no need to condition on W in the first place. Therefore, under Figure 12–2, the graphical criteria show that one should not adjust for maternal diabetes, lest one introduce bias where none was present to begin with. In this sense, adjustment for W would be one form of overadjustment (Chapter 15), and the traditional criteria were mistaken to identify W as a confounder.

Figure 12–2 illustrates why in Chapter 9 it was said that the traditional criteria do not *define* a confounder: While every confounder will satisfy them, Figure 12–2 shows that some nonconfounders satisfy them as well. In some cases, adjusting for such nonconfounders is harmless, but in others, as in the example here, it introduces a bias. This bias may, however, be removed by adjustment for another variable on the newly opened path.

The situation in Figure 12–2 is analogous to Berksonian bias if we focus on the part of the graph (subgraph) in which $Z_1 \rightarrow W \leftarrow Z_2$: Conditioning on the collider W connects its parents Z_1 and Z_2, and thus connects X to Y. Another way to describe the problem is that we have a spurious appearance of confounding by W if we do not condition on Z_1 or Z_2, for then W is associated with X and Y. Because W temporally precedes X and Y, these associations may deceive one into thinking that W is a confounder. Nonetheless, the association between W and X is due solely to the effects of Z_1 on W and X, and the association between W and Y is due solely to the effects of Z_2 on W and Y. There is no common cause of X and Y, however, and hence no confounding if we do not condition on W.

To eliminate this sort of problem, traditional criterion 2 (here, that W is a "risk factor" for Y) is sometimes replaced by

2′. The variable must affect the outcome under study.

This substitution addresses the difficulty in examples like Figure 12–2 (for W will fail this revised criterion). Nonetheless, it fails to address the more general problem that conditioning may introduce

bias. To see this failing, draw an arrow from W to Y in Figure 12–2, which yields Figure 12–4. W now affects the outcome, Y, and thus satisfies criterion 2′. This change is quite plausible, because having a mother with diabetes might lead some subjects to be more careful about their weight and diet, thus lowering their own diabetes risk. W is now a confounder: Failing to adjust for it leaves open a confounding path ($X \leftarrow Z_1 \rightarrow W \rightarrow Y$) that is closed by adjusting for W. But adjusting for W will open an undirected (and hence biasing) path from X to Y ($X \leftarrow Z_1 \rightarrow W \leftarrow Z_2 \rightarrow Y$), as just discussed. The only ways to block both biasing paths at once is to adjust for Z_1 (alone or in combination with any other variable) or both Z_2 and W together.

If neither Z_1 nor Z_2 is measured, then under Figure 12–4, we face a dilemma not addressed by the traditional criteria. As with Figure 12–2, if we adjust for W, we introduce confounding via Z_1 and Z_2; yet, unlike Figure 12–2, under Figure 12–4 we are left with confounding by W if we do not adjust for W. The question is, then, which undirected path is more biasing, that with adjustment for W or that without? Both paths are modulated by the same X–W connection ($X \leftarrow Z_1 \rightarrow W$), so we may focus on whether the connection of W to Y with adjustment ($W \leftarrow Z_2 \rightarrow Y$) is stronger than the connection without adjustment ($W \rightarrow Y$). If so, then we would ordinarily expect less bias when we don't adjust for W; if not, then we would ordinarily expect less bias if we adjust. The final answer will depend on the strength of the effect represented by each arrow, which is context-specific. Assessments of the likely relative biases (as well as their direction) thus depend on subject-matter information.

In typical epidemiologic examples with noncontagious events, the strength of association transmitted by a path attenuates rapidly as the number of variables through which it passes increases. More precisely, the longer the path, the more we would expect attenuation of the association transmitted by the path (Greenland, 2003a). In Figure 12–4, this means that the effects of Z_2 on W and Z_2 on Y would both have to be much stronger than the effect of W on Y in order for the unadjusted X–Y association to be less biased than the W-adjusted X–Y association. However, if the proposed analysis calls for stratifying or restricting on W (instead of adjusting for W), the bias within a single stratum of W can be larger than the bias when adjusting for W (which averages across all strata).

To summarize, expressing assumptions in a DAG provides a flexible and general way to identify "sufficient" sets under a range of causal structures, using the d-separation rules. For example, if we changed the structure in Fig 12–2 only slightly by reversing the direction of the relationship between Z_1 and W (so we have $X \leftarrow Z_1 \leftarrow W \leftarrow Z_2 \rightarrow Y$), then conditioning on W would be desirable, and any of Z_1, W, or Z_2 would provide a sufficient set for identifying the effect of X on Y. Modified versions of the conventional criteria for confounder identification have been developed that alleviate their deficiencies and allow them to identify sufficient sets, consistent with the graphical criteria (Greenland et al., 1999a). We do not present these here because they are rarely used and, in general, it is simpler to apply the graphical criteria.

GRAPHICAL ANALYSES OF SELECTION BIAS

Selection forces in a study may be part of the design (e.g., enrollment criteria, or hospitalization status in a hospital-based case-control study) or may be unintended (e.g., loss to follow-up in a cohort study, or refusals in any study). Selection forces can of course compromise generalizability (e.g., results for white men may mislead about risk factors in black women). As shown by the above examples and discussed in Chapters 7 through 9, they can also compromise the internal validity of a study.

Causal diagrams provide a unifying framework for thinking about well-known sources of bias and also illustrate how some intentional selection and analysis strategies result in bias in more subtle situations. To see these problems, we represent selection into a study as a variable, and then note that all analyses of a sample are conditioned on this variable. That is, we conceptualize selection as a variable with two values, 0 = not selected and 1 = selected; analyses are thus restricted to observations where selection = 1. Selection bias may occur if this selection variable (that is, entry into the study) depends on the exposure, the outcome, or their causes (whether shared or not).

BIAS FROM INTENTIONAL SELECTION

Even seemingly innocuous choices in dataset construction can induce severe selection bias. To take an extreme example, imagine a study of education (X) and Alzheimer's disease (Y) conducted

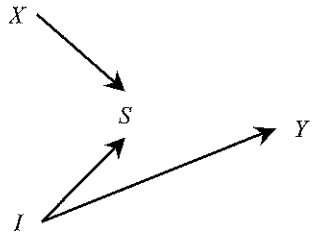


FIGURE 12–5 • A diagram with a selection indicator *S*.

by pooling two datasets, one consisting only of persons with college education (X = high), the other consisting only of persons diagnosed with impaired memory ($I = 1$). Within this pooled study, everyone without college education (X = low) has memory impairment ($I = 1$), which in turn is strongly associated with Alzheimer's disease because impairment is often a symptom of early, undiagnosed Alzheimer's disease (in fact, it is a precursor or prodrome). Likewise, any subject with no impairment ($I = 0$) has college education (X = high). Thus, in this study, college education is almost certainly negatively associated with Alzheimer's disease. This association would be completely spurious, induced by defining selection as an effect of both education (X) and memory impairment (I) as a result of pooling the two datasets. Graphing the relations in Figure 12–5, this association can be viewed as Berksonian bias: Selection S is strongly affected by both the exposure X and an independent cause of the outcome Y, hence is a collider between them. All analyses are conditioned on selection and the resulting collider bias will be large, greatly misrepresenting the population association between education and Alzheimer's disease.

This example parallels Berksonian bias in clinic-based and hospital-based studies, because selection was affected directly by exposure and outcome. Selection is often only indirectly related to exposure and outcome, however. Suppose we study how education affects risk for Alzheimer's disease in a study with selection based on membership in a high-prestige occupation. Achievement of high-prestige occupations is likely to be influenced by both education and intellect. Of course, many people obtain prestigious jobs by virtue of other advantages besides education or intelligence, but to keep our example simple, we will assume here that none of these other factors influence Alzheimer's disease.

There is evidence that intelligence protects against diagnosis of Alzheimer's disease (Schmand et al., 1997). Consider Figure 12–5 (relabeling the variables from the previous example), in which selection S (based on occupation) is influenced by education (X) and intellect (I), where the latter affects Alzheimer's disease (Y). Among the high-prestige job holders, people with less education (X = lower) are more likely to have high intellect (I = high), whereas those with lesser intellect (I = lower) are more likely to have advanced education (X = high), because most individuals had to have some advantage (at least one of X = high or I = high) to get their high-prestige job. In effect, X and I are compensatory, in that having more of one compensates somewhat for having less of the other, even if everyone in the study is above average on both.

The selection process thus biases the education–intellect association away from the association in the population as a whole. The strength of the spurious association will depend on the details of the selection process, that is, how strongly education and intellect each affect occupation and whether they interact in any way to determine occupation. Note, however, that if high-education subjects are less likely to have high intellect than low-education subjects, and high intellect protects against Alzheimer's disease, then high-education subjects will exhibit excess risk of Alzheimer's disease relative to low-education subjects even if education has no effect. In other words, whatever the true causal relation between education and Alzheimer's disease, in a study of high-prestige job holders, the association in the study will be biased downward, unless one can adjust for the effect of intellect on Alzheimer's disease.

Telling this story in words is complicated and prone to generating confusion, but analyzing a corresponding diagram is straightforward. In Figure 12–5, we can see that S is a collider between X and I, and so we should expect X and I to be associated conditional on S. Thus, conditional on S, we expect X and Y to be associated, even if X does not affect Y. Whether selection exacerbates or reduces bias in estimating a specific causal effect depends crucially on the causal relations among

variables determining selection. If we added an arrow from I to X in Figure 12–5 (i.e. if intellect directly affects education), I would be a confounder and the X–Y association would be biased before selection. If the confounding produced by I were upward, the bias produced by selection on S might counteract it enough to lessen the overall (net) bias in the X–Y association.

SURVIVOR BIAS

Survivor bias, and more generally bias due to differential competing risks or loss to follow-up, can be thought of as a special case of selection bias. In life-course research on early life exposures and health in old age, a large fraction of the exposed are likely to die before reaching old age, so survivor bias could be large. Effect estimates for early life exposures often decline with age (Elo and Preston, 1996; Tate et al., 1998). An example is the black–white mortality crossover: Mortality is greater for blacks and other disadvantaged groups relative to whites at younger ages, but the pattern reverses at the oldest ages (Corti et al., 1999; Thornton, 2004). Do such phenomena indicate that the early life exposures become less important with age? Not necessarily. Selective survival can result in attenuated associations among survivors at older ages, even though the effects are undiminished (Vaupel and Yashin, 1985; Howard and Goff, 1998; Mohtashemi and Levins, 2002). The apparent diminution of the magnitude of effects can occur due to confounding by unobserved factors that conferred a survival advantage.

Apart from some special cases, such confounding should be expected whenever both the exposure under study and unmeasured risk factors for the outcome influence survival—even if the exposure and factors were unassociated at the start of life (and thus the factors are not initially confounders). Essentially, if exposure presents a disadvantage for survival, then exposed survivors will tend to have some other characteristic that helped them to survive. If that protective characteristic also influences the outcome, it creates a spurious association between exposure and the outcome. This result follows immediately from a causal diagram like Figure 12–5, interpreted as showing survival (S) affected by early exposure (X) and also by an unmeasured risk factor (I) that also affects the study outcome (Y).

RESIDUAL CONFOUNDING AND BIAS QUANTIFICATION

Ideally, to block a back-door path between X and Y by conditioning on a variable or set of variables Z, we would have sufficient data to create a separate analysis stratum for every observed value of Z and thus avoid making any assumptions about the form of the relation of Z to X or Y. Such complete stratification may be practical if Z has few observed values (e.g., sex). In most situations, however, Z has many levels (e.g., Z represents a set of several variables, including some, such as age, that are nearly continuous), and as a result we obtain cells with no or few persons if we stratify on every level of Z. The standard solutions compensate for small cell counts using statistical modeling assumptions (Robins and Greenland, 1986). Typically, these assumptions are collected in the convenient form of a regression model, as described in Chapter 20. The form of the model will rarely be perfectly correct, and to the extent that it is in error, the model-based analysis will not completely block confounding paths. The bias that remains as a result is an example of *residual confounding*, i.e., the confounding still present after adjustment.

Causal diagrams are nonparametric in that they make no assumption about the functional form of relationships among variables. For example, the presence of open paths between two variables leads us to expect they are associated in some fashion, but a diagram does not say how. The association between the variables could be linear, U-shaped, involve a threshold, or an infinitude of other forms. Thus the graphical models we have described provide no guidance on the form to use to adjust for covariates.

One aspect of the residual confounding problem, however, can be represented in a causal diagram, and that is the form in which the covariates appear in a stratified analysis or a regression model. Suppose Z is a covariate, that when uncontrolled induces a positive bias in the estimated relationship between the exposure and outcome of interest. Stratification or regression adjustment for a particular form of Z, say $g(Z)$, may eliminate bias; for example, there might be no bias if Z is entered in the analysis as its natural logarithm, $\ln(Z)$. But there might be considerable bias left if we enter Z in a different form $f(Z)$, e.g., as quartile categories, which in the lowest category combines persons

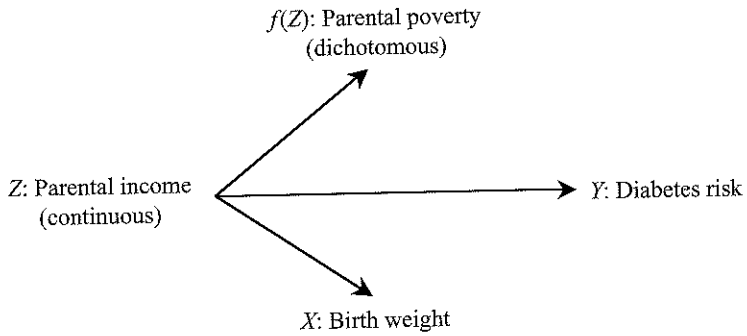


FIGURE 12–6 ● Diagram with residual confounding of the *X*–*Y* association after control of $f(Z)$ alone.

with very different values of $\ln(Z)$. Similarly, use of measures $f(Z)$ of Z that suffer from substantial error could make it impossible to adjust accurately for Z.

"Blocking the path at Z" involves complete stratification on the variables in a sufficient set, or anything equivalent, even if the resulting estimate is too statistically unstable for practical use. We can thus represent our problem by adding to the diagram the possibly inferior functional form or measurement $f(Z)$ as a separate variable. This representation shows that, even if Z is sufficient to control confounding, $f(Z)$ may be insufficient.

To illustrate, suppose that we are interested in estimating the effect of birth weight on adult diabetes risk, and that Figure 12–6 shows the true causal structure. We understand that parental income Z is a potential confounder of the relationship between birth weight and diabetes risk because it affects both variables. Suppose further that this relationship is continuously increasing (more income is better even for parents who are well above the poverty line), but, unfortunately, our data set includes no measure of income. Instead, we have only an indicator $f(Z)$ for whether or not the parents were in poverty (a dichotomous variable); that is, $f(Z)$ is an indicator of very low income—e.g., $f(Z) = 1$ if $Z <$ poverty level, $f(Z) = 0$ otherwise. Poverty is an imperfect surrogate for income. Then the association between birth weight and diabetes may be confounded by parental income even conditional on $f(Z)$, because $f(Z)$ fails to completely block the confounding path between parental income and diabetes. The same phenomena will occur using a direct measure of income that incorporates substantial random error. In both cases, residual confounding results from inadequate control of income.

BIAS FROM USE OF MISSING-DATA CATEGORIES OR INDICATORS

Many methods for handling missing data are available, most of which are unbiased under some assumptions but biased under alternative scenarios (Robins et al., 1994; Greenland and Finkle, 1995; Little and Rubin, 2002; see Chapter 13). In handling missing data, researchers usually want to retain as many data records as possible to preserve study size and avoid analytic complexity. Thus, a popular approach to handling missing data on a variable Z is to treat "missing" as if it were just another value for Z. The idea is often implemented by adding a stratum for $Z =$ "missing," which in questionnaires includes responses such as "unknown" and "refused." The same idea is implemented by adding an indicator variable for missingness to a regression model: We set Z to 0 when it is missing, and add an indicator $M_Z = 0$ if Z is observed, $M_Z = 1$ if Z is missing.

Missing indicators allow one to retain every subject in the analysis and are easy to implement, but they may introduce bias. This bias can arise even under the best-case scenario, that the data are missing completely at random (MCAR). MCAR means that missingness of a subject's value for Z is independent of every variable in the analysis, including Z. For example, if Z is sexual orientation, MCAR assumes that whether someone skips the question or refuses to answer has nothing to do with the person's age, sex, or actual preference. Thus MCAR is an exceedingly optimistic assumption, but it is often used to justify certain techniques.

Next, suppose that Figure 12–7 represents our study. We are interested in the effect of X on Y, and we recognize that it is important to adjust for the confounder Z. If Z is missing for some

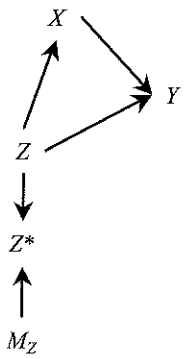


FIGURE 12–7 • Diagram with a missing-data indicator M_Z.

subjects, we add to the analysis the missing indicator M_Z. If Z is never zero, we also define a new variable, Z^*, that equals Z whenever Z is observed and equals 0 whenever Z is missing, that is, $Z^* = Z(1 - M_Z)$. There are no arrows pointing into M_Z in the diagram, implying that Z is unconditionally MCAR, but Z^* is determined by both Z and M_Z. Using the missing-indicator method, we enter both Z^* and M_Z in the regression model, and thus we condition on them both.

In Figure 12–7, the set $\{Z^*, M_Z\}$ does not block the back-door path from X to Y via Z, so control of Z^* and M_Z does not fully control the confounding by Z (and we expect this residual confounding to increase as the fraction with Z missing increases). Similarly, it should be clear from Figure 12–7 that conditioning only on Z* also fails to block the back-door path from X to Y. Now consider a complete-subject analysis, which uses only observations with Z observed—in other words, we condition on (restrict to) $M_Z = 0$. From Figure 12–7 we see that this conditioning creates no bias. Because we have Z on everyone with $M_Z = 0$, we can further condition on Z and eliminate all confounding by Z. So we see that instead of the biased missing-indicator approach, we have an unbiased (and even simpler) alternative: an analysis limited to subjects with complete data. The diagram can be extended to consider alternative assumptions about the determinants of missingness. Note, however, that more efficient and more broadly unbiased alternatives to complete-subject analysis (such as multiple imputation or inverse probability weighting) are available, and some of these methods are automated in commercial software packages.

ADJUSTING FOR AN INTERMEDIATE DOES NOT NECESSARILY ESTIMATE A DIRECT EFFECT

Once an effect has been established, attention often turns to questions of mediation. Is the effect of sex on depression mediated by hormonal differences between men and women or by differences in social conditions? Is the effect of prepregnancy body mass index on pre-eclampsia risk mediated by inflammation? Is the apparent effect of occupational status on heart disease attributable to psychologic consequences of low occupational status or to material consequences of low-paying jobs?

In considering exposure X and outcome Y with an intermediate (mediator) Z, a direct effect of X on Y (relative to Z) is an X effect on Y that is not mediated by Z. In a causal diagram, effects of X on Y mediated by Z, or "indirect effects," are those directed paths from X to Y that pass through Z. Direct effects are then represented by directed paths from X to Y that do not pass through Z. Nonetheless, because Z may modify the magnitude of a direct effect, the total effect of X on Y cannot necessarily be partitioned into nonoverlapping direct and indirect effects (Robins and Greenland, 1992).

The term *direct effect* may refer to either of two types of effects. The first type is the effect of X on Y in an experiment in which each individual's Z is held constant at the same value z. This has been termed the *controlled direct effect* because the intermediate is controlled. The magnitude of this direct effect may differ across each possible value of Z; thus there is a controlled direct effect defined for every possible value of Z. The second type is called a *pure* or *natural* direct effect and is the effect of X on Y when Z takes on the value it would "naturally" have under a single reference

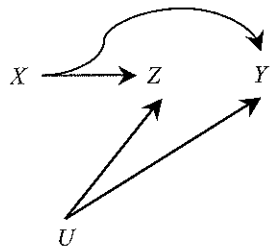


FIGURE 12–8 ● Diagram with an unconfounded direct effect and no indirect effect of X on Y.

value x for X. Thus there is one of these effects for each possible value of X. For each direct effect of X on Y, we can also define a contrast between the total effect of X on Y and that direct effect. This contrast is sometimes referred to as the "indirect effect of X on Y" relative to the chosen direct effect. There will be one of these contrasts for every controlled direct effect (i.e., for every level of Z) and one for every pure direct effect (i.e., for every level of X).

A causal diagram can reveal pitfalls in naive estimation procedures, as well as additional data and assumptions needed to estimate direct effects validly. For example, a standard method of direct-effect estimation is to adjust for (condition on) Z in the analysis—e.g., by entering it in a regression of Y on X. The Z-adjusted estimate of the X coefficient is taken as an estimate of "the" direct effect of X on Y (without being clear about which direct effect is being estimated). The difference in the X coefficients with and without adjustment for Z is then taken as the estimate of the indirect effect of X on Y (with respect to Z).

The diagram in Figure 12–8 shows no confounding of the total effect of X on Y, and no effect of Z on Y at all, so no indirect effect of X on Y via Z (all the X effect on Y is direct). Z is, however, a collider on the closed path from X to Y via U; thus, if we adjust for Z, we will open this path and introduce bias. Consequently, upon adjusting for Z, we will see the X association with Y change, misleading us into thinking that the direct and total effects differ. This change, however, only reflects the bias we have created by adjusting for Z.

This bias arises because we have an uncontrolled variable U that confounds the Z–Y association, and that confounds the X–Y association upon adjustment for Z. The bias could be removed by conditioning on U. This example is like that in Figure 12–3, in which adjusting for a seeming confounder introduced confounding that was not there originally. After adjustment for the collider, the only remedy is to obtain and adjust for more covariates. Here, the new confounders may have been unassociated with X to begin with, as we would expect if (say) X were randomized, and so are not confounders of the total effect. Nonetheless, if they confound the association of Z with Y, they will confound any conventionally adjusted estimate of the direct effect of X on Y.

As an illustration of bias arising from adjustment for intermediates, suppose that we are interested in knowing whether the effect of education on systolic blood pressure (SBP) is mediated by adult wealth (say, at age 60). Unfortunately, we do not have any measure of occupational characteristics, and it turns out that having a high-autonomy job promotes the accumulation of wealth and also lowers SBP (perhaps because of diminished stress). Returning to Figure 12–8, now X represents education, Y represents SBP, Z represents wealth at age 60, and U represents job autonomy. To estimate the effect of education on SBP that is not mediated by wealth, we need to compare the SBP in people with high and low education if the value of wealth were not allowed to change in response to education. Thus we might ask, if we gave someone high education but intervened to hold her wealth to what she would have accumulated had she had low education (but changed no other characteristic), how would SBP change compared with giving the person less education?

We cannot conduct such an intervention. The naive direct-effect (mediation) analysis described above instead compares the SBP of people with high versus low education who happened to have the same level of adult wealth. On average, persons with high education tend to be wealthier than persons with low education. A high-education person with the same wealth as a low-education person is likely to have accumulated less wealth than expected for some other reason, such as a low-autonomy job. Thus, the mediation analysis will compare people with high education but low

job autonomy to people with low education and average job autonomy. If job autonomy affects SBP, the high-education people will appear to be worse off than they would have been if they had average job autonomy, resulting in underestimation of the direct effect of education on SBP and hence overestimation of the indirect (wealth-mediated) effect.

The complications in estimating direct effects are a concern whether one is interested in mediator-controlled or pure (natural) direct effects. With a causal diagram, one can see that adjusting for a confounded intermediate will induce confounding of the primary exposure and outcome—even if that exposure is randomized. Thus confounders of the effect of the intermediate on the outcome must be measured and controlled. Further restrictions (e.g., no confounding of the X effect on Z) are required to estimate pure direct effects. For more discussion of estimation of direct effects, see Robins and Greenland (1992, 1994), Blakely (2002), Cole and Hernán (2002), Kaufman et al. (2004, 2005), Peterson et al. (2006), Peterson and van der Laan (2008), and Chapter 26.

INSTRUMENTAL VARIABLES

Observational studies are under constant suspicion of uncontrolled confounding and selection bias, prompting many to prefer evidence from randomized experiments. When noncompliance (nonadherence) and losses are frequent, however, randomized trials may themselves suffer considerable confounding and selection bias. Figure 12–9 illustrates both phenomena. In an observational study, U represents unmeasured confounders of the X–Y association. In a randomized trial, U represents variables that affect adherence to treatment assignment and thus influence received treatment X. In Figure 12–9, Z is called an *instrumental variable* (or *instrument*) for estimating the effect of X on Y.

Valid instruments for the effect of X on Y can be used to test the null hypothesis that X has no effect on Y. With additional assumptions, instrumental variable analyses can be exploited to estimate the magnitude of this effect within specific population subgroups. We will first review the assumptions under which a valid instrument can be used to test a null hypothesis of no causal effect, and then describe examples of additional assumptions under which an instrumental variable analysis identifies a specific causal parameter.

Under the assumptions in the DAG in Figure 12–9, assignment Z can be associated with Y only if Z affects X and X in turn affects Y, because the only open path from Z to Y is $Z \rightarrow X \rightarrow Y$. In other words, Z can be associated with Y only if the null hypothesis (that X does not affect Y) is false. Thus, if one rejects the null hypothesis for the Z–Y association, one must also reject the null hypothesis that X does not affect Y. This logical requirement means that, under Figure 12–9, a test of the Z–Y association will be a valid test of the X–Y null hypothesis, even if the X–Y association is confounded. The unconfoundedness of the Z–Y test, called the *intent-to-treat* test, is considered a "gold standard" in randomized trials: If Z represents the assigned treatment, Figure 12–9 holds if Z is truly randomized, even if the treatment received (X) is influenced by unmeasured factors that also affect the outcome Y.

In a DAG, a variable Z is an unconditionally valid instrument for the effect of X on Y if:

1. Z affects X (i.e., Z is an ancestor of X).
2. Z affects the outcome Y only through X (i.e., all directed paths from Z to Y pass through X).
3. Z and Y share no common causes.

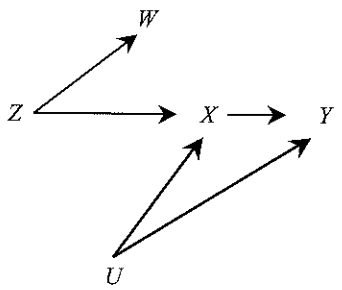


FIGURE 12–9 • Diagram with valid instruments Z, W for the X–Y effect.

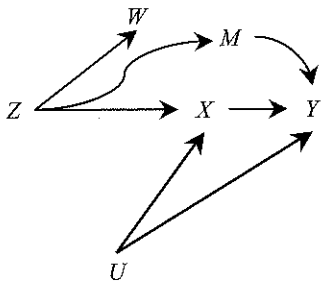


FIGURE 12–10 • Diagram for a confounded trial in which treatment assignment directly affects the outcome.

These assumptions are met in a well-conducted randomized trial in which Z is the randomized treatment-assignment variable. In Figure 12–10, assumption 2 is violated, and in Figure 12–11, assumption 3 is violated, and no unconditionally valid instrument is available in either case.

Most methods can be extended to allow use of certain descendants of Z (such as W in Figure 12–9) instead of Z itself to test whether X affects Y. Some authors extend the definition of instrumental variables to include such descendants. Note first that assumptions 2 and 3 imply that every open path from Z to Y includes an arrow pointing into X. This is a special case of a more general definition that W is an unconditional instrument for the $X \rightarrow Y$ effect in a DAG if (a) there is an open path from W to X, and (b) every open path from W to Y includes an arrow pointing into X. This definition extends to conditioning on a set of variables $\boldsymbol{S}$ that are unaffected by X: W is an instrument given $\boldsymbol{S}$ if, after conditioning on $\boldsymbol{S}$, (a) there is an open path from W to X, and (b) every open path from W to Y includes an arrow pointing into X (Pearl, 2000, section 7.4). For example, if W and Y share a common cause such as U_2 in Figure 12–11, but this common cause is included in $\boldsymbol{S}$, then W is a valid instrument for the effect of X on Y conditional on $\boldsymbol{S}$.

The assumptions for a valid instrument imply that, after conditioning on $\boldsymbol{S}$, the instrument–outcome association is mediated entirely through the X effect on Y. These assumptions require that $\boldsymbol{S}$ blocks all paths from W to Y not mediated by X. For example, conditioning on M in Figure 12–10 would render Z a valid instrument. Nonetheless, if $\boldsymbol{S}$ contains a descendant of W, there is a risk that conditioning on $\boldsymbol{S}$ may induce a W–Y association via collider bias, thus violating the conditional instrumental assumption (b). This collider bias might even result in an unconditionally valid instrument becoming conditionally invalid. Hence many authors exclude descendants of W (or Z) as well as descendants of X from $\boldsymbol{S}$.

Consider now a randomized trial represented by Figure 12–9. Although an association between Z and Y is evidence that X affects Y, the corresponding Z–Y (intent to treat or ITT) association will not equal the effect of X on Y if compliance is imperfect (i.e., if X does not always equal Z). In particular, the ITT (Z–Y) association will usually be attenuated relative to the desired $X \rightarrow Y$ effect because of the extra $Z \rightarrow X$ step. When combined with additional assumptions, however, the instrument Z may be used to estimate the effect of X on Y via special instrumental-variable (IV) estimation methods (Zohoori and Savitz, 1997; Newhouse and McClellan, 1998; Greenland, 2000b; Angrist and Krueger, 2001; Hernán and Robins, 2006; Martens et al., 2006) or related g-estimation methods (Robins and Tsiatis, 1991; Mark and Robins, 1993ab; White et al., 2002; Cole and Chu, 2005; Greenland et al., 2008; see also Chapter 21).

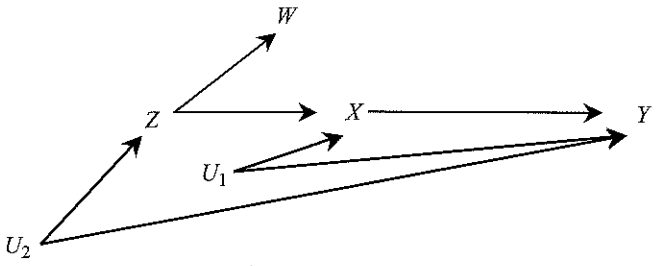


FIGURE 12–11 • Diagram for a confounded trial in which an unmeasured cause U_2 affects both treatment assignment Z and outcome Y.

Simple IV estimates are based on scaling up the Z–Y association in proportion to the Z–X association. An example of an assumption underlying these methods is *monotonicity* of the $Z \rightarrow X$ effect: For every member of the population, Z can affect X in only one direction (e.g., if increasing Z increases X for some people, then it cannot decrease X for anyone). Under monotonicity, IV estimates can be interpreted as the effect receiving the treatment had on those individuals who received treatment (got $X = 1$) precisely because they were assigned to do so (i.e., because they got $Z = 1$). Some methods use further assumptions, usually in the form of parametric models.

The causal structure in Figure 12–9 might apply even if the researcher did not assign Z. Thus, with this diagram in mind, a researcher might search for variables (such as Z or W) that are valid instruments and use these variables to calculate IV effect estimates (Angrist et al., 1996; Angrist and Krueger, 2001; Glymour, 2006a). Although it can be challenging to identify a convincing instrument, genetic studies (Chapter 28) and "natural experiments" may supply them:

- Day of symptom onset may determine the quality of hospital care received, but there is rarely another reason for day of onset to influence a health outcome. Day of symptom onset then provides a natural instrument for the effect of quality of hospital care on the outcome.
- Hour of birth may serve as an instrument for studying postpartum length of stay in relation to maternal and neonatal outcomes (Malkin et al., 2000).
- Mothers who deliver in hospitals with lactation counseling may be more likely to breast-feed. If being born in such a hospital has no other effect on child health, then hospital counseling (yes/no) provides an instrument for the effect of breastfeeding on child health.
- Women with relatives who had breast cancer may be unlikely to receive perimenopausal hormone therapy. If having relatives with breast cancer has no other connection to cardiovascular risk, having relatives with breast cancer is an instrument for the effect of hormone therapy on cardiovascular disease.

These examples highlight the core criteria for assessing proposed instruments (e.g., day of symptom onset, hour of birth). After control of measured confounders the instrument must have no association with the outcome except via the exposure of interest. In other words, if the exposure has no effect, the controlled confounders separate the instrument from the outcome.

A skeptical reader can find reason to doubt the validity of each of the above proposed instruments, which highlights the greatest challenge for instrumental variables analyses with observational data: finding a convincing instrument. Causal diagrams provide a clear summary of the hypothesized situation, enabling one to check the instrumental assumptions. When the instrument is not randomized, those assumptions (like common no-residual-confounding assumptions) are always open to question. For example, suppose we suspect that hospitals with lactation counseling tend to provide better care in other respects. Then the association of hospital counseling with child's outcome is in part not via breastfeeding, and counseling is not a valid instrument.

IV methods for confounding control are paralleled by IV methods for correcting measurement error in X. The latter methods, however, require only associational rather than causal assumptions, because they need not remove confounding (Carroll et al., 2006). For example, if Z is affected by X and is unassociated with Y given X, then Z may serve as an instrument to remove bias due to measurement error, even though Z will not be a valid instrument for confounding control.

BIAS FROM CONDITIONING ON A DESCENDANT OF THE OUTCOME

For various reasons, it may be appealing to examine relations between X and Y conditioning on a function or descendant Y^* of Y. For example, one might suspect that the outcome measurement available becomes increasingly unreliable at high values and therefore wish to exclude high-scoring respondents from the analysis. Such conditioning can produce bias, as illustrated in Figure 12–12. Although U affects Y, U is unassociated with X and so the X–Y association is unconfounded. If we examine the relation between X and Y conditional on Y^*, we open the $U \rightarrow Y \leftarrow X$ path, thus allowing a U–X association and confounding of the X–Y association by U.

Consider the effect of education on mental status, measuring the latter with the Mini-Mental Status Exam (MMSE). The MMSE ranges from 0 to 30, with a score below 24 indicating impairment (Folstein et al., 1975). Suppose we ask whether the effect of education on MMSE is the same for

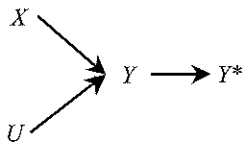


FIGURE 12–12 • Diagram illustrating effect of conditioning on an outcome variable.

respondents with MMSE ≥ 24 as for respondents with MMSE <24. If education does indeed affect MMSE score, we can apply Figure 12–12 with X = education, Y = MMSE score, and Y^* an indicator of MMSE score ≥ 24. U now represents unmeasured factors that affect MMSE score but do not confound the X–Y association. We should then expect to underestimate the association between education and MMSE score in both strata of Y^*. Among high-MMSE subjects, those with low education are more likely to have factors that raise MMSE scores, whereas among low-MMSE scorers, those with high education are less likely to have such factors. Thus, even if these unmeasured factors are not confounders to start, they will be negatively associated with education within strata of their shared effect, MMSE score.

This bias also occurs when there is an artificial boundary (ceiling or floor) on the measurement of Y and one deletes observations with these boundary values. It also can arise from deleting observations with extreme values of Y (outliers), although many might have to be deleted for the bias to become large. Such exclusions will condition the analysis on the value of Y and can thus introduce bias.

If X has no effect on Y, conditioning on Y^* will not open the $U \rightarrow Y \leftarrow X$ path in Figure 12–12. Thus, if there is no confounding of the X–Y relation and no effect of X on Y, the estimated effect of X on Y will remain unbiased after conditioning on Y^* (although precision of the estimate may be drastically reduced).

SELECTION BIAS AND MATCHING IN CASE-CONTROL STUDIES

Case-control studies are especially vulnerable to selection bias. By definition, case-control studies involve conditioning on a descendant of Y, specifically, the selection variable S. If we compute effect estimates from the case-control data as if there were no effect of Y on S—e.g., a risk difference—it will be severely biased. As discussed in Chapter 8, however, the bias produced by this conditioning will cancel out of the odds ratio from the study, provided S is associated with exposure only through Y (i.e., if Y separates S from X).

Suppose, however, that the situation is as in Figure 12–13. Here, W is not a confounder of the X–Y association if there is no conditioning, because it has no association with Y except through X. A case-control study, however, conditions on selection S. Because W is associated with exposure and affects selection, this conditioning results in a new association of W with Y via S. Thus $X \leftarrow W \rightarrow S \leftarrow Y$ is opened at S and so becomes a biasing path. To identify the effect of X on Y, this path must be blocked, for example, by conditioning on W. The same conclusion applies if Figure 12–13 is modified so that W is associated with X via a variable U with (say) $X \leftarrow U \rightarrow W$.

As discussed in Chapter 11, case-control matching on W means that W affects selection, and so Figure 12–13 can be taken to represent the situation in a case-control study matched on a nonconfounder associated with the exposure. Here, we see that the matching generated the W–S connection and thus necessitates control of W when no control would have been needed without matching. Thus, the figure illustrates a type of overmatching (Chapter 11).

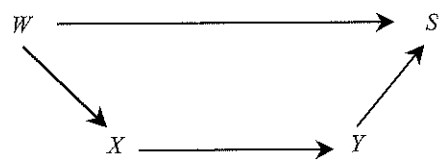


FIGURE 12–13 • Diagram showing potential selection bias in a case-control study with a cause of the exposure influencing selection into the study.

HOW ADJUSTING FOR BASELINE VALUES CAN BIAS ANALYSES OF CHANGE

Research often focuses on identifying determinants of change in a dynamic outcome, such as blood pressure, or depressive symptoms measured at start and end of follow-up, indicated by Y_1 and Y_2. Suppose we wish to estimate how much an exposure X, that was measured at baseline and preceded Y_1, affects the change in the outcome variable between times 1 and 2, measured with the *change score* $\Delta Y = Y_2 - Y_1$. An important issue is whether to adjust for (condition on) the baseline variable Y_1 when attempting to estimate the effect of X on change in the outcome. This conditioning may take the form of restriction or stratification on Y_1, or inclusion of Y_1 as a covariate in a regression of ΔY on X. Typically, X and Y_1 are associated. Indeed, this cross-sectional association may prompt researchers to investigate whether X similarly affects changes in Y.

A common rationale for baseline adjustment is that baseline "acts like a confounder" under the traditional confounder criteria: It is associated with X and likely affects the dependent variable (ΔY). This intuition can be misleading, however, in the common situation in which Y_1 and Y_2 are subject to measurement error (Glymour et al., 2005).

Suppose that our research question is whether graduating from college with honors affects changes in depressive symptoms after graduation. In a cohort of new college graduates, depressive symptoms are assessed with the Centers for Epidemiologic Studies—Depression scale at baseline (CES-D_1) and again after 5 years of follow-up (CES-D_2). The CES-D scale ranges from 0 to 60, with higher scores indicating worse depressive symptoms (Radloff, 1977). The dependent variable of interest is change in depressive symptoms, which we measure in our data using the CES-D change score $\Delta\text{CES-D} = \text{CES-D}_2 - \text{CES-D}_1$. The CES-D is a common measure of depressive symptoms, but it is known to have considerable measurement error. In other words, the CES-D score is influenced both by actual underlying depression and by randomly fluctuating events such as the weather and the interviewer's rapport with the subject. In a causal diagram, we represent this by showing arrows into CES-D score from underlying depression and from a summary "error" variable. The error is not measured directly but is defined as the difference between the CES-D score and the latent variable "Depression," so that

$$\text{CES-D} = \text{Depression} + \text{Error}$$

Bear in mind that we are actually interested in change in Depression (ΔDepression), rather than change in CES-D (ΔCES-D).

Now suppose that, at baseline, graduating with honors (X) is associated with lower CES-D scores, that is, there is an inverse association between X and Y_1, perhaps because graduating with honors improves mood, at least temporarily. These assumptions are shown in a DAG in Figure 12–14. In this figure, there is an arrow from Error_1 to ΔCES-D. This arrow represents a deterministic (inverse) relationship between ΔCES-D and Error_1, because

$$\begin{aligned}
\Delta\text{CES-D} &= \text{CES-D}_2 - \text{CES-D}_1 \\
&= \text{Depression}_2 + \text{Error}_2 - (\text{Depression}_1 + \text{Error}_1) \\
&= \text{Depression}_2 - \text{Depression}_1 + \text{Error}_2 - \text{Error}_1 \\
&= \Delta\text{Depression} + \text{Error}_2 - \text{Error}_1
\end{aligned}$$

Another assumption in Figure 12–14 is that Error_1 and Error_2 are independent. Positive association of these errors reduces the magnitude of the bias we will discuss, but this bias is not eliminated unless the errors are identical (and so cancel out). Under the conditions of Figure 12–14, honors degree has no effect on change in depression. Correspondingly, honors degree and ΔCES-D are unconditionally independent under the null hypothesis because the only path in the diagram connecting honors degree and change score is blocked by the collider CES-D_1. Thus, when not adjusting for CES-D_1, we obtain an unbiased estimate of the overall (i.e., total) effect of honors degree on change in depression.

Conditional on CES-D_1, however, honors degree and ΔCES-D are associated, because conditioning on CES-D_1 unblocks the path. This result can be explained as follows. Anyone with a high

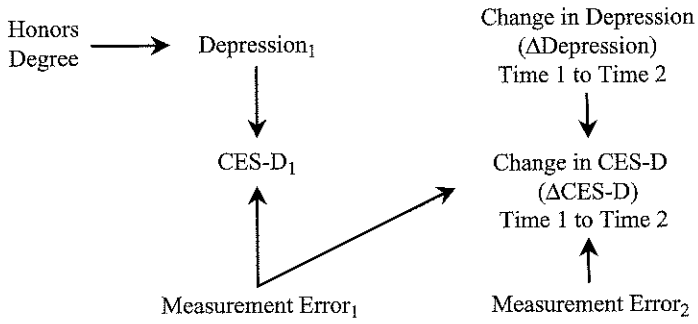


FIGURE 12–14 • An example in which baseline adjustment biases analyses of change.

CES-D_1 has either high Depression$_1$, or large positive measurement Error$_1$, or both. A nondepressed person with high CES-D_1 must have a positive Error$_1$, and a depressed person with low CES-D_1 must have a negative Error$_1$. Thus, *within levels of CES-D_1*, Depression$_1$ and Error$_1$ are inversely associated, and honors degree and Error$_1$ are therefore positively associated. Because Error$_1$ contributes negatively to ΔCES-D, ΔCES-D and Error$_1$ are negatively associated (this is an example of regression to the mean). Hence, conditional on CES-D_1, honors degree and ΔCES-D are inversely associated. Therefore the baseline-adjusted honors-degree association is inverse, making it appear that honors degrees predict declines in depression, even when receiving an honors degree does not affect changes in depression. The bias in the association is proportional to the error in the CES-D scores and the strength of the honors degree–Depression$_1$ association (Yanez et al., 1998).

To summarize the example, the unadjusted association of honors degree and ΔCES-D correctly reflects the effect of honors degree on change in actual depression (ΔDepression), whereas adjustment for baseline CES-D_1 biases the association downward in the direction of the cross-sectional association between CES-D_1 and honors degree.

Now consider baseline adjustment in a slightly different research question, in which we wish to estimate how much a baseline exposure X affects the change score ΔY over the follow-up period. In this case, we ignore measurement error and focus on identifying determinants of changes in CES-D score. We return to our example of the effect of graduating college with honors (X) and CES-D change scores. Figure 12–15 provides one model of the situation. There are confounding paths from X to ΔY via U and Y_1, which we can block by conditioning on baseline score Y_1. Thus, if U is unmeasured, it appears from this model that we ought to control for baseline score. This model for ΔY is fatally oversimplified, however, because there will always be other unmeasured factors that affect CES-D_1 (such as genetic risk factors), which influence both CES-D_1 and the rate of change.

If we expand Figure 12–15 to include such a factor, B, and B is unassociated with X, we obtain Figure 12–16. B does not appear to be a confounder, but it is a collider on a path between X and ΔY. Conditioning on baseline Y_1 opens the confounding path $X \leftarrow U \rightarrow Y_1 \leftarrow B \rightarrow \Delta Y$. Thus, adjusting for baseline is insufficient to eliminate bias in assessing the relation of X to the change score ΔY; after such adjustment, to ensure unbiasedness we would have to adjust for all shared causes of earlier and later scores—a daunting task to say the least.

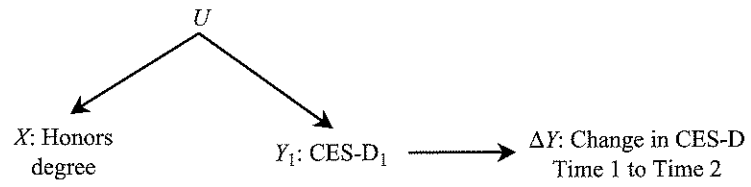


FIGURE 12–15 • An example in which baseline adjustment eliminates bias in analyses of change.

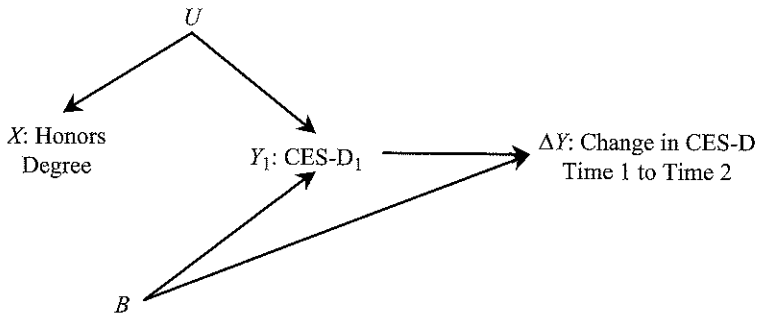


FIGURE 12–16 • An example in which baseline adjustment does not eliminate bias in analyses of change.

CAVEATS AND EXTENSIONS

For many if not most epidemiologic research questions, the data available or feasible to obtain are simply not adequate to identify the "right" answer. Given this reality, every method will be biased under conditions that we cannot rule out. Thus it is rarely enough to know that a particular approach is biased; we will want to estimate how biased it may be, especially relative to alternatives. Graphs alone, however, are silent with respect to the magnitude of likely biases.

Graphs can be augmented with signs on each arrow to indicate directions of effects, as is common with structural equation models (Duncan, 1975). Under the assumption of monotonic effects (causal monotonicity), the directions of associations and biases can be computed from these signs (VanderWeele and Robins, 2008b). More detail can be added to causal diagrams to indicate mechanism types under the sufficient-component cause model described in Chapters 2 and 5 (VanderWeele and Robins, 2007b). Nonetheless, causal diagrams as currently developed do not convey information about important aspects of causal relations and biases, such as the magnitudes or functional forms of the relations (e.g., effect size or effect-measure modification).

The g-computation algorithm or g-formula (Robins, 1986, 1987, 1997) can be used to quantify the size of effects and predict the consequences of interventions under an assumed causal structure (Pearl and Robins, 1995). The formula simplifies to ordinary standardization (Chapters 3 and 4) when the intervention variable is a fixed baseline characteristic (as opposed to a time-varying exposure) (Robins, 1987; Pearl, 1995, 2000). Applying the g-computation algorithm is often impractical, for the same reason that stratified analysis methods (Chapter 15) can be impractical with many covariates: There will rarely be enough observations for every combination of covariate levels. This problem is addressed by assuming parametric models for some or all of the relations among the covariates. This approach has a long history in the structural-equations literature (Duncan, 1975; Pearl, 2000). In the structural-equations model for a graph, each variable is represented as a function of its parents and a random error that represents effects of forces not shown in the graph. More advanced approaches such as *g-estimation* and *marginal structural modeling* estimate parameters using structural models only for the effect of interest, and use associational models to control confounding; see Chapter 21 for further description and references.

Modeling approaches allow comparison of bias magnitudes under various scenarios about causal relations. For example, assuming logistic models, Greenland (2003a) compared the bias left by failing to adjust for a variable that is both a collider and confounder, versus the bias introduced by adjusting for it, and found evidence that when the causal structure is unknown, adjustment is more likely to result in less bias than no adjustment. In many (if not most) situations, however, there will be insufficient information to identify the best strategy. In these situations, analyses under different assumptions (involving different diagrams or different structural equations under the same diagram) will be essential to get a sense of reasonable possibilities. For example, we can perform analyses in which a variable is not controlled because it is assumed to be an intermediate, and perform others in which it is treated as a confounder (Greenland and Neutra, 1980); and, in the latter case, we can vary the equations that relate the variable to exposure and disease. Such *sensitivity analyses* are described in Chapter 19.

A related difficulty is deciding whether a dubious causal relationship ought to be represented in a DAG. In typical epidemiologic examples, very weak relationships are unlikely to introduce large biases. Thus one heuristic for drawing DAGs would take the absence of an arrow between two variables to indicate that the direct causal relation between the variables is negligible. While such a heuristic can provide a useful perspective, we recommend starting with a DAG that shows all the arrows that cannot be ruled out based on available data or logic (like time order), to determine what assumptions are required in order to identify with certainty the causal parameter of interest with the available data.

CONCLUSION

Causal diagrams show how causal relations translate into associations. They provide a simple, flexible tool for understanding and discovering many problems, all using just a few basic rules. Rather than considering each type of bias as a new problem and struggling for the "right" answer, diagrams provide a unified framework for evaluating design and analysis strategies for any causal question under any set of causal assumptions. Nonetheless, drawing a diagram that adequately describes contextually plausible assumptions can be a challenge. To the extent that using diagrams forces greater clarity about assumptions, accepting the challenge can be beneficial. Although we may never know the "true" diagram, to the extent that we can specify the diagram, we will be able to identify key sources of bias and uncertainty in our observations and inferences.

Greenland & Brumback. An overview of relations among causal modeling methods. International Journal of Epidemiology 2002;31:1030–1037

An overview of relations among causal modelling methods

Sander Greenland[a] **and Babette Brumback**[b]

This paper provides a brief overview to four major types of causal models for health-sciences research: Graphical models (causal diagrams), potential-outcome (counterfactual) models, sufficient-component cause models, and structural-equations models. The paper focuses on the logical connections among the different types of models and on the different strengths of each approach. Graphical models can illustrate qualitative population assumptions and sources of bias not easily seen with other approaches; sufficient-component cause models can illustrate specific hypotheses about mechanisms of action; and potential-outcome and structural-equations models provide a basis for quantitative analysis of effects. The different approaches provide complementary perspectives, and can be employed together to improve causal interpretations of conventional statistical results.

Keywords Bias, causal diagrams, causality, confounding, data analysis, direct effects, epidemiological methods, graphical models, inference, instrumental variables, risk analysis, sufficient-component cause models, structural equations

Accepted 28 March 2002

Following a long history of informal use in path analysis, causal diagrams (graphical causal models) saw an explosion of theoretical development during the 1990s,[1–3] including elaboration of connections to other methods for causal modelling. The latter connections are especially valuable for those familiar with some but not all methods, as certain background assumptions and sources of bias are more easily seen with certain models, whereas practical statistical procedures may be more easily derived under other models. We provide here a brief overview of graphical causal models,[1–6] the sufficient-component cause (SCC) models of Rothman,[7,8 Ch. 2] the potential-outcome (counterfactual) models now popular in statistics, health, and social sciences,[9–15] and the structural-equations models long established in social sciences.[11–14] We focus on special insights facilitated by each approach, translations among the approaches, and the level of detail specified by each approach.

Graphical models

The following is a brief summary of terms and concepts of causal graph theory; see Greenland *et al.*[4] and Robins[5] for more detailed explanations. Figure 1 provides the graphs used for illustration below. An *arc* or *edge* is any line segment (with or without arrowheads) connecting two variables. If there is an arrow from a variable X to another variable Y in a graph, X is called a *parent* of Y and Y is called a *child* of X. If a variable has an arrow into it (i.e. it has a parent in the graph) it is called *endogenous*; otherwise it is *exogenous*.

A *path* between two variables X and Y is a sequence of arcs connecting X and Y. A *back-door path* from X to Y is a path whose

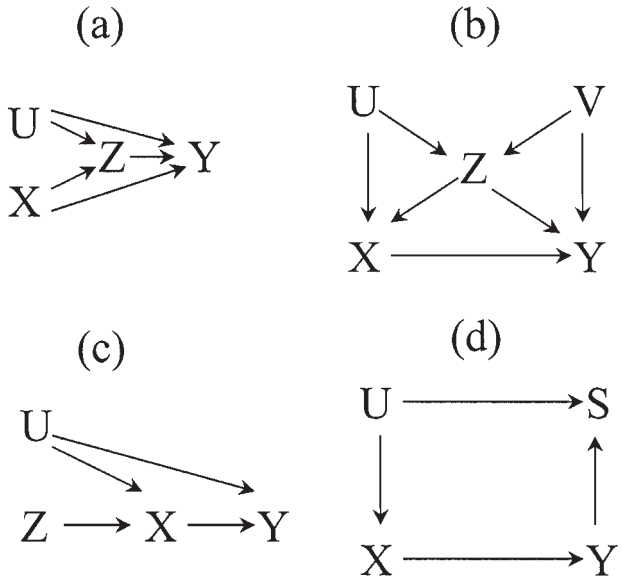


Figure 1 Four causal diagrams used in examples. In all four, X and Y are the exposure and outcome variables under study

[a] Department of Epidemiology, UCLA School of Public Health, Department of Statistics, UCLA College of Letters and Science, 22333 Swenson Drive, Topanga, CA 90290–3434, USA. E-mail: lesdomes@ucla.edu

[b] Department of Biostatistics, University of Washington, School of Public Health and Community Medicine, Seattle, WA 98195, USA.

first arc is an arrow pointing to X; there is no back-door path from X to Y in Figure 1a, whereas in Figure 1c the path X-U-Y is a back-door path from X to Y. A *blocked path* (or closed path) between X and Y is a path that passes from a parent to child and then back to another parent, i.e. there is a parent-child-parent sequence in the path; a path that has no such sequence is an *open path*.[2–4] In Figure 1b, the paths X-U-Z-Y and X-Z-V-Y are open, but the path X-U-Z-V-Y is not because U-Z-V is a parent-child-parent sequence.

A *directed path* is a sequence of arrows such that the child in the sequence is the parent in the next step. If there is a directed path from X to Y, X is called an *ancestor* of Y and Y is called a *descendant* of X. A graph is *directed* if all the arcs in it are arrows; a graph is *acyclic* if no directed path forms a closed loop (equivalently, if no variable is both an ancestor and descendant of another). A graph that is both directed and acyclic is a DAG; each graph in Figure 1 is a DAG.

A graph is *causal* if every arrow represents the presence of an effect of the parent (causal) variable on the child (affected) variable. In a causal graph, a directed path represents a causal pathway, and an X-to-Y arrow represents a direct effect of X on Y within the graph (an effect not mediated through any other variable in the graph). Each graph in Figure 1 summarizes causal relations within a population of individuals, and each variable represents the states or events among individuals in that population. For example, if X is a treatment variable, then the value of X for an individual is the level of treatment received by the individual. Absence of a directed path from X to Y in the graph corresponds to the causal null hypothesis that no alteration of the distribution of X could change the distribution of Y.

The 'population' might contain just one individual, in which case the graph is a model for effects on that individual. Furthermore, the 'individuals' in the population need not be persons; they may be administrative entities, natural groupings, or any other unit of interest. For example, in a study of the effect of state helmet laws on riding-accident mortality Y among motorcyclists, the individual units could be states, X could be helmet-law status, and Z could be helmet-law enforcement levels. One could also draw an accident-level graph in which X could be helmet-law status in the accident's locale, Z indicates whether the motorcyclist was wearing a helmet, and Y indicates whether the motorcyclist was killed.

An important result from graph theory is that if one stratifies (conditions) on a descendant Z of two variables U and X, and U and X are independent in the total population, then we should expect U and X to be associated within at least one stratum of Z (exceptions to this rule involve somewhat contrived cancellations of effects).[2,3 p. 17,4] To illustrate a consequence of this result, suppose in Figure 1a X represents a 6-month weight loss regimen that is randomly assigned within a cohort of cardiovascular patients, with X = 1 for regimen assigned and X = 0 for not assigned; Z represents a set of clinical CHD risk factors (serum lipids, blood pressure) measured at regimen completion; Y represents death within the year following completion; and U represents a set of unmeasured genes that affect death risk both directly and through the clinical factors Z. Although U affects Y, it is not a confounder of the X-Y association because it is independent of X.

A common approach to analysing effects of weight on health is to adjust for serum lipids and blood pressure. If weight affects serum lipids and blood pressure, such adjustment cannot be justified as confounding control because it removes that part of the weight effect mediated through serum lipids and blood pressure.[8 Ch. 4] It is often thought that such an analysis estimates the direct effect of weight, or of a weight-loss regimen. Using counterfactual models, however, it has been shown that this rationale fails if the intermediates were also affected by uncontrolled risk factors; it fails even if the treatment X is independent of the uncontrolled factors, so that there is no confounding of the crude X-Y association, as in Figure 1a.[16] Graph theory shows this fact more simply: Because Z is a child of both U and X, one should expect U and X to be associated within at least one stratum of Z; consequently, within strata of Z, U becomes a confounder, even though it was not one to begin with.[6] In general, one should expect control of an intermediate Z to generate confounding when Z and Y share causes other than X, as in Figure 1a; in such cases the association of Z with Y is confounded, and so the estimated indirect effect of X on Y being 'removed' by Z-adjustment is confounded.[17]

Figure 1b gives another example, which has a counter-intuitive quality and had to wait for graph theory for discovery. In this graph we ask, 'is it sufficient to stratify only on Z in order to unbiasedly estimate the effect of X on Y?' A common intuitive answer is 'Yes,' because physically preventing individual variation in Z would block the effects of U on Y and V on X and thus eliminate confounding by U and V (as well as confounding by Z). But in an observational study U and V would ordinarily be associated within some strata of Z, because they both affect Z. Within those strata, U would be associated with Y (through V) as well as with X, and V would be associated with X (through U) as well as with Y; consequently, both U and V would be confounders and one or the other would have to be controlled to remove the confounding.[2,4]

One can recognize the insufficiency of controlling Z alone given Figure 1b in more traditional ways: The association of Z with Y given X is confounded by V; because adjustment for Z alone depends on this confounded association, one might conclude correctly that such adjustment could mislead, and that adjustment for V as well as Z would remedy the problem. But graphical theory also shows that adjustment for U rather than V would also suffice: because the V-X association produced by Z-adjustment is mediated entirely through U, U-adjustment eliminates confounding by V within Z strata.

The preceding examples illustrate how causal graphs supply simple visual methods to check for confounding and for sufficiency of confounder adjustment. Some basic results are: (1) an open back-door path from X to Y can produce an association between X and Y, even if X has no effect on Y, and so can produce confounding; (2) adjustment for certain variables can produce open back-door paths, and so produce confounding; (3) the X-Y association will be unconfounded if the only open paths from X to Y are directed paths from X to Y (so that the only sources of X-Y association are effects of X on Y).[2,3 Ch. 3,4] These results lead to general criteria for identifying sets of variables sufficient for control of confounding given a graph.[2,3 Ch. 3,4]

Potential-outcome (counterfactual) models

Graphs display broad qualitative assumptions about causal directions and independencies in a population. Although it is

surprising how much can be deduced from such assumptions,[1–6] the deductions are only qualitative (e.g. confounding present or absent in a particular stratification). Usually, however, more precise deductions are needed, and such deductions require a quantitative model that specifies in detail what would happen under alternative possible patterns of intervention or exposure. One class of quantitative models originating with Neyman and Fisher in the early 20th century[9] are the *counterfactual* or *potential-outcome* models.[8 Ch. 4,9–12,15] These models formalize notions of cause and effect found in much of philosophy and epidemiology,[15,18,19] such as this passage from MacMahon and Pugh: '… an association may be classed presumptively as causal when it is believed that, had the cause [exposure] been altered, the effect [outcome] would have changed'.[19 p. 12] A key feature of this description is its *counterfactual* element: It refers to what would have happened if, contrary to fact, the exposure had been something other than what it actually was.

Suppose we have a population of individual units under study (e.g. mice, people, counties) indexed by i = 1,…,N, a treatment or exposure variable X with J + 1 levels (or actions) x_0, x_1,…,x_J, and an outcome variable of interest Y (such as an indicator for 'death by age 70'). The standard potential-outcome model for a non-contagious outcome assumes that:

(a) Each individual could have received any one of the treatment levels; this rules out (for example) having men in the population for an analysis of hysterectomy effects.

(b) For each individual i and treatment level x_j, at the time of treatment assignment the outcome that individual i would have if the individual gets treatment level x_j exists, even if the individual does not in fact get x_j; this value is called the *potential outcome* of individual i under treatment x_j.

The variable Y represents a generic variable for the actual outcome under the treatment actually given. Assumption (b) can be recast as stating that, for each individual i and each exposure level x_j, one can also define a potential-outcome (potential-response) variable Y_{ij} representing the outcome of the individual under that exposure. Thus, if Y is an indicator for 'death by age 70', Y_{ij} will be an indicator for 'death by age 70 of individual i if that individual is given treatment x_j'.

If individual i gets treatment x_j, Y_{ij} will equal the indicator for the actual outcome of individual i; but otherwise it may be quite different from that actual outcome. Such a difference is taken as the effect of actual treatment relative to treatment x_j. More generally, the choice of treatment is said to have had no effect on Y for individual i if $Y_{ij} = Y_{ik}$ for every possible pair of treatment levels x_j and x_k; otherwise, if $Y_{ij} \neq Y_{ik}$ for some pair of treatment levels x_j and x_k, treatment choice could have had an effect, or could have caused a change in the actual outcome of individual i (from Y_{ik} to Y_{ij}). Treatment choice is said to have had no effect on the population if it had no effect on any individual in the population.

In addition to (a) and (b), most applications also assume that the potential outcomes of each individual are independent of the treatments and outcomes of other individuals. This assumption is not always correct (e.g. in vaccine trials), but the model can be generalized to allow for violations.[20] A controversial aspect of assumption (b) is that it requires each potential outcome Y_{ij} remain a meaningful quantity even when individual i does not get treatment x_j. Even if one accepts this idea, the only Y_{ij} that can be observed for individual i is the one corresponding to the treatment actually received by that individual; the remaining Y_{ij} can only be estimated, not observed. People routinely estimate such quantities in day-to-day life (e.g. 'if I had only bought Microsoft stock when it was first issued, my net worth would be millions of dollars'). The problems attributed to modelling such quantities (such as the need for untestable assumptions in estimating causal effects) are in reality unpleasant intrinsic problems of causal inference that are obscured by other approaches; we believe it is a virtue of the counterfactual approach that it makes such problems explicit.[15,21]

Potential-outcome models are not inherently deterministic (as is often mistakenly claimed), because the potential outcomes (the Y_{ij}) may be parameters of probability distributions (e.g. expected age at death) rather than directly observable events (e.g. actual age at death).[21] This flexibility can be seen in the probabilistic notations based on the 'set' and 'do' operators in Pearl,[2,3] which can be used to represent effects in a single individual instead of a population. Furthermore, potential-outcome models are not limited to person-level analyses; for example, the 'individuals' in the model may be social units or aggregates (although the associations observed among these aggregates may be confounded by person-level effects).[22]

One way of summarizing the scope of potential-outcome models is that they represent the limit of what one could learn about individual causes and effects from perfect crossover trials. For example, if X and Y represent completely reversible exposure and outcome variables (e.g. as might occur with X indicating a nasal irritant and Y a sneezing probability), we could estimate an individual's Y_{i1} and Y_{i0} (sneezing probabilities when irritant present and absent) through a series of trials on the individual that alternated X = 1 with X = 0, provided there were no carry-over effects or temporal variations in the sneezing responses (as represented by the potential outcomes). When such trials cannot be performed, as is usual in human studies, we could still estimate the population distribution of Y_{ij} (the outcome when $X = x_j$) by treating a random sample from that population with x_j. By repeating such experiments for various treatment levels (or by randomizing a random sample to different treatment levels) we can estimate how the population outcome distribution would vary with treatment distribution.[9,11,15,21]

A practical aspect of potential-outcome models arising from assumption (b) is that any potential outcome Y_{ij} not observed (whether because treatment x_j was not given to i, or because of censoring) can be viewed as a quantity to be estimated or imputed from observed covariates and outcomes.[9,23] This idea underlies most methods of model-based standardization of effect estimates,[8, Ch. 21] and leads to numerous methods for confounder control based on the relation of actual treatment X to the potential outcomes predicted from various models.[23] Some effect measures do not require that assumptions (a) and (b) apply to all individuals in the study. For example, if unexposed (X = 0) individuals are used only to estimate the distribution of the Y_{i0} among the exposed (X = 1), as in many occupational studies, we need not assume that the unexposed could have been exposed or that Y_{il} is meaningful for the unexposed.[15]

Multifactorial causation and the sufficient-component cause model

The graphical and potential-outcome models can be used to portray the presence, though not the mechanics, of causal interactions. Consider for example the synergism between phenylketonuria (PKU)(X = 1) and significant phenylalanine consumption (SPC)(Z = 1) in inducing brain damage (Y = 1): In some people, these two factors together are necessary and sufficient to produce damage.[19] This synergism can be represented in basic graphs[1–6] by including a variable XZ that indicates their joint presence (XZ = 1 if X = Z = 1 are both present, XZ = 0 otherwise), then drawing an arrow from XZ to Y. To represent the synergism in a potential-outcome model, we may define four damage indicators Y_{ixz} for each individual i; the subscript x is 1 with PKU present, 0 with PKU absent, while z is 1 with SPC present, 0 with SPC absent. The synergism then corresponds to $Y_{i11} = 1$ but $Y_{i10} = Y_{i01} = Y_{i00} = 0$.[8 Ch. 18]

Because potential outcomes are quantities specific to individuals in the modelled population, they provide more detail than arrows in graphs. For example, the individuals affected by X and those affected by Z may be one and the same, or may not overlap at all. The graph $X \rightarrow Y \leftarrow Z$ would hold if the population were composed entirely of individuals with $Y_{i11} = Y_{i10} = Y_{i01} = 1$ and $Y_{i00} = 0$; in this case, if everyone had their actual X and Z equal to 0, everyone would be affected by changes in X or changes in Z. But the same graph would hold if the population was half individuals with $Y_{i11} = Y_{i10} = 1$, $Y_{i01} = Y_{i00} = 0$ (individuals affected only by changes in X) and half individuals with $Y_{i11} = Y_{i01} = 1$, $Y_{i01} = Y_{i00} = 0$ (individuals affected only by changes in Z). Like the graph, the potential-outcome models can be extended to include effects of X on Z as well as the effects on Y; doing so reveals many distinctions not captured by simply adding an arrow from X to Z in the graph.[16] Such examples show that potential-outcome models are logically finer (distinguish more situations) than graphical models of the same variables; this fineness leads to greater notational complexity.

Consideration of causal mechanisms leads to models that are logically finer than either potential-outcome models or graphs. Best known among epidemiologists is Rothman's sufficient-component cause (SCC) model.[7,8 Ch. 2] In this model, two factors are said to be causal *cofactors*, and have a (potential for) synergism, if they are components of the same causal mechanism; the presence of both cofactors is necessary for the mechanism to operate and so produce the outcome under study. This definition refers to mechanisms; thus, the basic units of analysis are the mechanisms that determine the potential outcomes of individuals, rather than individuals. Many different sets of mechanisms will lead to the same pattern of potential outcomes for an individual; hence, many different SCC models will lead to the same potential-outcome model.[8 Ch. 18,25] As with potential-outcome models, however, SCC models are not inherently deterministic, because the component causes may be random events[24] and because the outcome affected by the completion of a sufficient cause may be a probability parameter rather than an observable event.

The SCC model employs a pie-chart representation of causal mechanisms, in which each slice represents a necessary component of the mechanism.[7] To illustrate, suppose we are considering mechanisms for angiosarcoma induction in just one individual i. Figure 2 gives an illustration of two distinct SCC models for the disease-causing mechanisms within this individual. The U in the figure represent sets of unmeasured cofactors that would be present regardless of this individual's X or Z status. Model (a) posits that there are two mechanisms that can lead to disease in this individual, neither of which involve synergism of X levels and Z levels, while model (b) posits three such mechanisms, all of which show synergism of X levels and Z levels. Nonetheless, under both models this individual will get the disease unless X = 0 and Z = 0; in other words, under either SCC model the individual's potential outcomes would be $Y_{i11} = Y_{i10} = Y_{i01} = 1$ and $Y_{i00} = 0$. Thus, even if we could conduct a perfect crossover trial on the individual and so observe the individual's outcome under all four X-Z combinations, we would still be unable to determine which SCC model was correct.

As this example and more realistic ones[26–28] show, there are severe limits to the detail about causal mechanisms that can be distinguished using only ordinary ('black-box') randomized trials and epidemiological studies of exposure-disease relations.[26–29] Although discrimination among mechanisms can be important,[28,29] it will usually require direct observations of intermediate steps or of biomarkers for hypothesized mechanisms.

Structural-equations models

Informal use of graphs initially developed as an intuitive aid for structural-equations modelling (SEM), in which a web or network of causation is modelled by a system of equations and independence assumptions.[3 Ch. 1,13] Each equation shows how an individual response (outcome, affected, dependent) variable changes as its direct (parent) causal variables change. Again, the 'individual' may be any unit of interest, such as a person or aggregate. In the system, a variable may appear in no more than one equation as a response variable, but may appear in any other equation as a causal variable. A variable appearing as a response in the system is said to be endogenous (within the system); otherwise it is exogenous.

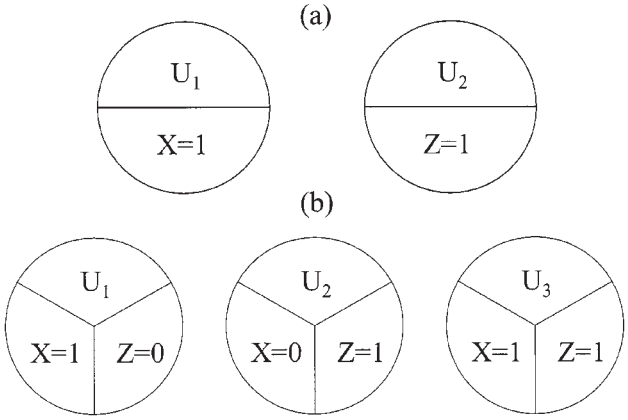


Figure 2 Two distinct sufficient-component cause (SCC) models for the set of mechanisms within an individual; each leads to the same potential-outcome model

A causal graph is a qualitative schematic for a class of structural-equations models. For example, Figure 1a is a schematic for the linear system

$$Z = \alpha_Z + \beta_{UZ}u + \beta_{XZ}x \quad (1a)$$

$$Y = \alpha_Y + \beta_{UY}u + \beta_{XY}x + \beta_{ZY}z \quad (1b)$$

in which u, x, z are specific values of U, X, Z, α_Z and α_Y are unmeasured individual-specific (random) disturbances of Z and Y, and α_Z, α_Y, U and X are assumed to be jointly independent of one another in the study population. Figure 1a is also a schematic for the very different system

$$Z = \alpha_Z + \beta_{UZ}u + \beta_{XZ}x + \beta_{UXZ}ux \quad (2a)$$

$$\ln(Y) = \alpha_Y + \beta_{UY}u + \beta_{XY}x + \beta_{ZY}z \quad (2b)$$

with α_Z, α_Y, U and X again assumed jointly independent. System 2 differs from system 1 in that a product term is added to the Z equation, and the Y equation is log-linear instead of linear. Nonetheless, both systems share the properties indicated by Figure 1a: U and X are the two exogenous variables (indicated by their lack of parents); U and X directly affect the two endogenous variables Z and Y, and Z directly affects Y (indicated by the arrows from U and X to Z and Y, and from Z to Y); and the exogenous variables and random disturbances are jointly independent of one another (indicated by the *absence* of connections among the variables other than the arrows just described).

Structural equations can be viewed as formulas for computing potential outcomes under various actions.[2,3 Ch. 1] For example, if X represents a treatment regimen, equation 1a asserts that the *potential* value of Z for any individual in the study population will trace out a linear function of X as the individual's values of X changes but U remains constant: an individual's Z will change by β_{XZ} units if we increase X by one unit while U remains constant (because α_Z is constant for the individual). Equation 1a also asserts that Z will not vary with Y if U and X remain constant. It is such within-individual causal interpretations that distinguish structural equations from ordinary regression equations (which represent only *associations* of actual outcomes with actual values of the covariates as one moves across individuals).[3 Ch. 5,8 Ch. 20] Structural-equations models (complete systems such as 1 and 2 above) combine potential-outcome models for the endogenous variables with independence assumptions about exogenous variables.

Structural equations with unknown parameters go beyond graphs by specifying the functional form of effects, but do not provide the exact values of effects; thus, they are algebraic but not fully quantified representations of causal relations. The equations can also be given a general non-parametric form that does not impose structure beyond that in the corresponding graph, and so is logically equivalent to that graph.[2,3 Ch. 1] For example, Figure 1a corresponds to

$$Z = f_Z(u,x,\alpha_Z) \quad (3a)$$

$$Y = f_Y(u,x,z,\alpha_Y) \quad (3b)$$

with α_Z, α_Y, U, X assumed jointly independent. The functions f_Z and f_Y are left unspecified, although statistical analysis will usually require some restrictions on f_Z and f_Y, such as smoothness of dose-response and additivity of effects on a particular scale. Equations 3a and 3b may be interpreted as alternative notations for the potential outcomes corresponding to Z and Y. For example, $f_Y(u,x,z,\alpha_Y)$ may be interpreted as the potential outcome y_{iuxz}, with the individual identifier i replaced by a 'random' source of inter-individual variation α_Y. Thus, non-parametric structural equation models provide a bridge between graphical and potential-outcome models.[2]

As with potential-outcome models, structural-equations models extend beyond deterministic outcomes, although the details of such extensions are rather technical. In the systems above, Z and Y may represent parameters of individual outcome distributions, rather than the observable outcome events. For example, Z and Y may represent expected values; the structural equations are then mixed models with random intercepts α_Z and α_Y. A common equivalent practice adds mean-zero 'random errors' ε_Z and ε_Y to the Z and Y equations; Z and Y then remain observable outcomes, but the random errors are not separable from α_Z and α_Y without repeated observations of all variables on each individual. It is also possible to treat the β coefficients as random.

Graphical versus algebraic representations

As an illustration of the differing insights obtained from graphical and algebraic representations of causation, Figure 1c diagrams a situation in which Z is an *instrumental variable* for estimating X effects: Z affects X, but is unassociated with the confounder U and is unassociated with Y except through X.[3 Sec. 7.4.5] Such variables occur in randomized trials, in which Z is the assigned (intended) treatment. Many patients do not fully comply, and instead take (or receive) a different level of treatment, X; this received-treatment variable is affected by unmeasured factors U that are also risk factors (or close correlates of risk factors) for the outcome under study. Standard intent-to-treat analyses examine only the Z association with Y and so are estimating the effect of treatment *assignment*, rather than a physiologic effect of received treatment X. Can we also estimate the latter effect? The answer is yes, provided we can make further (not necessarily unique) quantitative assumptions. The graph makes clear that we should not expect the crude X-Y association to equal the X-Y effect, because of confounding by U. The graph also shows, however, that there is no confounding of the Z effects on X or Y (as would be expected if Z was randomized); hence the crude Z-X and Z-Y associations will equal the Z-X and Z-Y effects. These facts alone can allow one to put bounds on the X-Y effect,[3, Sec. 8.4] although one or both bounds may be beyond any plausible range for the effect.[30]

Suppose we go beyond Figure 1c by assuming the linear structural relations

$$X = \alpha_X + \beta_{UX}u + \beta_{ZX}z \quad (4a)$$

$$Y = \alpha_Y + \beta_{UY}u + \beta_{XY}x \quad (4b)$$

with α_X, α_Y, U, Z jointly independent. As noted long ago by economists,[31] this model would allow us to unbiasedly estimate β_{XY} from the simple regressions of X on Z and Y on Z. First, because α_X, U, and Z are independent, there would be no

confounding of the simple β_{ZX} estimate obtained from regressing X on Z alone. Second, we can substitute 4a into 4b to get

$$\begin{aligned} Y &= \alpha_Y + \beta_{UY}u + \beta_{XY}(\alpha_X + \beta_{UX}u + \beta_{ZX}z) \\ &= (\alpha_Y + \alpha_X\beta_{XY}) + (\beta_{UY} + \beta_{UX}\beta_{XY})u + \beta_{ZX}\beta_{XY}z \\ &= \delta_Y + \delta_{UY}u + \delta_{ZY}z, \end{aligned} \quad (5)$$

where $\delta_Y \equiv \alpha_Y + \alpha_X\beta_{XY}$, $\delta_{UY} \equiv \beta_{UY} + \beta_{UX}\beta_{XY}$, and $\delta_{ZY} \equiv \beta_{ZX}\beta_{XY}$. Because of the independence assumptions, there would be no confounding of the simple δ_{ZY} estimate obtained from regressing Y on Z alone; therefore, the ratio of the simple δ_{ZY} and β_{ZX} estimates will consistently estimate

$$\delta_{ZY}/\beta_{ZX} = \beta_{ZX}\beta_{XY}/\beta_{ZX} = \beta_{XY}, \quad (6)$$

which is just the effect of X on Y in system 4. This ratio is an example of an *instrumental-variables estimate* of effect;[3 Sec. 3.5,30–32] one can also easily derive this estimate for binary X, Y, and Z by specifying potential outcomes directly.[30,32] In either approach, it is important to remember that equation (6) is based on the linearity assumptions seen in system 4, as well as on the directional assumptions in Figure 1c.

For instrumental variables, algebraic modelling led to discovery of assumptions (plausible in some settings) that are sufficient for estimating the effects of interest from the given data. Nonetheless, by focussing our attention on basic qualitative relations, graphs can help identify fallacies in causal inference. Some examples were given in our discussion of Figures 1a and 1b; as another example, some epidemiologists still believe (mistakenly) that an extraneous factor cannot induce selection bias unless it is a risk factor for disease. Consider a case-control study of magnetic-field exposure X and childhood leukaemia Y, with U representing socioeconomic factors and S selection. It has been argued (though disputed) that socioeconomic factors have little or no effect on childhood-leukaemia risk (as opposed to diagnosis or mortality); there is evidence, however, that those factors are associated with magnetic fields and with participation.[33,34] Because of the case-control design, leukaemia is also strongly associated with selection. Figure 1d summarizes this background. It shows that S is a descendant of both U and Y; hence, because the study data must be limited to those selected (the S = 1 stratum), we should expect U and Y to be associated in those data even if U has no effect on Y. Consequently, U would have to be controlled in order to ensure an unbiased estimate of the X-Y effect. Such control could not be accomplished if U were unmeasured or poorly measured. (Note however that if X itself affected selection, there would be no way to remove the resulting selection bias through control of a covariate.)

Discussion

What population should be modelled?

When using models in data analysis, it is essential to consider the distribution of exposure and confounders in the combined study population of all treatment (or exposure) groups that are under comparison, not in some specific target group of policy interest. Furthermore, in a population-based case-control study this population will be the source population of cases and controls, not just the subjects selected into the study.[8 Ch. 7] For example, a study of vinyl chloride effects may have as its target only workers actually exposed; nonetheless, to evaluate confounding one needs to include the unexposed group (as well as exposed group) in the population being modelled. Even though the target comprises only those exposed (X = 1), an unexposed population is needed for comparison, and whether or not an extraneous factor (indicated by U = 1) is a confounder depends on whether or not the factor is associated with the exposure in the entire (exposed plus unexposed) study population.[8 Ch. 8] This pivotal U-X association can only be represented in a model for relations in the entire population (among the exposed, X is always 1 and so cannot be associated with anything).

What is a causal variable?

A controversial issue in all theories of causation is whether a variable must be manipulable to be considered potentially causal. For modelling purposes, some authors would restrict the label 'causal' to variables that represent interventions or actions,[35] or at most allow only mutable variables (those susceptible to intervention) as potentially causal.[3] Such restrictions exclude as causal those variables regarded as immutable or defining characteristics of individuals, such as the birthdate and genetic sex of persons, but allow as causal such variables as perceived age and sex. Even when technology advances enough to allow alteration of a previously immutable characteristic (e.g. through genetic engineering), some authors would only label as 'causal' the intervention that alters the characteristic.[35]

In potential-outcome models, the levels of immutable variables may be represented by strata (i.e. subpopulations) but not by interventions (i.e. not by x_j). In graphical and structural models, immutable variables may appear as exogenous variables, and so are not distinguished from manipulable exogenous variables. This practice is more in accord with ordinary usage of 'causal'; it is useful because all the graphical rules for assessing bias sources and covariate control continue to apply when including immutable variables.[3 Ch. 3] The distinction between mutable and immutable variables remains important, however, as it leads to refinement of vague concepts like 'race' into multiple variables that have very different implications for health outcomes (e.g. mutable variables such as ethnic identification, and immutable variables such as ancestry).[36]

A more severe problem arises when variables that are not interventions are treated as interventions for planning purposes.[36] A common example is estimation of 'the effect' of eliminating a disease (e.g. lung cancer) on life expectancy. This effect is quite dependent on how the disease is eliminated; for example, if it is eliminated by chemoprevention or vaccination, there may be occasional fatal side effects, or there may be causal or preventive effects on other potentially fatal diseases. Careful consideration of the ambiguities inherent in 'disease elimination' should lead instead to estimation of the effect of specific interventions designed to reduce or eliminate the disease burden.[36]

Conclusions

Of the four causal modelling methods reviewed here, SCC models (the only ones originating in epidemiology) stand apart

in requiring specification of mechanisms within the individual units under study. There are rarely data to support such detailed specification, which may explain why SCC models have seen little use beyond teaching examples. Structural equations have seen extensive analytic application (especially in the social sciences[10,12,13,31]), and potential-outcome models have been used to derive permutation tests for randomized trials for 80 years.[9] Nonetheless, in epidemiology these models remain confined largely to the conceptual teaching realm (to the extent that they appear at all).[8,37] This confinement may be partly due to their absence from current training: unfamiliar techniques are rarely used. Furthermore, the most recent innovations based on potential outcomes (g-estimation[38,39] and marginal structural modelling[40]) are designed for longitudinal data on time-varying exposures and confounders, which precludes their use in many if not most studies; the techniques also require special programming.

Due to their qualitative form, graphical models have not led to as many analytic techniques as have algebraic models. On the other hand, they can be easily applied in any study to display assumptions of causal analyses, and to check whether covariates or sets of covariates are insufficient, excessive, or inappropriate to control given those assumptions.[1–6,14,17] When those assumptions are in doubt, one can still formulate a series of plausible graphs and conduct a corresponding series of analyses.[41] Constructing graphs to accompany conventional statistical analyses of effects can at least help avoid or spot common mistakes, such as control of intermediates as if they were confounders.[6,14,17]

Acknowledgements

The authors would like to thank Charles Poole, Judea Pearl, Katherine Hoggatt, and a referee for helpful comments on this paper.

KEY MESSAGES

- There are now at least four major classes of causal models in the health-sciences literature: Causal diagrams (graphical causal models), potential-outcome models, structural-equations models, and sufficient-component cause models.
- Causal diagrams can provide an easily understood depiction of qualitative assumptions behind a causal analysis, while potential-outcome and structural-equations models can depict more detailed quantitative assumptions about responses of units comprising the study population.
- Sufficient-component cause models differ from the other models in that they depict more elaborate qualitative assumptions about causal mechanisms within population units.

References

[1] Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search.* New York: Springer, 1993.

[2] Pearl J. Causal diagrams for empirical research (with discussion). *Biometrika* 1995;**82:**669–710.

[3] Pearl J. *Causality.* New York: Cambridge, 2000.

[4] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10:**37–48.

[5] Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;**12:**313–20.

[6] Hernán MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation. *Am J Epidemiol* 2002;**155:**176–84.

[7] Rothman KJ. Causes. *Am J Epidemiol* 1976;**104:**587–92.

[8] Rothman KJ, Greenland S. *Modern Epidemiology. 2nd Edn*. Philadelphia: Lippincott, 1998.

[9] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Ann Rev Public Health* 2000;**21:**121–45.

[10] Winship C, Morgan SL. Estimation of causal effects from observational data. *Ann Rev Sociol* 1999;**25:**659–706.

[11] Greenland S. Causal analysis in the health sciences. *J Am Statist Assoc* 2000;**95:**286–89.

[12] Sobel M. Causal inference in the social sciences. *J Am Statist Assoc* 2000;**95:**647–51.

[13] Heckman JJ, Vytlacil E. Econometric evaluations of social programs. In: Leamer E, Heckman JJ (eds). *Handbook of Econometrics, Vol. 6*. New York: Elsevier, 2003 (in press).

[14] Kaufman JS, Kaufman S. Assessment of structured socioeconomic effects on health. *Epidemiology* 2001;**12:**157–67.

[15] Maldonado G, Greenland S. Estimating causal effects (with discussion). *Int J Epidemiol* 2002;**31:**421–38.

[16] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;**3:**143–55.

[17] Cole S, Hernán M. Fallibility in estimating direct effects. *Int J Epidemiol* 2002;**31:**163–65.

[18] Levin ML. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* 1953;**9:**531–41.

[19] MacMahon B, Pugh TF. Causes and entities of disease. In: Clark DW, MacMahon B (eds). *Preventive Medicine*. Boston: Little Brown, 1967, pp. 11–18.

[20] Halloran ME, Struchiner CJ. Causal inference for infectious diseases. *Epidemiology* 1995;**6:**145–51.

[21] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999;**14:**29–46.

[22] Greenland S. Ecologic versus individual-level sources of confounding in ecologic estimates of contextual health effects. *Int J Epidemiol* 2001;**30:**1343–50.

[23] Robins JM. Causal inference from complex longitudinal data. In: Berkane M (ed.). *Latent Variable Modeling with Applications to Causality*. New York: Springer, 1997, pp. 69–117.

[24] Poole C. Positivized epidemiology and the model of sufficient and component causes. *Int J Epidemiol* 2001;**30:**707–09.

[25] Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. *Scand J Work Environ Health* 1988;**14:**125–29.

[26] Siemiatycki J, Thomas DC. Biological models and statistical interactions. *Int J Epidemiol* 1981;**10:**383–87.

[27] Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991;**44:**221–32.

[28] Beyea J, Greenland S. The importance of specifying the underlying biologic model in estimating the probability of causation. *Health Physics* 1999;**76:**269–74.

[29] Greenland S, Robins JM. Epidemiology, justice, and the probability of causation. *Jurimetrics* 2000;**40:**321–40.

[30] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *J Am Statist Assoc* 1996;**91:**444–72.

[31] Bowden RJ, Turkington DA. *Instrumental Variables.* Cambridge: Cambridge University Press, 1984.

[32] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;**29:**722–29.

[33] Bracken M, Belanger K, Hellenbrand K *et al.* Correlates of residential wiring code used in studies of health effects of residential electromagnetic fields. *Am J Epidemiol* 1998;**148:**467–74.

[34] Hatch EE, Kleinerman RA, Linet MS *et al.* Do confounding or selection factors of residential wiring codes and magnetic fields distort findings of electromagnetic field studies? *Epidemiology* 2000;**11:**189–98.

[35] Holland PW. Statistics and causal inference (with discussion). *J Am Statist Assoc* 1986;**81:**945–60.

[36] Greenland S. Causality theory for policy uses of epidemiological measures. In: Murray CJL, Mathers C, Salomon J, Lopez AD, Lozano R (eds). *Summary Measures of Population Health*. Cambridge, MA: Harvard, 2002, Ch. 6.2.

[37] Newman SC. *Biostatistical Methods in Epidemiology.* New York: Wiley, 2001.

[38] Witteman J, D'Agostino RB, Stijnen T *et al.* G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Study. *Am J Epidemiol* 1998;**148:**390–401.

[39] Joffe MM, Hoover DR, Jacobson LP *et al.* Estimating the effect of zidovudine on Kaposi's sarcoma from observational data using a rank-preserving structural failure-time model. *Stat Med* 1998;**17:** 1073–102.

[40] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effects on zidovudine on the survival of HIV-positive men. *Epidemiology* 2000;**11:**561–70.

[41] Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;**9:**361–67.

1.4 Graphs

We learned from Simpson's Paradox that certain decisions cannot be made on the basis of data alone, but instead depend on the story behind the data. In this section, we layout a mathematical language, *graph theory*, in which these stories can be conveyed. Graph theory is not generally taught in high school mathematics, but it provides a useful mathematical language that allows us to address problems of causality with simple operations similar to those used to solve arithmetic problems.

Although the word *graph* is used colloquially to refer to a whole range of visual aids—more or less interchangeably with the word *chart*—in mathematics, a graph is a formally defined

object. A mathematical graph is a collection of *vertices* (or, as we will call them, *nodes*) and edges. The nodes in a graph are connected (or not) by the edges. Figure 1.5 illustrates a simple graph. X, Y, and Z (the dots) are nodes, and A and B (the lines) are edges.

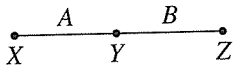


Figure 1.5 An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

Two nodes are *adjacent* if there is an edge between them. In Figure 1.5, X and Y are adjacent, and Y and Z are adjacent. A graph is said to be a *complete graph* if there is an edge between every pair of nodes in the graph.

A *path* between two nodes X and Y is a sequence of nodes beginning with X and ending with Y, in which each node is connected to the next by an edge. For instance, in Figure 1.5, there is a path from X to Z, because X is connected to Y, and Y is connected to Z.

Edges in a graph can be *directed* or *undirected*. Both of the edges in Figure 1.5 are undirected, because they have no designated "in" and "out" ends. A directed edge, on the other hand, goes out of one node and into another, with the direction indicated by an arrow head. A graph in which all of the edges are directed is a *directed graph*. Figure 1.6 illustrates a directed graph. In Figure 1.6, A is a directed edge from X to Y and B is a directed edge from Y to Z.

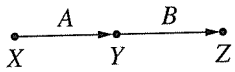


Figure 1.6 A directed graph in which node A is a parent of B and B is a parent of C

The node that a directed edge starts from is called the *parent* of the node that the edge goes into; conversely, the node that the edge goes into is the *child* of the node it comes from. In Figure 1.6, X is the parent of Y, and Y is the parent of Z; accordingly, Y is the child of X, and Z is the child of Y. A path between two nodes is a *directed* path if it can be traced along the arrows, that is, if no node on the path has two edges on the path directed into it, or two edges directed out of it. If two nodes are connected by a directed path, then the first node is the *ancestor* of every node on the path, and every node on the path is the *descendant* of the first node. (Think of this as an analogy to parent nodes and child nodes: parents are the ancestors of their children, and of their children's children, and of their children's children's children, etc.) For instance, in Figure 1.6, X is the ancestor of both Y and Z, and both Y and Z are descendants of X.

When a directed path exists from a node to itself, the path (and graph) is called *cyclic*. A directed graph with no cycles is *acyclic*. For example, in Figure 1.7(a) the graph is acyclic; however, the graph in Figure 1.7(b) is cyclic. Note that in (1) there is no directed path from any node to itself, whereas in (2) there are directed paths from X back to X, for example.

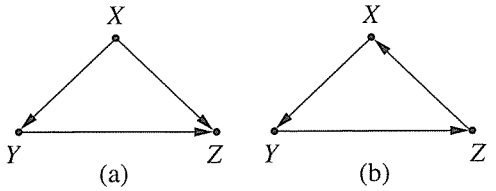


Figure 1.7 (a) Showing acyclic graph and (b) cyclic graph

Study questions

Study question 1.4.1

Consider the graph shown in Figure 1.8:

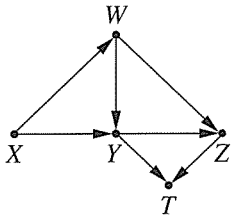


Figure 1.8 A directed graph used in Study question 1.4.1

(a) Name all of the parents of Z.
(b) Name all the ancestors of Z.
(c) Name all the children of W.
(d) Name all the descendants of W.
(e) Draw all (simple) paths between X and T (i.e., no node should appear more than once).
(f) Draw all the directed paths between X and T.

1.5 Structural Causal Models

1.5.1 Modeling Causal Assumptions

In order to deal rigorously with questions of causality, we must have a way of formally setting down our assumptions about the causal story behind a data set. To do so, we introduce the concept of the *structural causal model*, or SCM, which is a way of describing the relevant features of the world and how they interact with each other. Specifically, a structural causal model describes how nature assigns values to variables of interest.

Formally, a structural causal model consists of two sets of variables U and V, and a set of functions f that assigns each variable in V a value based on the values of the other variables in the model. Here, as promised, we expand on our definition of causation: A variable X is a *direct cause* of a variable Y if X appears in the function that assigns Y's value. X is a *cause* of Y if it is a direct cause of Y, or of any cause of Y.

The variables in U are called *exogenous variables*, meaning, roughly, that they are external to the model; we choose, for whatever reason, not to explain how they are caused. The variables in V are *endogenous*. Every endogenous variable in a model is a descendant of at least one exogenous variable. Exogenous variables cannot be descendants of any other variables, and in particular, cannot be a descendant of an endogenous variable; they have no ancestors and are represented as *root* nodes in graphs. If we know the value of every exogenous variable, then using the functions in f, we can determine with perfect certainty the value of every endogenous variable.

For example, suppose we are interested in studying the causal relationships between a treatment X and lung function Y for individuals who suffer from asthma. We might assume that Y also depends on, or is "caused by," air pollution levels as captured by a variable Z. In this case, we would refer to X and Y as endogenous and Z as exogenous. This is because we assume that air pollution is an external factor, that is, it cannot be caused by an individual's selected treatment or their lung function.

Every SCM is associated with a *graphical causal model*, referred to informally as a "graphical model" or simply "graph." Graphical models consist of a set of nodes representing the variables in U and V, and a set of edges between the nodes representing the functions in f. The graphical model G for an SCM M contains one node for each variable in M. If, in M, the function f_X for a variable X contains within it the variable Y (i.e., if X depends on Y for its value), then, in G, there will be a directed edge from Y to X. We will deal primarily with SCMs for which the graphical models are *directed acyclic graphs* (DAGs). Because of the relationship between SCMs and graphical models, we can give a graphical definition of causation: If, in a graphical model, a variable X is the child of another variable Y, then Y is a direct cause of X; if X is a descendant of Y, then Y is a potential cause of X (there are rare *intransitive cases* in which Y will not be a cause of X, which we will discuss in Part Two).

In this way, causal models and graphs encode causal assumptions. For instance, consider the following simple SCM:

SCM 1.5.1 (Salary Based on Education and Experience)

$$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$$
$$f_Z : Z = 2X + 3Y$$

This model represents the salary (Z) that an employer pays an individual with X years of schooling and Y years in the profession. X and Y both appear in f_Z, so X and Y are both direct causes of Z. If X and Y had any ancestors, those ancestors would be potential causes of Z.

The graphical model associated with SCM 1.5.1 is illustrated in Figure 1.9.

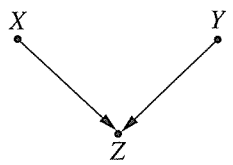


Figure 1.9 The graphical model of SCM 1.5.1, with X indicating years of schooling, Y indicating years of employment, and Z indicating salary

Because there are edges connecting Z to X and Y, we can conclude just by looking at the graphical model that there is some function f_Z in the model that assigns Z a value based on X and Y, and therefore that X and Y are causes of Z. However, without the fuller specification of an SCM, we can't tell from the graph what the function is that defines Z—or, in other words, *how* X and Y cause Z.

If graphical models contain less information than SCMs, why do we use them at all? There are several reasons. First, usually the knowledge that we have about causal relationships is not quantitative, as demanded by an SCM, but qualitative, as represented in a graphical model. We know off-hand that sex is a cause of height and that height is a cause of performance in basketball, but we would hesitate to give numerical values to these relationships. We could, instead of drawing a graph, simply create a partially specified version of the SCM:

SCM 1.5.2 (Basketball Performance Based on Height and Sex)

$$V = \{\text{Height, Sex, Performance}\}, \quad U = \{U_1, U_2, U_3\}, \quad F = \{f1, f2\}$$
$$\text{Sex} = U_1$$
$$\text{Height} = f_1(\text{Sex}, U_2)$$
$$\text{Performance} = f_2(\text{Height, Sex}, U_3)$$

Here, $U = \{U_1, U_2, U_3\}$ represents unmeasured factors that we do not care to name, but that affect the variables in V that we can measure. The U factors are sometimes called "error terms" or "omitted factors." These represent additional unknown and/or random exogenous causes of what we observe.

But graphical models provide a more intuitive understanding of causality than do such partially specified SCMs. Consider the SCM and its associated graphical model introduced above; while the SCM and its graphical model contain the same information, that is, that X causes Z and Y causes Z, that information is more quickly and easily ascertained by looking at the graphical model.

Study questions

Study question 1.5.1

Suppose we have the following SCM. Assume all exogenous variables are independent and that the expected value of each is 0.

SCM 1.5.3

$$V = \{X, Y, Z\}, \quad U = \{U_X, U_Y, U_Z\}, \quad F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = u_X$$
$$f_Y : Y = \frac{X}{3} + U_Y$$
$$f_Z : Z = \frac{Y}{16} + U_Z$$

(a) Draw the graph that complies with the model.
(b) Determine the best guess of the value (expected value) of Z, given that we observe $Y = 3$.
(c) Determine the best guess of the value of Z, given that we observe $X = 3$.
(d) Determine the best guess of the value of Z, given that we observe $X = 1$ and $Y = 3$.
(e) Assume that all exogenous variables are normally distributed with zero means and unit variance, that is, $\sigma = 1$.
  (i) Determine the best guess of X, given that we observed $Y = 2$.
  (ii) (Advanced) Determine the best guess of Y, given that we observed $X = 1$ and $Z = 3$. [Hint: You may wish to use the technique of multiple regression, together with the fact that, for every three normally distributed variables, say X, Y, and Z, we have $E[Y|X = x, Z = z] = R_{YX \cdot Z}x + R_{YZ \cdot X}z$.]
(f) Determine the best guess of the value of Z, given that we know $X = 3$.

1.5.2 Product Decomposition

Another advantage of graphical models is that they allow us to express joint distributions very efficiently. So far, we have presented joint distributions in two ways. First, we have used tables, in which we assigned a probability to every possible combination of values. This is intuitively easy to parse, but in models with many variables, it can take up a prohibitive amount of space; 10 binary variables would require a table with 1024 rows!

Second, in a fully specified SCM, we can represent the joint distributions of n variables with greater efficiency: We need only to specify the n functions that govern the relationships between the variables, and then from the probabilities of the error terms, we can discover all the probabilities that govern the joint distribution. But we are not always in a position to fully specify a model; we may know that one variable is a cause of another but not the form of the equation relating them, or we may not know the distributions of the error terms. Even if we know these objects, writing them down may be easier said than done, especially, when the variables are discrete and the functions do not have familiar algebraic expressions.

Fortunately, we can use graphical models to help overcome both of these barriers through the following rule.

Rule of product decomposition

For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(child|parents)$ over all the "families" in the graph. Formally, we write this rule as

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i|pa_i) \tag{1.29}$$

where pa_i stands for the values of the parents of variable X_i, and the product $\prod_i$ runs over all i, from 1 to n. The relationship (1.29) follows from certain universally true independencies among the variables, which will be discussed in the next chapter in more detail.

For example, in a simple chain graph $X \rightarrow Y \rightarrow Z$, we can write directly:

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

This knowledge allows us to save an enormous amount of space when laying out a joint distribution. We need not create a probability table that lists a value for every possible triple (x, y, z). It will suffice to create three much smaller tables for X, $(Y|X)$, and $(Z|Y)$, and multiply the values as necessary.

To estimate the joint distribution from a data set generated by the above model, we need not count the frequency of every triple; we can instead count the frequencies of each x, $(y|x)$, and $(z|y)$ and multiply. This saves us a great deal of processing time in large models. It also increases substantially the accuracy of frequency counting. Thus, the assumptions underlying the graph allow us to exchange a "high-dimensional" estimation problem for a few "low-dimensional" probability distribution challenges. The graph therefore simplifies an estimation problem and, simultaneously, provides more precise estimators. If we do not know the graphical structure of an SCM, estimation becomes impossible with large number of variables and small, or moderately sized, data sets—the so-called "curse of dimensionality."

Graphical models let us do all of this without always needing to know the functions relating the variables, their parameters, or the distributions of their error terms.

Here's an evocative, if unrigorous, demonstration of the time and space saved by this strategy: Consider the chain $X \rightarrow Y \rightarrow Z \rightarrow W$, where X stands for clouds/no clouds, Y stands for rain/no rain, Z stands for wet pavement/dry pavement, and W stands for slippery pavement/unslippery pavement.

Using your own judgment, based on your experience of the world, how plausible is it that $P(clouds, no\text{-}rain, dry\ pavement, slippery\ pavement) = 0.23$?

This is quite a difficult question to answer straight out. But using the product rule, we can break it into pieces:

$$P(clouds)P(no\ rain|clouds)P(dry\ pavement|no\ rain)P(slippery\ pavement|dry\ pavement)$$

Our general sense of the world tells us that $P(clouds)$ should be relatively high, perhaps 0.5 (lower, of course, for those of us living in the strange, weatherless city of Los Angeles). Similarly, $P(no\ rain|clouds)$ is fairly high—say, 0.75. And $P(dry\ pavement|no\ rain)$ would be higher still, perhaps 0.9. But the $P(slippery\ pavement|dry\ pavement)$ should be quite low, somewhere in the range of 0.05. So putting it all together, we come to a ballpark estimate of $0.5 \times 0.75 \times 0.9 \times 0.05 = 0.0169$.

We will use this product rule often in this book in cases when we need to reason with numerical probabilities, but wish to avoid writing out large probability tables.

The importance of the product decomposition rule can be particularly appreciated when we deal with estimation. In fact, much of the role of statistics focuses on effective sampling designs, and estimation strategies, that allow us to exploit an appropriate data set to estimate probabilities as precisely as we might need. Consider again the problem of estimating the probability $P(X, Y, Z, W)$ for the chain $X \rightarrow Y \rightarrow Z \rightarrow W$. This time, however, we attempt to estimate the probability from data, rather than our own judgment. The number of (x, y, z, w) combinations that need to be assigned probabilities is $16 - 1 = 15$. Assume that we have 45 random observations, each consisting of a vector (x, y, z, w). On the average, each (x, y, z, w) cell would receive about three samples; some will receive one or two samples, and some remain empty. It is very unlikely that we would obtain a sufficient number of samples in each cell to assess the proportion in the population at large (i.e., when the sample size goes to infinity).

If we use our product decomposition rule, however, the 45 samples are separated into much larger categories. In order to determine $P(x)$, every (x, y, z, w) sample falls into one of only two cells: $(X = 1)$ and $(X = 0)$. Clearly, the probability of leaving either of them empty is much lower, and the accuracy of estimating population frequencies is much higher. The same is true of the divisions we need to make to determine $P(y|x)$: $(Y = 1, X = 1), (Y = 0, X = 1), (Y = 1, X = 0)$, and $(Y = 0, X = 0)$. And to determine $P(z|y)$: $(Y = 1, Z = 1), (Y = 0, Z = 1), (Y = 1, Z = 0)$, and $(Y = 0, Z = 0)$. And to determine $P(w|z)$: $(W = 1, Z = 1), (W = 0, Z = 1), (W = 1, Z = 0)$, and $(W = 0, Z = 0)$. Each of these divisions will give us much more accurate frequencies than our original division into 15 cells. Here we explicitly see the simpler estimation problems allowed by assuming the graphical structure of an SCM and the resulting improved accuracy of our frequency estimates.

This is not the only use to which we can put the qualitative knowledge that a graph provides. As we will see in the next section, graphical models reveal much more information than is obvious at first glance; we can learn a lot about, and infer a lot from, a data set using only the graphical model of its causal story.

Study questions

Study question 1.5.2

Assume that a population of patients contains a fraction r of individuals who suffer from a certain fatal syndrome Z, which simultaneously makes it uncomfortable for them to take a life-prolonging drug X (Figure 1.10). Let $Z = z_1$ and $Z = z_0$ represent, respectively, the presence and absence of the syndrome, $Y = y_1$ and $Y = y_0$ represent death and survival, respectively, and $X = x_1$ and $X = x_0$ represent taking and not taking the drug. Assume that patients not carrying the syndrome, $Z = z_0$, die with probability p_2 if they take the drug and with probability p_1 if they don't. Patients carrying the syndrome, $Z = z_1$, on the other hand, die with probability p_3 if they do not take the drug and with probability p_4 if they do take the drug. Further, patients having the syndrome are more likely to avoid the drug, with probabilities $q_1 = P(x_1|z_0)$ and $q_2 = p(x_1|z_1)$.

(a) Based on this model, compute the joint distributions $P(x, y, z), P(x, y), P(x, z)$, and $P(y, z)$ for all values of x, y, and z, in terms of the parameters $(r, p_1, p_2, p_3, p_4, q_1, q_2)$. [Hint: Use the product decomposition of Section 1.5.2.]

(b) Calculate the difference $P(y_1|x_1) - P(y_1|x_o)$ for three populations: (1) those carrying the syndrome, (2) those not carrying the syndrome, and (3) the population as a whole.

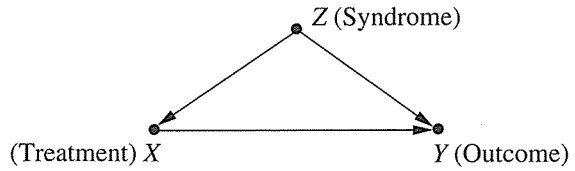


Figure 1.10 Model showing an unobserved syndrome, Z, affecting both treatment (X) and outcome (Y)

(c) *Using your results for (b), find a combination of parameters that exhibits Simpson's reversal.*

Study question 1.5.3

Consider a graph $X_1 \to X_2 \to X_3 \to X_4$ of binary random variables, and assume that the conditional probabilities between any two consecutive variables are given by

$$P(X_i = 1|X_{i-1} = 1) = p$$

$$P(X_i = 1|X_{i-1} = 0) = q$$

$$P(X_1 = 1) = p_0$$

Compute the following probabilities

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0)$$

$$P(X_4 = 1|X_1 = 1)$$

$$P(X_1 = 1|X_4 = 1)$$

$$P(X_3 = 1|X_1 = 0, X_4 = 1)$$

Study question 1.5.4

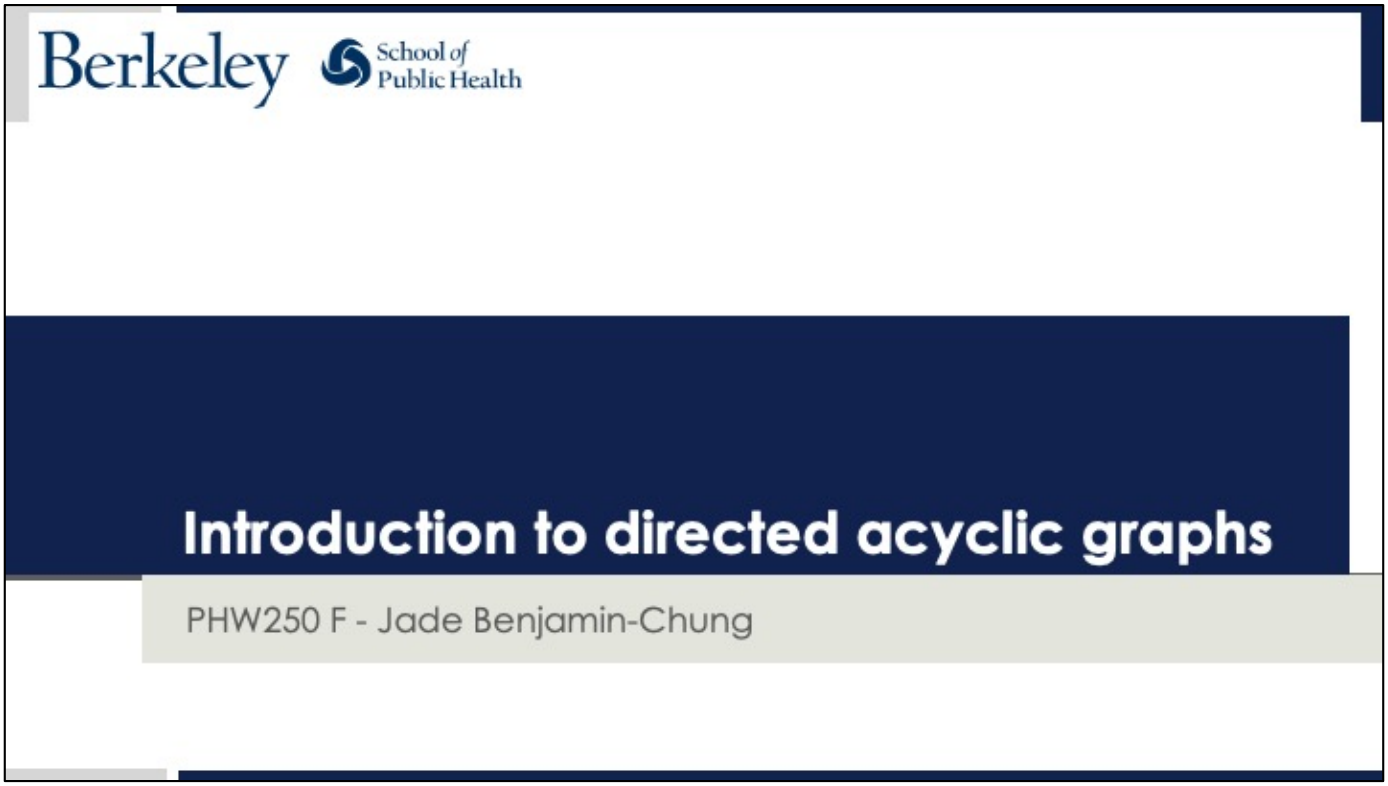
Define the structural model that corresponds to the Monty Hall problem, and use it to describe the joint distribution of all variables.

Bibliographical Notes for Chapter 1

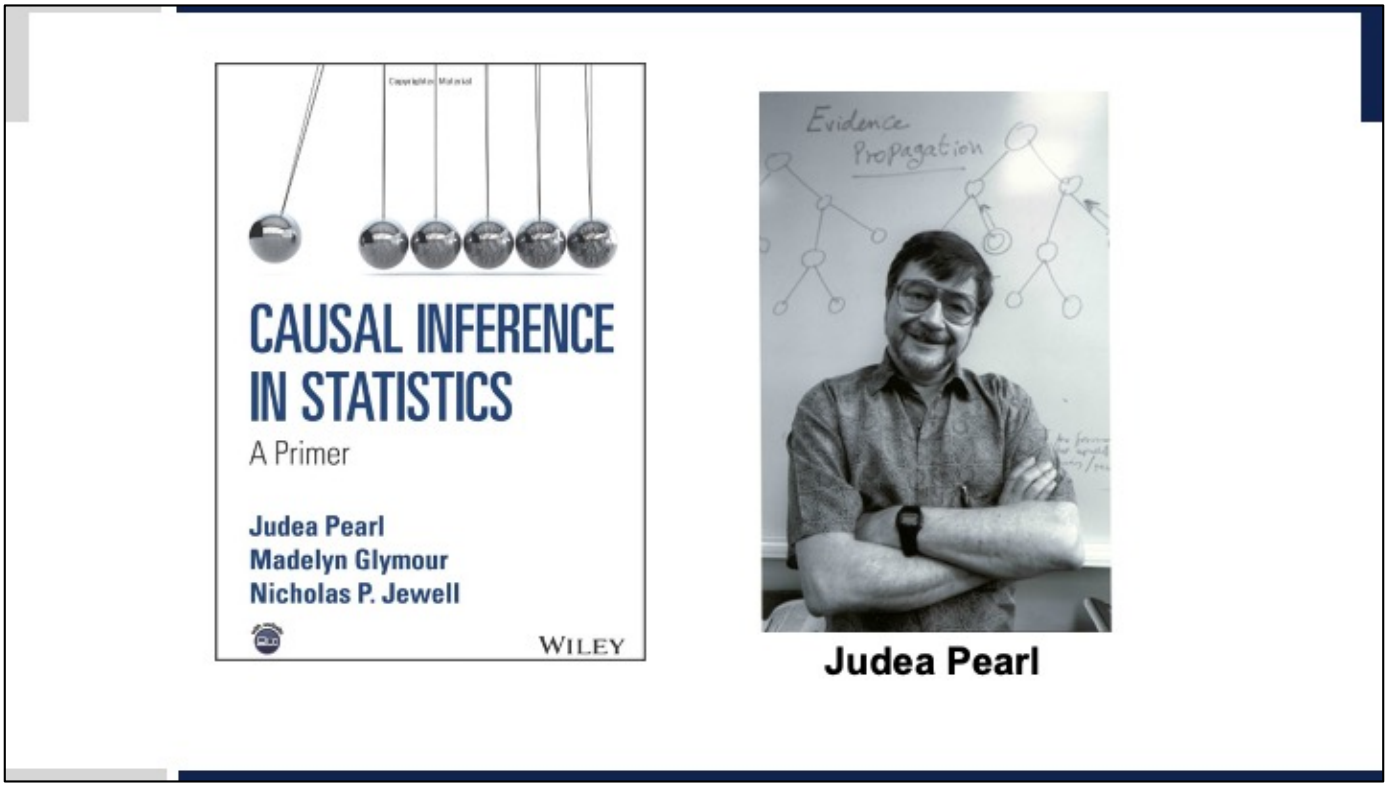
An extensive account of the history of Simpson's paradoxis given in Pearl (2009, pp. 174–182), including many attempts by statisticians to resolve it without invoking causation. A more recent account, geared for statistics instructors is given in (Pearl 2014c). Among the many texts that provide basic introductions to probability theory, Lindley (2014) and Pearl (1988, Chapters 1 and 2) are the closest in spirit to the Bayesian perspective used in Chapter 1. The textbooks by Selvin (2004) and Moore et al. (2014) provide excellent introductions to classical methods ofstatistics, including parameter estimation, hypothesis testing and regression analysis.

The Monty Hall problem, discussed in Section 1.3, appears in many introductory books on probability theory (e.g., Grinstead and Snell 1998, p. 136; Lindley 2014, p. 201) and is mathematically equivalent to the "Three Prisoners Dilemma" discussed in (Pearl 1988, pp. 58–62). Friendly introductions to graphical models are given in Elwert (2013), Glymour and Greenland (2008), and the more advanced texts of Pearl (1988, Chapter 3), Lauritzen (1996) and Koller and Friedman (2009). The product decomposition rule of Section 1.5.2 was used in Howard and Matheson (1981) and Kiiveri et al. (1984) and became the semantic basis of *Bayesian Networks* (Pearl 1985)—directed acyclic graphs that represent probabilistic knowledge, not necessarily causal. For inference and applications of Bayesian networks, see Darwiche (2009) and Fenton and Neil (2013), and Conrady and Jouffe (2015). The validity of the product decomposition rule for structural causal models was shown in Pearl and Verma (1991).

Lecture 2.2.1: Introduction to Directed Acyclic Graphs



ANDREW MERTENS: This video will introduce you to directed acyclic graphs. These are a tool that's increasingly used to assess causality in modern epidemiological studies.



I'll be drawing heavily from a few chapters in this book, Causal Inference in Statistics. The first author of this book is Judea Pearl, pictured here on the right, and he's really the father of this set of graphical techniques for representing statistics visually.

Graph theory

- A mathematical language that conveys statistical relationships in the data
- Allows you to explore causal relationships using a set of operations on graphs similar arithmetic operations
- In mathematics a graph is defined as a collection of **vertices** or **nodes** and **edges.**
- Statistical independence can be expressed visually using **directed acyclic graphs** (**DAGs**).
- We can use a **structural causal models** to predict patterns of interdependencies in the data based solely on the structure of a DAG.
- We will learn how to make causal models that represent the mechanism by which data in our studies was generated.

Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

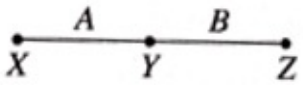
The tools that I'm going to present in this video really stem from graph theory. Graph theory is a mathematical language that can convey statistical relationships in the data. It allows you to explore causal relationships using a set of operations on graphs that are really similar to arithmetic operations that we would perform on numbers or variables.

A graph is defined as a collection of vertices or nodes and edges. So I'll show you a picture of what that looks like shortly. We can express statistical independence and also dependence visually using directed acyclic graphs, a certain type of graph.

I'm also going to be talking about structural causal models, which allow us to predict patterns of interdependence in data based solely on the structure of a DAG, or directed acyclic graph. And we'll talk about how to make causal models that represent the mechanism by which we believe our data was generated in our study.

Elements of graphs

- Two nodes are **adjacent** if there is an edge between them.
- A **path** between two nodes *X* and *Y* is a sequence of nodes beginning with *X* and ending with *Y*.
- In epidemiology, nodes in these graphs correspond to variables in our study.
- For example:
  - X = neighborhood of residence
  - Y = access to healthy food
  - Z = risk of obesity



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

At the bottom of the slide is a really simple graph. It has three nodes labeled X, Y, and Z. And they're depicted by these black circles or points. And then between these nodes, there are two paths in this graph, path A and path B.

We call two nodes adjacent if there's an edge between them, noting that an edge is another name for a path. And a path is just a line connecting two nodes or points.

X and Y is a sequence of nodes beginning with X and ending with Y. In epidemiology, the nodes in this graph are going to correspond to different variables in our study. So for example, X could be the neighborhood that someone resides in, and Y could be their access to healthy food. And Z could, say, be their risk of obesity.

Elements of graphs

- An edge is **directed** if an arrow indicates movement from one node to another.
- An edge is **undirected** if it connects two nodes without an arrowhead.
- A path is directed if it can be traced between two nodes along arrows.
  - No node has two edges on the path directed into it or two edges directed out of it.

Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

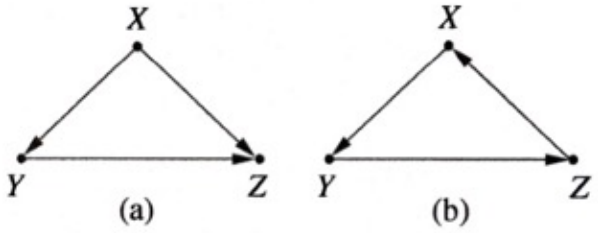
Here's another type of graph on the bottom right. It's the same as the one on the bottom left, except there's arrows from the edge A and the edge B. So an edge is called directed if it has an arrow that indicates movement from one node to another. And it's undirected if there's no arrowhead.

So A and B each label specific edges that are between nodes in this graph. And a path consists of multiple edges. So the path from X to Z contains the edges A and B.

A path is directed if it can be traced between two nodes along arrows. In these example graphs at the bottom, there's no node that has two edges on the path directed into it, or two edges directed out of it. This is just another way of saying if we look, for example, at the X to Y path, there's just one edge coming from X to Y.

Directed acyclic graphs (DAGs)

- When a directed path exists from a node to itself, the path is called **cyclic**.
- When a directed path has no cycles, it is called **acyclic**.
- In epidemiology, we use **directed acyclic graphs** because they allow us to visually represent statistical relationships that we believe exist in our data.



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

We're going to focus on a special type of graph called the directed acyclic graph. And moving forward, I'm going to refer to this as a DAG.

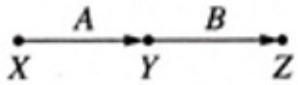
When a directed path exists from a node to itself, we can call that path cyclic. So graph (b) at the bottom right is a cyclic graph. We can see that if we start at X and then flow to Y, to Z, we end up back at X.

On the other hand, when a directed path has no cycles, it's called acyclic. So the graph (a) on the bottom of the slide, if we start at X, there's no way we can get back to X following the variables in the graph.

In epidemiology, we use directed acyclic graphs because they allow us to visually represent statistical relationships that we believe exist in our data. But later on in this course, you'll be drawing your own DAGs, and you need to make sure that they're truly DAGs, meaning that they're acyclic. They don't look like graph (b) on the slide.

Elements of DAGs

- A node that a directed edge starts from is called the **parent** of the node that the edge goes into.
- The node that the edge goes into is the **child** of the node it comes from.
- If two nodes are connected by a directed path, then the first node is the **ancestor** of every node on the path
- Every node on the path is the **descendent** of the first node.



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Let's define some elements of DAGs. We're going to define a parent, child, ancestor, and descendant.

A node that a directed edge starts from is called the parent of the node that the edge goes into. In the graph at the bottom, X is a parent of Y, and Y is a parent of Z. And the node that gets the edge into it is the child of the node it came from. Y is then the child of X, and Z is the child of Y.

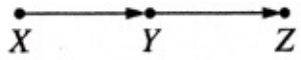
If two nodes are connected by a directed path, then the first node is the ancestor of every node on the path. So X is an ancestor of Y and Z. And every node on the path is a descendant of the first node. So both Y and Z are descendants of X.

Structural causal models (SCM)

- A model that describes how nature assigns values to variables of interest.
- A variable X is a direct cause of Y if X appears in the function that assigns Y's value.
- X is a cause of Z if it directly causes Z or any cause of Z.

Example:

- $X = f_X(U_X)$
- $Y = f_Y(X, U_Y)$
- $Z = f_Z(Y, U_Z)$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Now we'll define something called a structural causal model. It's a model that describes how nature assigns values to the variables of interest that we include in our DAG. We're going to define this model for this DAG here, where X goes into Y and Y goes into Z.

And this is how we define it. For each node, we have a formula, and the lowercase f here denotes a function. And then the subscript denotes the function of which variables we're referring to. Here at the top, X equals f of x. So that means we're defining X as a function of X with certain inputs. I'll define what this U of X means in a moment. And what U of Y and U of Z mean as well.

If we look at the formula for Y, it is equal to the function of Y with the inputs of the Y node, as well as U of Y. And this is because there's an arrow from X to Y. And so the value of Y will depend on the value of X.

From our example a few slides back, X is the neighborhood we live in, Y is our access to healthy food, and Z is the risk of obesity. And our function Y is unspecified. This is because we don't know the specific formula for how neighborhood residents affects access to food, but we do know that access to food depends on the neighborhood of residence.

We can say that a variable X is a direct cause of Y if X appears in the function that assigns Y's value. So that's what we were just talking about for X and Y. And X is a cause of Z if it directly causes Z or any cause of Z. We can say in our graph that X is a cause of Z because Z is a child of Y, and Y is a child of X.

Structural causal models (SCM)

Endogenous variables: variables internal to our model

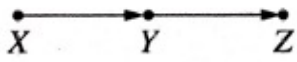
- We did measure these variables (usually)
- Every endogenous variable is the descendent of at least one exogenous variable.

Exogenous variables: variables external to our model

- We don't measure them because we can't directly measure them or we did not have the resources to in our study
- Exogenous variables cannot be descendents of other variables in our model.
- Denoted by "U" for unmeasured

Example:

- $X = f_X(U_X)$
- $Y = f_Y(X, U_Y)$
- $Z = f_Z(Y, U_Z)$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Here are a few more definitions. We have endogenous and exogenous variables. Endogenous variables are variables that are internal to our model. In epidemiology, this usually means that we measure these variables or that we intend to measure these variables. And every endogenous variable is a descendant of at least one exogenous variable.

An exogenous variable is a variable external to our model. We don't measure these variables usually because we can't directly measure them, or we didn't have the resources to measure them in our study. Exogenous variables cannot be descendants of other variables in our model. And we typically denote these exogenous variables by capital U to indicate that they're unmeasured.

In our example, X stands for our neighborhood of residence. And there was probably some other causes of the neighborhood of residence, but we haven't measured them. They're not included in our DAG. Our structural causal models function for X just contains U of X because we haven't measured any cause for the neighborhood of residence.

Similarly, there is unmeasured causes of Y that we haven't measured. We've only measured one cause of Y, which is X. The same goes for Z. So for each node in our DAG, the variables corresponding to it in the structural causal model has an unmeasured variable that's indicated by U within the subscript.

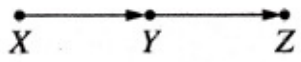
Structural causal models (SCM)

Every structural causal model is associated with a graphical causal model.

- **Nodes** in the graphical model represent **variables** in the structural causal model.
  - One node for each variable
- **Edges** in the graphical model represent **functions** in the structural causal model.
  - If a function f_Y for variable Y contains a variable X, then the graphical model will contain an edge from X to Y

Example:

- $X = f_X(U_X)$
- $Y = f_Y(X, U_Y)$
- $Z = f_Z(Y, U_Z)$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

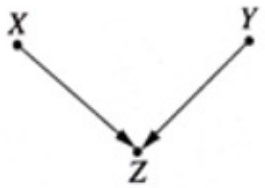
Every structural causal model is associated with the graphical causal model. And the nodes in the graphical model represent variables in the structural causal model, one node for each variable. And the edges in the graphical model represent functions in the structural causal model. If a function f of Y for variable Y contains a variable X, then the graphical model will contain an edge from X to Y. We can see that in the example to the right.

Example: DAG

- DAG for the salary (Z) an employer pays an employee with X years of schooling and Y years in the profession.

With just the DAG and no SCM, we can tell that X and Y cause Z, but we can't tell ***how*** X and Y cause Z.

<u>The structural causal model can tell us that.</u>



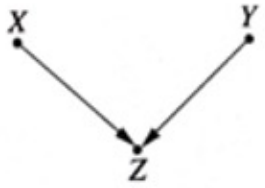
Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

In this simple example, we have a DAG for the salary Z an employer pays an employee with X years of schooling and Y years of experience in the profession. Our DAG on the right shows that X years of schooling affects the salary, as does Y, years in profession.

Currently, we just have a DAG, so we don't have a structural causal model. So we can tell that X and Y cause Z, but we can't tell how X and Y cause Z. The value of the structural causal model is that it may be able to tell us how X and Y cause Z.

Example: structural causal model

- Structural causal model for the salary (Z) an employer pays an employee with X years of schooling and Y years in the profession.
  - $Z = f_Z(X, Y)$
  - $X = f_X(U_X)$
  - $Y = f_Y(U_Y)$
  - $f_Z : Z = 2X + 3Y$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Here's an example of what the structural causal model could be for this DAG. I want to make it clear that there are many potential structural causal models that could explain this DAG, or could be associated with this DAG. This is just one example.

So here, we've defined Z as a function of Z that takes in X and Y as values. There's no U of Z here because this is a very simple example, and we're assuming that X and Y are the only two variables affecting Z.

In other words, there's nothing else we need to measure to determine the salary someone is paid. The salary someone is paid only depends on the years of schooling and their years in the profession. So this is pretty unrealistic, but let's just go with this for the example.

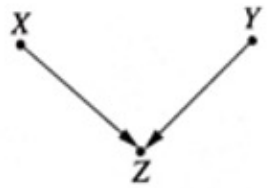
And then we didn't measure any causes of X and Y, so their functions just contain U of X and U of Y respectively.

In this particular structural causal model, we've defined the function of Z. Z equals 2X plus 3Y. So this is just a simple arithmetic formula. This is the how in the how of X and Y cause Z.

In this example, we can see that the structural causal model may provide more information than the DAG. But it's apparent immediately that in real life, it's pretty difficult to come up with this kind of formula for f of Z. There's almost always going to be too many variables involved, and we'll never know the exact arithmetic formula or other type of formula explaining how causes lead to their effect.

Why use DAGs if SCMs are more complete?

- In epidemiology, usually the knowledge we have about causal relationships is qualitative rather than quantitative.
- DAGs can effectively representing qualitative relationships.
- SCMs can represent either but are not more useful than DAGs when only qualitative information is available. For example we often only have the following information to accompany such a DAG:
  - $Z = f_Z(X, Y)$
  - $X = f_X(U_X)$
  - $Y = f_Y(U_Y)$
  - ~~$f_Z : Z = 2X + 3Y$~~



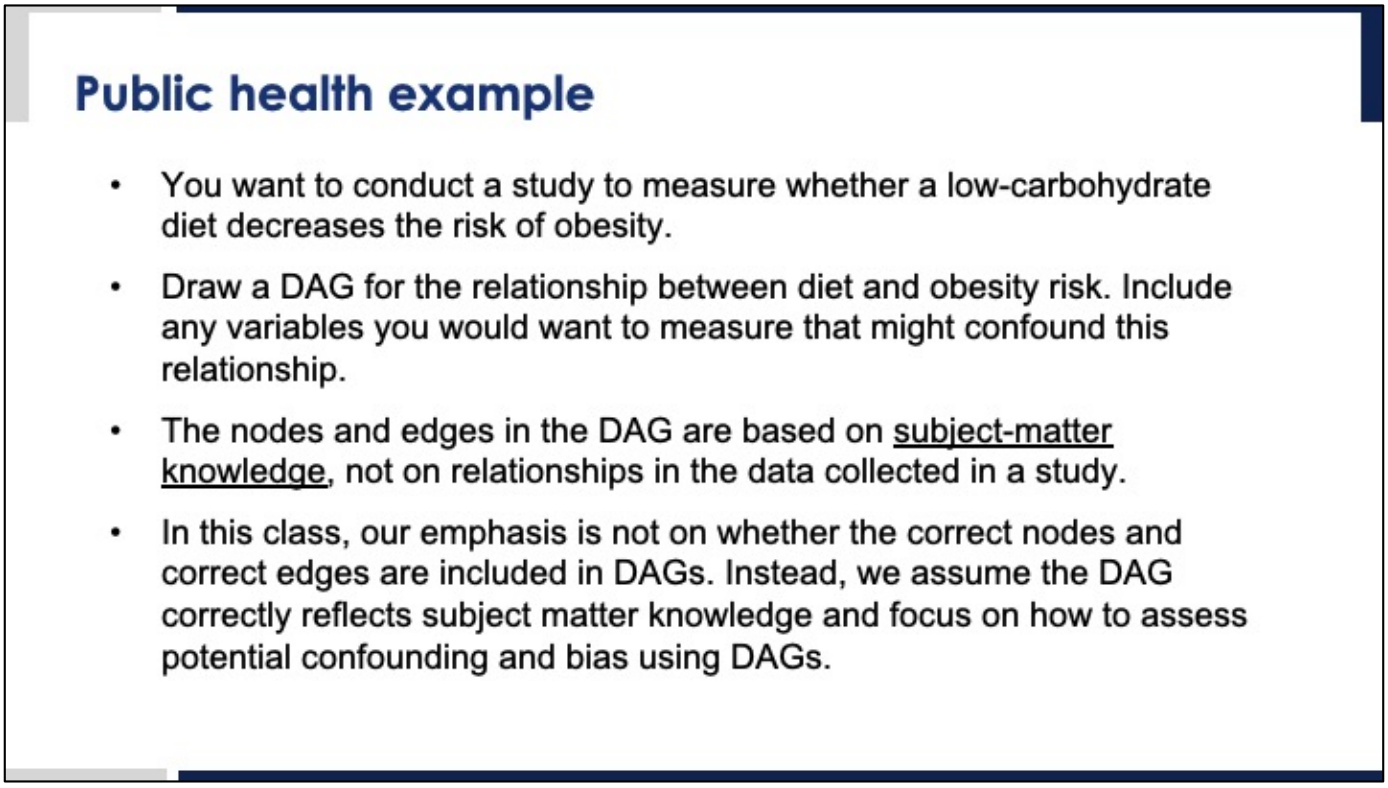
Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

You may be wondering, why use DAGs if structural causal models may be more complete. In epidemiology, usually the knowledge we have about causal relationships is qualitative rather than quantitative. This means that we have a sense of what causes another thing, but we don't know the exact level of effects for certain causes.

For example, we know that smoking causes lung cancer, but we can't say how many cigarettes smoked per person over the course of their life is needed to cause lung cancer. There's no one number, and it's definitely multifactorial when we think about the DAG for smoking and lung cancer.

DAGs are very effective at representing qualitative relationships, while structural causal models can represent either qualitative or quantitative relationships. But they're not more useful than DAGs when we only have qualitative information, which is usually the case in epidemiology.

So usually, our DAG will look like something like the DAG on the right, and will have the same structural causal model that I showed you on the last slide. But we won't have an f of Z. We'll just know that it's a function. Z is defined by a function of X and Y, but we don't know the specific values of that function.

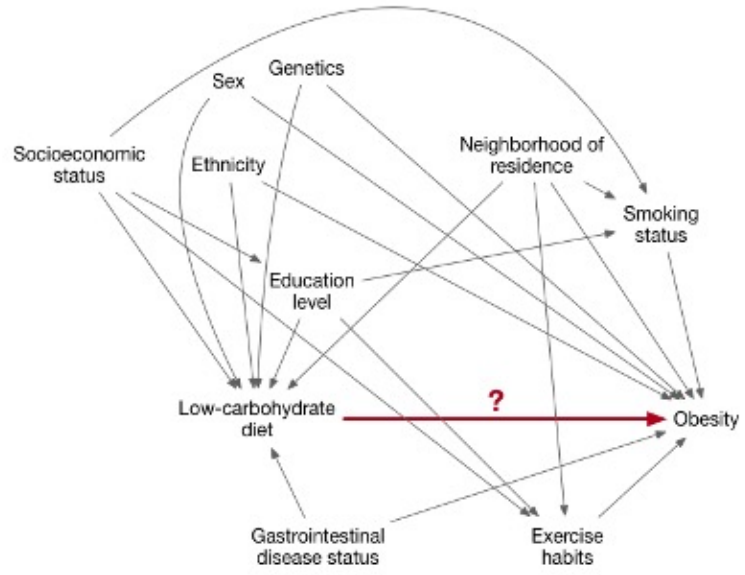


Let's go over a public health example. You conduct a study to measure whether a low-carbohydrate diet decreases the risk of obesity. We'll draw a DAG for the relationship between this low-carb diet and obesity risk. And we'll include any variables that we want to measure that might confound that relationship, or potentially serve as intermediates of that relationship.

The nodes and edges that we choose to include in the DAG are based on our subject matter knowledge. They're not based on relationships that we see in the data that we collect in a study. And in this class, our emphasis is not going to be on whether the correct nodes and the correct edges have been included.

I'll go over this example at a high level, but it's really more of a subject matter knowledge issue of what you're thinking about to include in the DAG. Instead, in this class, we'll assume that the DAGs that we draw correctly reflect our subject matter knowledge. And we'll focus our energy on how to assess potential confounding and bias using these DAGs.

Example: low-carbohydrate diet and obesity



All common causes of low-carb diet and obesity must be included in the DAG.

This DAG implies that only the variables for each node affect the relationship between low-carb diet and obesity.

A more realistic DAG would include U's for unmeasured variables.

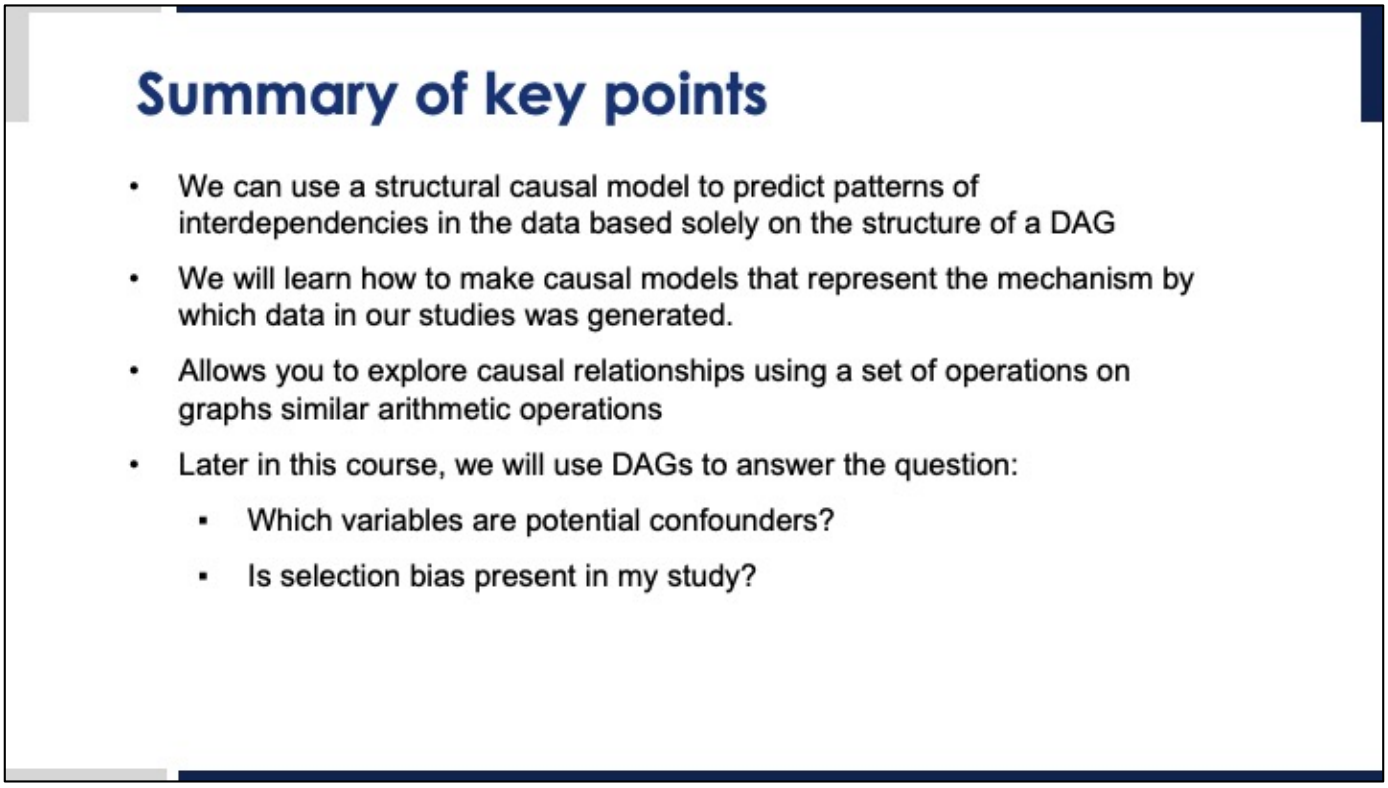
Later in this course, we will learn how to determine which variables are confounders and how to detect selection bias using a DAG.

Here's an example. Low-carb diet is right here, and our research question is, does a low-carb diet cause a decrease in obesity? I've included socioeconomic status, which can affect smoking status, which can in turn affect obesity risk. I've also included ethnicity, sex, genetics, and many other variables. And there's probably some variables that I should have included but I didn't include, and probably some edges I should have included but I didn't include.

So once again in our class, our focus is not really on whether the DAG is right. But rather, how do we analyze this DAG, and assess the risk of bias with this DAG?

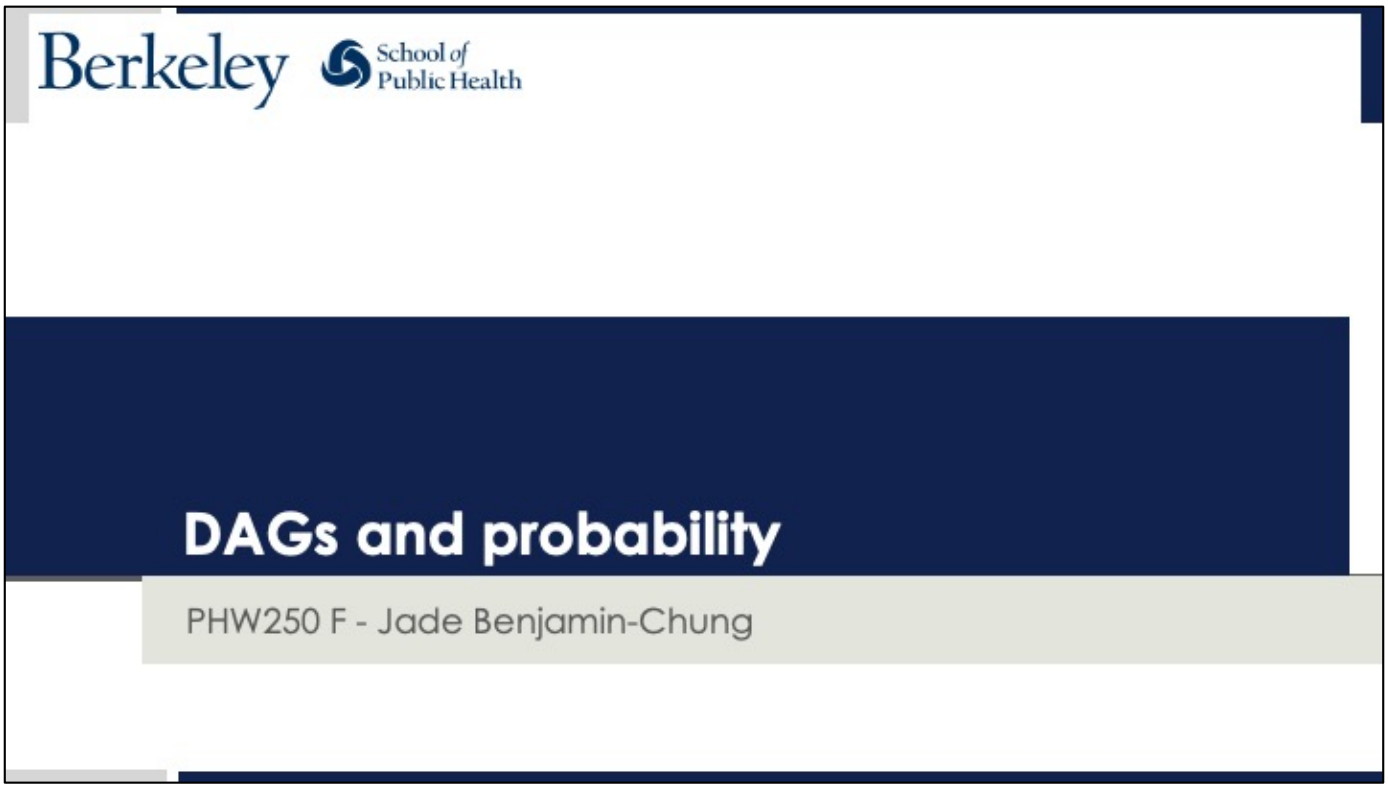
It's important to quickly mention that when we draw a DAG, all common causes of low-carb diet and obesity must be included in the DAG. The DAG implies that only the variables for each node affects the relationship between a low-carb diet and obesity. In other words, if there's another variable and we haven't included it in this DAG, we're implying that it does not affect the relationship between diet and obesity.

A more realistic DAG would include U's for all of the potential unmeasured variables. And there could potentially be a U going into almost all of these nodes. And later in the course, we'll come back to how to use this kind of DAG to determine which variables might be confounders, and how to detect selection bias and other kinds of bias using a DAG.



So to summarize, we can use a structural causal model to predict patterns of interdependencies in our data, based solely on the structures of a DAG. We're going to learn how to make causal models that represent the mechanism by which data in our studies was generated. These methods allow you to explore causal relationships using a set of operations on graphs that are similar to arithmetic operations. And then later in the course, we'll use DAGs to answer these questions-- which variables are potential confounders in my study, and is selection bias present in my study?

Lecture 2.2.2: DAGs and probability



ANDREW MERTENS: Now that you've learned what DAGs are, let's briefly talk about how DAGs are related to statistical probability theory. I want to emphasize, that for this course, you actually don't need to be able to do any of the things I going to show you, but this video is merely for your deeper level of understanding of how DAGs are related to statistics.

Link between DAGs and probability

- The **joint distribution** of the variables in the model is given by the product of the **conditional distributions** over all the "families" in the graph.

$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i|pa_i)$$

- **Example:** the joint distribution for the DAG $\mathrm{X} \rightarrow \mathrm{Y} \rightarrow \mathrm{Z}$ is:

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

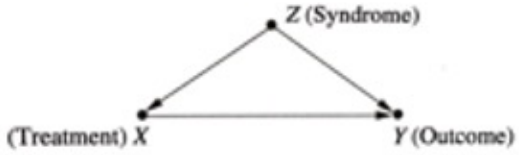
If we want to calculate the joint distribution of variables in a causal model, we can do so using the product of the conditional distributions over all of the families in the graph. So here's the notation for what that looks like. P of x1 and x2 and many other variables up to xn is denoted here on the left of the formula. So this is our joint distribution of all the variables in the model, and here's a grand product over the conditional probability of each variable conditional on its parents.

So a family is a parent group grouping within a DAG. For each node, we take the probability of that node conditional on all of its parents. That's our conditional probability, and we multiply the conditional probabilities of all of these different families in the graph together to get our joint distribution.

Let's go over a more specific example. For our DAG here, X leads to Y leads to Z. And we can write that joint distribution of X, Y, and Z as we've done right here on the left. And that's equal to the conditional probabilities for each family. So since nothing goes into X, we do the probability that X equals little x. For Y, since Y is conditional on X, we multiple-- we multiply times the probability of Y, conditional on the value for X. And since Y has an arrow going into Z, we also multiply it times the conditional probability of Z conditional on Y.

Example: DAG

- A population of patients contains a fraction *r* of individuals with fatal syndrome *Z*
  - $Z=z_1$ represents patients with the syndrome
  - $Z=z_0$ represents patients without the syndrome
- The syndrome makes it difficult for them to take treatment *X*
  - $X=x_1$ represents patients who take the drug
  - $X=x_0$ represents patients who do not take the drug
- Their survival status is indicated by *Y*
  - $Y=y_1$ represents patients who die
  - $Y=y_0$ represents patients who survive



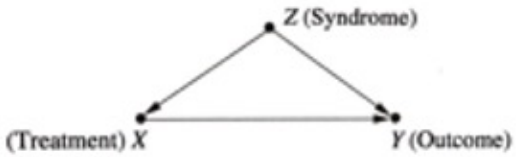
Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Let's go over a more concrete example. We have a population of patients that contains a fraction r of individuals with a fatal syndrome Z. The big Z is a random variable, and then little z subscript 1 or 0 indicates specific values of that syndrome. When I say a random variable, I just mean that when we use a capital letter for Z, X, or Y, it stands for any potential value of that variable-- so for Z whether the patient has the syndrome or doesn't have the syndrome. Z equals little z1 means that the patient has the syndrome, and little z0 means the patient doesn't have the syndrome.

In our example, the syndrome makes it difficult for people to take treatment X where little x1 represents patients who do take the drug, and little x0 represents patients who do not take the drug. The survival or death status of each patient is indicated by little y. Little y1 represents the patients who die, and little y0 represents the patients who survive. So here's our DAG. Having the syndrome affects whether you take that treatment, and whether you take that treatment affects whether you have the outcome, and the syndrome also affects whether you have the outcome.

Example: structural causal model

- $Z = f_Z(U_Z)$
- $X = f_X(Z, U_X)$
- $Y = f_Y(X, Z, U_Y)$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Here's a structural causal model for this DAG. There is no parents of Z in the DAG. The function for Z is merely just a function of U of Z, the unmeasured variable that causes Z. This could be genetics, for example, but it's unmeasured in our study. Z goes into X, so Z shows up in the function for X as well as U of X. And then there's two nodes that flow into Y-- X and Z-- and so both X and Z show up in the function for Y, as well as U of Y.

Example: probabilities

p q r

Assume the following:

- Patients without the syndrome who do not take the drug die with probability p_1
  - $P(Y \mid Z = z_0, X = x_0) = p_1$
- Patients without the syndrome who take the drug die with probability p_2
  - $P(Y \mid Z = z_0, X = x_1) = p_2$
- Patients with the syndrome who do not take the drug die with probability p_3
  - $P(Y \mid Z = z_1, X = x_0) = p_3$
- Patients with the syndrome who take the drug die with probability p_4
  - $P(Y \mid Z = z_1, X = x_1) = p_4$
- Patients with the syndrome are more likely to avoid the drug with the following probabilities:
  - $P(X = x_1 \mid Z = z_0) = q_1$
  - $P(X = x_1 \mid Z = z_1) = q_2$
- The probability patients have the syndrome:
  - $P(Z = z_1) = r$

Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Now, I'm going to use labels p, q, and r to indicate different types of probabilities in our causal model. So if we want the probability that someone dies conditional on not having the syndrome and not taking the drug, that probability as indicated by p1 in this equation here at the top. The probability that someone dies conditional on not having the syndrome and taking the drug is then indicated by p2 The probability that someone with the syndrome and who does not take the drug dies is indicated by a probability p3. And then the probability that patients with the syndrome who take the drug die is indicated by p4. And then, as I mentioned in the last slide, patients with the syndrome are more likely to avoid the drug, and then we can define these probabilities as follows.

The probability that someone takes the drug conditional on not having the syndrome q1. The probability that they take the drug conditional on having the syndrome is q2. And then, as I mentioned, the probability that patients have the syndrome is labeled as r. In reality, p, q, and r are all defined within 0 and 1 because they're probabilities.

Example: 2x2 table

- $P(Y \mid Z = z_0, X = x_0) = p_1 \quad = c_0 / (c_0 + d_0)$
- $P(Y \mid Z = z_0, X = x_1) = p_2 \quad = a_0 / (a_0 + b_0)$
- $P(Y \mid Z = z_1, X = x_0) = p_3 \quad = c_1 / (c_1 + d_1)$
- $P(Y \mid Z = z_1, X = x_1) = p_4 \quad = a_1 / (a_1 + b_1)$
- $P(X = x_1 \mid Z = z_0) = q_1 \quad = (a_0 + b_0) / N_0$
- $P(X = x_1 \mid Z = z_1) = q_2 \quad = (a_1 + b_1) / N_1$
- $P(Z = z_1) = r \quad = (a_1 + b_1 + c_1 + d_1) / N$

with

$Z=z_1$		
	$Y=y_1$	$Y=y_0$
$X=x_1$	a_1	b_1
$X=x_0$	c_1	d_1

without

$Z=z_0$		
	$Y=y_1$	$Y=y_0$
$X=x_1$	a_0	b_0
$X=x_0$	c_0	d_0

Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

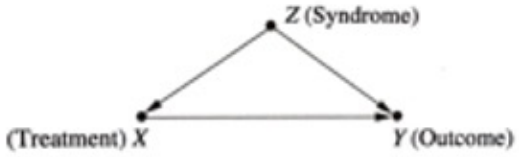
Since there's a lot of notation in this video, I wanted to briefly pause to just remind you of how these different probabilities correspond to values that we see in 2 by 2 tables. We can separate our data into data for patients with the syndrome-- z1-- and data for patients without the syndrome-- z0. And these two data sets each have their own 2 by 2 table, and within each table y is our outcome, which is represented in columns, and x is our exposure of taking the drug, which is represented in the rows.

So this is mainly for your reference, but I'll just go for one row of this to show you how these probabilities correspond to the 2 by 2 tables. For the first row here, the probability of death conditional on not having the syndrome, a.k.a. z0, and not taking the drug, a.k.a. x0, is equal to p1. So how do we find this in the 2 by 2 table? Well, first, we start by looking at the second 2 by 2 table because we're interested in Z equals z0. And then, we're also interested in X equals x0. So we want to look at the bottom row of the second table, and then we want to calculate the probability of death in the bottom row of the second table. So that's c0, which is the number of people who died who didn't have the syndrome and didn't take the drug, divided by the sum of all of the people who didn't have the syndrome and didn't take the drug. So that's c0 plus d0. So you can go over the remaining formulas in the slide if you're interested to see how these probabilities are linked to 2 by 2 tables.

Example: computing joint probabilities

What is $P(X = x_1, Y = y_1, Z = z_1)$ (the probability that $X = x_1$, $Y = y_1$, and $Z = z_1$)?

- $P(Y \mid Z = z_0, X = x_0) = p_1$
- $P(Y \mid Z = z_0, X = x_1) = p_2$
- $P(Y \mid Z = z_1, X = x_0) = p_3$
- $P(Y \mid Z = z_1, X = x_1) = p_4$
- $P(X = x_1 \mid Z = z_0) = q_1$
- $P(X = x_1 \mid Z = z_1) = q_2$
- $P(Z = z_1) = r$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

Now let's say we're interested in calculating the joint probabilities that X equals x1, Y equals y1, and Z equals z1. The blue box on the left reminds us how we're labeling each of these different probabilities, and the DAG for the relationship is on the right side. So we can use the formula from the very first slide to figure out how to calculate this probability and represent it with p's q's and r's.

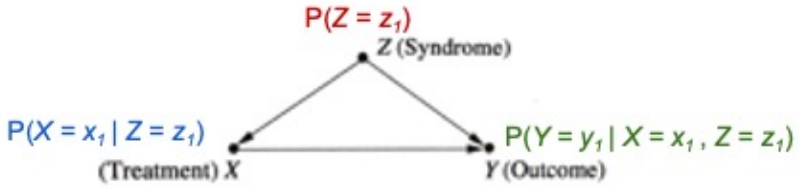
Example: computing joint probabilities

What is $P(X = x_1, Y = y_1, Z = z_1)$ (the probability that $X = x_1$, $Y = y_1$, and $Z = z_1$)?

$= P(Z = z_1)\, P(X = x_1 \mid Z = z_1)\, P(Y = y_1 \mid X = x_1, Z = z_1)$ (chain rule)

$= r * q_2 * p_4$

- $P(Y \mid Z = z_0, X = x_0) = p_1$
- $P(Y \mid Z = z_0, X = x_1) = p_2$
- $P(Y \mid Z = z_1, X = x_0) = p_3$
- $P(Y \mid Z = z_1, X = x_1) = p_4$
- $P(X = x_1 \mid Z = z_0) = q_1$
- $P(X = x_1 \mid Z = z_1) = q_2$
- $P(Z = z_1) = r$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

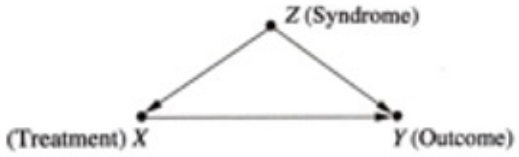
So remember, we need to take the grand product over the series of conditional probabilities for all families within the DAG. So let's look at each piece of the DAG. If we start with Z, there are no nodes going to Z. So we can just say that the probability of Z is equal to z1, because Z is essentially its own family. And this is written here in the formula up at the top of the slide. Now, if we focus on the X node, Z feeds into S. And so we need to say that the probability that X equals x1 is conditional on the probability that Z equals z1. And then, if we do the same thing for Y, there two nodes going into Y. So to calculate the probability that Y equals y1, it's conditional on X equals x1 and Z equals z1. And then, if you're wondering how we knew that this joint probability is equal to these conditional probabilities, you can go back to an intro stats textbook and see that this is something called the chain rule.

We can then represent these three pieces of the formula as r times q2 times p4, corresponding to the formulas above and defined in the blue box in the bottom left. And if we had values in our study that filled in our 2 by 2 table from the prior slide, we can use the information from both this slide and the prior slide to calculate this specific joint probability.

Example: computing conditional probabilities

Calculate the difference in the probability of death among those with the syndrome comparing those who did and did not take the drug.

- $P(Y \mid Z = z_0, X = x_0) = p_1$
- $P(Y \mid Z = z_0, X = x_1) = p_2$
- $P(Y \mid Z = z_1, X = x_0) = p_3$
- $P(Y \mid Z = z_1, X = x_1) = p_4$
- $P(X = x_1 \mid Z = z_0) = q_1$
- $P(X = x_1 \mid Z = z_1) = q_2$
- $P(Z = z_1) = r$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

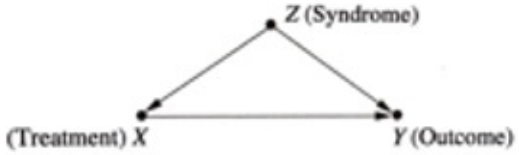
So now, let's calculate the difference and the probability of death among those with the syndrome comparing those who did and did not take the drug.

Example: computing conditional probabilities

Calculate the difference in the probability of death among those with the syndrome comparing those who did and did not take the drug.

$P(Y = y_1 \mid X = x_1, Z = z_1) - P(Y = y_1 \mid X = x_0, Z = z_1)$

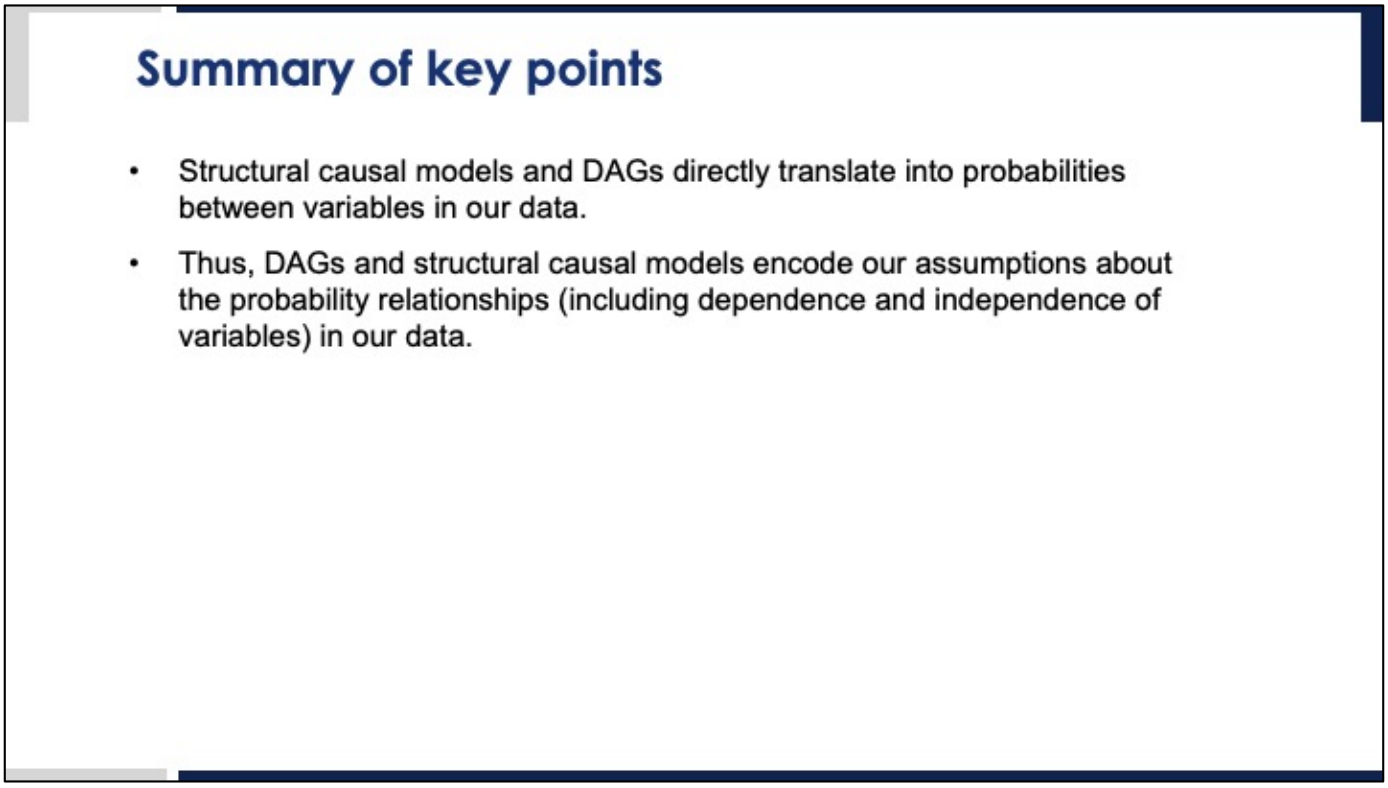
- $P(Y \mid Z = z_0, X = x_0) = p_1$
- $P(Y \mid Z = z_0, X = x_1) = p_2$
- $P(Y \mid Z = z_1, X = x_0) = p_3$
- $P(Y \mid Z = z_1, X = x_1) = p_4$
- $P(X = x_1 \mid Z = z_0) = q_1$
- $P(X = x_1 \mid Z = z_1) = q_2$
- $P(Z = z_1) = r$



Pearl et al., *Causal Inference in Statistics: A Primer*. Wiley 2016.

So this is the probability that Y equals y1, a.k.a. that someone died, conditional on X equals x1 and Z equals z1, a.k.a. that they took the drug and that they also had the syndrome, minus the probability that they died conditional on not taking the drug but having the syndrome. So this is pretty straightforward as we can just look at the probabilities in the blue box here. The first element of this equation, P of Y conditional and x1 and z1, is equal to p4. And the second element probability of Y conditional and x0 and z1 is equal to p3. So the answer is just p4 minus p3.

So the point of all this is just to show you that you're going to be using DAGs throughout this course to assess confounding and bias, and sometimes students wonder what do these DAGs really represent. It kind of seems like a really simplistic picture of the relationships in our data. And so the point of this video is to show you that it's not just a picture, it's actually linked to probability theory. And you can take more advanced courses in causal inference and learn all about the different operations that we can do on DAGs to get different statistical quantities of interest.



So to summarize, structural causal models and DAGs directly translate into probabilities between variables in our data. DAGs and structural causal models encode our assumptions about probability relationships, including dependence and independence of variables in our data.