

# THE EPISTEMOLOGICAL FUNCTION OF HILL'S CRITERIA

---

## Abstract

*Objective* This article outlines an epistemological framework for understanding how Hill's criteria may aid us in establishing a causal hypothesis (A causes B) in an observational study. *Method* We consider Hill's criteria in turn with respect to their ability or otherwise to exclude alternative hypotheses (B causes A; there is a common cause of A and B; there is no causal connection between A and B). *Results* We may classify Hill's criteria according to which of the alternative hypotheses they are able to exclude, and also on the basis of whether they relate to (a) evidence from within observational study or (b) evidence independent of that study. It is noted that no criterion is able to exclude the common cause hypothesis in a systematic way. *Conclusion* Observational studies are typically weaker than experimental studies, since the latter can systematically exclude competing hypotheses, whereas observational studies lack a systematic way of ruling out the common cause hypothesis.

*Keywords:* epidemiology; epistemology; causal inference; philosophy, medical

---

## 1. Introduction—knowledge of causation from observational studies

The focus of this paper is the observational study. The principal epistemological question is:

Q. Under what circumstances, if any, can an observational study give us knowledge of a causal relation between two variables?

While the corresponding question for experimental studies raises important epistemological issues also, the position for observational studies appears to be particularly confused. The well-known 'criteria' of causation articulated by Austin Bradford Hill (1965) aim so assist us in making causal inferences of the kind referred to in Q. Nonetheless those criteria are presented in an unsystematic manner, and subsequent discussion has focussed on the pros and cons of each individually rather than on seeking a general framework for their assessment. Below I articulate such a framework, the central idea of which is that a good criterion of causation is one that, when fulfilled, succeeds in eliminating potential error, i.e. it eliminates an alternative, false hypothesis.

## 2. Method—causal inference and the elimination of erroneous hypotheses

In both experimental and observational studies, the possibility of causation between factor A and health outcome B will be indicated by an association between A and B.

When such an association is observed there are four competing hypotheses that need to be considered:

- (T) A causes B (the target hypothesis);
- (R) B causes A (the reverse cause hypothesis);
- (C) A and B have a common cause (the common cause hypothesis);
- (N) There is no causal relation between A and B (the null hypothesis, that the association between A and B is pure chance).

The target hypothesis, that A causes B, (T), is the hypothesis we are testing. How do we know that (T) is true? I shall start from the premise that to know a hypothesis to be true we must at least have evidence that eliminates its plausible competitors. The importance of elimination of hypotheses has been emphasized by a number of authors (e.g. Popper 1959; Kitcher 1993; Papineau 1993; Bird 2005). In the best case we can come to know that (T) is true because we have eliminated each of (R), (C), and (N). This is known as *eliminative induction*. Consider the case where we accept (T), but competing hypotheses, e.g. (C), have not all been eliminated. Then we are open to possible error—(C) might turn out to be correct—and so our acceptance of (T) is imperfectly justified. A method or approach to causal inference can be evaluated by considering the extent to which it permits accurate, justified acceptance of (T) by providing reliable elimination of (R), (C), and (N).

Thus we may assess Hill's criteria and come to a more systematic answer to Q. by assessing each of Hill's criteria with respect to their ability to eliminate one or more of (R), (C), and (N).

### 3. Discussion—a systematic assessment of Hill's criteria

Hill lists nine criteria:

- Strength
- Consistency
- Specificity
- Temporality
- Biological gradient
- Plausibility
- Coherence
- Experiment
- Analogy

The criteria are intended to help us decide when we should interpret an observed association between A and B as showing causation from A to B, in the light of all our evidence. Some of that evidence will come from the observational study itself. This is *study-dependent* evidence. Other evidence comes from our background knowledge. This is *study-independent* evidence.

The criteria Strength, Consistency, Specificity, Biological gradient, Temporality all concern the relationship between a hypothesis and study-dependent evidence. Since these concern what the observational study itself tells us, they shall be the criteria of particular relevance to this paper.

The criteria Plausibility, Coherence, and Analogy all concern the relationship between the hypothesis study-independent evidence. That is, one can assess a hypothesis with respect to these three criteria without considering the nature of the evidence from the observational study in question. *Experiment* also concerns evidence independent

of the observational study, relating instead to independent evidence gained for example from an intervention in the environment, such as removing dust from the workshop, or from a randomized controlled trial. Since the intended contrast to Q. concerns the inferences that may be drawn from experimental studies, I shall not consider the criterion of Experiment further.

The first criterion, *Strength* can rule out the null hypothesis (N), when the degree of association is very great. Hill quotes his colleague Richard Doll (1964:333) in citing the statistic that chimney sweeps in the early twentieth century had a mortality rate from scrotal cancer some 200 times that of workers not exposed to a similar hazard. With such a strong correlation one can be confident that there is *some* causal connection between the occupation and the scrotal cancer—that is such a correlation can exclude (N). That kind of correlation does not come about by chance. It is true that one can calculate the probability that the observed statistic did come about by chance (if one knows the sample size), but Hill is clear that one does not need to make such a calculation in order justifiably to eliminate the null hypothesis, referring to Percivall Pott (1775) who first noted the connection and correctly drew a causal inference. In this case, as in the case of the causal connection between smoking and cancer one may be confident not only that there is a causal connection (i.e. (N) is eliminated), but also that the one knows the direction of causation (i.e. (R) is also eliminated). But that further knowledge is not afforded by satisfying the criterion of strength but by additional information, as we shall see.

The criterion of *Consistency* also works so as to rule out the null hypothesis (N), or at least to make it a poorer explanation of the observed association. Consistency, as Hill uses the term, is a matter of the same result being found by more than one study, especially when the studies are carried out in a variety of circumstances and by different observers. The simple fact of more studies and thus in effect a larger sample serves to make the same observed frequency more statistically significant, and so the probability of a chance association is reduced, i.e. reduces the probability of (N). That point is independent of the variety that Hill mentions. Consistency of results from a variety of observers makes less plausible certain sources of error that would produce a correlation even though the null hypothesis is true. Poor study design or implementation might produce such an outcome, for example, if the study permits selection bias. Repetition of the study by the same observers might reproduce the same errors, but this is less likely if association is reported by different observers. So the variety of observers reduces the chances of an error that would report causation even though (N) is true. Obtaining the same result under a variety of ‘places, circumstances and times’ can also help rule out (N) and so avoid the error of possibly ignoring a true null hypothesis. For example, one might observe a correlation between the introduction of anti-pollution measures, such as the U.K.’s Clean Air Act of 1956 and a reduction in mortality from respiratory disease. But that is consistent with that reduction being caused by some other factor that changed at around the same time. However, if major anti-pollution measures of a similar kind are introduced in different countries and at different times, we can regard such accidental (non-causal) sources of a correlation as less plausible.

*Specificity* also works so as to eliminate or reduce the plausibility of (N). Consider the case discussed by Hill (1962) (though not under the heading of specificity) concerning workers in a nickel refinery in South Wales. He and Richard Doll both found,

for differing periods, that workers in this refinery suffered from an unusual incidence of nasal cancer. Nasal cancer is rare: if the population under study had been representative of the population at large (adjusted for age), we would have expected less than one death from nasal cancer whereas in fact there were twenty-four. (Note that while Hill does not state that the local population from which the workers were drawn, excepting the workers themselves, had the same rate of cancer as the national population, that is implied by a fact he does cite: when the employees of the refinery were subdivided, it is only the workers in the chemical process who suffered from nasal cancer.) The null hypothesis in this case is that there is no causal connection between the employment in the nickel works and the incidence of nasal cancer. That absence of a connection would mean either (i) there is no connection between the cases of nasal cancer—this cluster is a pure chance aggregation of independent cases, or (ii) there is some connection between all or most of these cases, but that connection is causally independent of the circumstances of employment. (i) may be rejected on statistical grounds, but (ii) needs to be considered seriously. Here the idea is that there may be some common factor ‘X’ unusually prevalent among workers in the refinery and which disposes them to get nasal cancer; but it is chance that X is prevalent in the refinery: X is not caused by working in refinery factory, X does not cause people to work there, nor is there some common cause of X and working in the refinery. This is where specificity becomes relevant. Nasal cancer is rare in the general population, and so X must be correspondingly rare. But it is implausible that X should be very rare but also highly prevalent in the nickel refinery, but there be no causal connection between X and the refinery (in one direction or the other or a common cause). Furthermore, if there were such a factor, one might expect it to be apparent. Hence, where we have strong evidence of consistency we may be able to exclude the null hypothesis (N).

The most specific of Hill’s criteria is *Biological gradient*. Like several of the criteria it can provide useful evidence when present, but we should not expect it in every case. When present it can provide powerful evidence against chance correlation, i.e. against (N). In many studies we are presented with correlation between two-valued variables (receives therapy or not, recovers or not; gets water from Vauxhall and Lambeth Water Company or from Southwark Water Company, dies from cholera or not). Some such correlations in the data may come about by mere chance. But in Doll and Hill’s study the data was much more finely grained; they showed that there was a functional correlation between smoking and lung cancer—the more one smoked the more one is likely to get lung cancer. In effect this shows multiple correlations: between high consumption and high risk, between moderate consumption and moderate risk, and so forth. Such patterns of correlation are correspondingly less likely to arise together by chance.

The criterion of *Temporality* provides a basis for rejecting the reverse hypothesis, (R), if it is clear that A precedes B. For example, Germans born in 1920 or 1921 were shorter than those born before or after (as revealed by PoW reports). The natural hypothesis is that being born in these years of poverty is a cause of shorter height; its reverse hypothesis, that shorter height as an adult caused poverty while a baby or *in utero* is ruled out by the criterion of temporality. Fisher (1957) suggested that lung cancer might cause people to smoke, by soothing the irritation caused by the disease. But that hypothesis is ruled out by the fact that the data shows that the cancer victims

typically took up smoking decades before the disease afflicted them. Note that one researcher's (T) may be another's (R). This symmetry means that Temporality also states a necessary condition on (T).

*Plausibility* and *Coherence* are related. Plausibility is a positive quality—a proposed causal relationship is plausible if it is the sort of thing one might expect in the light of one's background knowledge. Coherence is negative—a matter of not conflicting with one's background knowledge. A hypothesis that is inconsistent with what one knows cannot be correct. Coherence is thus a necessary condition on any hypothesis. Hence it is an adequacy condition on (T); equally, it can also rule out (R). In this respect Coherence is like Temporality, with the difference that Coherence can be applied independently of (and so before) the study results whereas applying Temporality may depend on evidence revealed by the study.

Coherence must be applied carefully. For (T) cannot be ruled out just by being inconsistent with what we *believe* (as opposed to what we *know*). For example, Barry Marshall's hypothesis that peptic ulcer is often caused by a bacterium was held to be inconsistent with what people mistakenly thought they knew, that bacteria could not survive in the acid environment of the stomach. Of course, those critics did not *know* that the stomach is sterile (since it is not), and so Marshall's hypothesis was not ruled out by Coherence. But it might have so appeared to some commentators.

Unlike the other criteria so far discussed, Plausibility doesn't contribute to ruling out alternative hypotheses. Rather it makes the proposed hypothesis relatively more credible. And so although helpful, plausibility is not especially important. And as Hill points out that hypotheses that fail to be plausible can be correct. Semmelweis's hypothesis about the causation of puerperal fever seemed implausible to many of his contemporaries.

Hill's final criterion is *Analogy*. For example, just as smoking causes lung cancer, it may be that inhaling asbestos also causes lung cancer. The idea that a causal hypothesis is supported by the fact that an analogous causal relationship is known, can make that hypothesis a more credible explanation. The presence or lack of an analogy cannot eliminate a causal hypothesis or its competitors.

#### 4. Results—classifying Hill's criteria

The discussion of Hill's criteria in the forgoing section permits the following classification:

A. Concerning study-dependent evidence for (T):

- (i) *Strength, Consistency, Specificity, and Biological gradient* may all support (T) by helping to rule out (N).
- (ii) *Temporality* may support (T) by ruling out (R); since it could also rule (T), it is also a necessary condition on the acceptability of (T).

B. Concerning study-independent evidence for (T):

- (iii) *Coherence* may support (T) by ruling out (R); since it could also rule (T), it is also a necessary condition on the acceptability of (T).

- (iv) *Plausibility* and *Analogy* may make (T) more credible but cannot help prove it, since they do not contribute to the elimination of alternative hypotheses.

As discussed, in order to establish (T) we must eliminate (R), (C), and (N). But the above reveals that Hill's criteria only serve to rule out (R) and (N). They do not exclude (C). Hence the general problem with epidemiological studies is that they fail to exclude common cause hypotheses in a *systematic* manner.

Note that (C) is open ended—it states that there is *some* common cause. It is important to note that some of Hill's criteria may serve to eliminate specific common cause confounders, even if none can eliminate (C) altogether. For example, Fisher (1957) criticized Doll and Hill's (1950) study on the grounds that it failed to control for a possible confounding by a genetic predisposition both to enjoy smoking and to suffer from lung cancer. However, Biological Gradient makes this unlikely, for then the genetic factor is one that would have to be similarly gradated. But this seems implausible. Continuously varying biological characteristics, such as height, that are genetically dependent, are dependent on multiple genes, as Fisher (1930) himself hypothesized. It might be that a disposition to lung cancer could be induced by several genes; it is perhaps less plausible that a liking for tobacco could be the consequence of several genes operating in an additive way; but it is even less plausible that all or most of the tobacco-disposing genes should also be lung-cancer disposing. So the right data and the criterion of Biological Gradient can eliminate a specific way in which (C) might be true. Likewise the Consistency may show a specific common cause hypothesis to be false because the hypothetical common cause factor is not found in all the different times and places showing an association between A and B.

To eliminate (C) altogether requires that we are able to consider *all* the plausible common causes of the correlation and refute each of them. There can be uncertainty about whether we have achieved that—perhaps we cannot be sure that we have thought of all the plausible confounders. In such circumstances we may have shown (T) to be credible, but we will not have proved it. Nonetheless, in *some* studies where the where the data is especially rich and powerful, such as Doll and Hill's work on smoking and lung cancer, and in the light of study-independent background knowledge, all the potential common causes *can* be eliminated beyond all reasonable doubt.

Consequently, there is no systematic answer to Q, because there is no systematic elimination of all possible common causes. But that is consistent with certain particular studies being 'fortunate' enough in their evidence that they can give us knowledge of causation.

## 5. Conclusion

This study explains why certain of Hill's criteria are particularly useful in helping establish causation. These criteria can eliminate alternatives to our target hypothesis. However, none of Hill's criteria can eliminate the common cause hypothesis in a *systematic* way, although they may be able to eliminate specific confounders.

This contrasts with a well-conducted experimental study, e.g. a controlled clinical trial. In such a study (C) and (R) are straightforwardly ruled out, and the primary concern is the statistical analysis of the data to eliminate (N). Consequently, there is

a clear epistemological sense in which controlled trials are better than observational studies. The design of an experimental study such as a controlled trial will always permit the elimination of (C) but the design of an observational study cannot eliminate (C) in a systematic way.

That is not to say that the results of an actual controlled trial are *always* better than those from an observational study. A controlled trial may not have been properly carried out. Even when well-conducted, that study design cannot guarantee the absence of confounding in every case. And of course no study can guarantee that it will produce sufficiently strong evidence to come to a reliable conclusion. (See Worrall 2007 on the weaknesses of randomized studies.)

Nonetheless, the results above show that observational studies lack a feature that experimental studies possess, the ability to eliminate all common cause hypotheses; consequently this *disposes* (but does not guarantee) the latter to be of higher quality than the former. Hence the EBM hierarchy is correct to place experimental studies such as randomized controlled trials higher than observational studies. In epistemological terms, experimental studies will often permit an eliminative induction, whereas observational studies *typically* do not.

## 6. Conflict of Interest

The author declares that there is no conflict of interest.

## References

- Bird, A., 2005. Abductive knowledge and Holmesian inference, in: Gendler, T.S., Hawthorne, J. (Eds.), *Oxford Studies in Epistemology*. Oxford University Press, Oxford, pp. 1–31.
- Doll, R., 1964. Cancer, in: Witts, L.J. (Ed.), *Medical Surveys and Clinical Trials*. Oxford University Press, London, 2nd edition. pp. 333–49.
- Doll, R., Hill, A.B., 1950. Smoking and carcinoma of the lung: Preliminary report. *British Medical Journal* 2, 739–48.
- Fisher, R.A., 1930. *The Genetical Theory of Natural Selection*. Clarendon, Oxford.
- Fisher, R.A., 1957. Dangers of cigarette-smoking. *British Medical Journal* 2, 297–8.
- Hill, A.B., 1962. The statistician in medicine (Alfred Watson Memorial Lecture). *Journal of the Institute of Actuaries* 88, 178–191.
- Hill, A.B., 1965. Environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine* 58, 295–300.
- Kitcher, P., 1993. *The Advancement of Science*. Oxford University Press, New York.
- Papineau, D., 1993. *Philosophical Naturalism*. Blackwell, Oxford.
- Popper, K., 1959. *The Logic of Scientific Discovery*. Hutchinson, London.

Pott, P., 1775. Cancer scroti, in: *Chirurgical works of Percivall Pott*. L. Hawes, London, 2nd edition. pp. 63–8.

Worrall, J., 2007. Evidence in medicine and evidence-based medicine. *Philosophy Compass* 2, 981–1022.