Preventive Medicine 53 (2011) 235–238

Contents lists available at SciVerse ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Causality in medicine: Getting back to the Hill top

John Worrall

Department of Philosophy, Logic & Scientific Method, London School of Economics, London WC2A 2AE, UK

ARTICLE INFO

Available online 17 August 2011

Keywords:
Evidence
Evidence based medicine
Causality
Randomization
Austin Bradford Hill

ABSTRACT

Evidence from randomized controlled trials (RCTs) is almost universally regarded as setting the "gold standard" for medical evidence. Claims that RCTs carry special epistemic weight are often based on the notion that evidence from randomized studies, and only such evidence, can establish that any observed connection between treatment and outcome was *caused* by the treatment on trial. Any non-randomized trial, on the contrary, inevitably leaves open the possibility that there is some underlying connection independent of receiving the treatment between outcome and one or more differentiating characteristics between those in the experimental and control groups; and hence inevitably leaves open the possibility that treatment and an observed better outcome were "merely correlated" rather than directly causally connected. Here I scrutinize this argument and point towards a more tenable and more modest position by recalling some of the forgotten insights of the RCT pioneer, Austin Bradford Hill.

Within medicine evidence from randomised controlled trials (RCTs) is almost universally regarded as setting the "gold standard." Claims that RCTs carry special epistemic weight are often based on the notion that evidence from randomised studies, and only such evidence, can establish that any observed connexion between treatment and outcome was caused by the treatment on trial. Any non-randomised trial, on the contrary, inevitably leaves open the possibility that there is some underlying connexion independent of receiving the treatment between outcome and one or more differentiating characteristics between those in the experimental and control groups, and hence inevitably leaves open the possibility that treatment and an observed better outcome were "merely correlated" rather than directly causally connected. (From the philosophy of science and computer science literature see, e.g., Cartwright (1989), Papineau (1994), Pearl (2000)—and, for critical analysis, Worrall (2007).) By contrast, "In a randomised trial, the only difference between the two groups being compared is that of most interest: the intervention under investigation."[1] Here I scrutinise this argument and point towards a more tenable and modest position by recalling some of the forgotten insights of the RCT pioneer, Austin Bradford Hill. (Hill was principal investigator in the first ever RCT—that of streptomycin for the treatment of tuberculosis and was undoubtedly the most influential medical statistician of the 20th Century. Although his "criteria" for causation are frequently cited, I argue here that some of his insights have been missed by later proponents of evidence based medicine.)

Clearly the argument that only RCTs can establish causation is a re-expression of the argument that RCTs (and again only RCTs) "control for all confounders, known and unknown." If "background knowledge" suggests that factors other than treatment (e.g. age) may play a role in outcome, then this can, in principle, be deliberately controlled for, but, however much deliberate matching has been done, the spectre of the "unknown confounder" still haunts the scene: perhaps there is some further factor that is the "real cause" (or part of it) but has so far not been thought of. The suggestion is that by randomising you solve this problem. If all possible confounders really were controlled for by randomising, then every other possible explanation of a positive result in a given trial would be eliminated and it could inexorably be inferred that that result had to be caused by the intervention under investigation. It is not difficult to find definite claims to this effect in the literature. Aside from Clarke, Sir Michael Rawlins (2008)writes "The greatest strength of an RCT is that the allocation of the treatments is random so that the groups being compared are similar for baseline factors."

But the claim is false—as everyone, including Clarke and Rawlins, really allows. (Both eventually make it clear that they defend only a much weaker claim and would perhaps regard the quoted claims as harmless simplifications. However the simplifications have caught on and helped create a climate in which the RCT is taken to carry much more weight than it really can. This in turn has led to some very doubtful ethical decisions (Worrall, 2008)). No one really believes that, given a particular random division, the groups are bound to be equivalent in all other respects and hence that any difference in the outcome is automatically attributable to the difference in treatment. No one really believes that having randomised is sufficient to establish that any observed effect must be due to the treatment. In any particular randomised division, it is of course entirely possible that

E-mail address: J.Worrall@lse.ac.uk.

[1] Mike Clarke, http://209.211.250.105/docs/whycc.htm; accessed 18 December 2008.

0091-7435/$ – see front matter
doi:10.1016/j.ypmed.2011.08.009

some factor is unbalanced between the two groups and hence this lack of balance remains a possible rival explanation for the positive result.

This is implicitly conceded by most orthodox accounts of RCT methodology. In trials where no attempt has been made to match with respect to "known" prognostic factors, investigators are (usually) recommended, before drawing any evidential conclusions, to look at the particular division into control and experimental groups that randomization has yielded and check for "baseline imbalances"—inequalities between the two groups in some factor (age, sex, co-morbidities…) that background knowledge tells us might play a causal role. If such baseline imbalances are found then the recommendation is to re-randomise in the hope that this time no baseline imbalances will occur. But if an imbalance is possible in known factors, despite impeccable randomization, then it must equally be acknowledged that there may be an imbalance in unknown confounders. The difference being of course that investigators cannot check for imbalance in unknown confounders.

An amusing example is provided by Leibovici (2001). This study identified 3393 patients who had a bloodstream infection of some sort whilst inpatients at the Rabin Medical Centre during 1990–6. In July 2000 (so at least 4 years after these patients had been in hospital), a random number generator was used to divide them into two groups. Which of the two became the treatment group was decided by a coin toss. 1691 were randomised to the intervention group and 1702 to the control. A check was made for baseline imbalances with regard to main risk factors for death and severity of illness. None having been found, the names of those in the intervention group were given to a person "who said a short prayer for the well being and full recovery of the group as a whole."

Mortality, length of stay in hospital and duration of fever were then recorded from the hospital notes and compared in the two groups. Mortality was 28.1% in intervention group and 30.2% in the control group; this was "not significant" according to the usual significance testing methodology. However both length of stay in hospital and duration of fever were significantly shorter in the intervention group ($p = 0.01$ and $p = 0.04$)! The study concluded (tongue in cheek) that

> Remote, retroactive intercessory prayer said for a group is [causally] associated with a shorter stay in hospital and shorter duration of fever in patients with bloodstream infection and should be considered for use in clinical practice.

But even those who believe that god moves in mysterious ways are hardly likely to believe that they are mysterious as this! The reason why the result is not to be taken seriously reveals a further lesson: that we are all naturally (at least a little bit) Bayesian. As Leibovici (2002) himself subsequently wrote,

> If the pre-trial probability is infinitesimally low, the results of the trial will not really change it, and the trial should not be performed. This, to my mind, turns the article into a non-study, though the details provided (randomization done only once, statement of a prayer, analysis, etc.) are correct.

But again no one really believes (do they?) that randomization guarantees that the groups are similar in all other respects and hence that a positive result in a properly randomised trial is sufficient for a treatment to be declared effective. Well actually I think lots of people in medicine do believe this, because this is what they think they are being told by the experts. As we saw, many people, Mike Clarke included, certainly sometimes say that they believe it (even though they qualify it later), but I agree that the only claim that can seriously be defended is of some sort of probabilistic quasi-guarantee.

But what exactly could this amount to? The "guarantee" in fact seems to involve a slip from what is arguably true in the indefinite long run to a claim about what is true of a particular random allocation. An enormous amount of effort has gone into the attempt to make sense of single case probabilities on an objective view of probability. (This is in distinction to the "subjective" Bayesian view of probabilities as degrees of belief for which the "single case" presents no problem.) The only sustainable objectivist view seems to be the frequency interpretation, but then the claim that there is a high probability that the experimental and control groups are balanced with respect to some particular factor really amounts to the claim that if one were to take some group and divide them into two by some random procedure and if one were then to randomise again and then again … keeping a cumulative total for the relative frequencies of patients exhibiting this factor in the two groups (and forgetting about the fact that these different trials would not be independent!) then in the indefinite long run the limiting frequency of this factor within both the experimental and control groups would be the same. But we are never in the long run; medical researchers only randomise once, and in that one random allocation, the two groups can be as unbalanced with respect to the factor at issue as you like.

Of course this does not mean that randomization is of no use (on the contrary as we shall see in moment), but it does mean that it is a mistake to make a fetish of it—a positive result in an RCT does not establish causality: nothing can. Instead we need to examine every result whether from an RCT or any other study with what Austin Bradford Hill (1963) called "the fundamental question" in mind. That "fundamental question [is]—is there any other way of explaining the set of facts before us [in our current case the facts supplied by the results of some RCT], is there any other answer equally, or, more likely, than cause and effect?" The Leibovici case shows that sometimes the answer to this fundamental question will be—"there must be such a more likely explanation even though we presently cannot specify it!" Background knowledge tells us that there is no way that a prayer said for patients some years later can have had any effect on their recovery from bloodstream infections now, so no matter how perfectly randomised the trial, no matter how large the trial, no matter how "statistically significant" the result, we take the result of the trial as no sort of evidence for the effectiveness of the treatment.

Call this "exercising judgement" if you like (see Rawlins, op.cit), but it is surely not unanalysable judgement. We already know a lot about the world ahead of any particular trial and it would be folly indeed to ignore what we know (even accepting the ever present defeasibility of our knowledge). Fisher's insistence on not bringing any prior information into the assessment of the impact of a stochastic experiment in order to guarantee objectivity was an understandable, but egregious error. (Of course Bayesians have not helped by insisting on calling the extra factors "subjective", since introducing subjectivity was exactly what Fisher feared, but it is not merely a subjective opinion to hold that prayer can have no retroactive effect!)

We should use these reflections to try to build a more measured account of the evidential virtues of both randomised and non-randomised studies. Many aspects of this account are already to be found in Hill's work.

Keeping Hill's "fundamental question" in mind, it is easy to see that RCTs have one undoubted advantage: randomising means that the clinician has no control over which of the two groups in a trial any particular patient goes into. However this is not controlling for all confounders, but rather for a particular (possible) confounder which background knowledge gives us reason to think may play a role—that of selection bias (narrowly construed as the possible bias introduced through clinicians' selections of the two groups).

Indeed Hill (1971, p. 255) never claims that RCTs "control for all confounders"; their only virtue is elimination of selection bias—though he splits the one virtue into three.

> Faithfully adhered to [randomisation] offers three great advantages: (1) it ensures that our personal feelings or judgements,

applied consciously or unconsciously, have not played any part in building up the various treatment groups; from that aspect, and therefore, the groups are unbiassed;(2) it removes the very real danger, inherent in any allocation which is based upon personal judgements, that believing our judgements may be biassed, we endeavour to allow for that bias and in so doing may "lean over backwards" and thus introduce a lack of balance from the other direction; (3) having used such a random allocation we cannot be accused by critics of having set up personally biassed groups for comparison.

However, Hill also allows that selection bias is only a plausible rival explanation when the outcome effect is small. Where the outcome at issue is at all substantial then not only is randomisation unnecessary, so also is the use of any formal statistical test of significance. One of his investigations was into whether there was a causal connexion between working conditions in the card rooms of mid-20th century cotton mills and certain kinds of illness, and he records that he arrived at a very definite (positive) conclusion on the basis of the evidence.

Yet I cannot find anywhere I thought it necessary to use a test of significance. The evidence was so clear cut, the differences between the groups were mainly so large, the contrast between respiratory and non-respiratory causes of illness so specific, that no formal tests could really contribute anything of value to the argument. So why use them? (Hill, 1963, p. 297)

The mirror image of EBM's often exaggerated view of the epistemic virtues of RCTs is its overly pessimistic view of what can be reasonably evidentially established via non-randomised, and indeed non-experimental, studies. It became an article of faith within EBM (largely on the basis of some 1980s meta-studies) that non-randomised studies have a constant tendency to be more positive than "proper" (i.e. randomised) studies. But aside from the obvious circularity (the "real effect" is taken to be identified by the RCT!), this argument is based on particular observational studies that were obviously flawed—ones in which the historical control group patently failed to match the experimental one. Of course historically controlled studies must be analysed with Hill's fundamental question in mind—could it plausibly have been something else that explains the difference between the outcome with the new treatment and the historical outcome? Hill held that we could sometimes reasonably answer the question positively on the basis of such studies.

First, he enthusiastically endorsed Claude Bernard's view that there is no qualitative epistemic difference between experiment and (properly scientific) observation.

… it is imperative that we draw no precise line between observation and experiment. It is just 100 years since the great experimentalist Claude Bernard … wrote: "a physician observing a disease in different circumstances reasoning about the influence of these circumstances and deducing consequences which are controlled by other observations—this physician reasons experimentally even though he makes no experiments" (Hill, 1966, p.108).

Then he confronted the "fundamental question": we have the "facts before us"—whether it be an observed association between an environmental factor, like the working conditions in a particular part of a mid-20th century cotton mill, and increased prevalence of some disease amongst those working under those conditions, or the facts, say, about increased recovery rates amongst those given some new treatment relative to some earlier treated group or, indeed the result of an RCT—do these facts provide good evidence of a causal relationship between environmental factor/treatment and outcome or is the relationship more likely to be a "mere" association? Both possible answers are clearly consistent with the facts, so further evidential reasoning is needed if we are to be on as safe ground as possible. Of course if we have some rock solid background science that is relevant then no one would deny that this must be taken into account. If we had had for example a well-evidenced account of the biochemical pathways that lead from hot tarry smoke impinging on the lining of the lung to the development of tumours (subject to some initial conditions about the smoker and the amount he smokes) then there is a convincing causal link between cigarette smoking and lung cancer, independently of any statistical data. However such clear cut further evidence will seldom be available. This is where Hill (1963) articulates his famous "criteria" (though he was of course explicit that they were not in fact criteria and he himself used the term "viewpoints"). I take his main point to be simply that there may well be elements of background knowledge which at least assist us in supporting, of course always defeasibly, one answer or the other.

Background knowledge tells us, for instance, that many, though by no means all, causal relationships are linear (or reasonably close to it over reasonable ranges). So if the evidence is not only that more smokers develop lung cancer, but also that the heavier smokers develop more lung cancers than the lighter smokers, then this clearly strengthens the case for a causal connexion. This is Hill's "biological gradient""viewpoint."

What is the difference between the connexion between ashtray ownership and cancer (not Hill's example but it makes the point) and that between smoking and cancer? In each case we have of course a correlation, but background knowledge tells us that it is, to say the least, very unlikely that there is a (deterministic, non probabilistic) mechanism linking owning glass, pottery, plastic or stainless steel objects of various types and lung cancer; on the other hand, even in the absence of detailed confirmed "mechanisms" linking inhaling hot tarry smoke on a regular basis and the development of tumours in the lining of the lung, background knowledge does tell us that it seems plausible indeed that there might be such a link. This reasoning is a combination of Hill's "[biological] plausibility" and his "coherence" viewpoints.

As Hill is at pains to insist, reasoning based on his viewpoints is of course defeasible: there are well known cases where no mechanism was known and the link consequently regarded as "merely" associational, where a causal link was later found, and conversely cases in which background knowledge made a causal link plausible that was subsequently shown to be a case of association. This is why these are not criteria, why they cannot be regarded as "hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect" (ibid.), but that is the nature of science and, as Hill again perspicaciously points out, we cannot absolve ourselves from making decisions about the way in which the evidence points just because we may be wrong.

Many hard line EBM-ers will dislike the idea of bringing this sort of judgement based on background knowledge into what they would like to be the rules of evidence, but they are seeking the unattainable. As Michael Rawlins points out (op.cit) – unconsciously echoing Hill – judgement is inevitable. We can get some insights into how such judgement should be constrained from Hill's work: EBM needs to get back to the Hill top.

Conflict of interest statement
I declare that there are no conflicts of interest relevant to this article.

References

Cartwright, N.D., 1989. Nature's Capacities and their Measurement. Oxford University Press.
Hill, A.B., 1963. The environment and disease: association or causation? Proc Roy Soc Med 58, 295–300.

Hill, A.B., 1966. Reflections on the controlled trial. Ann. Rheum. Dis. 25, 107–113.
Hill, A.B., 1971. Principles of Medical Statistics, 9th Edition. The Lancet, London.
Leibovici, L., 2001. Effects of remote, retroactive, intercessory prayer on outcomes in patients with bloodstream infection. BMJ 323, 1450–1451.
Leibovici, L., 2002. Author's reply. BMJ rapid response. http://www.bmj.com/cgi/content/full/324/7344/1037#resp82002.
Papineau, D., 1994. The virtues of randomization. BJPS 45, 437–450.
Pearl, J., 2000. Causality—Models, Reasoning and Inference. Cambridge University Press.
Rawlins, M.D., 2008. De testimonio: on the evidence for decisions about the use of therapeutic interventions. The Harveian Oration 2008. Royal College of Physicians, London.
Worrall, J., 2007. Why there's no cause to randomize. BJPS. 58, 451–458.
Worrall, J., 2008. Evidence and ethics in medicine. Perspect. Biol. Med. 51, 418–431.