Causation and Causal Inference in Epidemiology

Kenneth J. Rothman, DrPH, Sander Greenland, MA, MS, DrPH, C Stat

Concepts of cause and causal inference are largely self-taught from early learning experiences. A model of causation that describes causes in terms of sufficient causes and their component causes illuminates important principles such as multicausality, the dependence of the strength of component causes on the prevalence of complementary component causes, and interaction between component causes.

Philosophers agree that causal propositions cannot be proved, and find flaws or practical limitations in all philosophies of causal inference. Hence, the role of logic, belief, and observation in evaluating causal propositions is not settled. Causal inference in epidemiology is better viewed as an exercise in measurement of an effect rather than as a criterion-guided process for deciding whether an effect is present or not. (*Am J Public Health.* 2005;95:S144–S150. doi:10.2105/AJPH.2004.059204)

What do we mean by causation? Even among those who study causation as the object of their work, the concept is largely self-taught, cobbled together from early experiences. As a youngster, each person develops and tests an inventory of causal explanations that brings meaning to perceived events and that ultimately leads to more control of those events.

Because our first appreciation of the concept of causation is based on our own direct observations, the resulting concept is limited by the scope of those observations. We typically observe causes with effects that are immediately apparent. For example, when one turns a light switch to the "on" position, one normally sees the instant effect of the light going on. Nevertheless, the causal mechanism for getting a light to shine involves more than turning a light switch to "on." Suppose a storm has downed the electric lines to the building, or the wiring is faulty, or the bulb is burned out—in any of these cases, turning the switch on will have no effect. One cause of the light going on is having the switch in the proper position, but along with it we must have a supply of power to the circuit, good wiring, and a working bulb. When all other factors are in place, turning the switch will cause the light to go on, but if one or more of the other factors is lacking, the light will not go on.

Despite the tendency to consider a switch as the unique cause of turning on a light, the complete causal mechanism is more intricate, and the switch is only 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 on the switch, and as a result we slip into the 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. These concepts of causation that are established empirically early in life are too rudimentary to serve well as the basis for scientific theories. To enlarge upon them, we need a more general conceptual model that can serve as a common starting point in discussions of causal theories.

SUFFICIENT AND COMPONENT CAUSES

The concept and definition of causation engender continuing debate among philosophers. Nevertheless, researchers interested in causal phenomena must adopt a working definition. We can define a cause of a specific disease event as 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 event is an event, condition, or characteristic that preceded the disease event and without which the disease event either would not have occurred at all or would not have occurred until some later time. Under this definition it may be that no specific event, condition, or characteristic is sufficient by itself to produce disease. This is not a definition, then, of a complete causal mechanism, but only a component of it. A "sufficient cause," which means a complete causal mechanism, can be defined as a set of minimal conditions and events that inevitably produce disease; "minimal" implies that all of the conditions or events are necessary to that occurrence. In disease etiology, the completion of a sufficient cause may be considered equivalent to the onset of disease. (Onset here refers to the onset of the earliest stage of the disease process, rather than the onset of signs or symptoms.) For biological effects, most and sometimes all of the components of a sufficient cause are unknown.[1]

For example, tobacco smoking is a cause of lung cancer, but by itself it is not a sufficient cause. First, the term smoking is too imprecise to be used in a causal description. One must specify the type of smoke (e.g., cigarette, cigar, pipe), whether it is filtered or unfiltered, the manner and frequency of inhalation, and the onset and duration of smoking. More importantly, smoking, even defined explicitly, will not cause cancer in everyone. Apparently, there are some people who, by virtue of their genetic makeup or previous experience, are susceptible to the effects of smoking, and others who are not. These susceptibility factors are other components in the various causal mechanisms through which smoking causes lung cancer.

Figure 1 provides a schematic diagram of sufficient causes in a hypothetical individual. Each constellation of component causes represented in Figure 1 is minimally sufficient to produce the disease; that is, there is no redundant or extraneous component cause. Each one is a necessary part of that specific causal

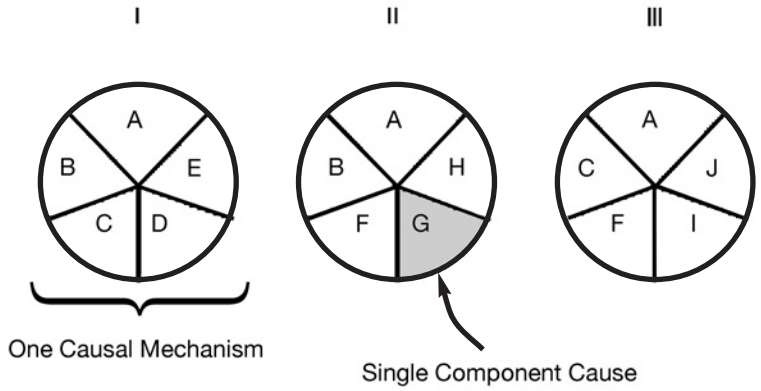

FIGURE 1—Three sufficient causes of disease.

mechanism. A specific component cause may play a role in one, two, or all three of the causal mechanisms pictured.

MULTICAUSALITY

The model of causation implied by Figure 1 illuminates several important principles regarding causes. Perhaps the most important of these principles is self-evident from the model: A given disease can be caused by more than one causal mechanism, and every causal mechanism involves 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, or the body weight of the person, among others. The complete causal mechanism involves a multitude of factors. Some factors, such as the person's weight and the earlier injury that resulted in the equilibrium disturbance, reflect earlier events that have had a lingering effect. Some causal components are genetic and would affect the person's weight, gait, behavior, recovery from the earlier trauma, and so forth. Other factors, such as the force of the wind, are environmental. It is a reasonably safe assertion that there are nearly always some genetic and some environmental component causes in every causal mechanism. Thus, 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.

The importance of multicausality is that most identified causes are neither necessary nor sufficient to produce disease. Nevertheless, a cause need not be either necessary or sufficient for its removal to result in disease prevention. If a component cause that is neither necessary nor sufficient is blocked, a substantial amount of disease may be prevented. That the cause is not necessary implies that some disease may still occur after the cause is blocked, but a component cause will nevertheless be a necessary cause for some of the cases that occur. That the component cause is not sufficient implies that other component causes must interact with it to produce the disease, and that blocking any of them would result in prevention of some cases of disease. Thus, one need not identify every component cause to prevent some cases of disease. In the law, a distinction is sometimes made among component causes to identify those that may be considered a "proximate" cause, implying a more direct connection or responsibility for the outcome.[2]

STRENGTH OF A CAUSE

In epidemiology, the strength of a factor's effect is usually measured by the change in disease frequency produced by introducing the factor into a population. This change may be measured in absolute or relative terms. In either case, the strength of an effect may have tremendous public health significance, but it may have little biological significance. The reason is that given a specific causal mechanism, any of the component causes can have strong or weak effects. The actual identity of the constituent components of the causal mechanism amounts to the biology of causation. In contrast, the strength of a factor's effect depends on the time-specific distribution of its causal complements in the population. Over a span of time, the strength of the effect of a given factor on the occurrence of a given disease may change, because the prevalence of its causal complements in various causal mechanisms may also change. The causal mechanisms in which the factor and its complements act could remain unchanged, however.

INTERACTION AMONG 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; so 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 is 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 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 biological basis for a concept of

Table 1–Hypothetical Rates of Head and Neck Cancer (Cases per 100 000 Person-Years) According to Smoking Status and Alcohol Drinking

	Alcohol Drinking	
Smoking Status	No	Yes
Nonsmoker	1	3
Smoker	4	12

interaction distinct from the usual statistical view of interaction.[3]

SUM OF ATTRIBUTABLE FRACTIONS

Consider the data on rates of head and neck cancer according to whether people have been cigarette smokers, alcohol drinkers, or both (Table 1). Suppose that the differences in the rates all reflect causal effects. Among those people who are smokers and also alcohol drinkers, what proportion of the cases is attributable to the effect of smoking? We know that the rate for these people is 12 cases per 100 000 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 100 000 person-years. If this difference reflects the causal role of smoking, then we might infer that 9 of every 12 cases, or 75%, are attributable to smoking among those who both 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 be able to attribute 8 of every 12 cases, or 67%, to alcohol drinking.

How can we attribute 75% of the cases to smoking and 67% to alcohol drinking among those who are exposed to both? We can because some cases are counted more than once. Smoking and alcohol interact in some cases of head and neck cancer, and these cases are attributable both to smoking and to alcohol drinking. One consequence of interaction is that we should not expect that the proportions of disease attributable to various component causes will sum to 100%.

A widely discussed (though 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 this claim.[4,5] One of the arguments used in rebuttal was as follows: *x* percent of cancer is caused by smoking, *y* percent by diet, *z* percent by alcohol, and so on; when all 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 both 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 thus has no upper limit.

A single cause or category of causes that is present in every sufficient cause of disease will have an attributable fraction of 100%. Much publicity attended the pronouncement in 1960 that as much as 90% of cancer is caused by environmental factors.[6] Since "environment" can be thought of as an all-embracing category that represents nongenetic causes, which must be present to some extent in every sufficient cause, it is clear on a priori grounds that 100% of any disease is environmentally caused. Thus, Higginson's estimate of 90% was an underestimate.

Similarly, one can show that 100% of any disease is inherited. MacMahon [7] cited the example of yellow shanks, [8] a trait occurring in certain strains of fowl fed yellow corn. Both the right set of genes and the yellow-corn diet are necessary to produce yellow shanks. A farmer with several strains of fowl, feeding them all only yellow corn, would consider yellow shanks to be a genetic condition, since 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 reality, yellow shanks is determined by both genes and environment; there is no reasonable way to allocate a portion of the causation to either genes or environment. Similarly, 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.

Many researchers have spent considerable effort in developing heritability indices, which are supposed to measure the fraction of disease that is inherited. Unfortunately, these indices only assess 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 having these genes is a cause of the disease. The two farmers in the example above 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.

If all genetic factors that determine disease are taken into account, whether or not they vary within populations, then 100% of disease can be said to be inherited. Analogously, 100% of any disease is environmentally caused, even those diseases that we often consider purely genetic. Phenylketonuria, for example, is considered by many to be purely genetic. Nonetheless, the mental retardation that it may cause can be prevented by appropriate dietary intervention.

The treatment for phenylketonuria illustrates the interaction of genes and environment to cause a disease commonly thought to be purely genetic. What about an apparently purely environmental cause of death such as death from an automobile accident? 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. Consider another more extreme environmental example, being killed by lightning. Partially heritable psychiatric conditions can influence whether someone will take shelter during a lightning storm; genetic traits such as

athletic ability may influence the likelihood of being outside when a lightning storm strikes; and having an outdoor occupation or pastime that is more frequent among men (or women), and in that sense genetic, would also influence the probability of getting killed by lightning. The argument may seem stretched on this example, but the point that every case of disease has both genetic and environmental causes is defensible and has important implications for research.

MAKING CAUSAL INFERENCES

Causal inference may be viewed as a special case of the more general process of scientific reasoning, about which there is substantial scholarly debate among scientists and philosophers.

Impossibility of Proof

Vigorous debate is a characteristic of modern scientific philosophy, no less in epidemiology than in other areas. Perhaps the most important common thread that emerges from the debated philosophies stems from 18th-century empiricist David Hume's observation that proof is impossible in empirical science. This simple fact is especially important to 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. Such a 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 irreproducible. The cold-fusion debacle demonstrates well that neither physical nor experimental science is immune to such problems.

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. Furthermore, such control is no guarantee against error. 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.

Testing Competing Epidemiologic Theories

Biological 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 are absent.

If the causal mechanism is stated specifically enough, epidemiologic observations under some circumstances might provide crucial tests of competing non-null causal hypotheses. On the other hand, many epidemiologic studies are not designed to test a causal hypothesis. For example, epidemiologic data related to the finding that women who took replacement estrogen therapy were at a considerably higher risk for endometrial cancer was examined by Horwitz and Feinstein, who 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.[9] The resulting diagnostic workup led to the detection of endometrial cancer at an earlier stage in these women, as compared with women not taking estrogens. Many epidemiologic observations could have been and were used to evaluate these competing hypotheses. The causal theory predicted that the risk of endometrial cancer would tend to increase with increasing use (dose, frequency, and duration) of estrogens, as for other carcinogenic exposures. The detection bias theory, on the other hand, predicted that women who had used estrogens only for a short while would have the greatest risk, since the symptoms related to estrogen use that led to the medical consultation tend to appear soon after use begins. Because the association of recent estrogen use and endometrial cancer was the same in both long-term 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.

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 noncausal hypotheses, in the sense of involving no causal connection between the study exposure and the disease. For example, hypotheses that amount to explanations of how specific types of bias could have led to an association between exposure and disease are the usual alternatives to the primary study hypothesis that the epidemiologist needs to consider in drawing inferences. Much of the interpretation of epidemiologic studies amounts to the testing of such noncausal explanations.

THE DUBIOUS VALUE OF CAUSAL CRITERIA

In practice, how do epidemiologists separate out the causal from the noncausal explanations? Despite philosophic criticisms of inductive inference, inductively oriented causal criteria have commonly been used to make such inferences. 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 forgotten: it would only be necessary to consult the checklist of criteria to see if a relation were causal. We know from philosophy 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.

Hill's Criteria

A commonly used set of criteria was proposed by Hill,[10] it was an expansion of a set of criteria offered previously in the landmark surgeon general's report on smoking and health,[11] which in turn were anticipated by the inductive canons of John Stuart Mill[12] and the rules given by Hume. [13]

Hill suggested that the following aspects of an association be considered in attempting to distinguish causal from noncausal associations: (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) biological gradient, (6) plausibility, (7) coherence, (8) experimental evidence, and (9) analogy. These criteria suffer from their inductivist origin, but their popularity demands a more specific discussion of their utility.

1. Strength. Hill's argument is essentially that strong associations are more likely to be causal than weak associations because, if they could be explained by some other factor, the effect of that factor would have to be even stronger than the observed association and therefore would have become evident. Weak associations, on the other hand, are more likely to be explained by undetected biases. To some extent this is a reasonable argument but, as Hill himself acknowledged, the fact that an association is weak does not rule out a causal connection. A commonly cited counterexample is the relation between cigarette smoking and cardiovascular disease: one explanation for this relation being weak is that cardiovascular disease is common, making any ratio measure of effect comparatively small compared with ratio measures for diseases that are less common.[14] Nevertheless, cigarette smoking is not seriously doubted as a cause of cardiovascular disease. Another example would be passive smoking and lung cancer, a weak association that few consider to be noncausal.

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 but noncausal 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 adjustment for the factor. These examples remind us that a strong association is neither necessary nor sufficient for causality, nor is weakness necessary or sufficient for absence of causality. Furthermore, neither relative risk nor any other measure of association is a biologically consistent feature of an association; as described above, such measures of association are characteristics of a given population that depend on the relative prevalence of other causes in that population. A strong association serves only to rule out hypotheses that the association is entirely due to one weak unmeasured confounder or other source of modest bias.

2. Consistency. 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 HIV infection 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 the 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. Furthermore, this fallacy is not eliminated by "standardizing" estimates.

3. Specificity. The criterion of specificity requires that a cause leads to a single effect, not multiple effects. This argument has often been advanced to refute causal interpretations of exposures that appear to relate to myriad effects—for example, by those seeking to exonerate smoking as a cause of lung cancer. Unfortunately, the criterion is invalid as a general rule. Causes of a given effect cannot be expected to lack all other effects. In fact, everyday experience teaches us repeatedly that single events or conditions may have many effects. Smoking is an excellent example; it leads to many effects in the smoker, in part because smoking involves exposure to a wide range of agents.[15,16] The existence of one effect of an exposure does not detract from the possibility that another effect exists.

On the other hand, Weiss[16] convincingly 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. Thus, specificity can come into play when it can be logically deduced from the causal hypothesis in question.

4. Temporality. Temporality refers to the necessity for a cause to precede an 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.

5. Biological gradient. Biological gradient refers to the presence of a unidirectional dose–response curve. We often expect such a

monotonic relation to exist. For example, more smoking means more carcinogen exposure and more tissue damage, hence more opportunity for carcinogenesis. Some causal associations, however, show a single jump (threshold) rather than a 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.

Alcohol consumption and mortality is another example. Death rates are higher among nondrinkers than among moderate drinkers, but 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.

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 biological gradient in its relation with disease. The noncausal relation between birth rank and Down syndrome mentioned in part 1 above shows a biological gradient that merely reflects the progressive relation between maternal age and Down syndrome occurrence.

These examples 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.

6. Plausibility. Plausibility refers to the biological plausibility of the hypothesis, an important concern but one that is far from objective or absolute. Sartwell, emphasizing this point, cited the 1861 comments of Cheever 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."[17] What was to Cheever an implausible explanation turned out to be the correct explanation, since it was indeed the vermin that caused the typhus infection. Such is the problem with plausibility: it is too often not based on logic or data, but only on prior beliefs. This is not to say that biological knowledge should be discounted when evaluating a new hypothesis, 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.[12] Nevertheless, the Bayesian approach cannot transform plausibility into an objective causal criterion.

7. Coherence. Taken from the surgeon general's report on smoking and health,[11] 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 gender, 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, presence of conflicting information may indeed refute a hypothesis, but one must always remember that the conflicting information may be mistaken or misinterpreted.[18]

8. Experimental evidence. It is not clear what Hill meant by experimental evidence. It might have referred to evidence from laboratory experiments on animals, or to evidence from human experiments. Evidence from human experiments, however, is seldom available for most epidemiologic research questions, and animal evidence relates to different species and usually to levels of exposure very different from those humans experience. From Hill's examples, it seems that what he had in mind for experimental evidence was the result of removal of some harmful exposure in an intervention or prevention program, rather than the results of laboratory experiments. The lack of availability of such evidence would at least be a pragmatic difficulty in making this a criterion for inference. Logically, however, experimental evidence is not a criterion but a test of the causal hypothesis, a test that is simply unavailable in most circumstances. Although experimental tests can be much stronger than other tests, they are often not as decisive as thought, because of difficulties in interpretation. For example, one can attempt to test the hypothesis that malaria is caused by swamp gas by draining swamps in some areas and not in others to see if the malaria rates among residents are affected by the draining. As predicted by the hypothesis, the rates will drop in the areas where the swamps are drained. As Popper emphasized, however, there are always many alternative explanations for the outcome of every experiment. In this example, one alternative, which happens to be correct, is that mosquitoes are responsible for malaria transmission.

9. 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 only reflects lack of imagination or experience, not falsity of the hypothesis.

Is There Any Use for Causal Criteria?

As is evident, the standards of epidemiologic evidence offered by Hill are saddled with reservations and exceptions. Hill himself was ambivalent about the utility of these "viewpoints" (he did not use the word criteria in the paper). On the one hand, he asked, "In what circumstances can we pass from this observed

association to a verdict of causation?" Yet despite speaking of verdicts on causation, he disagreed that any "hard-and-fast rules of evidence" existed by which to judge causation: This conclusion accords with the views of Hume, Popper, and others that causal inferences cannot attain the certainty of logical deductions. Although some scientists continue to promulgate causal criteria as aids to inference, others argue that it is actually detrimental to cloud the inferential process by considering checklist criteria.[19] An intermediate, refutationist approach seeks to transform the criteria into deductive tests of causal hypotheses.[20,21] Such an approach avoids 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.

CRITERIA TO JUDGE WHETHER SCIENTIFIC EVIDENCE IS VALID

Just as causal criteria cannot be used to establish the validity of an inference, there are no criteria that can be used to establish the validity of data or evidence. There are methods by which validity can be assessed, but this assessment would not resemble anything like the application of rigid criteria.

Some of the difficulty can be understood by taking the view that scientific evidence can usually be viewed as a form of measurement. If an epidemiologic study sets out to assess the relation between exposure to tobacco smoke and lung cancer risk, the results can and should be framed as a measure of causal effect, such as the ratio of the risk of lung cancer among smokers to the risk among nonsmokers. Like any measurement, the measurement of a causal effect is subject to measurement error. For a scientific study, measurement error encompasses more than the error that we might have in mind when we attempt to measure the length of a piece of carpet. In addition to statistical error, the measurement error subsumes problems that relate to study design, including subject selection and retention, information acquisition, and uncontrolled confounding and other sources of bias. There are many individual sources of possible error. It is not sufficient to characterize a study as having or not having any of these sources of error, since nearly every study will have nearly every type of error. The real issue is to quantify the errors. As there is no precise cutoff with respect to how much error can be tolerated before a study must be considered invalid, there is no alternative to the quantification of study errors to the extent possible.

Although there are no absolute criteria for assessing the validity of scientific evidence, it is still possible to assess the validity of a study. What is required is much more than the application of a list of criteria. Instead, one must apply thorough criticism, with the goal of obtaining a quantified evaluation of the total error that afflicts the study. This type of assessment is not one that can be done easily by someone who lacks the skills and training of a scientist familiar with the subject matter and the scientific methods that were employed. Neither can it be applied readily by judges in court, nor by scientists who either lack the requisite knowledge or who do not take the time to penetrate the work. ■

About the Authors

Kenneth J. Rothman is with the Boston University Medical Center, Boston, Mass. Sander Greenland is with the University of California, Los Angeles.

Requests for reprints should be sent to Kenneth J. Rothman, DrPH, Boston University School of Public Health, Department of Epidemiology, 715 Albany St., Boston, MA 02118 (e-mail: krothman@bu.edu).

This article was accepted November 18, 2004.

Contributors

Kenneth J. Rothman and Sander Greenland participated equally in the planning and writing of this article.

Acknowledgments

This work is largely abridged from chapter 2 of *Modern Epidemiology,* 2nd ed., by K.J. Rothman and S. Greenland, Lippincott, Williams & Wilkins, 1998, and chapter 2 of *Epidemiology—An Introduction* by K.J. Rothman, Oxford University Press, 2002.

References

1. Rothman KJ. Causes. *Am J Epidemiol.* 1976;104: 587–592.

2. Honoreé A. Causation in the Law. In: Zalta EN, ed. *Stanford Encyclopedia of Philosophy.* Winter 2001 ed. Stanford, Calif: Stanford University; 2001. Available at: http://plato.stanford.edu/archives/win2001/entries/causation-law.

3. Rothman KJ, Greenland S. *Modern Epidemiology.* Philadelphia, Pa: Lippincott; 1998: chap 18.

4. Higginson J. Proportion of cancer due to occupation. *Prev Med.* 1980;9:180–188.

5. Ephron E. *Apocalyptics: Cancer and the Big Lie—How Environmental Politics Controls What We Know about Cancer.* New York, NY: Simon and Schuster; 1984.

6. Higginson J. Population studies in cancer. *Acta Unio Internat Contra Cancrum* 1960;16:1667–1670.

7. MacMahon B. Gene-environment interaction in human disease. *J Psychiatr Res.* 1968;6:393–402.

8. Hogben L. *Nature and Nurture.* London, England: Williams and Norgate; 1933.

9. Horwitz RI, Feinstein AR. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *N Engl J Med.* 1978;299: 1089–1094.

10. Hill AB. The environment and disease: association or causation? *Proc R Soc Med.* 1965;58:295–300.

11. Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service. Washington, DC: US Department of Health, Education, and Welfare; 1964. Public Health Service Publication No. 1103.

12. Mill JS. *A System of Logic, Ratiocinative and Inductive.* 5th ed. London, England: Parker, Son and Bowin, 1862. Cited in Clark DW, MacMahon B, eds. *Preventive and Community Medicine.* 2nd ed. Boston, Mass: Little, Brown; 1981:chap 2.

13. Hume D. *A Treatise of Human Nature.* (Originally published in 1739.) Oxford University Press edition, with an Analytical Index by L.A. Selby-Bigge, published 1888. Second edition with text revised and notes by P.H. Nidditch, 1978.

14. Rothman KJ, Poole C. A strengthening programme for weak associations. *Int J Epidemiol* 1988;17(Suppl): 955–959.

15. Smith GD. Specificity as a criterion for causation: a premature burial? *Int J Epidemiol.* 2002;31:710–713.

16. Weiss NS:. Can the specificity of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology.* 2002;13:6-8.

17. Sartwell P. On the methodology of investigations of etiologic factors in chronic diseases—further comments. *J Chron Dis.* 1960;11:61–63.

18. Popper, KR. *The Logic of Scientific Discovery.* New York, NY: Harper & Row; 1959 (first published in German in 1934).

19. Lanes SF, Poole C. "Truth in packaging?" The unwrapping of epidemiologic research. *J Occup Med.* 1984;26:571–574.

20. Maclure M. Popperian refutation in epidemiology. *Am J Epidemiol.* 1985;121:343–350.

21. Weed D. On the logic of causal inference. *Am J Epidemiol.* 1986;123:965–979.