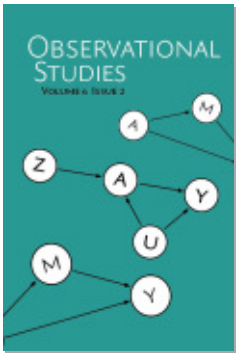
Following Bradford Hill

Mike Baiocchi



Observational Studies, Volume 6, Issue 2, 2020, pp. 11-16 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2020.0002*

Observational Studies 6 (2020) 11-16 Submitted 12/19; Published 1/20

Following Bradford Hill

Mike Baiocchi **baiocchi@stanford.edu**
Department of Epidemiology and Population Health
Stanford University
Stanford, CA 94305, USA

Abstract

In 1965, Sir Austin Bradford Hill offered his thoughts on: "What aspects of [an] association should we especially consider before deciding that the most likely interpretation of it is causation?" He proposed nine means for reasoning about the association, which he named as: strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy. In this paper, we look at what motivated Bradford Hill to propose we focus on these nine features. We contrast Bradford Hill's approach with a currently fashionable framework for reasoning about statistical associations – the Common Task Framework. And then suggest why following Bradford Hill, 50+ years on, is still extraordinarily reasonable.

Keywords: Causality, Bradford Hill Criteria, Common Task Framework

1. Reading with context

It feels odd writing about a paper that is more than 50 years old, particularly inside of a discipline that is currently undergoing extraordinary growth and innovation. But the "Bradford Hill criteria" (Bradford Hill, 1965) occupy a particularly prominent peak for those of us interested in making decisions that hinge on causal claims. To some, Bradford Hill laid out friendly signposts that suggest safer paths to achieving solid inferences about causal connections. To more, the "Bradford Hill criteria" are only stood up in order to be knocked right back down; the nine "criteria" are introduced and logical holes are punched through until students are left with the impression that hemming in causality with rules is a fool's errand. And yet, this paper persists. In fact, I was discussing this paper the other day with a colleague who told a fascinating story about when she served as an expert witness for a defense team in some legal case or another. The defense attorneys wanted her to work through the plausibility of each of the criteria. By report, it sounded like a deeply interesting (and lucrative) exercise in careful thinking. Her story got me curious so I dug into the legal world – and lo and behold – there are citations, and guides, and warnings about both deploying the Bradford Hill criteria in one's arguments before the court, as well as detailed guides on how to counter opposing counsel's expert witness's use of the criteria. These ideas seem to have sprouted legs and scurried out of our exclusive domain and into others, even while still kicking up heated exchanges in our own academic literature (Phillips and Goodman (2004), Höfler (2005), Phillips and Goodman (2006), Höfler (2006) – as you might be able to guess from the alternation of authors, that's a fun exchange). So what's going on here?

If you encountered Bradford Hill's ideas in a setting disconnected from the original manuscript then it may help unlock a bit more of his meaning by considering his audience. Bradford Hill first gave his remarks to the newly formed Section of Occupational Medicine. In context, these ideas were offered to medical practitioners charged with making complex decisions about health but with an eye toward bottom line economics, employment dynamics, and (to some degree) consumer tastes. His audience was not the usual data-analysts we think of, concerned with uncertainty intervals and p-values. Bradford Hill tells us this directly as he sets the table: "[Suppose] our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?" Using modern terminology, we might say Bradford Hill is quite a bit less interested in statistical inference and more interested in study design considerations. Though even that terminology does not get quite at the nub of his line of reasoning.

To get closer to the flow of his arguments, let's jump over all the important bits in the middle and pull from his concluding section: The Case for Action. Again, letting the man speak for himself: "Finally, in passing from association to causation I believe in 'real life' we shall have to consider what flows from that decision... In occupational medicine our object is usually to take action... While that is a commendable ambition it almost inevitably leads us to introduce differential standards before we convict." What follows is a discussion of balances – how does the strength of evidence enter into the decision to forbid/compel people to take actions? The size of the effect, levels of certainty, chains of consequences that may arise from our (in)actions – these all need to be weighed out. If he stopped his reasoning at that level of thought then this manuscript likely would have resolved into some kind of call for a better decision-theoretic framework. But that's not where Bradford Hill took the argument, and this is why this paper is still fascinating all these years later. As far as I can discern, there are two additional tensions he is tracking. The first is the tension he highlights quite a bit which is the need for action right now, which he contrasts with the academic's slower, more careful building of evidence toward solid, scientific conclusions. More interesting (at least to me and my contemporary eyes) is his focus on convincing people.

2. Reasoning with context

There are several ways to think about what we – those of us who are interested in making empirically rigorous, positive change in the world – are doing. Perhaps we are mathematizing the scientific method, providing crisp, quantified boundaries on what can be known and how best to empirically know it. Maybe we merely clear away the rubbish others bring into the Ivory Tower. These are recognizable roles we play in academic settings. But Bradford Hill is reminding us of a more fundamental role: we do all this to convince people. We may believe that rigorous evidence will compel, but it won't. Look at the absurdity of climate-change denial, or the rates of anti-vax. Change does not – exclusively – arise from rigorous empirical conclusions.

One way to understand the challenge of convincing people is to unpack why we tend to formalize and create decision rules. The first reason we formalize is to make discovery

more productive. When a new causal discovery emerges it often feels a bit shocking and it's natural to marvel at its departure from what has come before. For us, data-analysts, we tend to focus on the methods by which the discovery was achieved – which can be even more novel than the underlying causal discovery. But quickly, what used to seem novel and cutting-edge in science gets pulled into the core – what used to be "art" becomes codified and reproducible. This process is wildly productive, allowing many researchers to explore new directions that previously only the cleverest could. The second reason we tend to formalize requires some insight about how humans make decisions and assign responsibility: if we can formalize these kinds of causal-discovery methods, then we shift the burden of responsibility for declaring discovery outside of the individual (e.g., located in the idiosyncrasies of both the situation under investigation and the researcher making the claim) and into the general (e.g., rules that are recognizable across settings but also rules that have buy-in from the researchers or policy makers or stakeholders in our domain who will be impacted by our empirical investigations). Formalized decision rules that standardize discovery and regulate our claims on the strength of conclusions – in a manner that is much like laws – help communities set standards and make planning and settling disagreements easier and less arbitrary. In fact, looking at several of Bradford Hill's "criteria" with modern eyes, you can see how his insights have been formalized with statistical methodologies (to pick a few): (i) "strength of effect" looks quite a bit like the thinking behind modern sensitivity analysis addressing unobserved confounders; (ii) "consistency" looks like meta-analysis, replicability, and transportability of effect; (iii) even the initially surprising appeal to "analogy" that he suggests has found some formalization in the work on "transfer learning" in machine learning. These additions to our formalized tool set are great.

But, again, formalization is not what Bradford Hill is interested in. In fact, he takes some wonderfully cheeky shots at formalized decision making. He suggests that using statistical processes allows decision-makers to obscure and shirk their critical responsibilities. The tension that keeps Bradford Hill's argument fresh is the one that makes many of us excited about doing our jobs: figuring out how to bring new discoveries to the larger community. When a debate is vital and complex, when the stakes really matter then how do we reach solid conclusions that will be strong enough to win over our colleagues and those impacted by our conclusions? If you're in the position Bradford Hill was, talking to a room full of physicians interested in Occupational Medicine, then you'll understand the tension between the kinds of formalized rules that rigorous statistical analysis provides and the kinds of arguments used to convince and debate in the larger (less technical) community. A concrete example: given our analysis, we believe we should order a popular agricultural product removed from the market. If we decide to act on this belief then a new rule will come into existence. There will be many "losers" in this new regime, and a number of them will need to change their behaviors. How do we explain this rule? How do we get buy-in for this rule? The less familiar the logic used to create the rule is to the people on the receiving end then the less likely they will engage in the required change in behavior.

For a moment, let's pause this unpacking of Bradford Hill's manuscript and move forward to approximately contemporary time. One of the dominant modes of reasoning in contemporary data analysis – the Common Task Framework (CTF) – serves as an extraordinary contrast to Bradford Hill's ideas; it's worth exploring the CTF to better understand Bradford Hill.

3. Reasoning without context

The productivity, and explosive improvements, in statistical prediction ("machine learning") has rightly stood out in fields touched by data science. (If we're being honest then to some degree it has also caused some feelings of anxiety inside those of us less inclined to flashy prediction, and more enamored of the slower accumulation of information in fields interested in causal inference.) If, like me, you are less familiar with the field of prediction then you're likely even less familiar with the epistemological engine that has powered its growth. The Common Task Framework (Liberman (2015); Donoho (2017)) provides a fast, low-barriers-to-entry way for analysts to debate which algorithm performs "best" on a given data set. The CTF is an alternative way of assessing the suitability of an algorithm; it stands in contrast to the more traditional methods like mathematical theorems or simulations from a given data generating function. Even if the CTF name is unfamiliar you've likely heard of this dynamic; the NetFlix Prize (Bennett and Lanning, 2007) was an excellent example of this framework. The key features of the CTF are (slightly modified from Donoho (2017)):

1. A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and an outcome for that observation.

2. A set of enrolled competitors (analysts) whose common task is to infer a prediction rule from the training data.

3. A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule.

All competitors share the common task of creating prediction rules which will receive a good score; hence the phase common task framework. The performance metric provides an ordering that gives analysts permission to claim "this algorithm provides useful insights when used on this data set" – such claims are strongest when framed relative to other algorithms. This is where the "leaderboard" style of algorithmic development came from.

Obviously, predictive models are not causal models. But it is not hard to find colleagues who have become a bit too enamored of the predictive power of this or that fantastical black box – believing a bit too much in its ability to accurately describe all possible dynamics of the data. For these folks, it is a small leap of faith to using a model like this to try answering questions about what will happen if we intervene. (Dear reader, I assure you: it pains me too.) How? Perhaps they use something like predictive margins (see Graubard and Korn, 1999) – first, setting all the observational units level to unexposed, second setting all the observational units level to exposed, and then contrasting the two hypothetical groups' outcomes. Hidden behind almost all actions taken after consulting a black box is a confidence in its ability to faithfully describe all potential configurations of the data. But where did their confidence in the model come from?

The CTF allows fantastically complex, "black box" algorithms to be developed and (in a particular sense) evaluated. Without the CTF, complex algorithms – so complicated that they cannot be described mathematically – would have much weaker evidence to be trusted and thus deployed. With the CTF, we can see the performance of any algorithm vis-à-vis

any other algorithm on the same data set. The CTF has allowed algorithm developers to be extremely productive, principally by being able to avoid both slow moving math as well as the kind of deeper engagement with nuanced issues that gave rise to the data that traditional causal inference analysts do. Algorithms that grow up in the CTF aren't required to be accountable to slow-moving, tradition-bound, coherence-seeking people. In fact, there's a very explicit line of argument inside some data communities that "human experts need to be removed from the decision-making process" – rather, the machines should do the learning because they are more likely to produce the most optimal results. The thinking goes: Humans are slow. Humans are hard to understand. Let's remove humans from this process.

But here's the thing, when we stand with Bradford Hill, humans are the point. We're trying to convince humans to change. Rules that are (in a particular sense) "optimal" are not the same as rules that are useful for affecting change. In fact, it's easy to imagine that rules generated by "black box" algorithms (i.e., literally inexplicable) are less likely to be complied with than rules reached through consensus building and through reasoning accessible by those who are being asked to have their lives shaped by the rules. Do not mistake what I'm saying as arguing against well-reasoned, formal, quantified rules that come out of statistical analyses. Our rigorous statistical methods are the strong bones of the beast, but they don't provide the heart, mind, and muscles that animate and make these decisions human. We are better now that we have statistical procedures that formalize many of Bradford Hill's criteria. But these new statistical procedures do not solve the principal issue Bradford Hill was engaging, how to convince and change.

4. Following Bradford Hill

I didn't introduce the CTF to either bury or praise it, but rather because it is a perfect example of how one might reason about data in a way that is about as far removed as possible from the way Bradford Hill advocates we reason about data. The contrast here, I hope, helps illuminate the point that Bradford Hill was making. When I read this manuscript, I see someone making tough, impactful decisions in the presence of uncertainty. He is steeped in the particulars of the situation. While formalizing Bradford Hill's criteria is useful, and will produce better decision-making, it is also beside the point. (And, in the most extreme, can lead to a type of blindness about the role of experts, stakeholders, and consequences for our analyses.) The criteria are paths of reasoning about causality that resonate and reassure. In the kinds of questions epidemiologists, health policy researchers, economists, criminologists... engage, the ultimate audience is a community of people who our conclusions impact. Statistical reasoning can be like mathematics at times, but in answering these kinds of questions it is better to think of statistical reasoning as a form of rigorous, quantitative argumentation – meant to guide thought and shift beliefs.

I have a friend that keeps a copy of Bradford Hill's criteria pinned to his office wall. He uses it to remind himself of the paths he might take. I like that. If you follow Bradford Hill then I think you'll have an easier time reaching your audience.

Acknowledgments

I would like to Jordan Rodu for conversations about the CTF and the cheese plate.

References

Bennett, J. and Lanning, S. (2007). The netflix prize. *Proceedings of KDD cup and workshop*, 2(3):35.

Bradford Hill, A. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58(2):295–300.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 2(26):745–66.

Graubard, B. and Korn, E. (1999). Predictive margins with survey data. *Biometrics*, 55(2):652–9.

Höfler, M. (2005). The bradford hill considerations on causality: a counterfactual perspective. *Emerging Themes in Epidemiology*, 2(1).

Höfler, M. (2006). Getting causal considerations back on the right track. *Emerging Themes in Epidemiology*, 3(1).

Liberman, M. (2015). Reproducible research and the common task method. *Simmons Foundation Lecture*.

Phillips, C. and Goodman, K. (2004). The missed lessons of sir austin bradford hill. *Epidemiologic Perspectives & Innovations*, 1(1).

Phillips, C. and Goodman, K. (2006). Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. *Emerging Themes in Epidemiology*, 3(1).