Hill's Criteria for Causality

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 **associations** in **observational 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 [3, 6]. Nevertheless, lists of causal criteria have become popular, possibly because they seem to provide a road map through complicated territory.

A commonly used set of criteria was proposed by **Sir Austin Bradford Hill** [1]; 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 [5] and the rules of causal inference given by Hume [3]. Hill suggested that the following aspects of an association be considered in attempting to distinguish causal from noncausal associations: strength, consistency, specificity, temporality, biologic gradient, plausibility, coherence, experimental evidence, and analogy. The popular view that these criteria should be used for causal inference makes it necessary to examine them in detail:

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 (*see* **Cornfield's Inequality**). Weak associations, on the other hand, are more easily 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.

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 nor sufficient for absence of causality. In addition to these counterexamples, we have to remember that neither **relative risk** nor any other measure of association is a biologically consistent feature of an association; as described by many authors [4, 7], it is a characteristic of a study population that depends on the relative **prevalence** of other causes. 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.

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 when other conditions are met, such as presence of certain bacteria. Consistency is apparent only after all the relevant details of a causal mechanism are understood, which is to say very seldom. 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.

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03072

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, especially by those seeking to exonerate smoking as a cause of lung cancer. The criterion is wholly invalid, however. Causes of a given effect cannot be expected to lack other effects on any logical grounds. 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. The existence of one effect does not detract from the possibility that another effect exists. Thus, specificity does not confer greater validity to any causal inference regarding the exposure effect. Hill's discussion of this criterion for inference is replete with reservations, and many authors regard this criterion as useless and misleading [8, 9].

Temporality

Temporality refers to the necessity that the cause precede the effect in time. This criterion is unarguable, 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 shows 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.

Biologic Gradient

Biologic gradient refers to the presence of a monotone (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 carcinogenesis. Such an expectation is not always present, however. The somewhat controversial topic of alcohol consumption and mortality is an example. Death rates are higher among nondrinkers than among moderate drinkers, but ascend to the highest levels for heavy drinkers. Because modest alcohol consumption can have beneficial effects on serum lipid profiles, such a J-shaped dose–response curve is at least biologically plausible.

Conversely, 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 noncausal relation between birth rank and Down syndrome mentioned above shows a biologic gradient that merely reflects the progressive relation between maternal age and the occurrence of Down syndrome.

Thus the existence of a monotonic association is neither necessary nor sufficient for a causal relation. A nonmonotonic relation only conflicts with those causal hypotheses specific enough to predict a monotonic dose–response curve.

Plausibility

Plausibility refers to the biologic plausibility of the hypothesis, an important concern but one that is far from objective or absolute. Sartwell [9], emphasizing this point, cited the remarks of Cheever, in 1861, who was 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, 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

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03072

can also provide a means of testing those quantified beliefs against new evidence [2]. Nevertheless, the Bayesian approach cannot transform plausibility into an objective causal criterion.

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 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, presence of conflicting information may indeed undermine a hypothesis, but one must always remember that the conflicting information may be mistaken or misinterpreted [12].

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 that 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 [10]. 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 epidemiologic circumstances.

Although experimental tests can be much stronger than other tests, they are not as decisive as often 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.

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.

Conclusion

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 "standards" (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?*" (original emphasis). 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 [criteria] can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*.

Actually, the fourth criterion, temporality, is a *sine qua non* for causality: If the putative cause did not precede the effect, that indeed is indisputable evidence that the observed association is not causal (although this evidence does not rule out causality in other situations, for in other situations the putative cause may precede the effect). Other than this one condition, however, which may be viewed as part

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03072

of the definition of causation, there is no necessary or sufficient criterion for determining whether an observed association is causal.

Acknowledgment

This article is adapted from Chapter 2 of *Modern Epidemiology* 2nd Ed. [8], with permission from the publisher.

References

[1] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.
[2] Howson, C. & Urbach, P. (1993). *Scientific Reasoning. The Bayesian Approach*, 2nd Ed. Open Court, LaSalle.
[3] Hume, D. (1978). *A Treatise of Human Nature* (originally published in 1739). Oxford University Press edition, with an Analytical Index by L. A. Selby-Bigge, published 1888. 2nd Ed. with text revised and notes by P.H. Nidditch, published 1978.
[4] MacMahon, B. & Pugh, T.F. (1967). Causes and entities of disease, in *Preventive Medicine*, D.W. Clark & B. MacMahon, eds. Little, Brown & Company, Boston.
[5] Mill, J.S. (1862). *A System of Logic, Ratiocinative and Inductive*, 5th Ed. Parker, Son and Bowin, London.
[6] Popper, K.R. (1968). *The Logic of Scientific Discovery*. Harper & Row, New York.
[7] Rothman, K.J. (1976). Causes, *American Journal of Epidemiology* **104**, 587–592.
[8] Rothman, K.J. & Greenland, S. (1997). *Modern Epidemiology*, 2nd Ed. Lippincott, Philadelphia, Chapter 8.
[9] Sartwell, P. (1960). On the methodology of investigations of etiologic factors in chronic diseases – further comments, *Journal of Chronic Diseases* **11**, 61–63.
[10] Susser, M. (1988). Falsification, verification and causal inference in epidemiology: reconsiderations in the light of Sir Karl Popper's philosophy, in *Causal Inference*, K.J. Rothman, ed. Epidemiology Resources, Inc., Boston.
[11] US Department of Health, Education and Welfare (1964). Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service, *Public Health Service Publication No. 1103*. Government Printing Office, Washington.
[12] Wald, N.A. (1985). Smoking, in *Cancer Risks and Prevention*, M.P. Vessey & M. Gray, eds. Oxford University Press, New York, Chapter 3.

(*See also* **Causation**)

KENNETH J. ROTHMAN &
SANDER GREENLAND

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a03072