Philosophy Compass 2/6 (2007): 981–1022, 10.1111/j.1747-9991.2007.00106.x

Evidence in Medicine and Evidence-Based Medicine

John Worrall*
London School of Economics

Abstract

It is surely obvious that medicine, like any other rational activity, must be based on evidence. The interest is in the details: how exactly are the general principles of the logic of evidence to be applied in medicine? Focussing on the development, and current claims of the 'Evidence-Based Medicine' movement, this article raises a number of difficulties with the rationales that have been supplied in particular for the 'evidence hierarchy' and for the very special role within that hierarchy of randomized controlled trials (and meta-analyses of the results of randomized controlled trials). The point is not at all to question the application of a scientific approach to evidence in medicine, but, on the contrary, to indicate a number of areas where philosophers of science can contribute to a proper implementation of exactly that scientific-evidential approach.

1. Introduction

Unusually, this *Compass* article on philosophy of science is not a guide to the existing literature in some subfield of the discipline, but rather an attempt to point to a new area where philosophers of science could have enormous impact – both intellectual and (very unusually) practical – but have so far very largely not done so.[1] The area is that of the logic of evidence as applied to medicine. The study of evidence or confirmation theory has, of course, always been at the very centre of the discipline of philosophy of science. At *very general* root, the principles of evidence are – so I would argue – universal and common to all disciplines.[2] But of course the way that these very general principles are applied in a particular discipline may be highly dependent on particular features of that discipline; and undoubtedly, interesting specific issues about evidence arise in the area of medicine.

A suitable focus for the study of many (though by no means all) of those issues is provided by the relatively recent, and highly influential, movement called Evidence-Based Medicine.[3]

The basic idea underlying that movement – that medical science and medical practice should be based on evidence – is surely a 'no-brainer':

the rational person follows the evidence in *all* areas. *Obviously* we must apply proper standards of evidence to claims about medical science and about medical treatment – and, presumably, (most) medical practitioners always believed they were doing so. The fact that EBM is a new movement indicates that its founders believed that some, perhaps many, decisions were *in fact* being taken (usually of course implicitly rather than consciously) about what counts as evidence in medicine that were *normatively* mistaken. And they believed that there are forms of real evidence that carry great weight from a normative, scientific point of view but were not, sociologically speaking, being accorded the weight they deserve by the medical community as a whole at the time.[4]

And indeed EBM-ers initially seemed to many to be taking a very definite view about what counts as real evidence and why there was much that was wrong with the then current evidential practice: an individual clinician's 'clinical experience' should, they seemed to suggest, be pretty well entirely discounted as liable to be biased in any number of ways; 'patho-physiologic rationale' (that is, the 'basic science' sometimes underlying therapeutic claims) should at least be given less weight than it generally was given; what *really* counts are the results of properly conducted clinical trials. And, concerning the latter, the message was again initially taken by many to be the very sharp one – that a clinical trial was properly conducted and therefore its result carried true scientific weight if, and only if, the trial involved randomized controls. (In a randomized controlled trial (hereafter RCT), the study population is divided into an experimental group, members of which receive the treatment under test, and a control group, members of which receive something else, perhaps a placebo or currently accepted treatment, and that division is made by some random process.[5]) Thus Sackett et al. described the basic ideas of the new movement as follows:

> EBM de-emphasises intuition, unsystematic clinical expertise, and pathophysiologic rationale . . . and stresses the examination of evidence from clinical research. In 1960, the randomized trial was an oddity. It is now accepted that virtually no drug can enter clinical practice without a demonstration of its efficacy in [implicitly: randomized] clinical trials. Moreover the same randomized trial method is increasingly being applied to surgical therapies and diagnostic tests. ('Evidence-Based Medicine' 71)

As I already suggested, there surely is, at the underlying general level, nothing special about the role of evidence in medicine. Real evidence-based medicine results from applying the universal general principles of the logic of evidence to the particular case of medicine, and especially (though not of course exclusively) to claims about which treatments are and which are not genuinely therapeutic.[6] I have no doubt that the founders of EBM are in agreement with this. Although they talk, as seems almost obligatory nowadays, of EBM as a 'new paradigm',[7] there is not the slightest hint of relativism in their writings. They are not proposing the formation of a new group that, simply, as a matter of fact, decides, for example, to give great (perhaps overwhelming) significance to RCTs. Instead they argue (implicitly) that the principles that govern weight of evidence across the board in 'proper' science should be applied systematically in medicine too; and that the principles of evidence-based medicine that they advocate are exactly the results of this application.

The fundamental question, then, is whether they are correct.

Unsurprisingly, as the EBM view was further articulated and defended against criticism, it soon became a good deal less clear-cut. EBM-ers, as we shall see, no longer endorse the very strong claims about what does and does not count as evidence that I just recorded. Non-randomized trial could supply *some* evidence and the 'best clinical expertise' was to be incorporated rather than overridden – so for example Table 1 in Straus and McAlister lists the 'Steps involved in the practice of evidence-based medicine', step 4 being 'To integrate [the] appraisal [of the validity and importance of the evidence] with clinical expertise. [in order] to apply the results in clinical practice'.[8] Indeed EBM-ers deny that they ever did endorse those strong claims.

Although there may be some validity to their denials, they cannot legitimately feel aggrieved that it was the strong message that initially got across to the medical community. One can, for example, still read in what is often described as the movement's 'Bible' '[i]f the study was not randomized we'd suggest that you stop reading it and go on to the next article in your search' (Sackett et al., *Evidence-Based Medicine* 108). But whatever may be the truth about what they did or did not initially endorse, their current position, as we shall see, is altogether more guarded and nuanced. Far from making the methodological issues go away, however, that current more nuanced view raises still more, and still more challenging, methodological issues.

Instead of pursuing the forlorn hope of clarifying all those methodological issues, I restrict myself here to the following agenda. *First* (section 2) I sketch a brief history of the Evidence Based Medicine movement. This will show that the position it currently occupies involves (a) the view that different types of evidence do indeed carry some legitimate weight but (b) that nonetheless the results of randomized trials carry *very special* weight (to the extent that they should 'trump' other kinds of evidence). Point (a), I shall argue in section 3, leads to a number questions that cry out for the clarification that high-quality philosophy of science could provide: questions not only about the ranking of particular types of evidence in terms of the strength of support that they supply for (especially) therapeutic claims, but also about how different types of evidence (from clinical experience, clinical trials and biochemistry) should be combined to produce some overall judgement. The particular point (b) – the continuing insistence that the results of randomized trials carry at least very special epistemic weight – is examined in section 4. In that section, I investigate whether there is indeed any (successful) argument from 'first principles' for according any special weight to the results of randomized trials. This is the one topic that *has* been given some attention in the philosophy of science literature – in particular within the Bayesian part of that literature.[9] I shall show that – again – many questions remain open. I proceed here by identifying five different arguments that claim to establish that the results of randomized trials carry special epistemic weight. It should go without saying that, although some of my conclusions here will be negative, the object is certainly *not* to debunk all randomized trials, still less to oppose the application of scientific method to medicine. On the contrary, the object is to encourage the attempt *properly* to think through how to apply scientific method to medicine.

2. The Birth of Evidence-Based Medicine

An important part of the motivation for the EBM movement was the recognition that the individual practices of many medics were often not based on what might generally be judged as the currently best available evidence. This issued in a series of recommendations for improving practice through giving individual medics better access to that evidence: for example, by introducing measures to try to keep medics up to date rather than relying on the medicine they learned in Medical School, in some cases upwards of 40 years previously; or measures to make the results of clinical trials more readily widely known amongst doctors. I shall not be concerned here with this (undoubtedly important) disseminatory, educational or institutional aspect aimed at improving individual performance; instead I concentrate entirely on their underlying view as to what – objectively (or intersubjectively) – constitutes the best available evidence.

It is notorious that the history of medicine features a number of 'treatments' – such as blood-letting (for a variety of conditions some of which are now known to involve low blood pressure) – that were sworn-by for decades, if not for centuries, but that we now know to be ineffective, at best, and indeed not infrequently positively harmful. Presumably most (or many) physicians who relied on bloodletting, leeching and the rest were not cynics who were out to make money from their patients independently of whether or not their treatments helped those patients. Instead they genuinely believed that their treatments were – of course overall – beneficial, and no doubt believed that their practice, their experience with patients, supplied good evidence that those treatments were indeed beneficial. (No matter how soon and how painfully their patients died after receiving treatment, they might always have died sooner and more painfully!) Yet we now take ourselves to know, on the basis of better, wider evidence, that these treatments were ineffective at best and, not infrequently, outright harmful.

One of the central drivers of the Evidence Based Medicine movement, which first came to prominence in the early 1980s (and has spread to all parts of the globe from its original base at McMaster University in Canada), was the thought – itself inherited from A. L. Cochrane (1972/1989) – that there is no reason to think that this is merely an historical phenomenon: perhaps many of the 'therapies' accepted by modern medicine, just like blood-letting, have no real evidential basis and are ineffective (or worse). And indeed EBM-ers could point to, indeed were partly inspired by, a number of cases of treatments that had become standard, but which, when subjected to 'proper' scientific trial, were judged to be in fact ineffective.

One favourite example is grommets for glue ear. Glue ear is a condition of children produced by a build-up of fluid in the middle ear, itself caused by (earlier) infection. This fluid is unable to drain away because of pressure differentials maintained in the ear. The idea of the treatment is to insert a small grommet – a valve that lets air into the middle ear and hence equalizes the pressure. This would mean that the fluid would drain away down the Eustachian tube. In fact when a controlled trial was performed, it indicated that the insertion of grommets has no positive effect on the condition. It followed that in view of the (slight but non-zero) danger involved in the procedure it was better to let the condition clear up of its own accord – which the trial had indicated that it did (on average) just as quickly without the insertion of a grommet as with it.

Another example often cited in the EBM literature concerns a phenomenon called ventricular ectopic beats. After a myocardial infarction, the heart remains electrically unstable and sometimes throws off characteristic beats. Those patients who exhibited these ventricular ectopic beats showed a greater incidence of subsequent cardiac arrest than those who did not exhibit them. It seemed to make good sense therefore to suppress the beats in the expectation that this would reduce the risk of cardiac arrest. They could be fairly straightforwardly repressed by administering substances like encainide or flecainide (also used as local anaesthetics). This became standard treatment but when a randomized trial was performed it showed a *higher* rate of mortality from cardiac arrest amongst those treated for the suppression of the beats. And this treatment has now been abandoned.

A final example concerns routine foetal heart rate monitoring once the mother had been admitted to the maternity ward. This seemed like a good idea – babies in distress could be identified earlier and appropriate action taken, while surely the procedure, being entirely non-invasive for the foetus and seemingly negligibly invasive for the mother (the listening device is just strapped to the mother's abdomen), could at least do no harm. As indicated, this was routine treatment – obstetricians clearly felt that their experience established that it was an effective measure. However a randomized trial indicated that routine foetal heart monitoring has no positive effect in terms of infant mortality, but does lead (presumably via the effects of extra stress on the mother) to a greater number of interventions in labour – notably caesarian sections.

These cases (to which a *few* more could be added) reveal the two main initial targets for this aspect of the EBM movement – what they saw as over-reliance in medicine on clinical judgement and experience and what they saw as over-reliance on background theory. Obstetricians' judgement that routine foetal heart-rate monitoring was effective might be awry because of lack of a control group – all foetuses about to be born were monitored; background theory might assure us that the fluid created by infection would drain away once the grommet was inserted and that this would cure glue ear, or that flecainide would repress ventricular ectopoic beats, but it doesn't follow that these 'therapies' will work or be a good idea in practice.

Hence, as we already saw:

> EBM de-emphasises intuition, unsystematic clinical expertise, and pathophysiologic rationale . . . and stresses the examination of evidence from clinical research. In 1960, the randomized trial was an oddity. It is now accepted that virtually no drug can enter clinical practice without a demonstration of its efficacy in clinical trials. Moreover the same randomized trial method is increasingly being applied to surgical therapies and diagnostic tests. (Sackett et al. 'Evidence-Based Medicine')

This seemed to many at the time to amount to the very straightforward and challenging view that not only should all procedures in medicine, of course, be based on evidence, but that the *only really telling scientific* evidence came from RCTs, *and* from 'meta-analyses' or 'systematic reviews' that amalgamated the evidence from different RCTs (often themselves involving rather few patients) into one overall result (of which, more later).

It should, however, be a sobering thought for anyone (even half-way) inclined toward a 'no RCT, no real evidence' view that the number of therapies and procedures that continue to be sanctioned in modern medicine and on which no RCT has ever been performed *far* outweighs the number of cases that motivated EBM by indicating that hitherto accepted therapies might not be therapeutic at all. No one would seriously question that penicillin is a good treatment for pneumonia (of course it does not follow, given further innovations, that it remains the *best* treatment, nor that it is even a good treatment for *all* patients with pneumonia), that aspirin is a good treatment for mild headache (in those without particular gastric problems), that diuretics should be administered in cases of heart failure, that appendectomy is the right treatment for acute appendicitis, or cholecystectomy for gallstone disease, and the list goes on and on. No RCT has ever been performed on any of these treatments and none presumably ever will. (Think of the outcry if some patients with acute appendicitis were randomized to 'placebo surgery'!)

Those seriously committed to the unique scientific value of RCTs might take the heroic line here, and claim that we do indeed have no truly solid evidence in any of these cases; it is only the (allegedly) entirely separate issue of the 'ethical cost' of performing a trial that prevents trials being performed whose outcomes would (if positive) finally provide the missing real evidence. However, this is surely an unsustainable view. If the belief that appendectomy is a good treatment for appendicitis were based on nothing more than subjective opinion – if there were no good objective grounds at all for thinking that those with acute appendicitis when treated this way did better than if left untreated, then once this had been pointed out (and assuming that some doctors wished to continue using this 'treatment'), then surely reasonable medics would, far from resisting a trial, in fact *demand* one. In that (of course, highly counterfactual) case there would be no reasonable qualms about assigning a patient to the control group of such a trial. But no one can seriously believe this – there *are* good objective evidential grounds for believing in the efficacy of appendectomy for acute appendicitis, it is just that those grounds do not include the results of any RCTs. The (astronomical) 'ethical cost' of an RCT on appendectomy, if one were now performed, would result from the fact that, whatever the heroic EBM-er may claim, we already have strong (objective) evidence and hence good objective reason to believe that in assigning a patient to the control group of such a trial, we would be condemning him/her to a treatment that is (massively) sub-optimal. (It is an importantly under-appreciated fact that, in general, a judgement of the 'ethical cost' of a trial is *not* one that can be made independently of an *epistemological* judgement about the weight of evidence that we already have ahead of that trial; and hence that different epistemological views may underwrite quite different judgements about whether a particular trial is ethical.[10])

Again it might be (heroically) claimed that we don't *really* have positive evidence in these cases (because there has been no RCT!) but only *believe* that we do. But, given the strength of the conviction that all sensible people surely share in cases like the appendicitis one – that such a trial would recklessly endanger the lives of acutely ill patients – this seems, to say the least, a difficult view to sustain. It seems altogether more plausible to hold that the true account of evidence in medicine does not give any unique role to randomization and allows instead that proper, scientific evidence can be derived from other sources. Or at least that we should seriously investigate the possibility of articulating such an account before accepting such a counterintuitive suggestion as that we have no real evidence for the effectiveness of a (very) wide range of unquestioned treatments.

Although I have heard the heroic line taken in discussion (at least temporarily!), it was not the line taken by the influential players in EBM. While, as already noted, it was no surprise that medics initially took the message of EBM to be 'no RCT, no evidence', later articles with titles like 'EBM what it is, and what it isn't' (Sackett et al. 'Evidence-Based Medicine') denied this:

> **EBM is not restricted to randomized trials and meta-analyses.** . . . some questions about therapy do not require randomized trials (successful interventions for otherwise fatal conditions) or cannot wait for the trials to be conducted. And if no randomized trial has been carried out for our patient's predicament, we must follow the trail to the next best external evidence and work from there. (72)

Moreover, in the selection criteria for articles to be abstracted in the journal *Evidence-Based Medicine*, randomization is required only for *therapeutic* trials, while an explicitly more open policy is declared towards studies of (presumably disease) causation:

> **Criteria for studies of causation**: a clearly identified comparison group for those at risk for, or having, the outcome of interest (whether from randomized, quasi-randomized, or nonrandomized controlled trials; cohort-analytic studies with case-by-case matching or statistical adjustment to create comparable groups; or case-control studies. . . .) (*Evidence-Based Medicine* 1 (1995): 1, 2)

In Sackett et al. ('Inpatient General Medicine'), randomized trials are explicitly pronounced inessential for 'Group 2 interventions'. These are defined as follows

> Interventions with convincing non-experimental evidence – Interventions whose face validity is so great that randomized trials were unanimously judged by the team to be both unnecessary, and, if a placebo would have been involved, unethical. Examples are starting the stopped hearts of victims of heart attacks and transfusing otherwise healthy individuals in haemorrhagic shock. A self-evident intervention was judged effective for the individual patient when we concluded that its omission would have done more harm than good. (408–9)

Other 'clarifications' insisted that 'pathophysiologic rationale' was *not* to be discounted ('sometimes the evidence we need will come from the basic sciences such as genetics or immunology'); and that the 'best' clinical expertise was to be *incorporated* into the overall evidential picture, not entirely overridden.

So, outside the area of therapy, RCTs are not always even indicated. Therapeutic trials are always best done using randomization ('the randomized trial, and especially the systematic review of several randomized trials [of which more shortly], is so much more likely to inform us and so much less likely to mislead us [that] it has become the "gold standard" for judging whether a treatment does more good than harm'); but this is not *always* necessary (for therapies with great 'face validity') and we should explicitly *not* infer from 'no RCT' to 'no real scientific evidence' ('if no randomized trial has been carried out for our patient's predicament, we must follow the trail to the next best external evidence and work from there').

What does 'next best' mean here? EBM-ers have spent a good deal of time recently developing an *evidence hierarchy* (which comes with associated advice to medics on how to translate the rankings of evidence into practical action). In fact a number of these evidence hierarchies are available, though they share most essential characteristics. Readers can consult, for example, the – representative – hierarchy supplied by the Scottish Intercollegiate Guidelines Network (SIGN) at http://www.sign.ac.uk/guidelines/fulltext/50/index.html.

3. Ranking and Combining Evidence: Some Under-Investigated Methodological Issues

All this means that EBM correctly rejected a view (that only RCTs provide really telling scientific evidence for the efficacy of medical procedures) that is clear and crisp, and therefore can readily be seen to be untenable, only to replace it with a view that is not readily seen to be untenable, but only, perhaps, because it is unclear. Certainly any number of issues arise with the more complex view of evidence that now seems to underpin EBM (when interpreted carefully). In this section, I simply briefly identify and comment on a number of such issues. As experts on evidence, philosophers of science could of course make important contributions to all of them.

3(A). WHY SHOULD ONLY THERAPEUTIC CLAIMS REQUIRE RCTS?

Sackett et al. ('Evidence-Based Medicine') denied that 'EBM is all about RCTs' by, in the first place, allowing that other ways of garnering evidence are more appropriate for other hypotheses in medicine that are not about the efficacy of some proposed treatment: for example, hypotheses about the accuracy of a diagnostic test or about prognosis, given a certain range of symptoms. It is only in the case of claims about the effectiveness of proposed therapies that RCTs are definitely to be preferred: 'It is when asking questions about therapy that we should try to avoid the non-experimental approaches' and therefore randomize (72).[11] But *why* should randomization play such a specially weighty role in assessing claims about therapeutic effectiveness, when, far from playing any special role in the case of apparently quite similar kinds of hypothesis, randomized designs are explicitly contra-indicated in those cases? It is, relatedly, and as Bayesians have sometimes pointed out, rather strange that the RCT methodology should be thought of as *the* embodiment of correct scientific method when it comes to assessing the effectiveness of some proposed therapy and yet it seemingly plays no role at all in physics, unambiguously the most successful science we have.[12]

3(B). THERE'S NO SUCH THING AS 'FACE VALIDITY'

The 'concession' that RCTs are not necessary in the case of certain treatments, specifically those exhibiting 'face validity', is clearly aimed at the sort of example – like appendectomy for acute appendicitis – that I mentioned earlier. But what does it mean to exhibit 'face validity'? Nothing in therapeutic medicine surely is 'self evident' or *a priori*: judgements of 'face validity' too are based on some sort of evidence (in fact very extensive evidence). Moreover, since 'the team' that is 'unanimously to judge face validity' is surely not meant to be making such judgements out of nowhere (if they were, then their subjective views would be entirely irrelevant so far as the objective evidential situation is concerned), the evidence on which they rely may be implicit but surely must be objective. This particular 'concession' surely ought, then, to have forced a wider-ranging reappraisal of what exactly randomization is supposed to deliver in terms of epistemic weight. Surely – intuitively speaking – the claim that appendectomy is effective for acute appendicitis is as securely based on evidence (I would suggest altogether *more* securely based on evidence) than any currently accepted treatment that owes its acceptance to success in an RCT. And if this is conceded in the case of treatments that possess 'face validity', why exactly should it continue to be held that, in the case of treatments where such validity is not apparent, and where no RCTs are available, the process of looking for other types of evidence is inevitably looking for the 'next best'?

In fact, the advice I quoted earlier from Sackett et al. (*Evidence-Based Medicine*) to 'stop reading' an article if it reports a study that has not been randomized is followed by further advice about what to do as an Evidence-Based Medic should it turn out that there is not even one study relevant to the clinical issue you are considering that has been 'properly randomized'. One piece of advice is 'See whether the treatment effect is so huge that you can't imagine it could be a false-positive study' (108). But again (like 'face validity') what anyone can 'imagine' is, of course, neither here nor there unless that inability to imagine is firmly based on (objective) evidential considerations; hence this psychological statement can, when translated into 'objectivese', only mean that there is near-overwhelmingly strong evidence that the treatment is effective. There seems to be some suggestion both here and at other places in the EBM literature that we, so to speak, do not need very strong evidence when a treatment has a large effect. The underlying idea seems to be that RCTs supply the most sensitive tests, the ones most likely to reveal delicate differences, but that requiring an RCT to see whether appendectomy is better than placebo for acute appendicitis would be like using a micrometer to judge if a football is bigger than a golf ball. But aside from the issue (taken up in section 4) of whether it is indeed true that RCTs are more 'sensitive' than other forms of trial, this idea is surely incoherent. We don't of course know the effect size for any treatment, we have *theories* about it and, we hope, evidence for it. How big the effect size is and whether or not we have evidence for that size of effect are separate issues. Of course it is true that, certainly on standard Bayesian probabilistic accounts of evidence, if we have some evidence for a very large (of course implicitly positive) effect size, then we automatically have *stronger* evidence for the logically weaker claim that the treatment has *some* positive effect. In that sense evidence may be 'easier to get' in the case of treatments that appear to have large effects. Nonetheless it would be clearly wrong to think that we need less evidence in the case of treatments that appear to have large effects. And indeed, as already pointed out several times, it seems clear intuitively that we have evidence for effectiveness in the cases of some such treatments that is at least as strong as the evidence for the effectiveness of any other intervention whether gained from an RCT or not.[13] Moreover, not all the treatments of unquestioned effectiveness that have never been validated in RCTs are ones with 'huge' effect sizes (though of course the most dramatic are). Given the 'if it moves, RCT it' mentality of current medicine the list of such treatments is growing smaller by the day but there are some left: for example, so far as I am aware, no RCT has ever been done on aspirin for mild headache relief.

3(C). HOW ARE DIFFERENT TYPES OF EVIDENCE TO BE AMALGAMATED?

The clarification of EBM, then, as we have seen, explicitly calls for various different types of evidence – from 'the best clinical expertise', from biochemistry and from clinical trials – to be 'incorporated' into the overall judgements that determine clinical practice. But then we surely need guidance on how these different types of evidence are to be amalgamated and in particular on what to do when different types of evidence seem *prima facie* to clash. How are we to identify 'best' expertise? If the 'best' clinicians are the ones whose expertise always leads to judgements in agreement with the results of trials, then of course nothing is added. And if not, wasn't part of the central drive behind EBM the fear that *any* clinician's judgement might be radically biased? Moreover, if there is some independent way of identifying 'best' clinical expertise, then what should be done when it clashes with the results of some well-performed trial? The initial EBM suggestion – perhaps even the current suggestion – is that the trial result should always prevail, but this, as we shall see shortly, is surely not a sustainable view.

3(D). HOW SHOULD THE EVIDENCE FROM INDIVIDUAL TRIALS ON THE SAME TREATMENT BE COMBINED?

The SIGN hierarchy (see above, p. 9) ranks meta-analyses or systematic reviews of the results of RCTs as providing the best possible evidence alongside individual RCTs (at least those RCTs that show 'little risk of bias'). Other hierarchies place meta-analyses and systematic reviews in the number one spot, with (large) individual RCTs in second place.[14] Meta-analyses and systematic reviews both attempt – in somewhat different ways – to amalgamate the results of different trials that have investigated the 'same' treatment for the same condition. It is notorious that medical trials often involve such small numbers that it is impossible to draw any firm conclusion from them.[15] It therefore sounds like a good idea, given that larger trials pose practical and often ethical difficulties and given that these smaller trials have already been performed, to somehow or other amalgamate the results of different trials on the same proposed treatment into one 'overall result'. Surely biases that may have crept into the individual trials (intuitively all the more likely when these are small) will tend to be eliminated when the trials are all brought together. The principle, of course, is that it should be reasonable to view the amalgamated data as equivalent to that coming from a much bigger trial whose study population is the union of the study populations of all the individual trials. And it is common to all approaches to clinical trials (both Bayesian and orthodox frequentist – not to mention the commonsense approach) that the bigger the trial, then *ceteris paribus*, the stronger the evidence it provides. However there are major issues about how exactly this amalgamation is to be carried out – in view of the fact that there will almost always be differences between the individual trials (perhaps, for example, in the inclusion (and exclusion) criteria applied to patients admitted to trials, or in the exact protocol for treatment). Methods are provided for dealing with such problems in the meta-analysis literature,[16] but there are many issues about the underlying rationale of these methods. (Obviously the fact that some such amalgamation methods appear in treatments that are as a matter of fact regarded as authoritative is, in itself, no rationale.)

Where it seems obvious that there are differences between the different component trials, a *systematic review*, which does not issue in some overall, and arguably possibly bogus, statistic is recommended. But this too is something of a dark art: complex protocols (which often differ at least in part from account to account) for how to rank and combine and how to produce some overall 'result' concerning the efficacy of the treatment concerned are laid down, but their underlying rationale is unclear.

Furthermore, both systematic reviews *and* meta-analyses face the (important) problem of the 'grey literature'. Medical journal editors – seemingly fooled by the purely semantic error of inferring 'not significant' from 'not *statistically* significant'! – have tended to be biased against publishing accounts of the results of trials that were 'negative' – that is, in which the 'null hypothesis' remained unrefuted and therefore no positive outcome for the therapy under test declared. (Indeed it seems that clinical investigators themselves have often tended not even to bother to send their results off to journals if they were 'negative' and therefore 'insignificant'.) Needless to say this means that the 'sample' of trials forming the basis for meta-analyses or systematic reviews is itself (uniformly) biased. Again this problem is well-recognized and responses have been developed. It is however unclear whether any of these responses is cogent – except in the case of the purely forward-looking response that the problem is likely to become a decreasing one, because of new regulations in many countries requiring registration of all trials and requiring the declaration of raw data, whether or not published. Needless to say this – though an important step forward – has no impact on the reliance that we can reasonably place on the results of *current* meta-analyses and systematic reviews.

3(E). BIASED RCTS?

The SIGN hierarchy (see above p. 9), along with others, involves a distinction between RCTs that are, and those that are not, 'at risk of bias'. This may seem somewhat surprising in that the official doctrine of the randomizers – as we will see in detail later – is that an RCT is, *by definition*, unbiased. A trial is (objectively) biased if (whether or not we know it) there is some difference between the experimental and control groups (other of course than the fact that they are given different treatments) that might plausibly play a role in producing the trial outcome. But, as for example, Mike Clarke, the Director of the Cochrane Centre in the UK, writes on the Centre's web-site: 'In a randomized trial, the only difference between the two groups being compared is that of most interest: the intervention under investigation'.

Now in fact, no one (not even Mike Clarke himself as he makes clear later in his article) really believes this. And this is reflected in the fact that most guides to RCTs recommend checking the two groups, once created by – let's suppose – an impeccable randomization, for 'baseline imbalances'. That is, for differences in factors that plausibly might play a role in the outcome independently of the treatment. If such 'baseline imbalances' are discovered, then RCT-ers recommend re-randomizing until two groups are created that do not exhibit such imbalances. (This seems – at any rate from the point of view of principle – rather quixotic advice. One could (again in principle, though there might be practical difficulties) instead deliberately match in advance for such 'known confounders' to create two 'equal' blocks and then, if you liked, choose which block becomes the treatment group by some random process.)

So perhaps when the SIGN hierarchy includes a category of RCTs that are at 'substantial risk of bias', this means trials in which baseline imbalances were not in fact investigated, and where it seems plausible in the light of background knowledge that some significant difference/s between the two groups occurred. (It is however difficult to know how that judgement could be made on the basis of reading only published reports.) Moreover it is difficult to see how, once it has been admitted that RCTs do not necessarily control for known confounders (and the suggestion to check for 'baseline imbalances' clearly embodies that admission), how anyone can continue to maintain that we know (or even that we have good reason to believe) that randomization does control for 'unknown confounders' (that is confounders that in fact play a role in the outcome but which are entirely unsuspected). And yet, as we shall see in section 4, precisely that claim is the (sociologically speaking) most potent argument for the epistemic superiority of randomized trials.

3(f). External validity?

Somewhat hidden away in the SIGN document from which the evidence hierarchy we have been discussing is taken as the concession that 'considered judgment' is required if the available evidence is to be sensibly applied to underwrite decisions about particular patients' treatments. Part of 'considered judgment' involves taking a view about the 'generalisability' of the evidence. This is presumably an allusion to the crucial, but systematically underemphasized, issue of the 'external validity' of a trial.

A trial has 'internal validity' to the extent that its outcome truly measures the impact of the treatment on the outcome *so far as the trial population is concerned*. (Of course it is another issue how we *judge epistemically* that a trial is internally valid – but, as already mentioned, it is generally assumed, for reasons that will in fact be questioned later, that a trial is internally valid exactly to the extent that it has been properly randomized.) Suppose that a trial deemed to be internally valid has a positive outcome – the treatment is then deemed to be effective so far as the trial population is concerned. Does this straightforwardly entail that it is a good idea to use the treatment in regular medical practice? This is the question of whether or not the trial has 'external validity'? Does the trial result 'generalise' to the 'target population'? This latter notion is of course far from precise – but can vaguely be taken to cover the set of people that are likely to be candidates for treatment by the medical community with the therapy at issue if that therapy is approved.

Raising the issue of external validity is definitely *not* a reflection of the general inductive scepticism inherent in 'Hume's Problem'. There is no sense in which the initial trial population can be considered a random sample from any population, and certainly not the 'target population'. Explicit (and sometimes not so explicit!) 'exclusion criteria' are applied in the selection of patients to be involved in trials (often based on ethical rather than epistemic considerations). All trial patients must express their 'informed consent' and there may, at least in some cases (those where psychological factors may play an important role), be a serious question as to whether those who agree to be involved in trials are representative of those who may eventually be treated.

That generalizability is 'more than a philosophical quibble' is – unsurprisingly in view of the above considerations – reflected in the fact that there are concrete cases in which a treatment that was 'validated' in an RCT and was subsequently adopted in medical practice and therefore given to a wider group than those involved in the original trial proved to be noxious. (One example is the drug benoxaprofen (trade name: Opren), a nonsteroidal antiflammatory treatment for arthritis and musculo-skeletal pain. This passed RCTs (explicitly restricted to 18 to 65 year olds) with flying colours. It is however a fact that musculo-skeletal pain predominantly afflicts the elderly. It turned out that, when the (on average older) 'target population' were given Opren, there were a significant number of deaths from hepato-renal failure and the drug was withdrawn.)

4. Why Should Randomization Deliver Higher Evidential Weight? An Analysis of the Arguments

The issues raised in section 3 about the evidence hierarchy are, then, undoubtedly open and interesting ones that philosophers of science could certainly help to clarify. However, the central feature of the hierarchy, and the one that I will concentrate on for the rest of this paper, is this. The hierarchy continues to give the most significant evidential role to randomized trials (and meta-analyses thereof). RCTs (and 'amalgamations' thereof) are definitely ranked highest in terms of evidential weight provided and definitely as carrying more such weight than, for example, even the most carefully performed 'case-control' study on the same treatment. This is so even if, for example, the case-control study involved many more patients than the RCT. Indeed the way in which the hierarchy is to be applied explicitly requires that an RCT result automatically *trump* that of any study based on an alternative design:

> The best evidence to use in decisions is then the evidence highest in the hierarchy. Evidence from a lower level should be used only if there is no good randomized controlled trial to answer a particular clinical question. (Barton 255)

Admittedly 'small inadequate [RCTs] do not automatically trump any conflicting observational study' but 'If high quality randomized controlled trials exist for a clinical question then they [*do*] trump any number of observational studies' (256). This is presumably so even if the observational studies are themselves of high quality (and so in particular every effort has been made to use historical controls that are carefully matched in terms of 'known confounders' with the patients involved in testing the new treatment); and it is presumably so even if the total number of patients involved in the observational studies is altogether greater than the number involved in the single RCT.

EBM-ers believe that there are solid arguments that establish that the results of randomized trials carry greater objective epistemic weight. EBM-ers have very largely convinced the medical profession that this is true. If I had a pound for every time that the phrase 'gold standard' occurs in medical journals in the past decade or so in connection with RCTs then I could readily pay off my mortgage. And the standing of RCTs within medicine is further reflected in the following comment from the editors of *The British Medical Journal* (2001); emphasis supplied: 'Britain has given the world Shakespeare, Newtonian physics, the theory of evolution, parliamentary democracy – *and the randomized controlled trial*' (1438). It almost seems as if the medical community assumes that the result of an RCT is *obviously* more telling, 'more scientific' than that of any other type of evidence. But it is, of course, *not* obvious – it needs to be argued, and clearly, then, argued from a more basic perspective: from the basis of the general principles of scientific evidence. Five such arguments for the special epistemic role of randomization can be found in the literature – each widely advocated. Is any of these arguments compelling?[17]

4(A). FISHER'S ARGUMENT FROM THE LOGIC OF SIGNIFICANCE TESTING

R. A. Fisher was the real inventor of randomized trials (though it was Austin Bradford Hill who first applied the methodology in the assessment of medical therapies). Fisher's own initial argument for randomization (see, e.g. *Design of Experiments*) was that it is necessary in order to underwrite the logic of his significance tests: only if the division between those in the experimental and those in the control group has been made by some random process can his significance test methodology be coherently applied.

Laying aside issues about whether or not it is a good idea to apply Fisher's methodology at all (and certainly the medical community has long been convinced that it is, while Bayesian critics have always insisted that it is not), what exactly was his argument for randomization as an essential underpinning for that methodology? And is that argument telling *within its own terms*?

Fisher's method, as is well-known, is essentially the following:

1. Set up the 'null hypothesis' – in the clinical case, this will state that the therapy at issue has no effect on the outcome under investigation (or rather that the therapy has *no extra effect* compared to whatever the control treatment is – placebo or conventional treatment).
2. The 'null' is immediately identified with 'the' chance hypothesis: this states that the probability that any given patient exhibits the outcome under investigation (remission of certain symptoms, let's suppose) is the same whether they were given the treatment under test or the alternative, and this in turn implies a particular probability distribution for the measured difference in those responding positively to the treatment in the two groups that is symmetrical about a mean value of zero.
3. In a one tail test (where the assumption is implicitly made that at least the treatment under test won't make things worse), the value of that difference V is then identified such there is only a 5% chance that such a difference (or a bigger one) would be observed if the chance hypothesis

were indeed correct; the set of values consisting of V and bigger differences being called the 'significance region'.

4. If, when the trial is performed, an outcome is observed that falls in the significance region, then the null hypothesis is rejected, and hence in effect the hypothesis is accepted that the treatment is effective (although Fisher was, as is well known, a hard-line falsificationist and for him acceptance of the effectiveness of the treatment is just synonymous with, means no more than, the fact that an outcome has been observed which has only 5% chance or less of happening if the null is correct, despite which his method is routinely used *positively to endorse* treatments).

As Colin Howson points out in an exceptionally clear treatment (48–51), the argument underlying Fisher's methodology is a version of the 'No Miracles Argument' – an argument that has attracted a good deal of attention in more recent debates about scientific realism. Fisher's methodology clearly embodies the claim that if something happens that would be extremely improbable if some hypothesis H (here the assumption that the treatment has some positive extra effect compared to placebo or conventional treatment) were false, then we are entitled to assume that H is in fact not false but true. Of course, as Fisher recognized, you can always be unlucky – maybe the null is true (and therefore the hypothesis of the treatment's positive effect false) even though you reject it: indeed, using a 5% significance region, you would expect to make this (so-called type I) error 5% of the time on average. This is why Fisher was always cagey about 'acceptance'. He wrote that the force of the inference from a 'statistically significant' outcome to the falsity of the null is

> logically that of a simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution [the chance hypothesis] is not true [and therefore we should reject the 'null' and 'accept' the theory that the treatment is effective]. (Fisher, *Statistical Methods* 39, italics in original)

There has of course been endless discussion of Fisher's overall approach. I agree with the Bayesians that it seems difficult to see any good reason to introduce acceptance and rejection rules here at all and that one should instead rest content with probabilities (the probability that one hypothesis rather than another is true) in this acknowledgely probabilistic area. And I agree with Colin Howson and other Bayesians that anything like a sensible version of the No Miracles Argument (when modelled probabilistically) requires some assumption about 'prior probabilities' (not, in my view, 'subjective' but instead essentially encoding the impact on the hypothesis of the *other* relevant evidence aside from that directly involved in the test under consideration). But these issues are not relevant to our present concern – which is to find out what role Fisher saw randomization as playing in his proposed methodology and whether the argument he presents for its playing that role is convincing when considered entirely *within* Fisher's own framework.

We need to concentrate on point 2 of the above outline of Fisher's methodology. It is easy to accept unquestioningly the identification of the null hypothesis (initially, remember, just the hypothesis that the treatment is not effective) with the chance hypothesis (binomial distribution of the difference in positive outcomes between the two groups about a mean of zero). This means taking it that the chance hypothesis just *is* the null hypothesis. But this is a paradigm-dependent judgement and Fisher himself, who of course invented the paradigm, knew better.

Fisher's No-Miracles-style argument requires the premise that some event has occurred that is extremely unlikely to have occurred if the *hypothesis H* (again in our case: the hypothesis that the treatment has at least some positive effect compared to control) *is false.* Suppose that some positive result has been observed in a clinical trial: one explanation of this positive result would of course be that the therapy on trial is indeed effective, i.e. that H holds. We want to be able to infer H (or something like 'the rational presumability of H') if an event occurs which has a very low probability on the assumption that H is false. But the negation of H, just like the negation of any (reasonably) precise assertion, as it stands, makes an extremely weak claim and is therefore consistent with a whole range of more definite possibilities. Certainly we cannot, without further ado (that is, without further assumptions), simply identify the negation of H with *any* determinate hypothesis of the chance of a specific patient having a positive outcome, let alone with the very particular such hypothesis that that probability of recovery is the same independently of whether the patient was in the experimental or the control group.

There are clearly many possible alternative ways in which it might be false that the treatment is effective aside from the probability of recovery being independent of treatment. It might for example be that the treatment is ineffective but all the younger healthier patients happen to be in the experimental group, so that the probability of recovery is higher in that experimental group. It might be that the treatment is ineffective but clinicians involved actually selected the patients with the best prognoses, whatever treatment they were given, to form the experimental group. It might be that the treatment is ineffective, but the clinicians knew which patients were in the experimental group and – hoping for a 'significant' outcome to the trial – lavished more attention on them. The list of possible scenarios in which H is false but the probability of recovery is not independent of which group you are in can clearly be extended as long as your imagination (or, more pertinently, patience) lasts.

As noted, Fisher was fully aware of this and realized therefore that some further assumption was needed in order to justify the identification of the bare assumption that H is false – the treatment has no effect – with *any* determinate probabilistic hypothesis and ultimately, of course, with 'the null hypothesis' of binomial distribution of the difference in positive outcomes between the two groups about a zero mean. Fisher himself introduced the assumption at issue and his argument for it in the context of his celebrated 'tea lady' example.

Fisher's 'tea lady' claimed to be able to tell by taste whether the milk had been added to a cup of tea before, or after, the tea. Here the hypothesis H that we want to test is that the tea lady does at least have *some* direct powers of discrimination between the two ways of serving the tea. (This is, of course, somewhat vague – but the vagueness reflects that in the standard clinical trial where the question at issue is whether the treatment has *some* positive effect.) Suppose we have decided on some specific test of H – in Fisher's example, the test involves her being presented with 8 cups of tea, 4 of which are to be milk-first (M) and 4 tea-first (T). Again on reflection it is obvious that the negation of H cannot just peremptorily be identified with any determinate assumption about the probability that she makes a certain number of correct identifications. There are again endless ways in which H might be false (that is, she in fact has no special ability based on tasting the tea to distinguish M from T cups) – she might have happened to sample the M cups before T ones and, soon becoming bored with tea (though this is unlikely given that she was, it seems, English!), had a more positive attitude to the tea she sipped earlier in the tasting; the M teas might, in this particular test, happen to have been the ones in thicker cups and she responds differently to tea in thicker as opposed to thinner cups; she might have an informant who signals to her with some, perhaps variable, degree of accuracy how the tea was produced, and so on. Many of these will seem implausible, but that is not (yet) the point: the *logic* of the argument is what is at issue.

Fisher saw, therefore, that an extra condition has to be satisfied before the claim that she has no powers of discrimination could be identified with one particular probabilistic hypothesis. It is exactly at this point that randomization comes in, and, he believed, must come in, to save the day. He argued that it is legitimate to identify what might be called the 'true null' (i.e. the simple negation of the claim that, in this case, she can distinguish M from T by taste) with the 'probabilistic null' (in this case, she is simply guessing each time with a half chance of success for each cup) *if, but only if*, the order in which the individual tea cups is presented has been decided by some random process – most 'expeditiously from a published collection of random sampling numbers' ('Statistical Tests' 474). This, and only this, condition will guarantee that all the possible combinations (in fact 70 of them) of tea-first and milk-first cups in this 8-cup test have the same chance of occurring – in the frequency sense that each such combination will occur in the limit, with the same frequency of 1/70. Randomization thus implies, and only if the allocation was randomized does it imply, claimed Fisher, that if the 'tea lady' in fact has no powers of discrimination, then the probability that she will, for example, get all 8 cups right on the particular occasion at issue is simply the frequency with which the particular combination that has occurred in this individual test will occur in the long run – that is, 1/70.

The argument does have a strange intuitive appeal – otherwise it would be difficult to explain why Fisher has managed to befuddle the medical trial community for over 70 years. But as Bayesians from Savage, through Good to Howson have insisted, it does not bear critical scrutiny (for references again see Howson).

Indeed, Fisher's suggestion does not even begin to meet the initial objection. Even given a randomized assignment, there are *still* indefinitely many ways in which the tea lady may fail to have the powers of discrimination she professes, all of which are incompatible with what Fisher now claims is definitively singled out as *the* null hypothesis – that is, the hypothesis that she is guessing in every case in a way that makes it just as likely she will be right as wrong. Suppose, to make it as sharp as possible, that the test outcome is that she made the correct identification in all 8 cases. One explanation of this result, of course, is that she does have the ability she professes to discriminate M from T by taste, but it is certainly not true that the randomized allocation means that the *only* alternative explanation is that she guessed in each case and got lucky (to the tune of pulling off a 1/70 shot). She may have a confederate who signals to her; she may have noticed earlier that there is a slight physical difference to the appearance of the tea when the milk has been put in first and simply applied this rule to this case; she may have been able slyly to peak through a window while the tea cups were being prepared; again this list could be extended indefinitely.

Moreover, if we concentrate not on what would happen in the long run (and why should the impact of the evidence that we have now from the one trial depend on what might or might not happen if the trial were repeated indefinitely?), then various other practical alternatives present themselves. Not knowing that a random method has been used, the tea lady (assumed, for the sake of this part of the argument, not in fact to possess any real ability to discriminate by taste) may, for example, know that the investigator in charge of the trial is a very methodical person and hence be quite convinced that the cup-preparations will strictly alternate; and it might happen that *this particular* random allocation does indeed produce MTMTMTMT – this order is after all just as likely as any other particular (fully specified) one; in such a situation rather than 1/70 being the appropriate probability of guessing correctly, the tea lady, given her false but nonetheless real belief, needs to guess correctly only in the case of the first cup and so the appropriate probability (arguably) is 1/2. And so on, and so on.

This may seem pettifogging since many of these alternative explanations of any apparent success she may display seem far-fetched. This is not in dispute, but the point, one we will need to emphasize again in other contexts later, is that randomization does not free us from having to think about alternative explanations for particular trial outcomes and from assessing the plausibility of these in the light of 'background knowledge'. And the direct immediate point is that Fisher's argument for randomization – that it is necessary to underwrite the validity of his significance test methodology – fails even in its own terms. (Moreover, as Bayesians have argued, even had it succeeded, the response might well have been 'who cares, given that in any case significance testing is so obviously and so deeply problematic?')

4(B). RANDOMIZATION CONTROLS FOR ALL 'CONFOUNDERS' – KNOWN AND UNKNOWN

It is a second argument for randomization that has, as a matter of empirical fact, exercised far and away the strongest impact on the medical profession. In order to explain the idea behind this argument, the first question to consider is 'Why "control" a trial at all?'.

The answer, clearly, has to do with seeking evidence that any positive effect observed in a clinical trial is genuinely attributable to the treatment under investigation rather than to some other factor/s. Suppose, to take a hoary old example, we are interested in the effect of giving patients suffering from common colds regular doses of vitamin C. We take a bunch of patients with colds, give them vitamin C, and then record, say, how many of the patients recover within a week. Suppose they *all* recover. It would be an obvious mistake to infer from this that vitamin C is an effective treatment for colds. The first reason is the possibility that all those patients would have recovered from their colds within a week even had they not been given the vitamin C. If that were true, then to infer that vitamin C is an effective treatment would be to commit a particularly egregious version of the famous – and sadly ubiquitous – '*post hoc ergo propter hoc*' fallacy.

From the epistemological point of view, it would surely be ideal if we could know what would have happened to *those particular* patients had they not been given the vitamin C. But of course we have no evidential access to these counterfactual histories – all we know about those patients is that they *were* given the vitamin C and *did* recover within a week. The best we can do, it seems, is take a *different* bunch of patients, also suffering from colds, and treat them differently – say in this case by doing nothing to them (in the jargon of the trade they would then form a 'natural history control group').

Suppose that none of the patients in this 'control group' recovers from his or her cold within the week. Given that those in the 'experimental group' who were given vitamin C *did* recover, it would be tempting to infer that this (now) controlled trial had shown that vitamin C is effective. But a moment's thought shows that this too is premature. Suppose that those in the control ('natural history') group were all suffering from much heavier colds than those in the experimental group, or suppose that all those in the control group were older people suffering from other conditions as well, whereas all those in the experimental group were, aside from their colds, young, fit and healthy. We intuitively want the two groups to be 'equal (at least on average) in all other (relevant) regards'. Only to the extent that we have evidence that the two groups were equal in other relevant respects, would we be justified in claiming that we have evidence from the trial outcome that the only difference between the two groups is the difference in treatment. And so only to that extent would we be justified in claiming that we have good evidence that any measured improvement in the experimental group relative to the control is due to the treatment.

Now we can of course – at least in principle (matters are much more complicated when it comes to the *practice* of clinical trials) – ensure equality in the case of any particular factor that commonsense (or 'background knowledge') tells us might well be relevant. This could be achieved by deliberately matching the two groups relative to the factor at issue. So, continuing with our hoary example, given that the severity of the cold is clearly relevant, we could *match* the two groups in this respect: either by ensuring that everyone in the trial had colds of (at least approximately) the same severity or by ensuring that the proportions of those with severe as opposed to mild colds (on some agreed scale) is the same in both groups. Similarly since age and general health and fitness seem likely to be factors possibly relevant to early recovery, we can match the two groups so as to ensure that these factors are similarly distributed within the experimental and control groups.

Suppose we have matched the two groups with respect to every factor that background knowledge makes plausible might be relevant to recovery. And suppose we again find a substantially higher proportion of recoverers within the group of those given vitamin C. Surely *now* we have telling evidence that vitamin C aids recovery from colds?

Well, still 'maybe not' and this for two separate reasons. *First* it may be (though we can't of course know it – a fact that we shall need to reflect on at length later) that the two groups now are indeed 'equal' with respect to *all* non-treatment-dependent factors relevant to recovery from colds (though this would actually be something of a miracle). In that case, we would indeed be justified in concluding that the observed difference in outcome was produced by the treatment rather than being due to some difference between the two groups. But it wouldn't follow that it was the fact that the treatment involved *giving vitamin C* that caused the improved outcome. It is conceivable – and there is some evidence (albeit disputed by some commentators) that it is in fact the case – that a patient's *expectations*, fired by being treated by an authority figure, play a role in recovery from, at any rate relatively minor, complaints. In the vitamin C case, as we are now envisaging it, those in the control group receive no treatment at all

– they are just the 'otherwise equal in all respects' comparators. But the very fact that they receive no treatment at all is obviously a difference and one that *might* be relevant.

This is the reason why the control groups in medical trials will generally be given either placebos or conventional treatment (sometimes there are two control groups in a trial: a placebo control *and* a conventional treatment control). That is, those in the control group will not be left untreated, but will instead be treated either with the currently accepted treatment for the condition at issue or with a substance 'known' (of course from non-trial studies) to have no specific biochemical effect on the condition, but which is intended to be indistinguishable so far as the patient is concerned from the 'active' drug under test. This feature of clinical trials permits a further feature with undoubted methodological merits – namely that the trials can, again at least in principle, be performed 'double blind'. The two treatments ('active' drug and placebo (or conventional treatment)) can be sorted into packets marked simply, say, 'A' and 'B', by someone not involved in seeing patients, and then delivered to the patient in ways that are indistinguishable: hence neither the patient herself nor the clinician involved knows whether that particular patient has received the 'active' drug or not. Indeed if the trial is not double blind then even though the two groups may start out equal in terms of all other possibly relevant causal factors, a possibly relevant difference might be introduced during the course of treatment: for example, a doctor who knew that a particular patient was in the experimental group might treat that patient more assiduously; or the placebo effect *might* be diluted in a patient who knew that the treatment she was being given was a placebo, while those patients who knew they were taking the 'active' treatment would, speaking intuitively, get both the effect of the 'characteristic features' of that treatment (if any) *plus* the full placebo effect.[18]

But even after matching with respect to factors that background knowledge implies may plausibly play a role and even after using placebo or conventional treatment controls to try to ensure that no other difference is introduced into what otherwise might have been 'equal' groups just by the way the two groups are treated in the trial, there is still a *further* problem. This is that the list of factors that *might* make a difference to treatment outcome is of course endless. The groups in our hoary common cold example may have been deliberately matched with respect to obvious factors such as severity of symptoms, sex, age, general level of health and so on, but what if your rate of recovery from colds depends significantly on whether you were breast- or bottle-fed as a child, or on whether you have previously been infected with, and recovered from, some particular non-cold virus, or . . .? This is the problem of '*unknown* (better: unsuspected) *factors*' – by definition the groups cannot be deliberately matched with respect to unknown factors, so it is certainly *possible* that even groups perfectly matched on known factors are significantly different in respect of some unknown one, and it is therefore possible that evidence provided by the observed improved outcome in the experimental group in favour of the effectiveness of vitamin C is spurious.

The strongest argument for the epistemic superiority of randomized trials ('strongest' in the sociological sense that it is the one that has convinced most people in medicine) is precisely that RCTs are alleged to solve the problem of 'unknown factors': a randomized trial is – allegedly – controlled for all factors known *and unknown*.

This is a claim that is often made in the literature about clinical trials. As noted earlier, Mike Clarke, the Director of the Cochrane Centre in the UK, for example, states on the Centre's Web site: 'In a randomized trial, the only difference between the two groups being compared is that of most interest: the intervention under investigation'.

This seems clearly to constitute a categorical assertion that by randomizing, *all other* factors – both known and unknown – are equalized between the experimental and control groups; hence the only remaining difference is exactly that one group has been given the treatment under test, while the other has been given either a placebo or conventional therapy; and hence any observed difference in outcome between the two groups in a randomized trial (but *only* in a randomized trial) *must* be the effect of the treatment under test.

Clarke's claim is repeated many times elsewhere and is widely believed. It is admirably clear and sharp, but it is clearly unsustainable (as indeed Clarke himself allows later in his article). Clearly the claim taken literally is quite trivially false: the experimental group contains Mrs Brown and not Mr Smith, whereas the control group contains Mr Smith and not Mrs Brown, *etc*. Some restriction on the range of differences being considered is obviously implicit here; and presumably the real claim is something like that the two groups have the same means and distributions of all the [causally?] relevant factors. Although this sounds like a meaningful claim, I am not sure whether it would remain so under analysis (for reasons hinted at later). And certainly, even with respect to a given (finite) list of potentially relevant factors, no one can really believe that it automatically holds in the case of any *particular* randomized division of the subjects involved in the study. Although many commentators often seem to make the claim (and although many medical investigators unquestioningly following the 'approved' methodology may believe it), no one seriously thinking about the issues can hold that randomization is a *sufficient* condition for there to be no difference between the two groups that may turn out to be relevant.

Here is one amusing counterexample to the sufficiency claim. A study by L. Leibovici and colleagues was published in the *British Medical Journal* in 2001 entitled 'Effects of Remote, Retroactive, Intercessory Prayer on Outcomes in Patients with Bloodstream Infection: Randomized Controlled Trial'. The study looked at 3393 inpatients at the Rabin Medical Centre in Israel during 1990–96 who had been admitted because suffering from various bloodstream infections. In July 2000 (so, note, between 4 and 10 years *after* they had suffered these infections), a random number generator was used to divide these patients into two groups – which of the two became the 'treatment (or intervention) group' was then decided by a coin toss. One thousand six hundred and ninety one patients were, so it turned out, randomized to the 'intervention' group and 1702 to the control group. A careful check was made for 'baseline imbalances' with regard to main risk factors for death and severity of illness. ('Baseline imbalances', as noted earlier, are differences between the two groups in respect of known[19] prognostic factors produced in a 'purely' randomized trial. In such a trial no effort is made in advance to balance the groups in terms of any given factor, but these imbalances may be identified *post hoc* by inspection of the groups produced by the randomization.) But no significant baseline imbalances were found. The names of those in the intervention group were then presented to a person 'who said a short prayer for the well being and full recovery of the group as a whole'. Then, but only then, were the medical records of all the patients checked for (i) those patients' mortality; (ii) their length of stay in hospital; and (iii) the duration of the fevers they had suffered. The trial was not only impeccably randomized but was clearly 'double blind' – *at the time when the outcomes were produced* neither the patient nor the doctors treating them could possibly know which arm of the trial they were in!

The results were that mortality was 28.1% in the 'intervention' group and 30.2% in the control group, a difference that orthodox statistical methodology declares (narrowly) 'non-significant'; however both length of stay in hospital and duration of fever were significantly shorter in the intervention group ($p = 0.01$ and $p = 0.04$ respectively).[20] Leibovici and colleagues drew the conclusion that

> Remote, retroactive intercessory prayer said for a group is 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. ('Effects' 1451)

Although it ought to have been clear that the authors were writing with tongues firmly in cheeks (for example they remark, po-facedly, that 'no patients were lost to follow-up'!), the paper produced a heated discussion on the BMJ website, which showed that some commentators clearly were ready to take the result seriously. But even the most religiously minded are surely unlikely *really* to believe that the mysterious ways in which god sometimes allegedly moves include predicting at time *t* that some prayer will be said on behalf of some patients some time between *t* + 4 years and *t* + 10 years, and intervening in the course of nature at *t*, on the basis of that prediction, to give those patients a better (average) outcome!

Leibovici himself fully agreed with this as he made clear in the subsequent discussion:

> If the pre-trial probability [of the eventual 'result'] 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. ('Author's Reply' 1037)

The sentiment, entirely in line with Bayesian, as opposed to classical statistical, methodology of course, is that we need to take into account not only the 'improbability' of a particular outcome if the 'null hypothesis' is correct (that is, there is no difference between the two groups despite the remote intercessory prayer 'intervention', and that consequently any observed difference in outcome between the two groups is due to chance), but also the prior probability of the 'non-null' (here the hypothesis that the prayer really did have a retroactive effect).

But although Leibovici may not have intended the study to be taken seriously as a basis for 'treatment', it surely *is* to be taken seriously as a criticism of orthodox statistical methodology and in particular of the suggestion that a properly randomized study *always* produces real evidence of effectiveness. Leibovici insisted, note, that 'the details provided (randomization done only once, statement of a prayer, analysis, *etc*) are correct'. So the fact is that this was a properly randomized study (in fact a commendably large one) that happened to produce what we take ourselves to know *must be* the 'wrong' result. Obviously what *must* have happened here is that although the division into 'intervention' and control groups was done impeccably and although the double blinding was equally impeccable(!), 'by chance' some unknown confounder/s *were* unbalanced (though of course we were, by definition, unaware of this), and it was this 'unknown' imbalance (or much more likely compound of imbalances) that produced the difference in outcome.

In sum, despite what is often said and written, no one can seriously believe that having randomized is a *sufficient* condition for a trial result to be reasonably supposed to reflect the true effect of some treatment. Is randomizing a *necessary* condition for this? That is, is it true that we cannot have real evidence that a treatment is genuinely effective unless it has been validated in a properly randomized trial? Again, some people in medicine sometimes talk as if this were the case, but again no one can seriously believe it. Indeed, as pointed out earlier, modern medicine would be in a terrible state if it *were* true. As already noted, the overwhelming majority of all treatments regarded as unambiguously effective by modern medicine today – from aspirin for mild headache through diuretics in heart failure and on to many surgical procedures (appendectomy, cholecystectomy, etc., etc.) – were never (and now, let us hope, never will be) 'validated' in an RCT.

The above criticism – particularly of the argument for the alleged sufficiency of randomization to provide solid evidence that a treatment is effective – will be regarded by some as an attack on a straw man. Maybe this straw man produces real writing, but if so, it is of self-consciously simplified accounts aimed at medical practitioners (or perhaps those involved with the administration of research) with no knowledge of, or taste for, statistical niceties. The serious claim is, *not* that in a randomized trial all other factors aside from the treatment are automatically equal in the two groups, but rather that this is highly *probable*. A positive result in a randomized test, because the two groups are *probably* equal in all other respects, gives us, not of course foolproof, but still *the best* evidence of treatment effectiveness that we could possibly have. We do not eliminate entirely the possibility of 'bias' by randomizing, but we do 'eliminate' it 'in some probabilistic sense'.

The problem with this suggestion is that – for all its seeming plausibility and indeed for all its widespread acceptance and therefore immense practical significance – it seems difficult to make anything like full intuitive sense of it on the basis of the orthodox approach to statistics. The latter (officially) refuses to deal in the probability of hypotheses at all, but only in the acceptance or rejection of hypotheses that attribute some probability to the values of some random variable. In order even to begin to make sense of the claim, we would need to be able to show that, for any particular (potentially) prognostic factor aside from which treatment a patient is given, it is probable that that extra factor is identically (or near identically?) distributed in the two groups – treatment and control. Any plausibility that such a claim might appear to have seems to depend, however, on confusing what can reasonably be asserted in the case of a single random division with what might reasonably be asserted about an *indefinite number of repetitions* of the random division.

What might it mean to claim that it is improbable that factor X is identically distributed between the two groups? Assuming the classical, non-Bayesian, frequentist approach to probability (and there is no *direct* role for randomization according to the Bayesian approach[21]), it can only amount to a claim about an *indefinite series of repetitions of the trial*: if you were to take a population (or perhaps a series of 'equivalent populations' whatever that exactly means), randomly divide it in two lots and lots of times and record the cumulative relative frequency of positive values of X in the two groups (assume for simplicity that X is a two-valued random variable), then in the indefinite long run that frequency would be the same in the experimental and control groups, and in fact would be the same as the actual frequency of positive values of X in the study population as a whole. But medical researchers involved in some particular trial do not make a random division indefinitely often, they do it *once*![22] In that one trial, factor X *may* be as substantially unbalanced between the two groups as you like, and there seems just to be no way to quantify what the 'probability' of a substantial imbalance is: 'single case probabilities' not being properly defined on this approach. Once you further take into account the fact that, by definition, the list of possible 'unknown' factors is indefinitely long, then matters become even murkier. Even if one wanted to insist that despite the lack of any adequate formal analysis it was somehow 'intuitively' clear that for any single factor X, it is 'improbable' that it is significantly maldistributed between the two groups in a single randomization, it would not of course follow even 'intuitively' that it is improbable that *there is* no factor relative to which the single randomization is unbalanced – because of the lack of any real grasp of the list of potential other factors and of how large it is, this just seems to be, even intuitively, undefined.[23]

I repeat only because this is, if correct, of such great significance: the argument that has convinced the great majority of the medical community that RCTs supply the 'gold standard' is without real foundation.

4(C). 'SELECTION BIAS'

A third argument for the value of randomized controls is altogether more down-to-earth. If the clinicians running a trial are allowed to determine the arm to which a particular patient is assigned then, whenever they have views (even subconscious ones) about the comparative merits and comparative risks of the two treatments, there is room for those clinicians to affect the outcome of the trial.

As I use the term, then, a trial suffers from 'selection bias' if there are differences between the two groups resulting from the selections made by the researchers involved concerning which patients are assigned to which groups (and of course the trial *may* suffer from such bias if researchers' decisions might have resulted in significant differences). It should be noted, though, that this term is sometimes used in other, quite different senses in the clinical trials literature.[24] There are several ways in which selection bias in my sense might conceivably be actualized. Clinicians might, for example, – no doubt subconsciously – predominantly direct those patients they think are most likely to benefit to the new treatment or, in other circumstances, they might predominantly direct those whom they fear may be especially badly affected by any side-effects of the new treatment to the control group (which will generally mean that those patients who are judged frailer will be overrepresented in the control group, and that is of course likely to *overestimate* the effectiveness of the therapy under test). Or, since how the eligibility criteria for a trial apply to a particular patient may be open to interpretation, if a clinician is aware of the arm of trial that a given patient will go into, then there is room for that clinician's views about whether or not one of the therapies is likely to be more beneficial to affect whether or not that patient is declared eligible.

If the investigators are able to choose the arm of the trial that a particular patient joins then this means – or at any rate usually means – that the trial is at best single blind: that is, only the patient and not the clinician is (explicitly) unaware of which arm of the trial they are in. But this further opens up the possibility of other factors not directly linked to the treatment being differentially brought to bear. For example it opens up the possibility that the doctor's expectations about likely success or failure may subconsciously play a role in affecting the patient's attitude toward the treatment s/he receives, which may in turn affect the outcome – especially where the effect expected is 'subjective' and/or comparatively small. It may also mean that such patients receive better levels of ancillary treatment, independently of whatever is going on in the trial. Finally, performing the trial single-blind also means that the doctor knows which arm the patient was on when coming to assess whether or not there was any benefit from whichever treatment was given – the doctor's own prior beliefs may well affect this judgement whenever the outcome measure is at any rate partially subjective (some pain relief, some improvement of mood, etc).

It is undeniable that selection bias may 'confound' a trial. Because it provides an alternative explanation for positive outcomes (at any rate for small positive effects[25]), we need to control for such bias before declaring that the evidence favours the efficacy of the treatment. One way to control is by standard methods of randomization – applied after the patient has been declared eligible for the trial. The arm to which a given patient is assigned will then be determined by the toss of a coin (or more usually a random number table) and not by any clinician.

This is surely a cast-iron argument for randomization: far from facing methodological difficulties, it is underwritten by the simple but immensely powerful general principle that one should test a hypothesis against plausible alternatives before pronouncing it well-supported by 'favourable' evidence. The theory that any therapeutic effect – whether negative or positive – observed in the trial is caused (or 'caused to some significant degree') by selection bias is always – again at least in the case of small apparent positive effects – a plausible alternative to the theory that the effect is produced by the characteristic features of the therapeutic agent itself. This is the one argument for the importance of randomization that is explicitly endorsed by Bayesians as well as their classical opponents.[26]

Notice however that randomization as a way of controlling for selection bias is very much a means to an end, rather than an end in itself. The important methodological point is that control of which arm of the trial a particular patient ends up in is taken away from the experimenters – randomization (as normally performed) is simply one method of achieving this.

4(d). observational studies are 'known' to exaggerate treatment effects

A fourth influential argument for the virtue of randomizing is that, no matter how the epistemological normative niceties about randomization play out, it is just *an empirical matter of fact* that other forms of trial have proved less reliable and shown themselves much more likely to produce a positive result than 'properly randomized' studies.

What forms of trials are there other than RCTs? Well, although the RCT is often unthinkingly identified with 'the experimental method', there is in fact surely no reason why a trial could not be performed that is fully experimental (in the sense that the clinicians intervene to create the two groups) but in which formal randomization plays no role. Suppose, for example, that clinicians conducting a pharmaceutical trial, and using a double blind technique in which the treatments are simply labelled A and B, produced some ordering of all the patients in the trial, say by day of the month of their birthdays (ranking ties alphabetically by surname), and finally gave treatment A to those at odd numbered places on that ordering, and treatment B to those at even-numbered places. Or suppose the trial was to consist of a large number of patients all attending a clinic at once, the patients might be asked to line up at the door of the clinic and patients alternately assigned to treatments A and B. It is difficult for me to see any reason why such protocols could be considered any less telling than one involving a random number table. But this 'haphazard trial' methodology (as it might be called) is not – sociologically speaking – considered a real rival to the RCT. (For one thing it is clearly at least as simple to randomize as to go through either of the above rigmaroles.) The main *de facto* rival is the historically controlled trial (also often called, especially by critics, one form of 'observational study').

The idea in such trials is that all the patients who are *newly* involved are given the (usually new) treatment under investigation; while the controls are supplied by previous patients treated by the previously preferred method. Speaking again from the point of view of intuitive scientific method, whether or not such a trial is telling will clearly depend a great deal on how carefully the patients in the trial have been matched with the controls – of course, by definition, such matching can only occur with respect to *known* (possible) confounders. Such trials are sometimes called 'observational studies' because they do not involve the investigators actively separating out treatment and control groups and hence are deemed non-experimental (perhaps a little strangely) and hence (?) 'observational'. Indeed, although often written up for publication as trials, what happens is often more naturally described as just the substitution of one new method of treatment for another, followed by a systematic comparison of the results achieved with those achieved (on what are hoped to be 'equivalent' patients) using the earlier treatment. It seems to me that the term 'historically controlled trial' is more accurate and suggestive, and I shall use that term below.

Some EBM-ers (inconsistently, I think) allow that historically controlled trials may be attractive from an ethical point of view in circumstances in which it already seems likely that a new treatment is effective – because in such a trial all active patients are, of course, given that new treatment.

(The consistent line, at least for the heroic EBM-er, would surely be that any such indication of increased effectiveness can only be based on subjective conviction since there is no real evidence of efficacy ahead of a properly conducted trial and hence that, whatever may be believed, there is no objective reason to think that patients would do worse on the control arm of a 'proper' RCT.) But, whatever the right view about the ethical attractions of such trials EBM-ers hold that they are, epistemologically flawed (or at any rate very much less than ideal) and the flaws manifest themselves in the fact that such trials are known to routinely lead to false positive conclusions about efficacy.

The evidence for this latter claim comes from 'meta-level' studies done in the 70s and 80s[27] which looked at cases where some single treatment had been assessed using *both* randomized *and* non-randomized trials – the latter nearly always involving 'historical controls'. These studies found that, in the cases investigated, the historically controlled trials tended to produce more 'statistically significant' results and more highly positive point-estimates of the effect of the treatment under test than did RCTs on the same treatment.

One point to be noted immediately is that, even if we accept the results of these meta-studies at face value, there is a clear circularity involved in the argument from those results to the claim that historically controlled trials 'routinely lead to false positive conclusions'. What these meta-studies found was that more of the historically controlled trials that they looked at had positive outcomes than did the RCTs that they looked at on the same treatment. However it follows from this finding the historically controlled trials 'exaggerate' the 'true effect', only if it is further assumed that the RCTs reveal (or at least are more likely to reveal) that true effect. This is of course *de facto* routinely taken to be the case in medicine (the RCT provides the 'gold standard' after all!), but in the current debate it is precisely the *normative* point at issue. Without this premise, the data from these meta-studies is equally consistent with the claim that RCTs consistently *underestimate* the true effect. (This is not just a philosopher's logic-chopping point. There have been serious suggestions in the literature – see for example Black– that there may be good reasons to suppose that RCTs will underestimate the 'true effect': at least if this is identified, not with the outcome in some artificial trial involving strict protocols, but rather with the effect that can be expected amongst real patients treated – 'in the wild' – by empathic doctors. Again this is fertile ground for analysis by philosophers of science.)

Moreover, whether or not it is reasonable to infer from these meta-studies that there is a *general* tendency for historically controlled trials to produce more positive results than RCTs is – at least – going to need some premise to the effect that the historically controlled trials and the RCTs considered in them arguably constitute a representative sample of trials performed using those methodologies (of course there is no formal sense in which the trials studied were drawn at random from some population). In fact there are clear suggestions that the particular historically controlled trials considered were comparatively poorly done – in that there were obvious ways in which the selected historical controls did not match the patients being given the new treatment: ways that it is plausible to believe were relevant to outcome. Indeed Chalmers et al. themselves suggest that the control and experimental groups in the historically controlled trials they investigated were patently 'maldistributed' with respect to a number of plausible prognostic factors. But how would the comparison look if only methodologically more sophisticated historically controlled trials were considered – ones were all the factors that background knowledge makes plausible might play the role of confounders have been considered and controlled for via suitable selection of the historical controls? More recent studies of newer research in which some therapeutic intervention has been assessed using both RCTs and 'observational' (non-randomized, historical) trials have suggested answers quite different from those arrived by Chalmers et al.

Kunz and Oxman, for example, looking again at a range of such cases where different types of trial had been made on the same intervention found that

> Failure to use random allocation and concealment of allocation were associated with relative increases in estimates of effects of 150% or more, relative decreases of up to 90%, inversion of the estimated effect and, in some cases, no difference. (1185)[28]

More significantly still, Benson and Hartz, comparing RCTs to 'observational' trials with concurrent but non-randomly selected control groups, found 'little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials' (1878).

And they suggest that the difference between their results and those found earlier by Chalmers et al. may be due to the more sophisticated methodology underlying the observational studies that they investigated compared to the ones that Chalmers et al. studied: 'Possible methodologic improvements include a more sophisticated choice of data sets and better statistical methods. Newer methods may have eliminated some systematic bias' (1878).

In the same issue of the *New England Journal of Medicine*, Concato, Shah and Horwitz argue that 'The results of well-designed observational studies . . . do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic' (1887).

They explicitly point out that their findings 'challenge the current [EBM-based] consensus about a hierarchy of study designs in clinical research'. The 'summary results of RCTs and observational studies were remarkably similar for each clinical topic [they] examined'; while investigation of the spread of results produced by single RCTs and by observational studies on the same topic revealed that the RCTs produced much greater variability. Moreover, the different observational studies despite some variability of outcome none the less all pointed in the same direction (treatment effective or ineffective); while, on the contrary, the examination of cases where several RCTs had been performed on the same intervention produced several 'paradoxical results' – that is, cases of individual trials pointing in the opposite direction to the 'overall' result (produced by techniques of meta-analysis).[29]

This last point is in line with the result of the 1997 study by Lelorier et al. who found – contrary at least to what clinicians tend to believe when talking of RCTs as the 'gold standard' – that 'the outcomes of . . . large randomized, controlled trials that we studied were not predicted accurately 35% of the time by the meta-analyses published previously on the same topics' (536).

The results of Concato et al. and of Benson and Hartz have in turn been criticized on methodological grounds by Pocock (who has consistently and strongly argued for the extra epistemic virtues of randomization) and Elbourne. This is largely on the grounds that both the randomized and non-randomized studies chosen by these analysts may have been unrepresentative in important ways. But again the only reason for thinking so seems to be a prior commitment to the idea that randomization (if properly done) is bound to be epistemically more telling.

There are certainly, then, issues about this argument that remain to be clarified. But it does, however, seem safe to claim that, as things stand, this 'reliabilist'-style point has not been shown to provide any solid independent reason for thinking that randomization has automatic extra epistemic weight.

4(E). RANDOMIZATION AND PROBABILISTIC CAUSALITY

The final argument for the special epistemic power of randomization comes from the burgeoning literature that attempts to articulate a defensible notion of 'probabilistic causality'.

We all do, it seems, happily accept that there are true claims of the form 'X causes Y' where X and Y are generic events ('Smoking' and 'Lung Cancer' form a favourite example), and where the alleged connection fails to be deterministic. In cases like 'the cause of this initially stationary 3-kg mass's being accelerated at 3 m/sec^2 is that a constant total force of 9 newtons was applied to it', the cause (the total force) inexorably brings about the outcome (the acceleration) given the initial conditions (initial velocity zero, mass 3 kg); but of course a's smoking tobacco (even heavily) does not inexorably bring about a's contracting lung cancer – yet we still want to say that smoking tobacco does cause lung cancer.

Clearly this non-deterministic causal claim has something to do with smoking increasing your chance of developing lung cancer. However, the causal claim smoking (X) causes lung cancer (Y) is clearly not captured (or not fully captured) by the claim that $\text{Prob}(Y|X) > \text{Prob}(Y)$ (or, equivalently, $\text{Prob}(Y|X) > \text{Prob}(Y|\neg X)$). First of all, increase in probability is symmetric: $\text{Prob}(Y|X) > \text{Prob}(Y)$ if and only if $\text{Prob}(X|Y) > \text{Prob}(X)$; while 'cause' is asymmetric – smoking causes lung cancer, but lung cancer does not of course cause smoking. Moreover, there are arguably at least two ways in which $\text{Prob}(Y|X)$ may be higher than $\text{Prob}(Y)$ without it being true *either* that X causes Y *or* that Y causes X. First, X and Y might 'just happen' to be correlated – they could for example be two variables that have just happened to increase together over time for two quite separate sets of reasons. (Elliott Sober's favourite example (see his (2001)) is the price of bread in London and the water level in Venice.)

More often, and more pertinently for us, $\text{Prob}(Y|X)$ might be higher than $\text{Prob}(Y)$, but rather than X causing Y they are both the effects of some underlying cause. So for example, the feature Z (owning more than 5 ashtrays) is such that $\text{Prob}(Y|Z) > \text{Prob}(Y)$, where Y is, again, developing lung cancer; however, ashtray ownership is clearly not a cause of lung cancer but is 'merely associated' with it: Y and Z are (in a rather stretched sense of cause and effect at least in the case of Z) both effects of the 'common cause' X – smoking tobacco. Reichenbach's celebrated 'common cause' principle says that C is a common cause of the two effects X and Y just in case C 'screens off' X from Y, that is, just in case the probabilistic dependence between X and Y disappears once we conditionalise on C: although $\text{Prob}(Y|X) > \text{Prob}(Y)$, $\text{Prob}(Y|X \,\&\, C) = \text{Prob}(Y|C)$. So for example the fact that smoking is a common cause of both lung cancer and ashtray ownership (and hence that there is no causal connection between ashtray ownership and cancer) is revealed by the fact that, although Prob(Lung Cancer|Ashtray ownership) > Prob(Lung Cancer), Prob(Lung Cancer|Ashtray ownership & smoking) = Prob(Lung Cancer|smoking). Various attempts to develop a full account of probabilistic causality, while differing in important details, all incorporate the common cause principle.

Several important contributors to the topic – notably Nancy Cartwright, David Papineau and Judea Pearl[30] – have explicitly claimed that it follows from their accounts that randomizing in a clinical trial is the vital ingredient in underwriting the claim that there is a *genuinely causal* connection between treatment and outcome, rather than a merely associational one (on the assumption, of course, that the outcome of the RCT is positive).

Since I have treated this final argument in detail in my recent article ('Why There's No Cause'), I shall here be brief. Although there are interesting differences between the accounts of Cartwright, Papineau and Pearl, they all are driven in their advocacy of randomization by the need to guard against the possibility that a positive result in a trial may issue from the fact that there is a 'common cause' of both the positive outcome and the treatment. Of course it is (as often) a little forced to say that there are 'causes' of treatment – but the idea is that there may be other factors that are related to whether or not patient *a* recovers that are *also* connected to whether or not *a* received the treatment (that is, the 'experimental treatment' rather than control). Being below 40 may play a causal role in good outcome and so if the relative numbers of those below 40 are markedly higher in the experimental compared to the control group, it may be that being below 40 is a 'common cause' of being treated and of having good outcome.

It is not surprising then that, despite the differences in their approaches and in the details of their arguments, Cartwright, Papineau and Pearl are all in effect presenting the third argument for the special power of randomization considered above – that is, the argument that randomization controls for all possible confounders known *and unknown* – though they present it in somewhat different guises. If this analysis – developed at length in 'Why There's No Cause' – is correct, then all these probabilistic causality arguments fail for the same reason that the earlier argument failed. No one can seriously believe that a single randomized experiment (all that you ever have in reality) can be guaranteed to involve groups that are balanced with respect to all possible unknown confounders; and the claim that you at least make this more probable by creating the groups via some randomizing process rests – to say the least – on very shaky grounds.

Conclusion

I have argued, then, that there are a great many issues of a traditional philosophy of science kind that arise when thinking through the question of how best to base medicine on evidence. In particular, it still remains to be seen if the belief, almost universally held within medicine, that RCTs provide the strongest, scientifically most telling kind of evidence can be underwritten. Clearly the practical consequences of resolving that issue alone would be enormous.

My own, more positive, slant on these issues is that applying scientific method properly to issues of therapeutic effectiveness in the end just throws us back on the 'scientific commonsense' underlying, for example, Mill's methods or Popper's ideas about testing; and in particular on the obvious (but in practice immensely powerful) idea that really telling evidence for any claim is evidence that at the same time tells against plausible rival alternative hypotheses. In the case of therapeutic claims in medicine, this means evidence that tells in favour of the claim that the treatment at issue (or rather the 'characteristic features' of that treatment over and above any placebo effect) is responsible for the observed outcome, but that at the same time tells against plausible rival explanations of that observed outcome. This obviously requires that the two groups in any clinical trial must be balanced with respect to 'known (possible) confounders' – that is factors, such as age, previous medical history, associated pathology and so on that background knowledge makes plausible might well play a role in recovery. If the control group received no treatment at all (making it a ‘natural history’ group) then – depending, I would suggest, on the nature and size of the observed difference in average outcome – the alternative hypothesis may be plausible that the difference is a placebo effect; that is why clinical trials invariably involve a control group that is given either placebo or conventional treatment. Clearly even in such a controlled trial, if a positive outcome is observed but the patients in the control group had, on average, poorer levels of general health, then that difference in level is a plausible rival theory to the theory that the improvement was produced by the (characteristic features) of the treatment. Whether this balance is brought about by randomizing and then checking *post hoc* for ‘baseline imbalances’ (and re-randomizing if such imbalances are noted) *or* by deliberate *ante hoc* matching of the two groups seems epistemically unimportant. (Although it is clear that the former method may well be much easier to apply in practice.)

Suppose the experimental and control groups have identical distributions of all factors that background knowledge makes plausible might play a role in outcome (aside of course from the treatment on trial). The central question is then whether anything further is achieved if the division into the two groups has been produced via some random process. Suppose we are dealing with an experimental double blind study where the two groups are indeed created from some initial population, both sets are treated contemporaneously and outcomes eventually recorded. It is then difficult indeed to see any such advantage for randomized over non-randomized divisions. The driving force behind the claim that there *is* such an advantage is of course that by randomizing one can somehow or other control not just for known but for ‘unknown’ factors (better *unsuspected* factors – Donald Rumsfeld’s ‘unknown unknowns’). But to seek to control for all possible factors known or unknown is surely to chase a will o’ the wisp. As argued above, no one can seriously believe that making a random division *guarantees* that all possible confounders are dealt with (as the instruction to check, after dividing randomly, for ‘baseline imbalances’ in fact concedes). And the tempting idea that making the division randomly makes it at least *more probable* that all ‘unknown’ factors are balanced also fails to withstand analysis. (Or at least it fails to withstand analysis in any sense of improbability that could be of any practical importance. Of course one can, with respect to *single* such ‘unknowns’ X at least, simply *define* it to be improbable that there is a significant imbalance in X between the two groups if there would be no such imbalance in cumulative averages produced by the indefinite long run of repetitions of the random division. But again it is difficult to see why the improbability as thus defined should be of any practical consolation in the single case.) The best epistemic state you can be in is if there is no plausible evidential reason to think that the observed outcome of a trial can be attributed to any other cause (or set of causes) aside from the treatment under trial. You are in such an epistemic state if the two groups (however created) are balanced with respect to all factors that background knowledge tells you may plausibly play a role. There is simply no effectively attainable superior epistemic state in which to be.

Now consider the case of 'observational studies' or 'historically controlled trials'. Here the controls are supplied by patients treated under some earlier regime and all the patients actually treated in the trial are given the treatment under test. Again there seems to be no reason – at least in principle – why the controls cannot be selected so as to mimic the trial patients in respect of all the factors that background knowledge indicates may be relevant to outcome. However, there are bound to be some differences. For example, there might have been improvements in the general levels of ancillary care since the time that the historical controls are selected from. More significantly there is bound, as discussed earlier, to be an element of 'selection bias' since the clinicians involved in the trial know to which group the patients they are treating belong: they are all in the experimental group. Being especially interested in those patients, because they are perhaps keen to show that the new treatment is effective, the clinicians may well give those patients unrepresentatively high levels of attention and care. Moreover, because the patients in the trial standardly know they are in the experimental group (while the control patients were just being given the previous conventional treatment without being thought of as in a trial), and because both they and their doctors may be excited about the prospects of the new treatment they are being given, there is some room (more or less depending on the type of condition at issue) for what might be called an 'enhanced placebo effect' – the control (historical) patients were being given a treatment supposed to be active so will, plausibly, have benefited from *some* placebo effect, but it could be argued that the placebo effect induced by a new treatment, as opposed to a conventional perhaps long-established one, *may* be greater.

These concerns cannot of course be dismissed, but once again we surely need to appeal to scientific commonsense: it is implausible that, assuming carefully selected controls, any such factors could produce any *large* difference between those newly treated and the historical controls. Hence if some large effect is observed in such a trial then we surely have strong evidence that the new treatment is indeed effective. (I stress this obvious point only because it has often been ignored and clinicians have insisted that 'proper' RCTs be performed even when there seems to have been already overwhelming evidence from historically controlled trials for the efficacy of some new treatment.[31]) Too little attention, as noted earlier, has been paid to (of course, strictly, *apparent*, effect size) as opposed to merely whether or not the result is statistically 'significant'. Again to reiterate, those leading proponents of the virtues of randomization, Richard Doll and Richard Peto, acknowledge this point when writing that selection bias

> cannot plausibly give rise to a *tenfold* artefactual difference in disease outcome . . . [but it may and often does] easily give rise to *twofold* artefactual differences. Such twofold biases are, however, of critical importance, since most of the really important therapeutic advances over the past decade or so have involved recognition that some particular treatment for some common condition yields a *moderate but important* improvement in the proportion of favourable outcomes. (44)

It seems then that, arguing from the 'first principles' of intuitive scientific commonsense rather than on the basis of currently received dogma, there may be – depending on proper judgement of circumstances – extremely strong evidence for the effectiveness of a treatment from non-randomized trials. (As noted, the vast majority of treatments that no one seriously supposes are anything other than highly effective were validated by what in effect was a historically controlled trial: they were introduced and clearly worked better than the earlier treatments.) RCTs do not, because they cannot, guard against all possible confounders, though they are useful – if only in cases where the treatment involved is likely to produce only at best a small improvement over currently available ones – because they control for the 'known' confounder of selection bias (interpreted in the sense just indicated). Then, however, as I argue in 'Evidence and Ethics in Medicine', if the effect of some treatment is likely to be so small as to need such refined controls, serious questions may arise as to whether that treatment is worth having.

Acknowledgments

I am grateful to Dr Jennifer Worrall whose practical knowledge of clinical trials and their impact has been of great value to me; to the *Compass* Editor, Craig Callender, for asking me to take this on and for very welcome encouragement; and to Jeremy Howick for research assistance. I particularly want to acknowledge the constructively critical comments of an anonymous referee.

Short Biography

John Worrall's current research covers two main areas: basic issues in the philosophy of science concerned with theory-change, the rationality of science and scientific realism, and more specific issues concerning the methodology of medicine. He has published a number of major articles in the first of these areas and is currently completing a book for Oxford University Press entitled *Reason in 'Revolution': A Study of Theory-Change in Science*. The general thesis of this book is that, following Kuhn and others, the alleged 'revolutionary' discontinuities in science have been greatly exaggerated and hence so have the threats that theory-change poses to the 'old-fashioned' view that science is a rational enterprise – and one that can reasonably to taken to be leading us toward the truth about the universe. He has also recently published a number of articles on evidence in medicine – the most recent being 'Why There's No Cause to Randomize' *British Journal for the Philosophy of Science* (September 2007). He is Professor of Philosophy of Science in the Department of Philosophy, Logic and Scientific Method at the London School of Economics, where he has made his whole career – having written his Ph.D. there under the supervision of Imre Lakatos (whose *Proofs and Refutations* and two volume *Philosophical Papers* he later edited).

Notes

* Correspondence address: Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London, N8 8RB, UK. Email: J.Worrall@lse.ac.uk.

[1] There naturally are exceptions, some of which will be noted as we go along – thus I do try to produce at least a partial guide to what literature there is. One soon-to-be-published example of incipient interest in the field by philosophers of science is Kincaid and McKitrick.

[2] For a counter to the popular line developed most forcefully by Larry Laudan – that the rules of scientific method themselves are subject to change over time – see my articles, 'Values of a Fixed Methodology'; 'Fix it and be Damned'.

[3] See, e.g. Sackett et al., *Evidence-Based Medicine.*

[4] They also believed that as a matter of sociological fact there were barriers to medics obtaining access to the best evidence. I shall ignore this (undoubtedly important) institutional and educational issue here and concentrate on EBM-ers' philosophical-epistemological views about what counts as best evidence (to which medical access should be especially facilitated). (Some other general issues of the above mixed social and epistemological kind are also raised in Goodman.)

[5] It is also, sometimes implicitly, required that, in order to count as a 'real' RCT with fully scientifically telling result, the trial be performed (at least) 'double blind' – that is, in such a way that neither the participant nor the investigating clinician knows to which arm of the trial any particular participant has been assigned.

[6] I am not of course asserting that it is easy to discern what these general principles are – as centuries of work on the problem in philosophy of science attest. I do, though, believe that there is pretty well universal agreement if only you go general and abstract enough. For example, the idea that a theory does not obtain very impressive empirical support simply by being consistent with some datum, or even by entailing it, if there are plausible rival theories that also entail the data is common to a whole range of proposals that differ in detail but agree on this very general point. The idea is, for example, one of the bases for Mill's methods, of course, and also for Popper's more intuitive claims about genuine tests. It lies at the basis of the whole idea of 'controlling' clinical trials.

[7] See, for example Haines.

[8] There has, unsurprisingly, been an enormous amount of discussion in the *medical* literature about the virtues of evidence-based medicine and the brief article by Straus and McAlister, which lists, and responds on behalf of EBM, to a list of criticisms, is a good starting point and source of references. As indicated however, almost none of this literature involves any systematic attempt to use insights from philosophy of science – and, where it does, as for example in Harari or Ashcroft, it either sheds more postmodernist philosophical darkness than light (Harari) or gives undue weight to standard 'S knows that p' epistemology (Ashcroft), which in my view is entirely irrelevant for any sort of science.

[9] See for example Howson and Urbach 259–79.

[10] See Worrall, 'Evidence and Ethics in Medicine'.

[11] So many claims, so many philosophy of science issues: the (definitely implicit) identification of 'experimental' with 'randomized' is also a questionable assumption as I shall indicate in more detail below.

[12] This is of course because physicists generally believe that, at least ideally, they can fully control an experiment, by blocking all relevant perturbing factors – a feature itself connected to the power of theory in physics. Clearly lacking such powerful theories, there is no doubt that clinical trialists are in a more difficult position – constantly plagued by the possibility that some 'unknown confounder' may be uncontrolled-for. So obviously I am not claiming that because they are seldom, if ever, used in physics, RCTs should not be used in medicine, but am simply pointing to the oddity that a methodology that is routinely heralded in medical circles as *the* expression of the scientific method, plays no role in our best science.

[13] Nor do I go along with the rather grudging nature of the implicit concession in Sackett et al. (*Evidence-Based Medicine*) that 'huge' treatment effects do not need to be validated in RCTs: they say that 'this option is very rare', but while it may be rare amongst currently-investigated experimental treatments, it is certainly not rare amongst *all* treatments.

[14] See, for example, the Centre for Evidence-Based Medicine hierarchy <http://www.cebm.net/index.aspx?o=1025>, accessed 1 July 2007.

[15] See for example Doll and Peto.

[16] See for example Egger et al.

[17] I raised this question in '*What* Evidence'; subsections 4(b), 4(c) and 4(d) of the current article represent major extensions and reworkings of the two arguments that I concentrated on there; the argument in subsection 4(a) was merely mentioned but not considered in '*What* Evidence'; in section 4(e) I analyze an argument based on probabilistic causality that I did not raise in '*What* Evidence' but have considered in detail in 'Why There's No Cause'.

[18] There are in fact more issues about 'double blinding' than might meet the eye. The active drug will invariably have noticeable side-effects while the traditional bread- or sugar-pill will have none (or at least none worth the name in most cases: you wouldn't want to be too liberal with your sugar pills if diabetics were involved in your trials, nor with bread-pills if it involved some patients with gluten-intolerance). But characteristic side-effects make it easy for the 'masking' to be 'broken' certainly for the clinicians involved and also – perhaps to a lesser extent – for the subjects themselves. In recognition of this problem, there are now attempts to perform trials using 'active placebos' – substances that again are 'known' to have no characteristic effect on the condition but which do mimic the side-effects of the (allegedly) active treatment under trial (see, e.g. Moncrieff et al.). There are also some claims that there is evidence that, for some patients and some conditions at least, a placebo treatment may continue to exert a beneficial effect even when the patient is authoritatively assured that s/he is receiving a placebo.

[19] A 'known' confounder, as noted earlier, is really a factor that background knowledge tells you *may* plausibly play a causal role. A sensible trial on the effectiveness of a drug would clearly control for previous health history, even if it was objectively the case that the effectiveness of the drug is – surprisingly – independent of previous health history.

[20] These 'p values' mean that there was only a 1% chance of observing such a large difference in length of stay in hospital (or a still larger one) if the 'null hypothesis' (of exactly the same probability in the two groups of staying in hospital for any given period) were correct; and only a 4% chance of observing such a large difference in duration of fever (or a still larger one) if the corresponding null hypothesis were correct.

[21] See for example Seidenfeld and Kadane.

[22] And of course even if they repeat the trial, the two together still just constitute one random trial (with a study population that is the union of the two individual populations).

[23] See for example Lindley.

[24] For example in Bowling's *Research Methods*, selection bias is defined (392) very generally (and vaguely) as 'bias in the sample obtained' (presumably *however* that bias happened to originate).

[25] Doll and Peto claim that selection bias, which they seem to mean in my sense, 'cannot plausibly give rise to a *tenfold* artefactual difference in disease outcome [but it may and often does] easily give rise to *twofold* artefactual differences. Such twofold biases are, however, of critical importance, since most of the really important therapeutic advances over the past decade or so have involved recognition that some particular treatment for some common condition yields a *moderate but important* improvement in the proportion of favourable outcomes'.

[26] See Urbach 1985.

[27] See in particular Chalmers, Matta, Smith and Kunzler; Chalmers, Celano, Sacks and Smith.

[28] Kunz and Oxman take themselves to be looking at the variety of 'distortions' that can arise from not randomizing (and concealing). They explicitly concede, however, that 'we have assumed that evidence from randomized trials is the reference standard to which estimates of non-randomized trials are compared'. Their subsequent admission that 'as with other gold standards, randomized trials are not without flaws and this assumption is not intended to imply that the true effect is known, or that estimates derived from randomized trials are always closer to the truth than estimates from non-randomized trials' leaves their results hanging in thin air. Indeed their own results showing the variability of the results of randomized and non-randomized on the same intervention seems intuitively to tell strongly against their basic assumption. (They go on to make the interesting suggestion, echoing Black, that 'it is possible that randomized controlled trials can sometimes underestimate the effectiveness of an intervention in routine practice by forcing healthcare professionals and patients to acknowledge their uncertainty and thereby reduce the strength of the placebo effect'.)

[29] See, for example Sox.

[30] See especially Cartwright; Papineau; Pearl.

[31] See, for example, the insistence on performing an RCT on ECMO as a treatment for persistent pulmonary hypertension of the new born discussed in my article, 'Why There's No Cause'.

Works Cited

Ashcroft, R. E. 'Current Epistemological Problems in Evidence Based Medicine'. *Journal of Medical Ethics* 30 (2004): 131–5.

Barton, S. 'Which Clinical Studies Provide the Best Evidence? The Best RCT still Trumps the Best Observational Study'. *British Medical Journal* 321.7256 (2000): 255–6.

Benson, K. and A. J. Hartz. 'A Comparison of Observational Studies and Randomized, Controlled Trials'. *New England Journal of Medicine* 342.25 (2000): 1878–86.

Black, N. 'Why We Need Observational Studies to Evaluate the Effectiveness of Health Care'. *British Medical Journal* 312 (1996): 1215–18.

Bowling, A. *Research Methods in Health, Investigating Health and Health Services*. Buckingham, PA: Open UP, 1997.

Cartwright, N. *Nature's Capacities and their Measurement*. Oxford: Oxford UP, 1989.

Centre for Evidence Based Medicine. 'Levels of Evidence'. *Centre for Evidence Based Medicine, EBM Tools*. 11 July 2007. <http://www.cebm.net/index.aspx?o=1025>.

Chalmers, T. C., R. J. Matta, H. Smith, Jr. and A. M. Kunzler. 'Evidence Favouring the Use of Anticoagulants in the Hospital Phase of Acute Myocardial Infarction'. *New England Journal of Medicine* 297 (1977): 1091–7.

——, P. Celano, H. S. Sacks and H. Smith, Jr. 'Bias in Treatment Assignment in Controlled Clinical Trials'. *New England Journal of Medicine* 309 (1983): 1358–61.

Clarke, M. 'Systematic Reviews and the Cochrane Collaboration'. 2004. 1 July 2007. <http://www.cochrane.org>.

Cochrane, A. L. *Effectiveness and Efficiency. Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust, 1972. Reprinted in 1989 in association with the BMJ.

Concato, J., M. P. H. Shah and R. I. Horwitz. 'Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs'. *New England Journal of Medicine* 342.25 (2000): 1887–92.

Dawid, P. 'Causal Inference without Counterfactuals'. *Journal of the American Statistical Association* 95 (2000): 407–48.

Doll, R. and R. Peto. 'Randomized Controlled Trials and Retrospective Controls'. *British Medical Journal* 280 (1980): 44.

'Editorial'. *Evidence-Based Medicine* 1.1 (1995): 2.

Egger, M. et al., eds. *Systematic Reviews in Health Care, Meta-analysis in context*. London: BMJ Books, 2001.

Ellis, J. et al. 'Inpatient General Medicine is Evidence Based. A-Team, Nuffield Department of Clinical Medicine'. *The Lancet* 346.8972 (1995): 407–10.

Fisher, R. A. *The Design of Experiments*. 4th ed. Edinburgh: Oliver and Boyd, 1947 [1926].

——. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd, 1956.

——. 'Statistical Tests'. *Nature* (1935): 136, 474.

Goodman, K. W. *Ethics and Evidence-Based Medicine – Fallibility and Responsibility in Clinical Science*. Cambridge, MA: UP, 2003.

Haines, B. 'What Kind of Evidence is it that Evidence-Based Medicine Advocates Want Health Care Providers and Consumers to Pay Attention to?'. *BMC Health Services Research* 2 (2002): 3.

Harari, E. 'Whose Evidence? Lessons from the Philosophy of Science and the Epistemology of Medicine'. *Australia and New Zealand Journal of Psychiatry* 35.6 (2001): 724–30.

Howson, C. *Hume's Problem, Induction and the Justification of Belief.* Oxford: Clarendon Press, 2000.

—— and P. Urbach. *Scientific Reasoning, a Bayesian Approach*. 2nd ed. Chicago: Open Court, 1993.

Kincaid, H. and J. McKitrick, eds. *Establishing Medical Reality*. Dordrecht: Springer, 2007.

Kunz, R. and A. D. Oxman. 'The Unpredictability Paradox, Review of Empirical Comparisons of Randomized and Non-Randomized Clinical Trials'. *British Medical Journal* 317.7167 (1998): 1185–90.

Laudan, L. *Science and Values*. Berkeley, CA: U of California P, 1984.

Leibovici, L. et al. 'Author's Reply'. *British Medical Journal Rapid Responses*. 1037 (2002). 20 August 2007 <http://www.bmj.com/cgi/content/full/324/7344/1037#resp8>.

——. 'Effects of Remote, Retroactive, Intercessory Prayer on Outcomes in Patients with Bloodstream Infection, Randomized Controlled Trial'. *British Medical Journal* 323.7327 (2001): 1450–1.

LeLorier, J. et al. 'Discrepancies between Meta-Analyses and Subsequent Large Randomized, Controlled Trials'. *New England Journal of Medicine* 337.8 (1997): 536–42.

Lindley, D. V. 'The Role of Randomization in Inference'. *PSA 1982* 2 (1982): 431–46.

Moncrieff, J. et al. 'Active Placebos versus Antidepressants for Depression'. *The Cochrane Database of Systematic Reviews* 1 (2004).

Papineau, D. 'The Virtues of Randomization'. *British Journal for the Philosophy of Science* 45 (1994): 437–50.

Pearl, J. *Causality, Models, Reasoning, and Inference*. New York: Cambridge UP, 2000.

Pocock, S. J. and D. R. Elbourne. 'Randomized Trials or Observational Tribulations'. *New England Journal of Medicine* 25 (2000): 1907–9.

Sackett, D. L. et al. *Evidence-Based Medicine. How to Practice and Teach EBM*. 2nd ed. Edinburgh and London: Churchill Livingstone, 2000.

——. 'Evidence-Based Medicine, What It is and What It isn't'. *British Medical Journal* 312 (1996): 71–2.

——. 'Inpatient General Medicine is Evidence Based'. *The Lancet* 346 (1995): 407–10.

Scottish Intercollegiate Guidelines Network (SIGN). 'Sign 50, A Guideline Developer's Handbook'. 11 July 2007. <http://www.sign.ac.uk/guidelines/fulltext/50/index.html>.

Seidenfeld, T. and J. Kadane. 'Randomization in a Bayesian Perspective'. *Journal of Statistical Planning and Inference* 25 (1990): 329–45.

Smith, R. and I. Chalmers. 'Britain's Gift, a 'Medline' of Synthesised Evidence'. *British Medical Journal* 323 (2001): 1437–8.

Sober, E. 'Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause'. *British Journal for the Philosophy of Science* 52 (2001): 1–16.

Sox, H. 'Screening Mammography for Younger Women'. *Annals of Internal Medicine* 137 (2002): 361–2.

Straus, S. E. and F. A. McAlister. 'Evidence-Based Medicine, a Commentary on Common Criticisms'. *Canadian Medical Association Journal* 163.7 (2000): 837–41.

Urbach, P. 'Randomization and the Design of Experiments'. *Philosophy of Science* 52 (1985): 256–73.

Worrall, J. 'The Value of a Fixed Methodology'. *British Journal for the Philosophy of Science* 39 (1988): 263–75.

——. 'Fix It and be Damned, a Reply to Laudan'. *British Journal for the Philosophy of Science* 40 (1989): 376–88.

——. '*What* Evidence in Evidence Based Medicine?'. *Philosophy of Science*, 69 Supplement (2002): S316–30.

——. 'Why There's No Cause to Randomize'. *British Journal for the Philosophy of Science* (September 2007): 451–88.

——. 'Evidence and Ethics in Medicine'. Forthcoming (2008).