2

What Is Causation?

The acquired wisdom that certain conditions or events bring about other conditions or events is an important survival trait. Consider an infant whose first experiences are a jumble of sensations that include hunger pangs, thirst, color, light, heat, cold, and many other stimuli. Gradually, the infant begins to perceive patterns in the jumble and to anticipate connections between actions such as crying and effects such as being fed. Eventually, the infant assembles an inventory of associated perceptions. We can imagine that the concept slowly develops that some of these phenomena are causally related to others that follow. Along with this growing appreciation for specific causal relations comes the general concept that some events or conditions can be considered causes of other events or conditions.

Thus, our first appreciation of the concept of causation is based on our own observations. These observations typically involve causes with effects that are immediately apparent. For example, when one changes the position of a light switch on the wall, one can see the instant effect of the light going on or off. There is more to the causal mechanism for getting the light to shine than turning the light switch to the "on" position, however. Suppose the electric lines to the building are down in a storm. Turning on the switch will have no effect. Suppose the bulb is burned out. Again, the switch will have no effect. One cause of the light going on is having the switch in the proper place, but along with it we must include a supply of power to the circuit, a working bulb, and wiring. When all other factors are already in place, turning the switch will cause the light to go on, but if one or more of the other factors is not playing its causal role, the light will not go on when the switch is turned. There is a tendency to consider the switch to be the unique cause of turning on the light, but in reality we can define a more intricate causal mechanism, in which the switch is one component of several. The tendency to identify the switch as the unique cause stems from its usual role as the final factor that acts in the causal mechanism. The wiring can be considered part of the causal mechanism, but once it is put in place, it seldom warrants further attention. The switch, however, is often

the only part of the mechanism that needs to be activated to obtain the effect of turning on the light. The effect usually occurs immediately after turning the switch, and as a result we slip into a frame of thinking in which we identify the switch as a unique cause. The inadequacy of this assumption is emphasized when the bulb goes bad and needs to be replaced.

The Causal Pie Model

Causes of disease can be conceptualized in the same way as the causes of turning on a light. A helpful way to think about causal mechanisms of disease is depicted in Figure 2–1.[1] Each "pie" in the diagram represents a theoretical *causal mechanism* for a given disease, sometimes called a "sufficient cause." There are three pies, to illustrate that there are multiple mechanisms that cause any type of disease. Each individual instance of disease will occur through a single mechanism or sufficient cause. A given causal mechanism requires the joint action of many component factors, or *component causes.* Each component cause is an event or condition that plays a necessary role in the occurrence of some cases of a given disease. For example, the disease might be cancer of the lung and, in the first mechanism in Figure 2–1, factor C might be cigarette smoking. The other factors would include genetic traits or other environmental exposures that play a causal role in cancer of the lung. Some component causes would presumably act in many different causal mechanisms.

Implications of the Causal Pie Model

Multicausality

The model of causation showed in Figure 2–1 illuminates several important principles regarding causes. Perhaps the most important of these principles is self-evident from the model: every causal mechanism in-

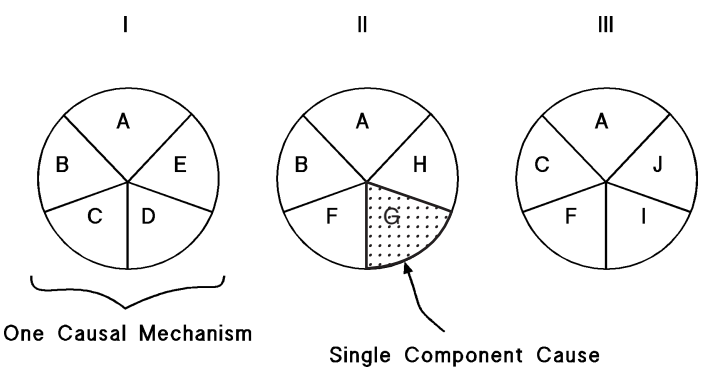

Figure 2–1. Three sufficient causes of a disease.

Genetic versus environmental causes

It is a strong assertion that every case of every disease has both genetic and environmental causes. Nevertheless, if all genetic factors that determine disease are taken into account, then essentially 100% of disease can be said to be inherited, in the sense that nearly all cases of disease have some genetic component causes. What would be the genetic component causes of someone who gets drunk and is killed in an automobile after colliding with a tree? It is easy to conceive of genetic traits that lead to psychiatric problems such as alcoholism, which in turn lead to drunk driving and consequent fatality. Analogously, one can also claim that essentially 100% of any disease is environmentally caused, even those diseases that we often consider to be purely genetic. Phenylketonuria, for example, is considered by many to be purely genetic. Nonetheless, if we consider the disease that phenylketonuria represents to be the mental retardation that may result from it, we can prevent the disease by appropriate dietary intervention. Thus, we can say that the disease has environmental determinants. Although it may seem like an exaggeration to claim that 100% of any disease is environmental and genetic at the same time, it is a good approximation. It may seem counterintuitive because most of the time we cannot manipulate many of the causes and the ones that can be controlled tend to be either environmental or genetic but usually not both.

volves the joint action of a multitude of component causes. Consider as an example the cause of a broken hip. Suppose that someone experiences a traumatic injury to the head that leads to a permanent disturbance in equilibrium. Many years later, the faulty equilibrium plays a causal role in a fall that occurs while the person is walking on an icy path. The fall results in a broken hip. Other factors playing a causal role for the broken hip could include the type of shoe the person was wearing, the lack of a handrail along the path, a strong wind, and the weight of the person. The complete causal mechanism involves a multitude of factors. Some factors, such as the earlier injury that resulted in the equilibrium disturbance and the weight of the person, reflect earlier events that have had a lingering effect. Some causal components of the broken hip are genetic. Genetic factors would affect the person's weight, gait, behavior, and recovery from the earlier trauma. Other factors, such as the force of the wind, are clearly environmental. It is reasonably safe to assert that there are nearly always some genetic and some environmental component causes in every causal mechanism. (Here, we use *environmental* to mean simply nongenetic.) Apparently, even an event such as a fall on an icy path leading to a broken hip is part of a complicated causal mechanism that involves many component causes.

Strength of Causes

It is common to think that some component causes play a more important role than others in the causation of disease. One way this concept is expressed is by the strength of a causal effect. Thus, we say that smoking has a strong effect on lung cancer risk because smokers have about 10 times the risk of lung cancer as nonsmokers. On the other hand, we say that smoking has a weaker effect on myocardial infarction because the risk of a heart attack is only about twice as great in smokers as in nonsmokers. With respect to an individual case of disease, however, every component cause that played a role in bringing that case into existence was necessary to the occurrence of that case. According to the causal pie model, for a given case of disease, there is no such thing as a strong cause or a weak cause. There is only a distinction between factors that were causes and factors that were not causes.

To understand what epidemiologists mean by *strength* of a cause, we need to shift from thinking about an individual case to thinking about the total burden of cases occurring in a population. We can then define a strong cause to be a component cause that plays a causal role in a large proportion of cases, whereas a weak cause would be a causal component in a small proportion of cases. Because smoking plays a causal role in a high proportion of the lung cancer cases, we call it a strong cause of lung cancer. For a given case of lung cancer, smoking is no more important than any of the other component causes for that case; but on the population level, it is considered a strong cause of lung cancer because it causes such a large proportion of cases.

The strength of a cause, defined in this way, necessarily depends on the prevalence of other causal factors that produce disease. As a result, the concept of a strong or weak cause cannot be a universally accurate description of any cause. For example, suppose we say that smoking is a strong cause of lung cancer because it plays a causal role in a large proportion of cases. Exposure to ambient radon gas, in contrast, is a weaker cause because it has a causal role in a much smaller proportion of lung cancer cases. Now imagine that society eventually succeeds in eliminating tobacco smoking, with a consequent reduction in smoking-related cases of lung cancer. One result is that a much larger proportion of the lung cancer cases that continue to occur will be caused by exposure to radon gas. It would appear that eliminating smoking has strengthened the causal effect of radon gas on lung cancer. This example illustrates that what we mean by strength of effect is not a biologically stable characteristic of a factor. From the biologic perspective, the causal role of a factor in producing disease is neither strong nor weak: the biology of causation corresponds simply to the identity of the component causes in a causal mechanism. The proportion of the population burden of disease that a factor causes, which we use to define the strength of a cause, can change from population to population and over time if there are changes

in the distribution of other causes of the disease. In short, the strength of a cause does not equate with the biology of causation.

Interaction between Causes

The causal pie model posits that several causal components act in concert to produce an effect. "Acting in concert" does not necessarily imply that factors must act at the same time. Consider the example above of the person who sustained trauma to the head that resulted in an equilibrium disturbance, which led years later to a fall on an icy path. The earlier head trauma played a causal role in the later hip fracture, as did the weather conditions on the day of the fracture. If both of these factors played a causal role in the hip fracture, then they interacted with one another to cause the fracture, despite the fact that their time of action was many years apart. We would say that any and all of the factors in the same causal mechanism for disease interact with one another to cause disease. Thus, the head trauma interacted with the weather conditions as well as with the other component causes, such as the type of footwear, the absence of a handhold, and any other conditions that were necessary to the causal mechanism of the fall and the broken hip that resulted. One can view each causal pie as a set of interacting causal components. This model provides a biologic basis for the concept of interaction that differs from the more traditional statistical view of interaction. We discuss the implication of this difference later, in Chapter 9.

Sum of Attributable Fractions

Consider the data in Table 2–1, which shows the rate of head-and-neck cancer according to smoking status and alcohol exposure. Suppose that the differences in the rates reflect causal effects. Among those who are smokers and alcohol drinkers, what proportion of cases of head and neck cancer that occur is attributable to the effect of smoking? We know that the rate for these people is 12 cases per 10,000 person-years. If these same people were not smokers, we can infer that their rate of head-and-neck cancer would be 3 cases per 10,000 person-years. If this difference reflects the causal role of smoking, then we might infer that 9 out of every 12 cases, or 75%, are attributable to smoking among those who smoke and drink alcohol. If we turn the question around and ask what proportion of disease among these same people is attributable to alcohol drinking, we would attribute 8 out of every 12 cases, or 67%, to alcohol drinking.

Can we attribute 75% of the cases to smoking and 67% to alcohol drinking among those who are exposed to both? The answer is yes, because when we do so, some cases are counted more than once as a result of the interaction between smoking and alcohol. These cases are attributable both to smoking and to alcohol drinking, because both factors played a causal role in producing those cases. One consequence of inter-

Table 2–1. Hypothetical rates of head-and-neck cancer (cases per 10,000 person-years) according to smoking status and alcohol drinking

Smoking Status	Nondrinker	Drinker
Nonsmoker	1	3
Smoker	4	12

action is that we should not expect that the proportions of disease attributable to various component causes will sum to 100%.

A widely discussed but unpublished paper from the 1970s written by scientists at the National Institutes of Health proposed that as much as 40% of cancer is attributable to occupational exposures. Many scientists thought that this fraction was an overestimate and argued against the claim.[2,3] One of the arguments used in rebuttal was as follows: x% of cancer is caused by smoking, y% by diet, z% by alcohol, and so on; when all of these percentages are added up, only a small percentage, much less than 40, is left for occupational causes. But this rebuttal is fallacious because it is based on the naive view that every case of disease has a single cause and that two causes cannot contribute to the same case of cancer. In fact, since diet, smoking, asbestos, and various occupational exposures, along with other factors, interact with one another and with genetic factors to cause cancer, each case of cancer could be attributed repeatedly to many separate component causes. The sum of disease attributable to various component causes in reality has no upper limit.

Induction Time

Because the component causes in a given causal mechanism do not act simultaneously, there will usually be a period of time between the action of a component cause and the completion of a sufficient cause. The only exception is the last component cause to act in a given causal mechanism. The last-acting component cause completes the causal mechanism, and we can say that disease begins concurrently with its action. For earlier-acting component causes, we can define the *induction period* as the period of time beginning at the action of a component cause and ending when the final component cause acts and the disease occurs. For example, in our illustration of the fractured hip, the induction time between the head trauma that resulted in an equilibrium disturbance and the later hip fracture was many years. The induction time between the decision to wear nongripping shoes and the hip fracture may have been a matter of minutes or hours. The induction time between the gust of wind that triggered the fall and the hip fracture might have been seconds or less.

In an individual instance, we would not be able to learn the exact length of an induction period, since we cannot be sure of the causal mechanism that produces disease in an individual instance or when all of the relevant component causes in that mechanism acted. With research data, however, we can learn enough to characterize the induction period that relates the action of a single component cause to the occurrence of disease in general. 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 the ages of 15 and 30 years. Since the causal exposure to DES occurs during gestation, there is an induction time of about 15 to 30 years for its carcinogenic action. During this time, other causes presumably operate; some evidence suggests that hormonal action during adolescence may be part of the mechanism.[4]

The causal pie model makes it clear that it is incorrect to characterize a disease itself as having a lengthy or brief induction time. The induction time can be conceptualized only in relation to a specific component cause. Thus, we say that the induction time relating DES exposure to clear cell carcinoma of the vagina is 15 to 30 years, but we cannot say that 15 to 30 years is the induction time for clear cell carcinoma in general. Since each component cause in any causal mechanism can act at a time different from the other component causes, each will have its own induction time. For the component cause that acts last, the induction time always equals 0. 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 than DES. Thus, induction time characterizes a specific cause–effect pair rather than just the effect.

In carcinogenesis, the terms *initiator* and *promotor* are used to refer to 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 and, by definition, the induction time will always be 0 for the last component cause to act.

After disease occurs, its presence is not always immediately apparent. If it becomes apparent later, the time interval between disease occurrence and its subsequent detection, whether by medical testing or by the emergence of symptoms, is termed the *latent period*.[5] The length of the latent period can be reduced by improved methods of disease detection. The induction period, however, cannot be reduced by early detection of disease, because there is no disease to detect until after the induction period is over. Practically, it may be difficult to distinguish between the induction period and the latent period, because there may be no way to

establish when the disease process began if it is not detected until later. Thus, diseases such as slow-growing cancers may appear to have long induction periods with respect to many causes, in part because they have long latent periods.

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 focusing on causes that are more proximal to the disease occurrence. Such biomarkers may reflect the effects on the organism of agents that acted at an earlier time.

Is a catalyst a cause?

Some agents may have a causal action by shortening the induction time of other agents. Suppose that exposure to factor A leads to epilepsy after an interval of 10 years, on the average. It may be that exposure to a drug, B, would shorten this interval to 2 years. Is B acting as a catalyst or as a cause of epilepsy? The answer is both: a catalyst *is* a cause. Without B, the occurrence of epilepsy comes 8 years later than it comes with B, so we can say that B causes the onset of the early epilepsy. It is not sufficient to argue that the epilepsy would have occurred anyway, so B is not a cause of its occurrence. First, it would not have occurred at that time, and the time of occurrence is considered part of the definition of an event. Second, epilepsy will occur later only if the person survives an additional 8 years, which is not certain. Therefore, agent B determines when the epilepsy occurs and it can determine whether it occurs at all. For this reason, we consider any agent that acts as a catalyst of a causal mechanism, shortening the induction period for other agents, to be a cause. Similarly, any agent that postpones the onset of an event, drawing out the induction period for another agent, we consider to be a preventive. It should not be too surprising to equate postponement with 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.

The Process of Scientific Inference

Much of epidemiologic research is aimed at uncovering the causes of disease. Now that we have a conceptual model for causes, how do we go about determining whether a given relation is causal? Some scientists refer to checklists for causal inference, and others focus on complicated statistical approaches, but the answer to this question is not to be found either in checklists or in statistical methods. The question itself is tanta-

mount to asking how we apply the scientific method to epidemiologic research. This question leads directly to the philosophy of science, a topic that goes well beyond the scope of this book. Nevertheless, it is worthwhile to summarize two of the major philosophical doctrines that have influenced modern science.

Induction

Since the rise of modern science in the seventeenth century, scientists and philosophers alike have puzzled over the question of how to determine the truth about assertions that deal with the empirical world. From the time of the ancient Greeks, deductive methods have been used to prove the validity of mathematical propositions. These methods enable us to draw airtight conclusions because they are self-contained, starting with a limited set of definitions and axioms and applying rules of logic that guarantee the validity of the method. Empirical science is different, however. Assertions about the real world do not start from arbitrary axioms, and they involve observations on nature that are fallible and incomplete. These stark differences from deductive logic led early modern empiricists, such as Francis Bacon, to promote what they considered a new type of logic, which they called *induction* (not to be confused with the concept of induction period, discussed above). *Induction* was an indirect method used to gain insight into what has been metaphorically described as the fabric of nature.

The method of induction starts with observations on nature. To the extent that the observations fall into a pattern, the observations are said to induce in the mind of the observer a suggestion of a more general statement about nature. The general statement could range from a simple hypothesis to a more profound natural law or natural relation. The statement about nature will be either reinforced by further observations or refuted by contradictory observations. For example, suppose an investigator in New York conducts an experiment to observe the boiling point of water and observes that the water boils at 100°C. The experiment might be repeated many times, each time showing that the water boils at about 100°C. By induction, the investigator could conclude that the boiling point of water is 100°C. The induction itself involves an inference beyond the observations to a general statement that describes the nature of boiling water. As induction became popular, it was seen to differ considerably from deduction. Although not as well understood as deduction, the approach was considered a new type of logic, inductive logic.

Although induction, with its emphasis on observation, represented an important advance over the appeal to faith and authority that characterized medieval scholasticism, it was not long before the validity of the new logic was questioned. The sharpest criticism came from the philosophical skeptic David Hume, who pointed out that induction had no

logical force. Rather, it amounted to an assumption that what had been observed in the past would continue to occur in the future. When supporters of induction argued for the validity of the process because it had been seen to work on numerous occasions, Hume countered that the argument was an example of circular reasoning that relied on induction to justify itself. Hume was so profoundly skeptical that he distrusted any inference based on observation, for the simple reason that observations depend on sense perceptions and are therefore subject to error.

Refutationism

Hume's criticisms of induction have been a powerful force in modern scientific philosophy. Perhaps the most influential reply to Hume was offered by Karl Popper. Popper accepted Hume's point that in empirical science one cannot prove the validity of a statement about nature in any way that is comparable with a deductive proof. Popper's philosophy, known as *refutationism*, held that statements about nature can be corroborated by evidence but that corroboration does not amount to logical proof. On the other hand, Popper asserted that statements about nature can be refuted by deductive logic. To grasp the point, consider the example above regarding the boiling point of water. The refutationist view is that the repeated experiments showing that water boils at 100°C corroborate the hypothesis that water boils at this temperature, but do not prove it.[6] A colleague of the New York researcher who works in Denver, a city at high altitude, might find that water there boils at a much lower temperature. This single contrary observation carries more weight regarding the hypothesis about the boiling point of water than thousands of repetitions of the initial experiment at sea level.

The asymmetrical implications of a refuting observation, on the one hand, and supporting observations, on the other hand, are the essence of the refutationist view. This school of thought encourages scientists to subject a new hypothesis to rigorous tests that might falsify the hypothesis, in preference to repetitions of the initial observations that add little beyond the weak corroboration that replication can supply. The implication for the method of science is that hypotheses should be evaluated by subjecting them to crucial tests. If a test refutes a hypothesis, then a new hypothesis needs to be formulated, which can then be subjected to further tests. Thus, after finding that water boils at a lower temperature in Denver than in New York, one must discard the hypothesis that water boils at 100°C and replace it with a more refined hypothesis, one that will explain the difference in boiling points under different atmospheric pressures. This process describes an endless cycle of *conjecture and refutation.* The conjecture, or hypothesis, is the product of scientific insight and imagination. It requires little justification except that it can account for existing observations. A useful approach is to pose competing hypotheses to explain existing observations and to test them against one

another. The refutationist philosophy postulates that all scientific knowledge is tentative in that it may one day need to be refined or even discarded. Under this philosophy, what we call scientific knowledge is a body of as yet unrefuted hypotheses that appear to explain existing observations.

How would an epidemiologist apply refutationist thinking to his or her work? If causal mechanisms are stated specifically, an epidemiologist can construct crucial tests of competing hypotheses. For example, when toxic shock syndrome was first studied, there were two competing hypotheses about the origin of the toxin. Under one hypothesis, the toxin responsible for the disease was a chemical in the tampon, so women using tampons were exposed to the toxin directly from the tampon. Under the other hypothesis, the tampon acted as a culture medium for staphylococci that produced the 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 risk of toxic shock. Under the hypothesis of a chemical intoxication, more frequent changing of the tampon would lead to more exposure to the toxin and possible absorption of a greater overall dose. This hypothesis predicted that women who changed tampons more frequently would have a higher risk of toxic shock syndrome than women who changed tampons infrequently. The culture-medium hypothesis predicts that the women who changed tampons frequently would have a lower risk than those who left the tampon in for longer periods, 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, which showed that infrequent changing of tampons was associated with greater risk of toxic shock, refuted the chemical theory.

Causal Criteria

Earlier, we said that there is no simple checklist that can determine whether an observed relation is causal. Nevertheless, attempts at such checklists have appeared and merit comment here. Most of these lists stem from the canons of inference described by John Stuart Mill.[5] The most widely cited list of causal criteria, originally posed as a list of standards, is attributed to Hill,[7] who adapted them from the U.S. Surgeon General's 1964 report on smoking and health.[8] The "Hill criteria," as they are often described, are listed in Table 2–2, along with some problems relating to each of them.

Although Hill did not propose these criteria as a checklist for evaluating whether a reported association might be interpreted as causal, many others have applied them in that way. Admittedly, the process of causal inference as described above is difficult and uncertain, making the appeal of a simple checklist undeniable. Unfortunately, this checklist, like

Table 2–2. "Causal criteria" of Hill

Criterion	Problems with the criterion
1. Strength	Strength depends on the prevalence of other causes and, thus, is not a biologic characteristic; could be confounded
2. Consistency	Exceptions are understood best with hindsight
3. Specificity	A cause can have many effects
4. Temporality	It may be difficult to establish the temporal sequence between cause and effect
5. Biologic gradient	Could be confounded; threshold phenomena would not show a progressive relation
6. Plausibility	Too subjective
7. Coherence	How does it differ from consistency or plausibility?
8. Experimental evidence	Not always available
9. Analogy	Analogies abound

all others with the same goal, fails to deliver on the hope of clearly distinguishing causal from noncausal relations. Consider the first criterion, strength. It is tempting to believe that strong associations are more likely to be causal than weak ones, but as we have seen above from our discussion of causal pies, not every component cause will have a strong association with the disease that it produces; strength of association depends on the prevalence of other factors. Some causal associations, such as that between cigarette smoking and coronary heart disease, are weak. Furthermore, a strong association could be noncausal, a confounded result stemming from the effect of another risk factor for the disease that is strongly associated with the one under study. For example, birth order is strongly associated with the occurrence of Down syndrome, but it is a confounded association that is completely explained by maternal age. If weak associations can be causal and strong associations can be noncausal, it does not appear that strength of association can be considered a criterion for causality.

The third criterion, specificity, suggests that a relation is more likely to be causal if the exposure is related to a single outcome rather than myriad outcomes. This criterion is misleading: it implies, for example, that the more diseases with which smoking is associated, the greater the evidence that smoking is not causally associated with any of them. Now consider the fifth criterion, biologic gradient. It is often taken as a sign of a causal relation, but it can just as well result from confounding or other biases as from a causal connection. The relation between Down syndrome and birth order mentioned above, for example, shows a biologic gradient despite it being completely explained by confounding from maternal age. Other criteria from Hill's list either are vague (consistency,

plausibility, coherence, and analogy) or do not apply in many settings (experimental evidence). The only criterion on the list that is truly a causal criterion is temporality, which implies that the cause comes before the effect. This criterion, which is part of the definition of a cause, is useful to keep in mind, although it may be difficult to establish the proper time sequence for cause and effect. For example, does stress lead to overeating or does overeating lead to stress? In general, it is better to avoid a checklist approach to causal inference and instead to consider approaches such as conjecture and refutation. Checklists lend a deceptive and mindless authority to an inherently imperfect and creative process. In contrast, causal inference based on conjecture and refutation fosters a highly desirable critical scrutiny.

Generalization in Epidemiology

A useful way to think of scientific generalization is to consider a generalization to be the elaboration of a scientific theory. A given study may test the viability of one or more theories. Theories that survive such tests can be viewed as general statements about nature that tell us what to expect in people or settings that were not studied. Because theories can be incorrect, scientific generalization is not a perfect process. Formulating a theory is not a mathematical or statistical process, so generalization should not be considered a statistical exercise. It is really no more nor less than the process of causal inference itself.

It is curious that many people believe that generalizing from an epidemiologic study involves a mechanical process of making an inference about a target population of which the study population is considered a sample. This type of generalization does exist, in the field of survey sampling. In survey sampling, researchers draw samples from a larger population to avoid the expense of studying the entire population. In survey sampling, the statistical representativeness of the sample is the main concern for generalizing to the source population.

Nevertheless, while survey sampling is an important tool for characterizing a population efficiently, it does not always share the same goals as science. Survey sampling is useful for problems such as trying to predict how a population will vote in an election or what type of laundry soap the people in a region prefer. These are characteristics that depend on attitudes and for which there is little coherent biologic theory on which to base a scientific generalization. For this reason, survey results may be quickly outdated (election polls may be repeated weekly or even daily) and do not apply outside of the populations from which the surveys were conducted. (Disclaimer: I am not saying that social science is not science or that we cannot develop theories about social behavior. I am saying only that surveys about the current attitudes of a specific group of people are not the same as social theories.) Epidemiologic re-

sults, in contrast, seldom need to be repeated weekly to see if they still apply. A study conducted in Chicago that shows that exposure to ionizing radiation causes cancer does not need to be repeated in Houston to see if ionizing radiation also causes cancer in people living in Houston. Generalization about ionizing radiation and cancer is based on an understanding of the underlying biology rather than on statistical sampling.

It may be helpful to consider the problem of scientific generalization about causes of cancer from the viewpoint of a biologist studying carcinogenesis in mice. Most researchers study cancer, whether it be in mice, rats, rabbits, hamsters, or humans, because they would like to understand better the causes of human cancer. But if scientific generalization depended on having studied a statistically representative sample of the target population, researchers using mice would have nothing to contribute to the understanding of human cancer. They certainly do not study representative samples of people; they do not even study representative samples of mice. Instead, they seek mice that have uniformly similar genes and perhaps certain biologic characteristics. In choosing mice to study, they have to consider mundane issues such as the cost of the mice. Although researchers using animals are unlikely to worry about whether their mouse or hamster or rabbit subjects are statistically representative of all mice or hamsters or rabbits, they might consider whether the biology of the animal population they are studying is similar to (and in that sense representative of) that of humans. This type of representativeness, however, is not statistical representativeness based on sampling from a source population; it is a biologic representativeness based on scientific knowledge. Indeed, despite the absence of statistical representativeness, no one seriously doubts the contribution that animal research can make to the understanding of human disease.

Of course, many epidemiologic activities do require surveys to characterize a specific population, but these activities are usually examples of applied epidemiology as opposed to the science of epidemiology. In applied epidemiology, we use general epidemiologic knowledge and apply it to specific settings. In epidemiologic science, just as in laboratory science, we move away from the specific toward the general: we hope to generalize from research findings, a process based more on scientific knowledge, insight, and even conjecture about nature than on the statistical representativeness of the actual study participants. This principle has important implications for the design and interpretation of epidemiologic studies, as we shall see in Chapter 5.

Questions

1. Criticize the following statement: The cause of tuberculosis is infection with the tubercle bacillus.

2. A trait in chickens called yellow shank occurs when a specific genetic strain of chickens is fed yellow corn. Farmers who own only this strain of chickens observe the trait to depend entirely on the nature of the diet, that is, whether they feed their chickens yellow corn. Farmers who feed all of their chickens only yellow corn but own several strains of chicken observe the trait to be genetic. What argument could you use to explain to both kinds of farmer that the trait is both environmental and genetic?
3. A newspaper article proclaims that diabetes is neither genetic nor environmental but multicausal. Another article announces that half of all colon cancer cases are linked to genetic factors. Criticize both messages.
4. Suppose a new treatment for a fatal disease defers the average time of death among those with the disease for 20 years beyond the time that they would have otherwise died. Is it proper to say that this new treatment reduces the risk of death, or does it merely postpone death?
5. It is typically more difficult to study an exposure–disease relation that has a long induction period than one that has a short induction period. What difficulties ensue because the exposure–disease induction period is long?
6. Suppose that both A and B are causes of a disease that is always fatal so that the disease can only occur once in a single person. Among people exposed to both A and B, what is the maximum proportion of disease that could be attributed to either A or B alone? What is the maximum for the sum of the amount attributable to A and the amount attributable to B? Suppose that A and B exert their causal influence only in different causal mechanisms so that they never act in the same mechanism. Would that change your answer?
7. Adherents of induction claim that we all use this method of inference every day. We assume, for example, that the sun will rise tomorrow as it has in the past. Critics of induction claim that this knowledge is based on belief and assumption and is no more than a psychologic crutch. Why should it matter to a scientist whether scientific reasoning is based on induction or on a different approach, such as conjecture and refutation?
8. Give an example of competing hypotheses for which an epidemiologic study would provide a refutation of at least one.
9. Could a causal association fail to show evidence of a biologic gradient (Hill's fifth criterion)? Explain.
10. Suppose you are studying the influence of socioeconomic factors on cardiovascular disease. Would the study be more informative if (1) the study participants had the same distribution of socioeconomic factors as the general population or (2) the study participants were recruited so that there were equal numbers in each category of the socioeconomic variable(s)? Why?

References

1. Rothman, KJ: Causes. *Am J Epidemiol* 1976;104:587–592.
2. Higginson, J: Proportion of cancer due to occupation. *Prev Med* 1980; 9:180–188.
3. Ephron, E: *The Apocalyptics. Cancer and the Big Lie.* New York: Simon and Schuster, 1984.
4. Rothman, KJ: Induction and latent period. *Am J Epidemiol* 1981;114:253–259.
5. Mill, JS: *A System of Logic, Ratiocinative and Inductive,* 5th ed. London: Parker, Son and Bowin, 1862.
6. Magee, B: *Philosophy and the Real World. An Introduction to Karl Popper.* La Salle, IL: Open Court, 1985.
7. Hill, AB: The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295–300.
8. US Department of Health, Education and Welfare. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service,* Public Health Service Publication 1103. Washington, D.C.: Government Printing Office, 1964.