

The Messed Lessons of Sir Austin Bradford Hill

Carl V. Phillips, MPP PhD
University of Texas School of Public Health
1200 Pressler St, Houston, TX, 77225
cphillips@sph.uth.tmc.edu, 713/500-9190
(corresponding author)

and

Karen J. Goodman, PhD
University of Texas School of Public Health
1200 Pressler St, Houston, TX, 77225
kgoodman@sph.uth.tmc.edu, 713/500-9268

revised November 2003
(relatively unchanged since September 2001)

Abstract

Austin Bradford Hill's famous 1965 paper contains several important lessons for the current conduct of epidemiology. Unfortunately, it is almost exclusively cited as the source of the "Bradford-Hill criteria" for inferring causation when association is observed, despite Hill's explicit statement that cause-effect decisions cannot be based on a set of rules. Overlooked are Hill's important lessons about how to make decisions based on epidemiologic evidence. He advised epidemiologists to not over-emphasize statistical significance testing given the possibility that systematic error is often greater than random. His compelling and intuitive examples point out the need to consider many costs and benefits when making decisions about health-promoting interventions. These lessons are as needed today as they were when Hill presented them, and they offer ways to dramatically increase the value of the science.

Keywords: epidemiologic methods; health policy; causal inference

One of the most significant and cited papers in public health research is Austin Bradford Hill's "The Environment and Disease: Association or Causation?",¹ Hill's 1965 Presidential Address to the Section of Occupational Medicine of the Royal Society of Medicine, where he presented what are now commonly called the "Bradford-Hill criteria." This paper ironically gains its fame as a checklist for inferring causation, though Hill had no such intention. Meanwhile, largely ignored are its potential contributions to critical methodological and policy issues.

Despite widely distributed and clearly stated advice to the contrary,² Hill's nine considerations for judging whether an observed association is a causal relationship are still frequently taught in epidemiology and discussed as "causal criteria." At a time when the discussion of the nature of causality is reaching new levels of sophistication in epidemiology,³⁻⁶ this is particularly unfortunate. Hill never used the term "criteria" and explicitly stated that he did not believe that any hard-and-fast rules of evidence could be laid down, stating that his nine "viewpoints"^{11.p. 299} were neither necessary nor sufficient for causation. His suggestions about intuiting causation are almost completely lost when his address is distilled into a list. (Interestingly, there are more extreme cases of a scholar's name being immortalized for something contrary to his beliefs. The "Coase Theorem" in economics, from the most cited article in that field,⁷ is usually invoked to make worldly claims, but much of Coase's work (including that paper) focuses on how the premise of the theorem is never true in the real world.)

The "criteria" are an intriguing subject for the history of science, including the question of why Hill's list seems more popular than others,⁸⁻¹⁰ and whether causal conclusions that explicitly appealed to the criteria have proven more likely to be borne out by subsequent evidence. (To our knowledge, there is no such validation study of causal criteria.) But the main

purpose of this commentary is not to join the extensive discussion of the history and merits of causal criteria. Suffice to say that Hill's list was a useful contribution to a young science that surely needed it at the time, but years after it should have become part of the historical foundation as an early rough cut, it is still being recited by many as almost natural law. Appealing in our teaching and epistemology to the untested "criteria" of a great luminary from the past is reminiscent of the "scientific" methods of the Dark Ages. Judging from his own caveats, we believe Hill would likely agree.

Our purpose is to call attention to the seldom-cited last page and a half of the article, which presents lessons that remain overlooked today.

Hill memorably warned about overemphasis on statistical significance testing, writing "the glitter of the t table diverts attention from the inadequacies of the fare."^{1, p. 299} The mistake of drawing conclusions from inadequate samples had been replaced with the mistake of treating statistical significance as necessary and sufficient for action. Despite the education of several new generations of researchers, there has been little improvement,¹¹ and a deluge of "significant" results confuses the public and decision makers.

One implication of Hill's advice is well understood. Emphasis on the p-value (let alone dichotomous statements of significance) has been soundly denounced for decades.^{12,13} Point estimates with confidence intervals are the preferred method in current textbooks¹⁴ and are generally reported, though in practice confidence intervals tend to be interpreted as mere tests of statistical significance.

Further inadequacy of the fare is less well appreciated, stemming not from the question of p-values vs. confidence intervals, but from systematic errors. No statistical test of random sampling error informs us about the possible impacts of measurement error, confounding, and

selection bias, and thus other methods are needed. Hill hints at this when he notes that one of his own studies,¹⁵ like most, had great potential for selection bias (though he does not use this term). In effect, he asks "why would I bother to do an exaggeratedly precise statistical test when I know that the other sources of error are likely so large?" Rather than emphasize low p-values, he concluded that simple cell counts made both random chance and plausible systematic error unlikely. We can do better than this now, with modern methods for quantifying multiple sources of error,¹⁶⁻²⁰ but Hill's approach stands as a warning about mistaking statistical precision (let alone significance) for validity.

Even as modern epidemiologic analysts become less dazzled by the t-table there is still a tendency to completely overlook Hill's other important insight – what to do with causal inferences once made. In his last few paragraphs, he offers an important commentary on the policy recommendations that flow from decisions regarding cause and effect in epidemiology. Since "our object is usually to take action",^{1,p.300} policy considerations are central to the importance of the science. While epidemiology has its roots in very specific policy questions ("can we do something about cholera outbreaks?"), epidemiologists have ambivalent attitudes towards the policy decisions associated with their research.²¹ Health researchers frequently justify expensive research based on immediate practical benefits, but go on to deny the need to assess the policy contributions, defending the value of science for its own sake (sometimes even as they issue press releases calling for policy responses).

Even when policy implications are presented explicitly, they are seldom carefully analyzed. One journal famously goes so far as to forbid the commonly tacked-on policy recommendations that conclude original research reports because policy analysis is too complicated and too serious to be an afterthought by researchers whose expertise lies

elsewhere.^{22,23} Judging from Hill's comments, he might prefer more careful policy analysis attached to epidemiologic research, rather than none at all (though it is not clear he could solve the challenge of fitting it into the standard 3000-word, single-result health research paper).

In order to take policy action, Hill argues, we ought to pay attention to the absolute costs and benefits of what we are studying. While it would clearly be reading too much into the text to suggest that he had a prescient vision of modern probability-weighted cost-benefit analyses²⁴ (he never even used those terms), he did call for "differential standards before we convict."^{1,p. 300} Moving another step beyond statistical testing, we need to consider more than the degree of certainty that there is *some* health hazard, and act based on the likely gains and losses, with or without statistical certainty. Hill points this out (in an example sufficiently ill-chosen that it may have contributed to his important message being ignored):

on relatively slight evidence we might decide to restrict the use of a drug for early-morning sickness in pregnant women. If we are wrong in deducing causation from association no great harm will be done. The good lady and the pharmaceutical industry will doubtless survive.^{1,p. 300}

Setting aside the impolitic dismissal of women's preferences and the unsupported assertion that there is no great harm at stake (as well as the irony of the rise, fall, and possible rehabilitation of the morning sickness drug, Bendectin) the underlying point might be the most important lesson for current health policy. Hill goes on to strengthen his argument:

On fair evidence we might take action on what appears to be an occupational hazard, e.g. we might change from a probably carcinogenic oil to a noncarcinogenic oil in a limited environment and without too much injustice if we are wrong. But we should need very

strong evidence before we made people burn a fuel in their homes that they do not like or stop smoking the cigarettes and eating the fats and sugar that they do like.^{1,p.300}

Hill clearly stated that the science and data analysis should not be influenced by what is at stake. But public health researchers should recognize that the stakes matter, and incorporate a consideration of them into their work. Our subsequent experience has made clear that the alternative is to leave the weighing of costs and benefits to the unreliable post-science political process.

The observation that the costs and benefits matter, despite being fairly obvious, is frequently – indeed, typically – overlooked in public health discussions. The popular decongestant phenylpropanolamine was banned on slight evidence without regard to the high cost to consumers; dietary recommendations are made without considering absolute benefits, let alone the cost to people of avoiding their favorite foods; and regulations are tremendously uneven in their cost effectiveness, to cite just three examples. The "policy recommendations" paragraph found in many health research papers sometimes quantifies medical costs, but almost never analyzes lifestyle, psychological, or productivity costs. It is even rare to find quantification of the absolute aggregate benefit that would result from a policy or behavioral change.

Making a good decision does not depend on having studies with confidence intervals that exclude the null. A best decision can be based on whatever information we have now, and indeed a decision will be made – after all, the decision to maintain the status quo is still a decision.^{20,24} Hill offered his clearest condemnation of over-emphasizing statistical significance testing, not when he discussed p-values, but when he concluded by saying,

All scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time.^{1,p. 300}

The pursuit of the low p-value (or confidence interval that excludes the null) leaves our society postponing apparently useful policy choices while we do more research to try to show what we already believe to be true. It also creates the incentive to use dubious methods (e.g., unstated multiple hypothesis testing, choosing models or transforming data to get "better" results) in order to squeeze out significant results. Those same methods can be used by those who would prefer to find no association to make real causal relationships disappear below the $p=.05$ horizon. Making the best of the knowledge we have would reduce such temptations. If epidemiologists help empower policy makers to ban an easily-replaced chemical when we believe there is about a 50-50 chance that it is a health hazard (based on an honest assessment of all uncertainty), then the payoff for fiddling with the data to show the certainty is a bit higher or a bit lower would be eliminated.

This would release us from the trap of letting ignorance trump knowledge. Regulators often fail to act because we have not yet statistically "proven" an association between an exposure and a disease, even when there is enough evidence to strongly suspect it. There is a growing movement to escape this mistake by making a similar mistake in the other direction: adopting precautionary principles, which typically call for regulation until we have "proven" a lack of causal association – a decision based on ignorance that merely reverses the default. If we can escape from the false dichotomy of "proven vs. not proven" that statistical testing creates,

and from the notion that causality can be definitively inferred from a list of criteria, we can make decisions based on what we do know rather than what we don't.

It would clearly overinterpret Hill's short paper to find modern cost-benefit analysis and uncertainty quantification, just as it does to find causal criteria. Several generations of advancement in epidemiology and policy analysis provide much deeper exposition of his points. But Hill still offers timeless insightful analysis of uncertainty and costs in epidemiology, and the lessons are particularly interesting because they remain inexplicably overlooked in what is probably the best known paper in the field.

REFERENCES

1. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 1965; 58:295-300.
2. Rothman KJ, Greenland S. Causation and causal inference. Rothman KJ, Greenland S. Modern Epidemiology. 2nd edition. Philadelphia: Lippencott-Raven, 1998: 7-28.
3. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol 2002;31(2):422-429.
4. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10(1):37-48.
5. Greenland S. Causality theory for policy uses of epidemiologic results. Murray CJL, Mathers C, Salomon J, Lopez AD, Lozano R. Summary Measures of Population Health. Cambridge: Harvard University Press/World Health Organization, in press.
6. Pearl J. Causality. Cambridge: Cambridge University Press, 2000.
7. Coase R. The problem of social cost. Journal of Law and Economics 1960; 3(1):1-44.

8. Surgeon General's Advisory Committee on Smoking and Health. Smoking and health: 1964. DHEW publication no. (PHS) 1103). Rockville, MD: United States Public Health Service, 1964.
9. Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am J Epidemiol* 1991; 133(7):635-48.
10. Weed DL. Causal and preventive inference. Greenwald P, Kramer BS, Weed DL, editors. *Cancer Prevention and Control*. Marcel Dekker, Inc., 1995: 285-302.
11. Greenland S. Preface to Reprint of "Hill AB. The environment and disease: association or causation?". Greenland S, Editor. *Evolution of Epidemiologic Ideas*. Chestnut Hill, Massachusetts: Epidemiology Resources Inc., 1987: 14.
12. Rothman KJ. A show of confidence. *N Engl J Med* 1978; 299:1362-3.
13. Poole C. Beyond the confidence interval. *Am J Public Health* 1987; 77:195-9.
14. Rothman KJ, Greenland S. Approaches to statistical analysis. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd edition. Philadelphia: Lippencott-Raven, 1998: 343-58.
15. Hill AB. Sickness amongst operatives in Lancashire spinning mills. *J Inst Actu* 1962; 88:178.

16. Phillips CV. Quantifying and reporting uncertainty from systematic errors. Under review.
17. Greenland S. Basic methods for sensitivity analysis and external adjustment. Rothman KJ, Greenland S. Modern Epidemiology. 2nd edition. Philadelphia: Lippencott-Raven, 1998: 343-58.
18. Phillips CV, Maldonado G. Using Monte Carlo methods to quantify the multiple sources of uncertainty in studies. 149. 1999:S17.
19. Phillips CV. Applying fully articulated probability distribution calculations. 152. 2000:S41.
20. Phillips CV. Our estimates are uncertain, but that is OK. 151. 2000:S41.
21. Greenland S. Probability versus Popper: elaboration of the insufficiency of current Popperian approaches for epidemiologic analysis. Rothman KJ, Editor. Causal Inference. Chestnut Hill, Massachusetts: Epidemiology Resources Inc., 1988: 95-104.
22. Rothman KJ. Policy recommendations in Epidemiology research papers. Epidemiology 1993; 4:94-5.

23. The Editors. Our policy on policy. *Epidemiology* 2001; 12(4):371-2.

24. Phillips CV. The economics of 'more research is needed'. *Int J Epidemiol* 2001; 30:771-6.