

(VERY EARLY) DRAFT: MOST FOOTNOTES AND REFERENCES MISSING. NOT FOR CITATION IN ITS CURRENT FORM

Causality in Medicine: Getting Back to the Hill Top

JOHN WORRALL

Some patients suffering from a particular medical condition are given a new intervention T which may prove to be therapeutic for that condition; and a number of those patients subsequently exhibit an amelioration of their symptoms. Other patients who are not given T exhibit fewer cases of amelioration and the difference in positive outcomes between the two groups is declared ‘statistically significant’. Did T *cause* the better outcome in the experimental group? According to some accounts we have real scientific evidence for a positive answer to this question, that is, evidence for causality, if and only if the trial was randomized, that is if and only if the division between those given T (the experimental group) and those not given it (the control group) was made using some randomizing device.¹ This is one of several arguments that allegedly establish the special epistemic power of randomized trials.²

In the first section of this paper I outline one account of ‘probabilistic causality’ that professedly endorses this conclusion about randomized controlled trials (RCTs) – that of Nancy Cartwright. And I argue that that account fails.³ I will then, in section 2, use this failure to motivate an account of the testing of causal claims that (a) involves a much more measured view of the role and epistemic power of RCTs; and (b) involves a much more optimistic view than is common nowadays of our ability to derive solid evidence for causal claims from non-randomized studies (and indeed “non-experimental” studies). Most of the elements of this view are, as we shall see, to be found already in the writings of Austin Bradford Hill. Hill is widely regarded as the first to introduce Fisher’s randomized trial methodology into medicine. But he held a much more nuanced and balanced position on the role of RCTs and of other kinds of clinical evidence than many of his more recent followers who have marched under the banner of ‘Evidence-Based Medicine’ (EBM). (For example EBM seemed to many initially to advocate the very strong view that *only* randomised trials really count as evidence, while Hill wrote⁴: “Any belief that the controlled trial is the only way would mean not that the pendulum had swung too far, but that it had come right off its hook.”)

1. Nancy Cartwright’s Analysis of Probabilistic Causality and ‘Ideal’ RCTs

In his famous paper “The Environment and Disease: Association or Causation?” Hill makes no pretence of even attempting to develop a “philosophical account of causation”. But it is tempting to think that we must first try to make sense of what is

¹
²
³
⁴

meant by ‘probabilistic causal’ claims like ‘cigarette smoking causes lung cancer’ before being able to arrive at a defensible view of how evidence impinges on such claims and of the circumstances under which we can justifiably declare them evidence-based. Several attempts have been made to say what such claims mean; here I concentrate on Nancy Cartwright’s attempt (though the conclusions I draw would be the same whatever account was considered⁵).

The relationship between smoking and cancer is clearly not deterministic: many people who smoke 50 cigarettes a day throughout their adult life fail to develop lung cancer. It is widely agreed that the central idea must be that a (probabilistic) cause raises the probability of its effect: smoking cigarettes doesn’t of course ensure that you will develop lung cancer but it ‘increases your chances’ of doing so.

But everyone also agrees that the fact that the $\text{Prob}(\text{lung cancer} / \text{smoke cigarettes}) > \text{Prob}(\text{lung cancer})$ (or equivalently $\text{Prob}(\text{lung cancer} / \text{smoke cigarettes}) > \text{Prob}(\text{lung cancer} / \text{don't smoke cigarettes})$) cannot on its own represent a full analysis of probabilistic causation. We need to distinguish, as has long been recognised and as Hill’s famous article suggests, between (‘real’) cause and (‘mere’) association.

There are lots of cases (ones that might trap the unwary into committing the *post hoc ergo propter hoc* fallacy), where two factors A and B are associated or ‘correlated’ (more accurately probabilistically dependent) but not causally connected. For example, owning one or more ashtrays is (no doubt) associated with developing lung cancer (the probability that you will develop lung cancer is greater given that you own ashtrays), but ashtray ownership doesn’t cause lung cancer.

How do we know? Well partly because background knowledge tells us what sorts of things are likely to be causes of cancer and owning glass, pottery, plastic or stainless steel objects of a certain kind is not among them. This will be an important theme later, but there is a much more direct answer to be gleaned just from the statistics: there is a “common cause” of owning ashtrays and developing lung cancer – namely being a smoker. Smoking makes you both more likely to get cancer and more likely to own ashtrays and the fact that smoking is a common cause will be revealed by the fact that smoking (S) “screens off” ashtray ownership (A) and lung cancer (C). Intuitively if we look just at smokers, then the proportion of non-ashtray owners among them who develop cancer will be the same as the proportion of ashtray owners who develop cancer. (Presumably the former will simply have dirtier carpets.) More formally, although

$$\text{Prob}(C/A) > P(C/\neg A)$$

$$\text{Prob}(C/A\&S) = \text{Prob}(C/\neg A\&S).$$

5

The probabilistic dependence between A and C disappears once S is taken into account.

Although this is an old idea, it was first articulated in terms of common causes and “screening off” by the philosopher Hans Reichenbach.⁶ Nancy Cartwright generalises Reichenbach’s account to give an attempted characterisation of probabilistic causality: fundamentally, C probabilistically causes E if and only if $\text{Prob}(E/C) > \text{Prob}(E)$ and there is no common cause of C and E. This is captured in her principle ‘CC’⁷:

‘C causes E iff
 $P(E/C \pm F_1 \pm \dots \pm F_n) > P(E/\neg C \pm F_1 \pm \dots \pm F_n)$, where $\{F_1, \dots, F_n\}$ is a complete causal set for E.’ ([1989], p.56)

A number of obvious idealisations are involved here: for example, it is assumed, that all causal factors are binary, that is, either ‘on’ or ‘off’, whereas normally they are quantitative variables that can take on a range of values. But laying that aside, the idea seems straightforward: C causes E just in case it raises E’s probability once the other causes are fixed and fixed, moreover, at whatever values you please. C must raise E’s probability in *every* cell of the partition of the population at issue into “causally homogeneous sub-populations”. If there is some combination of the $\pm F$ s conditional on which $P(E/C) = P(E/\neg C)$, then that combination screens off C and E, and hence shows that C is not a cause of E.

Laying aside the much discussed issue of whether this condition is too strong,⁸ CC’s most obvious problem seems to be its circularity. Whether or not C is a cause of E depends on what the *complete* set of causes of E is! Cartwright explains that ‘To be a complete causal set for E means, roughly, to include all of E’s causes.’ I do not know what she had in mind by the qualifier ‘roughly’ except that her formulation of CC needs, of course, to be tidied up a little to state instead that $\{F_1, \dots, F_n\}$ is the set of all E’s causes *except perhaps for C*: if C is indeed a cause and $\{F_1, \dots, F_n\}$ were a genuinely complete set of causes then C would of course have to be identical to one of the F_i , say F_j , but then clearly for any ‘state description’ S of the form $\pm F_1 \pm \dots \& \neg F_j \pm \dots \pm F_n$, $P(E/C \& S)$ is not defined since C&S is a contradiction, while for any S’ of the form $\pm F_1 \pm \dots \& F_j \pm \dots \pm F_n$, C obviously does not raise the probability of E against the background of S’ since it is already included in that background.

But even once we remove in this obvious way the vicious circularity that would be involved if C was being asserted to be a cause of E on the basis that it is in the complete set of E’s causes, a less direct but still troubling circularity remains. The fact is that the notion of cause appears both on the left and on the right hand sides of

⁶

⁷ She does, however, as we shall later note, eventually reject this principle in favour of a modified version.

⁸

CC. These days it is fashionable to claim that ‘reductive’ definitions (aka definitions!) are an impossible ideal and that there is nothing wrong with ‘circular characterisations’. This seems to me to amount to surrendering the attempt to provide any analysis of the notion concerned (just as ‘externalist epistemology’ seems to me to give up on the whole enterprise of epistemology). But even if one were to swallow this, there remains an obvious problem if the aim is to give an account of what probabilistic causality means in order to help us characterise how a particular probabilistic causal claim is to be tested and perhaps accepted on the basis of evidence. It sounds as if, when it comes to ‘applying’ Cartwright’s CC account evidentially to underwrite causal claims, we are going to need to know already what the causes of E are before deciding that the evidence indicates that C is among them.

As a good empiricist Cartwright does indeed set out to find a methodological counterpart to CC – in the hope that the methodological counterpart is not vitiated by the circularity of the original metaphysical account. But she does not tread the treacherous path from metaphysics to methodology carefully enough: in particular she seems to fail to recognise that when it is *evidence for* a causal claim that is at stake, as opposed to what that causal claim *means*, then we need evidence that the putative cause raises the probability of the effect relative not just to actual but also to potential causes – factors that background knowledge makes it plausible may play a role. Even if it should turn out that, for example, the number and extent of any co-morbidities that may afflict a person have, in ‘objective fact’, no effect on her response to some treatment T for some condition C from which she happens also to suffer, we would surely want to have controlled for co-morbidities before claiming that we have good evidence in favour of the effectiveness of T. This is because background knowledge tells us that co-morbidities do play a role in responsiveness to some treatments and hence that it is plausible that they may play a role in the case of T: should those with significant co-morbidities dominate in the control group of some trial, then that dominance might be an alternative explanation of a positive result that is equally plausible as the explanation that it is caused by T. We should have evidence against the plausible alternative before declaring that we have evidence for the effectiveness of T.

More of this later, but even overlooking this confusion of actual and conjectured causes, there is still the circularity problem. As Cartwright puts it herself:

It seems that a method [for “applying” CC to “discover”, or, better, provide evidence for, causes] that requires that you know all the other [N.B.] causes of a given effect before you can establish any one of them [better: ‘that some other possible cause is also in the complete set of actual causes’] is no method at all. (*op.cit.*, p. 62)

However Cartwright sees *this* problem as being precisely solved by the randomized experiment:

... the method does not literally require one to know all the other causes. Rather what you must know are some facts about what the probabilities are in populations that are homogeneous with respect to all these other causes, and that you can sometimes find out without first having to know what all those causes are. That is the point of the randomized experiment ... (*ibid.*)

So Cartwright's account underwrites a special, indeed unique, role for RCTs in the following sense: the only evidence that entitles us to claim a causal connection between, for example, treatment and outcome is evidence from an RCT, since randomizing and only randomizing allows us to sidestep the circularity in her account of (probabilistic) causality. Cartwright goes on to make clear, however, that a randomized experiment will perform its magic in this way *only if* it is 'ideal'. To count as 'ideal' an experiment involving treatment and control groups must satisfy 'two related conditions'. The first is that 'all the other causes that bear on the effect in question must have the same probability distribution in both groups' (p. 64). The second is that 'the assignment of individuals to either the treatment or the control group should be statistically independent of all other causally relevant features that an individual has or will come to have.' (*ibid.*)

Using a randomizing device to divide the subjects into the two groups 'is ... a help to both ends', and then other 'clever and well-known devices' - such as double-blinding - also kick in to 'ensure that the results of the real experiment will be as close as possible to the results of an ideal experiment.' (*ibid.*)

There are however a number of difficulties here. Consider the first condition that an experiment is supposedly to satisfy if it is to count as 'ideal': that 'all the other causes that bear on the effect in question must have the same probability distribution in both groups'. But to apply the notion of probability to the result of a *single* experiment – even an 'ideal' one! – is, at least on the frequency interpretation, a category mistake. Probabilities are defined not on single, but only on *repeatable*, events. It is not merely that, as Cartwright herself allows, experimental results are about relative frequencies not probabilities (see, e.g., *op. cit.*, pp. 65-6) but moreover that these probabilities are arguably not defined at all relative to single experiments.

It seems reasonable to conjecture that what Cartwright really intends is that the 'ideal' experiment should involve a division between experimental and control group that has the following characteristic: where the number of patients in the study group as a whole satisfying potentially causally relevant characteristic F_i is f_i , then, for all i , the number of those satisfying F_i in the experimental group *equals* the number of those satisfying F_i in the experimental group *equals* $f_i/2$. (Or perhaps that that those pairs of numbers are within some small, and presumably empirically insignificant, interval.)

But for a real randomized experiment to be ideal in this sense requires a miracle that casts the ‘case’ of the loaves and fishes into the shade. Given that we are talking, as Cartwright admits following Fisher, of ‘innumerable’ unknown (possible) causes, it would clearly be a miracle if *all* of those factors just happened to be balanced in the two groups (even if we adopt a slacker notion of balance than outright equality).

As for Cartwright’s talk of ‘ensur[ing] that the results of the real experiment will be as close as possible to the results of an ideal experiment’ (p. 64), this seems to amount to an empty tease. We can of course balance (or attempt to balance) *known* factors - as investigators do (or rather attempt to⁹) in the case of patients’ and clinicians’ expectations via double-blinding. It is indeed in this precise connection that Cartwright talks of making the result of the real experiment ‘as close as possible’ to the ideal result. But we can of course balance known factors without randomizing (indeed, although there are practical difficulties, in principle it is much more efficient to do so by deliberate matching). But Cartwright’s remark makes it sound as if she holds that we can have at least some intuitive measure of how close a real experiment is to one that is ideal in respect of *all* ‘other causes’ known *or unknown*. Clearly, however, there can be no estimate of how closely balanced a particular real trial is with respect to any unknown factor – by definition since the unknown factor is unknown!

Nancy Cartwright may in the end agree with this. She ends the section of her book on randomized experiments as follows:

The point of discussing [randomized experiments] here is to recall that the demand for total information that can seem to follow from CC is not necessarily fatal. Sometimes we can find out what would happen were all the other causes held fixed without even knowing what the factors are that should be held fixed. It is important to keep in mind, however, that it takes an ideal experiment to do this, and not a real one. For, as with Principle CC itself, the connection between causality and regularity is drawn already well above the level of real data and actual experiment. It is not frequencies that yield causes, but probabilities; and it is not results in real experiments, where subjects are assigned to groups by a table of random numbers, but rather in ideal experiments where randomization is actually achieved. (pp. 65-66)

I say only that she *may* agree with me, because the first two sentences of this passage do not cohere with the final three. Since no real experiment is ideal and since we have no way of telling how near to the ideal a real experiment is (except for known ‘alternative causes’ for which we could deliberately control), how can we then take ourselves to ‘find out what would happen were all the other causes held fixed without even knowing what the factors are that should be held fixed’ by performing a randomized experiment? Indeed as already indicated it is not clear to me that the idea of ‘achieving’ randomization is a coherent one.

9

In any event, it seems safe to conclude that *either* Nancy Cartwright agrees with me that her analysis of probabilistic causality gives no further practical reason to insist on randomized experiments as especially telling from an epistemic point of view *or* if she *is* claiming that her analysis does supply such a reason, then she has provided no sustainable justification for that claim.¹⁰

Another way of looking at what an RCT would achieve were it to be ‘ideal’ in Cartwright’s sense is that it would control all at once for all possible confounders – not just the known possible confounders (really those that background knowledge gives reason to think *might* play a role in the outcome at issue) but also the ‘unknown’ ones (again ones that background knowledge gives us no reason to think might play a role in the outcome). In the ‘ideal’ randomized experiment everything else would be balanced in the two groups and the only difference would be in the interventions given to the two groups. This is why in the end the argument for randomization stemming from recent work in philosophy of science on probabilistic causality amounts in the end to the argument that randomizing removes all confounders – “known” and “unknown”.

If by randomizing we really did control for all possible confounders then every other possible explanation of a positive result would be eliminated and we could inexorably infer that that result (at least in this particular test) was due to the intervention under trial. This is the argument that has, sociologically speaking, undoubtedly carried the most weight within the medical community. And no wonder: it is a very powerful claim and one that is often voiced. Thus if you log on to the UK Cochrane Centre website you will be told by its Director Mike Clarke that “In a randomised trial, the *only* difference between the two groups being compared is that of most interest: the intervention under investigation.”¹⁴ Or Sir Michael Rawlins head of NICE 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.¹¹

¹⁰ Cartwright later in her [1989] book argues that her principle CC is in fact in need of modification to deal with what she sees as cases of ‘contrary capacities’ and ‘interactions’. There are again many issues here. (In particular I think that the talk of ‘contrary capacities’ is based on a failure to understand the underlying physiology in cases like Hesslow’s often-discussed (1976) example of the contraceptive pill which, it is alleged, both causes thrombosis and prevents it (by inhibiting pregnancy which is itself a cause of thrombosis).) But, again, there is no need to go into these for present purposes. Her view is essentially that, because they ‘dovetail’, the ‘ideal’ randomized experiment yields the truth about causality just in cases where CC applies. In other cases where ‘contrary capacities’ or ‘interactions’ are involved the ideal randomized experiment will not give the correct answers, although the modified version of principle CC (CC*) does give the correct answers. Hence, in so far as she argues that randomization plays a significant role at all, it is restricted to the simpler cases where CC as a matter of fact is, for her, true; and so, the later modifications are not relevant to the issue discussed above.

¹¹ Harveian Oration. This is a slip by Rawlins who on the whole has a very good grasp of the evidential issues.

It is not clear to me that the ‘similarity’ claim (let alone the identity claim implicit in Mike Clarke’s remark) makes any sense once we are talking about *all possible* extraneous factors that might play a role. But in any event, the claim, assuming it makes sense at all, is simply false (as everyone including Clarke and Rawlins *really* knows). Or at least, as we saw in analysing Cartwright’s argument, it is false about any real RCT (as opposed to the entirely metaphysical ‘ideal’ one envisaged by her).

No one really believes that, given a *particular* random division, the groups are bound to be equal in all other respects and hence any difference in the outcome is automatically attributable to the difference in treatment. That is, no one (not even Mike Clarke, as he indicates later in his article) really believes that having randomized is a *sufficient* condition for establishing that any observed effect must be due to the treatment and to the treatment alone. In any *particular* randomized division, it is of course entirely possible that some factor is unbalanced between the two groups and hence this lack of balance remains a possible rival explanation for the ‘positive’ result, a rival explanation not ruled out by the trial.

This is in fact quietly (and only implicitly) conceded by most orthodox treatments of RCT methodology. At least in those trials (the majority) where no attempt has been made to match with respect to ‘known’¹² prognostic factors, investigators are recommended, before drawing any evidential conclusions, to look at the particular division into control and experimental groups that randomization has given them and check for ‘baseline imbalances’. That is, investigators should check that the two groups are not in fact unbalanced with respect to any factor that background knowledge tells us might play a causal role – age, sex, co-morbidities and so on. If such baseline imbalances are found then the recommendation – clearly for practical rather than epistemic reasons – is to re-randomize in the hope that this time no baseline imbalances will occur.¹³

But if it is admitted, as of course it must be, that, in the absence of deliberate matching, 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, factors which do in fact play a role but which background knowledge supplies no reason to suspect do so. The difference between this and the (known) “baseline imbalance” case is of course that, by definition, investigators cannot check for imbalance in “unknown” confounders.

An amusing example that shows that randomization cannot guarantee that the two groups are equal in all relevant respects is provided by an article published in the *British Medical Journal* in 2001 by Leibovici and co-workers.¹⁵ 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 four years after* these

¹² I will henceforth drop the scare quotes intended to remind the reader that this really means a factor that background knowledge indicates *may* play a role.

¹³

patients had been in hospital), a random number generator was used to divide them into two groups. Which of these two became the treatment group was decided by a coin toss. 1691 were randomized to the intervention group and 1702 to the control. A check was actually made in this case for ‘baseline imbalances’ with regard to main risk factors for death and severity of illness – that is, whether this pure randomised division without any prior matching had in fact produced groups that were significantly unbalanced with respect to ‘known confounders’. 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. The results were as follows. Mortality was 28.1% in intervention group and 30.2% in the control group; this was “not significant” according to the usual significance testing/ null hypothesis methodology. However *both* length of stay in hospital *and* duration of fever were significantly shorter in the intervention group ($p = 0.01$ and $p = 0.04$)! Leibovici concluded - perfectly properly in accordance with accepted methodology in medicine - 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.

I take it that it goes without saying this is not to be taken seriously¹⁴: even those who believe that god moves in mysterious ways are hardly likely to believe that they are mysterious as this! The reason why it is not to be taken seriously reveals a further aspect of this simple but striking case: that we are all naturally Bayesian. As Leibovici himself wrote¹⁶:

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.

Before taking up this important remark, and before I am drowned by cries of ‘Strawman!’: no one really believes, do they?, that randomization inevitably guarantees groups that are similar in all other respects and hence guarantees that a positive result in a properly randomized 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. And as we saw, many people, Mike Clarke included, certainly sometimes *say* that they believe it. But I agree that it cannot be seriously believed. In so far as anyone seriously believes that there is some guarantee of equivalence or similarity between the two groups in a randomized trial it is not belief in a sure-fire guarantee but rather in some sort of

¹⁴ Though surprisingly some contributors to the BMJ website discussion managed to do so!

probabilistic quasi-guarantee – clearly, in what is admitted on all sides to be a stochastic domain, we could not reasonably expect any better.

But what exactly does such a ‘probabilistic guarantee’ amount to? Surprisingly many people take what seems to me a surprising amount of solace in the phrase ‘either the groups are equal or a chance event has occurred’. But in an area where it is acknowledged that we do not have control over all the factors that might play a role, then a chance event has always occurred – it just as ‘chancy’ if randomization produces equal groups as if it doesn’t! This often repeated mantra seems to make the mistake of assuming that if a fair coin is tossed 10 times and 9 heads result then this is a ‘chance event’ while if it produces 5 heads it is not.

Looking at it from the perspective of basic philosophy of science, there seem to me to be two reasons to be suspicious of the credentials of any such ‘probabilistic quasi-guarantee’.

The first and more significant is that the reasoning underpinning these ‘credentials’ involves a slip from what is arguably true in the indefinite long run to a claim about what is true of a *particular random allocation* (analogous to, indeed a more rigorous version of, Cartwright’s slip from what an ‘ideal’ RCT might tell to what a real RCT does). An enormous amount of effort in the philosophy of science literature 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.) My own view is that no real sense can be made of the notion of single case objective probabilities. I cannot hope to argue this here; but I can indicate one central difficulty in supposing that there are such single case probabilities. 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 group would be the same and would be the same as the frequency with which that factor is exhibited in the experimental population as a whole. But we are never in the long run, we never randomise indefinitely often, 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 – as the Leibovici study establishes, but which is in any event obvious.

The second problem with the reasoning behind this probabilistic quasi-guarantee was pointed out by the Bayesian statistician Dennis Lindley. There is more than a hint of a quantifier fallacy here.¹⁷ There seems to be some confusion between imbalance with

respect to a *particular* factor and an *overall* imbalance. Even if one were to try to make some single-case probability argument work, it would be an argument that there is a very low probability that *some particular* factor (say age, or co-morbidity) is unbalanced between the two groups. But how is one to weigh this against the assumption driving this whole issue that the list of possible unknown factors is indefinite? In such circumstances, it seems that even if we suppose that there is a definite probability that the groups are unbalanced with respect to some particular specified factor, the ‘probability that the groups are unbalanced with respect to *some* possibly confounding factor’ is unquantifiable.¹⁵

2. Evidence for causality in medicine reassessed

So there are no guarantees, because RCTs do not, cannot, control for ‘all possible confounders known and unknown’. As for quasi-guarantees, these are not worth more than the quasi-paper they are written on. As I hope will be clear, by criticising exaggerated claims about RCTs, I am not criticising all RCTs nor at all denying that sometimes randomizing does some epistemic good. But the Leibovici case is not exceptional – every RCT result should be critically examined with the fundamental question always uppermost in our minds: to quote Hill, 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 answer even though we have at present no way of specifying 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 randomized 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. And surely correctly.

Call this “exercising judgment” if you like,¹⁶ but it is surely not unanalysable judgment. 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 not merely a subjective opinion to hold that prayer can have no retroactive effect!)

¹⁵

¹⁶ See Rawlins op cit

We should use these reflections to try to build a more measured account of the evidential virtues of both randomized and non-randomized studies. Keeping the ‘fundamental question’ in mind, it is easy to see that RCTs have one undoubted advantage: randomizing means that the clinician has no control over which of the two groups in a trial,¹⁷ experimental or control, any particular patient goes into. And there are trials whose results do seem to have been rendered unreliable because the division made by clinicians’ selections was (consciously or unconsciously) biased. But this is not controlling for all confounders, it is controlling for a particular (possible) confounder – that of selection bias (though notice that this phrase is used in different ways in the clinical trials literature) – which background knowledge gives us reason to think may play a role. Background knowledge tells us that in a non-randomized trial an alternative explanation of a result to the explanation of a causal connection between treatment and outcome is, *if the difference in outcomes between the experimental and control groups though statistically significant was small*, that the two groups were made unequal in some other way because of selection bias.

Consistently with this view, Hill’s advocacy of RCTs is not expressed in terms of ‘controlling for all counfounders’ but in terms of its elimination of the specific possible confounder of selection bias¹⁸:

“Faithfully adhered to [randomizing] offers three great advantages: (1) it ensures that our personal feelings or judgments, applied consciously or unconsciously, have not played any part in building up the various treatment groups; from that aspect, therefore, the groups are unbiased; (2) it removes the very real danger, inherent in any allocation which is based upon personal judgments, that believing our judgments may be biased, we endeavour to allow for that bias 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 biased groups for comparison.”

However, as indicated, Hill makes it clear in a number of places in his writings, that selection bias is only a plausible rival explanation when the outcome effect is small. Where the outcome at issue is large or at all substantial then not only is randomization irrelevant *so also is the use of any formal statistical test of significance*. He writes of one of his investigations – into whether there was a connection between working in the card room in cotton mills and developing certain kinds of illness – that he arrived at a very definite conclusion on the basis of the evidence (that there was indeed a causal connection):

‘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

¹⁷ At least not if patients have already been declared eligible for the trial before randomization and/or the trial is also double blind

¹⁸

no formal tests could really contribute anything of value to the argument. So why use them?’ (8)

It took EBM-ers until 2007 and the paper by Glaziou et al to (partially) rediscover this lesson.¹⁹

Hill also points out, in his equally famous paper ‘Reflections on the Controlled Trial’, a number of ways in which RCTs can be misleading – not necessarily because they fail to provide good evidence of a causal connection, but because they may provide evidence for the *wrong* causal connection – one which has no real significance in the practice of medicine.

For example he argues that double-blinding is very much a mixed blessing: it removes one possible source of bias (selection bias) but ‘in some situations [double blinding] may be inexpedient and, indeed, injurious to the trial ... Such situations arise when it is important, for the sake of a realistic trial, that the doctor in charge of the patient be able to adjust the dose of a drug according to the patients’ reactions and according to his judgement of the patient’s requirements.’ Investigators must ask in planning the trial ‘which is the more important – for the doctor to be ignorant of the treatment and unbiased in his judgement or for him to know what he is doing and to be able to adjust what he is doing ..’ (110) Hill is therefore also firmly against the still widespread assumption that only fixed dose trials are scientifically valid. This is because the practising physician is not interested in the issue of the average effect of a drug administered according to an entirely fixed regime. Instead trials should be addressed to the question “if competent physicians in charge of defined types of patients used drug X in varying amounts and for such varying durations of time, and so forth, as they think advisable for each patient, what happens?” (ibid) Finally he is, again against much current thinking, very much in favour of ‘subgroup analysis’ within trials – of course not mindless ‘data mining’ but asking sensible questions looking for differences in outcome between groups where, in effect, background knowledge informs us that there may well indeed be a difference.²⁰

The mirror image of EBM’s (I would argue) exaggerated view of the epistemic virtues of RCTs is its (I would argue) overly pessimistic view of what can be reasonably evidentially established via non randomized, and indeed non experimental, studies. It became an article of faith within EBM (largely on the basis of studies by Chalmers and others) that non-randomized studies have a constant tendency to be more positive than ‘proper’ (i.e. randomized) studies. But aside from the obvious circularity (the ‘real effect’ is taken to be identified by the RCT!), Chalmers et al based their claims on ‘observational’ (mostly historically controlled) studies that were obviously flawed – that is ones in which no account had been taken

¹⁹

²⁰ All interesting and bear further investigation but about the right causal connection not evidence for cause in itself

of factors that background knowledge gave reason to think might play a role. Of course historically controlled studies must be analysed with Hill's fundamental question in mind – could it have been something else that explains the difference between the outcome with the new treatment and the historical outcome? But if it has been thus analysed then there is no reason to distrust the result. This is exactly the line that Hill took.

First, he enthusiastically endorsed Claude Bernard's surely correct 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 (1865) 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's famous paper ‘The Environment and Disease: Association or Causation?’ is, as the title suggests, explicitly concerned with the issue of evidence for a causal connection between some environmental variable and an increased incidence of disease. It is in connection with this issue that he articulated his famous, causal ‘criteria’ (though he was quite explicit that they are *not* in fact criteria at least in the usual sense of the term; he himself called them ‘viewpoints’). Various remarks here and in his other papers, as well as his endorsement of Bernard, show, however, that he also took the same line in connection with non-randomized trials.

Hill listed a number of criteria/viewpoints: strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy. Here is not the place to give a detailed analysis of particular ‘criteria’, about which there are a number of points to be made.²¹ But I do want to articulate more fully what I take to be the underlying evidential project.

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. The ‘fundamental question’ then is whether these facts provide good evidence of a causal relationship between environmental factor/ treatment and outcome. Both possible answers are certainly deductively consistent with the facts. So further evidence is needed if we are to be on as safe ground as we can be. 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

²¹

lung to the development of tumours (subject to some biophysiological initial conditions and the amount of smoke) then no one would have needed the statistical data to convince themselves of a causal link between cigarette smoking and lung cancer.²² But, as this indicates, such clear cut further evidence will not be available in cases of interest. I take Hill's main point to be that we should by no means give up at this juncture. There may well be elements of background knowledge which at least assist us in facing the 'fundamental question'. (And it is important to keep on remembering, as EBM aficionados appear to have tried to forget, that we still face this fundamental question even if the 'facts of the case' are the results of RCTs. Recall Leibovici.)

Background knowledge tells us, for instance, that many, though by no means all, causal relationships are linear (or reasonably close to it): the larger the cause, the larger the effect. So if the facts tell us not only that more smokers develop lung cancer, but moreover that the heavier smokers develop more lung cancers than the lighter smokers, then this clearly strengthens the case for a causal connection. This is Hill's 'biological gradient' [dose-response] "viewpoint". Similarly while background knowledge tells us that it is, to say the least, very unlikely that there is a genuinely (deterministic, non probabilistic) mechanism linking owning glass, pottery, plastic or stainless steel objects of various types and lung cancer, 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 awfully plausible that there will be such a link. This is a combination of Hill's "[biological] plausibility" and his 'coherence' "viewpoints".

Of course as Hill is at pains to insist, everything here is defeasible: there are well known cases where no mechanism was known and some investigators treated a statistical link as 'merely' associational, for which a mechanical link was later found. This is why these are not criteria in the normal sense, why they cannot be regarded as 'hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect' (or as he clearly also believes where if they are satisfied we can infallibly infer cause and effect!) 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 hate the idea of bringing this sort of judgment based on background knowledge into what they would like to be the *rules* of evidence. But they are deluded. As Michael Rawlins points out in his recent excellent Harveian Oration - unconsciously echoing Hill - judgment is inevitable. Rawlins does not mention the Leibovici study but he does list 22 drugs that were sanctioned in the UK on the basis of positive results in RCTs which had later to be withdrawn because they did not in practice operate as the RCT suggested. We will not have an adequate view

²² Disappointing work by Williamson and collaborators

of evidence in medicine until we have incorporated Hill's many insights. EBM needs to get back to the Hill Top.²³

²³ Recently accepted (Howick, Aranson, Glaziou) though somewhat halfheartedly and confusedly