LSE CPNSS

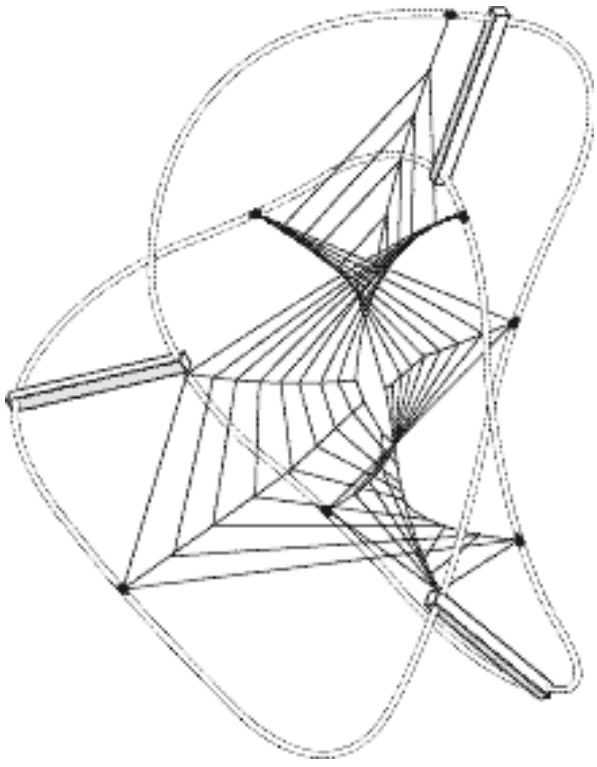
Centre for Philosophy of Natural and Social Science

Causality: Metaphysics and Methods

Technical Report 01/03

What *Evidence in Evidence-Based Medicine?*

John Worrall



Editor: Julian Reiss

What Evidence in Evidence-Based Medicine?

John Worrall*

Department of Philosophy, Logic and Scientific Method
London School of Economics
Houghton Street
London WC2A 2AE, England

September, 2002

*I am indebted to Brian Haynes and Ken Schaffner for comments and discussion. Haynes is one of the leaders of the EBM program worldwide. I am especially indebted to Dr Jennifer Worrall, for extensive discussions of the issues raised in this paper. Unlike myself, Dr Worrall has first hand experience both of clinical trials and of the problems of trying to practice medicine in an evidence-based way—and hence her contributions to this paper have been invaluable. Finally, I am also greatly indebted to my former colleague, Dr Peter Urbach who first brought me to question the orthdoxy on the randomization issue (see for example his 1993), whose arguments against the necessity for randomization form the starting point for my treatment, and who supplied me with a number of references.

Abstract

Evidence-Based Medicine is a relatively new movement that seeks to put clinical medicine on a firmer scientific footing. I take it as uncontroversial that medical practice should be based on best evidence - the interesting questions concern the details. This paper tries to move towards a coherent and unified account of best evidence in medicine, by in particular exploring the EBM position on RCTs (randomised controlled trials) in assessing causal claims from clinical medicine (especially of course concerning the efficacy of various drug therapies). I argue that even the qualified endorsement of RCTs that a more detailed and sympathetic reading of the EBM literature provides is not clearly based on solid epistemological grounds. I do this by examining four arguments that have been given that claim to show the special 'validity' of data obtained from RCTs. Finally, I discuss a case (involving a new treatment ECMO for neonates suffering from a particular condition) that shows how closely intertwined are ethical judgments and epistemological judgments about the power of certain types of trial.

1 Introduction

The usual reaction from outside observers on being told that there is a (relatively) new movement called "Evidence-Based Medicine" is "What on earth was medicine based on before?" The underlying idea of EBM is characterised by its leading proponents as follows:

> Evidence-based medicine ... is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. (Sackett *et al.* 1996)

Telling clinicians that they ought to operate in accordance with EBM sounds, then, about as controversial as telling people that they ought to operate in accordance with virtue. But the more obviously true the general idea of EBM is *normatively* speaking, the more important that idea is if, *descriptively* speaking, it has not, or has not uniformly, been put into clinical practice. And there can be no doubt that the general thrust of EBM has played a significant role, at least in the UK and no doubt elsewhere, in the recent increased awareness of clinicians of the need to keep up-to- date, to appraise their own performances critically in comparison to that of others, and to think about the principles of scientific method, of evidence, of rational decision- making and generally of clinical effectiveness.

The general idea of EBM is practically vitally important but evaluatively uncontroversial. However, just as everyone agrees that people should act in accordance with virtue, but disagrees about what virtue precisely is, so in the case of EBM disagreements soon emerge once we get to the details. We all presumably accept that clinicians ought to base practice on best evidence, but what counts as best evidence? How persuasive are different kinds of evidence (or rather how persuasive ought they to be)? What happens when different kinds of evidence point in opposite directions? What evidential role, if any, is played by 'clinical experience' or 'clinical expertise'? EBM needs a fully coherent, articulated and detailed account of the correct relationship between the evidence and various therapeutic and causal claims that would answer questions such as these from general first principles. I do not believe that it has such an account at present; and this seems to me an area where philosophers of science can, for once, be of real practical value. After all, the topic of the relationship between theory and evidence in general, and that of

the relationship between causal claims and evidence in particular, have, of course, long been recognised as central issues in philosophy of science. There could be no more practically important area in which philosophers could exercise their expertise than in clinical medicine.

There are, I believe, two main areas in which EBM has yet to produce a fully defensible account of the view of evidence that it recommends. The first concerns the role and evidential power of randomization; the second the role and evidential power of clinical judgment and expertise. In the present paper I concentrate exclusively on the first of these.

2 EBM and RCTs

It is widely believed in the medical profession that the only really scientifically "valid" evidence to be obtained from clinical trials is that obtained from trials employing randomized controls. This view derives from the statisticians. Tukey (1977) for example asserted that : "the *only* source of reliable evidence about the usefulness of almost any sort of therapy ... is that obtained from well-planned and carefully conducted randomized ... clinical trials." While Sheila Gore writes "Randomized trials remain *the* reliable method for making specific comparisons between treatments." (Gore, 1982). And quotations like this could be multiplied more or less indefinitely.

While it is often supposed that EBM endorses this view, in fact closer attention to the EBM literature reveals a much more qualified account.[1] For example, the 1996 attempt to clarify the position ("EBM what it is and what it isn't") is quite explicit that:

> **EBM is not restricted to randomised trials and meta-analyses.** It involves tracking down the best external evidence

[1]In fact the advocates of RCTs in general, whether explicit EBM-ers or not, tend to hide a much more guarded view behind slogans like those just quoted. The 'fine print view' tends to be that, at least under some conditions, some useful and "valid" information can sometimes be gleaned from studies that are not randomized; but that randomized trials are undoubtedly epistemically *superior*. So, Stuart Pocock for example writes: "it is now generally accepted that the *randomized controlled trial* is the most reliable method of conducting clinical research." (1983, p. 5) Or Grage and Zelen (1972) assert that the randomized trial "is not only an elegant and pure method to generate reliable statistical information, but most clinicians regard it as the most trustworthy and unsurpassed method to generate the unbiased data so essential in making therapeutic decisions" (p. 24).

> with which to answer our clinical questions. To find out about the accuracy of a diagnostic test, we need to find proper cross sectional studies of patients clinically suspected of harbouring the relevant disorder, not a randomised trial. For a question about prognosis, we need proper follow up studies of patients assembled at a uniform, early point in the clincial course of their disease. And sometimes the evidence we need will come from the basic sciences such as genetics or immunology. It is when asking questions about therapy that we should try to avoid the non-experimental approaches, since these routinely lead to false positive conclusions about efficacy. Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the "gold standard" for judging whether a treatment does more good than harm. However some questions about therapy do not require randomised trials (successful interventions for otherwise fatal conditions) or cannot wait for the trials to be conducted. And if no randomised 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." (Sackett *et al.*, 1996, p. 72)

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

> **Criteria for studies of causation:** a clearly identified comparison group for those at risk for, or having, the outcome of interest (whether from randomised, quasi-randomised, or nonrandomised controlled trials; cohort-analytic studies with case-by-case matching or statistical adjustment to create comparable groups; or case-control studies); masking of observers of outcomes to exposures (assumed to be met if the outcome is objective [e.g. all-cause mortality or an objective test]); observers of exposures masked to outcomes for case-control studies OR masking of subjects to exposure for all other study designs. (*Evidence-Based Medicine*, vol **1**, number 1, p. 2)

Finally in a 1995 Lancet article "Clinical Practice is evidence-based", randomised trials are explicitly deemed inessential for "Group 2 interventions". These are defined as follows

> **Intervention with convincing non-experimental evidence**. Interventions whose face validity is so great that randomised 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)

In sum,

(1) RCTs are not required except for trials of therapy.

(2) Even in the case of therapy, RCTs are sometimes unnecessary—for example, in "successful interventions for otherwise fatal conditions" (notice, by the way, that this seems clearly to imply that this is not just a pragmatic matter, we can properly judge an intervention "successful" independently of an RCT).

(3) Moreover, even presumably outside of such cases, RCTs may be deemed—presumably again *properly* deemed—unnecessary in the case of interventions with "convincing non-experimental evidence" defined to be those whose "face-validity" is agreed on unanimously by "the [presumably unusually competent] team".

(4) In the case of therapy, the RCT undoubtedly represents the "gold standard", while other non-randomized trials "routinely lead to false positive conclusions about efficacy"; but, despite this, and in general (and so in particular in the case of trials of therapy), "no RCT" should not be taken to entail "no scientific evidence"—instead "we must follow the trail to the next best external evidence and work from there". (And, of course, EBM supplies a hierarchy of strength of evidence—starting always with RCTs as the highest and working down towards clinical experience.)

No one, of course, disputes the need for a more qualified account of evidence in medicine—the claim that the only real scientific evidence is that obtained from an RCT may be clear, clean and crisp, but then it is clearly

untenable. The problem is not that the position is qualified, but that the qualifications are not explained. Several justificatory questions jump out once an attempt is made to think through the various claims and concessions. These include:

(i) What exactly does the view on the special power of randomization amount to once it is agreed that, even for therapeutic claims, non-randomized controlled evidence can sometimes be (effectively) conclusive? (Note that everyone, even the staunchest advocate of the virtues of randomization,.in the end admits this—even if only in the small print.[2] After all, everyone agrees that there is no doubt that aspirin is effective for minor headache, that penicillin is effective for pneumonia, that appendectomy may be beneficial in the case of (accurately diagnosed) acute appendicitis and so on and so on—yet none of these therapies has, of course, ever been subjected to an RCT. There is, by the way, no hint of "second-best" here—the effectiveness of these therapies is regarded, surely correctly, as at least as well established as that of therapies that have been successfully subjected to RCTs.)

(ii) The effectiveness of no therapy is "self-evident" of course. (Whatever the apparent strength of the positive evidence we may always be in the grip of a massive *post hoc ergo propter hoc* fallacy.). In calling a therapy's effectiveness "self evident" what is presumably meant is that that effectiveness is properly established by evidence we already have. But then, since that is, by definition, pre-RCT evidence, this in fact again concedes that other evidence may be, at least to all intents and purposes, compelling. So, again, why such an emphasis on RCTs now?

(iii) Why, if randomization is not specially privileged in the case of studies of causation, should it have this "highly preferred, if not strictly necessary" status concerning trials of therapy?

(iv) What justifies the hierarchy of evidence involved in EBM and just how far down that hierarchy are scientific clinicians supposed to go in search of the "next best" evidence—presumably there should be some point at which we ought to admit that there is no real evidence at all, but only unjustified opinion?

Contrary—perhaps—to certain fashionable views in philosophy of science about the inevitable (and welcome) "disunity" of methods, it must surely be a good idea to at least attempt to find some unified, general "first principles"

[2]See for example Doll and Peto (1980)

perspective from which to answer these questions. And hence to supply some sort of general rationale for the complex position summarised in points 1 to 4 (or, more likely, for some modified version of that position). This is, of course, a very tall order and I make no pretence to meet it fully here. But some important first steps can be made. These stem from reexamining the main arguments for the special power of RCTs. We shall see that at least some of the tensions in the complex EBM position on evidence may result from a continuing overestimation of the epistemic power of the RCT.

3 Why Randomize?

There have, so far as I can tell, traditionally been three answers to this question—to which, as we'll see, a fourth answer of a reliabilist kind was added later.

(3a) The Fisherian argument from the logic of significance testing

Fisher argued that the logic of the classical statistical significance test requires randomization. Fisher wrote that it is only "[t]he full method of randomization by which the validity of the test of significance can be guaranteed" (1947, p.19) And Fisher's claim is repeatedly echoed by classical frequentist statisticians. Byer *et al.* 1976 for example assert, explicitly in connection with RCTs, that "randomization guarantees the statistical tests of significance that are used to compare the treatments".[3]

An argument that some observed outcome of a trial was "statistically significant" at, say, the 95between control and experimental groups was made by some random process so that any given individual in the trial had the same probability of landing in either group. Only then might the observed data imply that an outcome has happened that has only a 5between the two treatments (standard and experimental or placebo and experimental) involved in the trial.

[3]Or, in the same paper. "It is the process of randomization that generates the significance test. See also, amongst any number of other classical statisticians, Kempthorne (1979, pp.125-6): "Only when the treatments in the experiment are applied by the experimenter using the full randomization procedure is the chain of inductive inference sound; it is only under these circumstances that the experimenter can attribute whatever effects he observes to the treatment and to the treatment only."

I shall not consider this often-examined argument in any detail here (it is in any event not the one that has carried most persuasive force sociologically speaking). I just report *first* that it is not in fact clear that the argument is convincing even on its own terms;[4] and *secondly* that there are, of course, many—not all of them card-carrying Bayesians—who regard the whole of classical signficance-testing as a broken-backed enterprise and hence who would not be persuaded of the need for randomization even if it *had* been convincingly shown that the justification for a significance test presupposes randomization.

(3b) Randomization "controls for all variables, known and unknown"

The *second* traditional argument for the power of randomization is, I guess, the one that has chiefly persuaded the medical community (albeit at—statistical—second-hand) that RCTs supply the "gold standard" for therapeutic evidence.

The basic logic behind *controlling* trials is, at least superficially, clear. First the *post hoc ergo propter hoc* fallacy must be avoided—the fact that, to take a hackneyed example, a large group of people suffering from a cold all recovered within a week when given regular vitamin C would constitute no sort of evidence for the efficacy of vitamin C for colds without evidence from a "control group" who were given some other treatment (perhaps none) and whose colds proved more tenacious. But secondly, not just any control group will do—the effects of the factor whose effect is being investigated must be "shielded" from other possible confounding factors. Suppose those in the experimental group taking vitamin C recovered from their colds much better on average than those in the control group—this would still constitute no sort of evidence for the efficacy of vitamin C if, say, the general state of health of the members of the experimental group were considerably better than that of members of the control group. The control and experimental groups could be deliberately *matched* relative to some features, and, despite the qualms of some avid randomizers, surely ought to be matched with respect to factors that there is some good reason to think may play a role in recovery from, or amelioration of the symptoms of, the condition at issue. But even laying aside issues about the practicality of matching with respect to any reasonable

[4]See for example Lindley 1982 and Howson and Urbach 1993, chapter 11.

number of factors, it is of course *in principle* impossible to match for all *possible* "confounding" factors. At most we can match for all the "known" (possibly-)confounding factors. (This really means all those it is reasonable to believe, on the basis of background knowledge, might play a role.) There are, however, clearly an indefinite number of unknown factors that *might* play a causal role. Even a pair of experimental and control groups matched perfectly with respect to all "known" confounding factors or "nuisance variables" might of course be significantly skewed with respect to one or more unknown factors. Thus the possibility is inevitably open that any observed positive effect might be due, not to the treatment or (alleged) causal factor at issue, but to the greater representation of patients with unknown factor X within one or the other group. This is where, according to this very influential line of reasoning, randomization, and randomization alone, can come to the rescue. It is often supposed that by dividing the patients in a study into experimental and control groups by some random process *all* possible confounding factors—both known *and unknown*—are controlled for at once.

Ron Giere says exactly as much: randomized groups "are automatically controlled for ALL other factors, even those no one suspects." (1979, p.296). And so did Fisher (1947, p.19):

> The full procedure of randomization [is the method] by which the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated.

Here by "eliminated", Fisher meant "deliberately controlled for ahead of the randomization".

This claim is of course—taken literally—trivially unsustainable. It is perfectly possible that a properly applied random process might "by chance", "unluckily" produce a division between control and experimental groups that is significantly skewed with respect to some prognostic factor which in fact plays a role in therapeutic outcome but which was not initially controlled for. Giere, and above all Fisher, of course knew this and so presumably what they meant, despite what they say, is something weaker—only that the randomization controls for all factors, known or unknown, "in some probabilistic sense".

And in fact most of those who advocate RCTs choose their words more carefully in line with this weaker formulation. But what exactly might that weaker claim amount to? Schwartz *et al.* suggest that

> Allocating patients to treatments A and B by randomization produces two groups which are alike as possible with respect to all their characteristics, both known and unknown. (1980, p.7; emphasis supplied).

While Byer *et al.* in their highly influential 1976 paper claim that (the emphases are mine)

> randomization *tends to* balance treatment groups in covariates (prognostic factors), whether or not these variables are known. This balance means that the treatment groups being compared will in fact tend to be truly comparable.

And Sheila Gore (1981) talks of randomization as supplying a "long run insurance against possible bias".

Presumably what is being claimed here is that if the division between experimental and control group is made at random then, *with respect to any one given possible unknown prognostic factor*, it is improbable that its distribution between the two groups is very skewed compared to the distribution in the population as a whole, where the improbability grows with the degree of skewedness and with the size of the trial. Hence if the randomization were performed indefinitely often, the number of cases of groups skewed with respect to that factor would be very small. The fact, however, is that a given RCT has not been performed indefinitely often but only once. Hence it is of course possible that—"unluckily"—the distribution even of the one unknown prognostic factor we are considering is significantly skewed between the two groups. And indeed, the advice given by even the staunchest supporters of RCTs is that, should it be noticed after randomization that the two groups are unbalanced with respect to a variable that may indeed, on reflection, play a role in therapeutic outcome, then one should either re-randomize or employ some suitable adjustment technique to control for that variable *post hoc*. (This of course again gives the lie to the idea, not seriously held but nonetheless highly influential, that randomization *guarantees* comparability of experimental and control groups. It also seems to me to render even more doubtful the advice that one quite often hears from advocates of RCTs that one should, in the interests of simplicity and pragmatic efficiency, explicitly control for few—if indeed any—variables—relying on the randomization

to control for all variables. Surely any telling trial *must* be deliberately controlled for all factors that it seems plausible to believe—in the light of background knowledge—may well play a role in therapeutic outcome.[5])

But, moreover, whatever may be the case with respect to *one* possible unknown "confounder", there is, as the Bayesian Dennis Lindsey amongst others has pointed out,[6] a major difficulty once we take into account the fact that there are indefinitely many possible confounding factors. Even if there is only a small probability that an individual factor is unbalanced, given that there are indefinitely many possible confounding factors, then it would seem to follow that the probability that there is some factor on which the two groups are unbalanced (when remember randomly constructed) might for all we know be high. *Prima facie* those frequentist statisticians who argue that randomization "tends" to balance the groups in all factors commit a simple quantificational fallacy.

(3c) Selection bias

The third argument for the value of randomized controls is altogether more down to earth (and tends to be the one cited by the advocates of RCTs when their backs are against the wall[7]). If the clinicians running a trial are allowed to determine which of the arms a particular patient is assigned to then, whenever they have views about the comparative merits and comparative risks of the two treatments—as they standardly will—then, given that they have their own patients' interests at heart, there is a good deal of leeway for those clinicians to affect the outcome of the trial. They may for example—perhaps subconsciously—predominantly direct those patients

[5]Peto *et al.* 1976, hold that stratification (a form of matching) "is an unnecessary elaboration of randomization". Stuart Pocock (1983), holds that while—to the contrary—in some circumstances "stratification would seem worthwhile", this is not true for larger trials ("If the trial is very large, say several hundred patients, ... then stratification has little point") and that , even for smaller trials, the extra complexity involved may, practically speaking, introduce errors that outweigh any theoretical gain ("If the organizational resources for supervising randomization are somewhat limited then the increased complexity of stratification may carry a certain risk of errors creeping in, so that simpler methods may prove more reliable".)

[6]See especially Lindley (1982).

[7]For example Doll and Peto *op.cit.* write that the main objection to historically controlled trials and the main reason why RCTs are superior is "that the criteria for selecting patients for treatment with an exciting new agent or method may differ from the criteria used for selecting the control patients."

they think are most likely to benefit to the new treatment or, in other circumstances, they may 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 standardly mean that those patients who are frailer will be over-represented in the control group, which is of course likely to overestimate the effectiveness of the therapy under test). Moreover, since the eligibility criteria for a trial are always, of course, open to interpretation, there is room—if a clinician is aware of the arm of trial that a given patient will go into (as they will for example in unblinded studies)—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. (Remember that anyone declared ineligible for the trial will, presumably, automatically receive "standard treatment".)

The fact that the investigator chooses the arm of the trial that a particular patient joins also means—or at any rate usually means - that the trial is at best single blind. This in turn opens up the possibility that the doctor's expectations about the 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 of course where the effect expected is comparatively small. 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 judgment whenever the outcome measure is at any rate partially subjective.

No one can deny that selection bias may operate. Because it provides an alternative explanation for positive outcomes (at any rate for small positive effects[8]), 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.

This seems to me a cast-iron argument for randomization which is not

[8]Doll and Peto *op.cit.* claim 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."

subject to any methodological difficulties—indeed the argument is underwritten by the simple but immensely powerful general injunction that one should test a hypothesis against plausible alternatives before pronouncing it well supported by the evidence. The theory that any therapeutic effect—whether negative or positive—observed in the trial is caused (or "largely caused") by selection bias is always a plausible alternative theory to the theory that the effect is produced by the characteristic features of the therapeutic agent itself. 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. What does the methodological work here is really the blinding—randomization is simply one method of achieving this.

(3d) Observational studies are "known" to exaggerate treatment effects

Brian Haynes, while insisting that EBM is not committed to the RCT outside of therpeutic claims and not exclusively committed to the RCT even for testing such claims, nonetheless also records that

> randomized allocation of participants to an intervention and control group is held to be better for controlling bias in intervention [therapeutic] studies than non-random allocation. This is not merely a matter of logic, common sense or faith: non-random allocation usually results in more optimistic differences between intervention and control groups than does random allocation.[9]

My analysis so far suggests that the preference for randomized groups may *not even* be a matter of logic or of common sense, but—and laying faith aside—what of the further reliabilist-style claim: that as a matter of fact the "track-record" of RCTs is better than so-called observational studies because the latter standardly give unduly optimistic estimates of treatment effects?

This claim stems from work in the 70s and 80s[10] which looked at cases where some single treatment had been assessed using *both* randomized *and*

[9]PSA presentation, 2000 (unpublished). Recall also that the main reason, given in the 1996 clarification of the EBM position, for downgrading "observational studies" at least when it comes to therapeutic studies was this: "It is when asking questions about therapy that we should try to avoid the non-experimental approaches, *since these routinely lead to false positive conclusions about efficacy.*" (emphasis supplied)

[10]See in particular Chalmers T.C., Matta R.J., Smith H, Jr, and Kunzler, A.M. (1977) and Chalmers T.C., Celano P, Sacks H.S., Smith H Jr (1983)

non-randomized trials—in fact the latter usually involved "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 than RCTs on the same intervention.

This issue merits careful examination, but let me here just make a series of brief and bald points:

(1) As is obvious when you think about it, the claim that these studies, even if correct, show that historically controlled trials exaggerate the true effect (and this was certainly the impression that you were intended to take away) follows *only* if the premise is added that RCTs measure that true effect (or at least can be reliably assumed to come closer to it than trials based on other methods). Without that premise, the data from these studies is equally consistent with the claim that RCTs consistently *underestimate* the true effect.[11]

(2) It is of course possible that the particular historically controlled trials that were compared to the RCTs were comparatively badly done; and indeed Chalmers *et al.* argue that the control and experimental groups in the trials they investigated were "maldistributed" with respect to a number of plausible prognostic factors. (Notice that if these maldistributions can be recognised, then it is difficult to see any reason why they should not be controlled for *post hoc*, by standard adjustment techniques.)

(3) And indeed, more recent studies of newer research in which some therapeutic intervention has been assessed using both RCTs and "observational" (non- randomized) trials have come to quite different conclusions from that of Chalmers *et al.* Kunz and Oxman (1998) found that

> Failure to use random allocation and concealment of allocation were associated with relative increases in estimates of effects of 150more, relative decreases of up to 90and, in some cases, no difference.[12]

[11] And indeed there are now some claims that they do exactly that—see *e.g.* Black (1996).

[12] 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

More significantly still, Benson and Hartz 2000, 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 (p. 1872).

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" investigated:

> Possible methodologic improvements [of the studies looked at by Benson and Harz compared to those looked at by Chalmers *et al.*] include a more sophisticated choice of data sets and better statistical methods. Newer methods may have eliminated some systematic bias (*ibid.*)

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 RCTs on the same topic. (Concato, Shah and Horwitz, 2000, p. 1887)

They explicitly point out that their findings "challenge the current [EBM-based] consensus about a hierarchy of study designs in clinical research". The

---

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 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.")

"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).

(4) 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, as opposed perhaps to their more sophisticated statistical colleagues, 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 35published previously on the same topics (Lelorier *et al.*, 1997, p. 536).

Unless I've missed something, these more recent (meta-)results have completely blown the reliabilist argument for RCTs out of the water.[13]

4 Towards a unified account of the evidential weight of therapeutic trials

So, recall the overall project: I embarked on a critical examination of the arguments for randomization, not for its own sake, but with a view to finding a general explanation from first principles of the complex account of clinical evidence given by EBM, or of some modified version of it. Here is the "first principles" view that *seems* to be emerging from that critical examination.

[13]It does seem to me that on this point Brian Haynes clearly fails to think through fully the consequences of concessions he explicitly makes. While conceding (p. 9) both that RCTs "often" disagree with one another and that "it has been shown that the findings of observational studies agree more often than not with the findings of allegedly more potent RCTs" he continues to insist (p. 3) that "non-random allocation usually results in more optimistic differences between intervention and control group than does random allocation".

Of course clinical practice should be based on best evidence. Best evidence for the positive effect of a therapeutic intervention comes when, but only when, plausible alternative explanations for the difference in outcomes between experimental and control groups have been eliminated. This means controlling for plausible alternatives. There are, of course, indefinitely many possible alternative causal factors to the characteristic features of the intervention under test. But "background knowledge" indicates which of these are *plausible* alternatives. It is difficult to see how we can do better than control—whether in advance or *post hoc*—for all plausible alternatives. The idea that randomization controls all at once for known and unknown factors (or even that it "tends" to do so) is a will-o'-the-wisp. The only solid argument for randomization appears to be that standard means of implementing it have the side- effect of blinding the clinical experimenter and hence controlling for a known factor—selection bias. But if selection bias can be eliminated or controlled for in some other way then randomization is inessential.

The main outcome of this critical survey of arguments for it should not be a more negative attitude towards randomization, but a more positive attitude towards the results of carefully conducted (*i.e.* carefully controlled) non-randomized studies.[14] If something like this is correct, then we do indeed move towards a more unified overall account of clinical evidence than that summarised in points 1 to 4 considered earlier (above pp. 4-5). For one (significant) thing, the same criteria would then be applied to evidence in causation studies as is applied in intervention studies.

5 Are the Methodological Issues Important? Methodology and Ethics

One, entirely understandable, fear—expressed by Clark Glymour during the course of the discussion at the Symposium at which an earlier version of

[14]Right from the beginning EBM-ers seemed to been ready to move from consideration of some spectactularly flawed non-randomized study or studies (where alternative explanations for the observed effect leap out at you) to the conclusion that such studies are *inherently* flawed (and therefore we need an alternative and the alternative is the RCT). A notable example is provided by Cochrane, very much a father figure to the movement, 1972, chapter 4.

this paper was delivered—is that the above analysis stands in danger of "driving out the good in searching for the best". Methodological horror stories concerning non-randomised studies in the clinical sciences abound (see for example Andersen 1990)—this of course was a major motivation of course for the EBM movement. Randomization rules out certain flaws; and nothing in my analysis shows, of course, that it can do any methodological harm. What *can* do practical harm (and I believe has done practical harm), however, is the attitude on the flipside of the "RCT as gold standard" coin—that is, the automatic (even if only partial) downgrading of even well-conducted "observational" studies (such as trials involving "historical controls").

Pursuing this issue will also allow me to engage with another theme in Brian Haynes's remarks. He expressed the view that the evidential issues associated with EBM are rather less important than other aspects that philosophers might also usefully engage with. He wrote:

> In weighing the philosophical scientific and moral issues raised by EBM, I believe that the priority target should not be about EBM's postulates about ways of knowing... Rather, it is the ethical issues that are of highest concern. (PSA presentation, 2000)

And elsewhere he expresses the fear that we may becoming "overly concerned with purity of scientific premises" at the expense of ethical (and related educational) issues.

But in fact, I don't believe you can be overly concerned with scientific purity, even if your primary concern is ethical. Ethical issues are often inextricably intertwined with methodological issues—as the following case graphically illustrates.

The ECMO case[15]

A mortality rate of more than 80 per cent had been observed historically in neonates experiencing a condition called persistent pulmonary hypertension (PPHS). A new method of treatment—"extracorporeal membraneous oxygenation" (ECMO)—was introduced in the late 1970s and Bartlett and colleagues at Michigan found, over a period of several years, mortality rates of less than 20 per cent in infants treated by ECMO. (See Bartlett, Andrews

[15]It was Peter Urbach who first drew my attention to this case.

et al. 1982.) These researchers felt forced to perform an RCT ("... we were compelled to conduct a prospective randomised study") despite the fact that this experience had already given them a high degree of confidence in ECMO ("We anticipated that most ECMO patients would survive and most control patients would die ..") They felt compelled to perform a trial because their claim that ECMO was significantly efficacious in treating PPHS would, they judged, carry little weight amongst their medical colleagues unless supported by a positive outcome in such a trial. These researchers clearly believed that, in effect, the long established mortality rate of more than 80 per cent on conventional treatment provided good enough controls—that babies treated earlier at their own and other centres with conventional medical treatment provided sufficiently rigorous controls; and hence that the results of more than 80 per cent survival that they had achieved with ECMO showed that ECMO was a genuinely efficacious treatment for this dire condition. But, because historically controlled trials are generally considered to carry little or no weight compared to RCTs, these researchers went ahead and conducted the trial.

They reported its outcome in 1985 (Barlett, Roloff *et al.* 1985). Babies suffering from PPHS were allocated to ECMO treatment or to the "control group" (which received the then conventional medical therapy—CT) using a modified protocol called "randomised play the winner". This protocol involves assigning the first baby to treatment group purely at random—say by selecting a ball from an urn which contains one red (ECMO) and one white (CT) ball; if the randomly selected treatment is a success (here: if the baby survives) then an extra ball corresponding to that treatment is put in the urn, if it fails then an extra ball corresponding to the alternative treatment is added. The fact that this protocol, rather than pure randomisation, was used was clearly itself a compromise between what the researchers saw as the needs of a scientifically convincing trial and their own convictions about the benefits of ECMO.

As it turned out, the first baby in the trial was randomly assigned ECMO and survived, the second was assigned CT and died. This of course produced a biassed urn, which became increasingly biassed as the next 8 babies all happened to be assigned ECMO and all turned out to survive. The protocol, decided in advance, declared ECMO the winner at this point, though a further two babies were treated with ECMO (officially "outside the trial") and survived. So the 1985 study reported a total of 12 patients, 11 assigned to

ECMO all of whom lived and 1 assigned to CT who died. (Recall that this is against the background of a historical mortality rate for the disease of around 80 per cent.)

Ethics and methodology are fully intertwined here. How the ethics of undertaking the trial in the first place are viewed will depend, amongst perhaps other things, on what is taken to produce scientifically significant evidence of treatment efficacy. If it is assumed that the evidence from historical (and contemporary) study was already good enough to give a high degree of confidence that ECMO was better than CT, then the ethical conclusion might seem to follow that the death of the infant assigned CT in the Bartlett study was unjustified.

But if, on the other hand, it is taken that

> ... the *only* source of reliable evidence about the usefulness of almost any sort of therapy .. is that obtained from well-planned and carefully conducted randomized .. clinical trials (Tukey 1977; emphasis supplied)

then you're likely to have a different ethical view, even perhaps that

> the results [of the 1985 study] are not... convincing... Because only one patient received the standard therapy, ... (Ware and Epstein, 1985)

Many commentators in fact took this latter view and concluded that

> Further randomized clinical trials using concurrent controls and .. randomization .. will be difficult but remain necessary. (*ibid.*)

Those taking this second view held that neither the "historically controlled" results (*i.e.* the comparison of the mortality rates achieved with ECMO with the historical mortality rate achieved with conventional treatment) nor the results from this initial "randomised play the winner" trial had produced any reliable, scientifically-telling information. The Michigan trial had not produced any real evidence because—in deference to the researchers' prior convictions—it had not been "properly randomized". Indeed, they even imply that such trials and their "historically controlled" antecedents, have, by encouraging the belief that the new treatment was effective in the absence

of proper scientific validation, made it more difficult to perform a "proper" RCT: both patients and more especially doctors find it harder subjectively to take the objectively dictated line of complete agnosticism ahead of proper evidence. Some such commentators have therefore argued that historical and non-fully randomised trials should be actively discouraged. (Of course since historical trials simply happen when some new treatment is tried instead of some conventional treatment, this really amounts to the suggestion that no publicity should be given to a new treatment and no claims made about its efficacy ahead of an RCT.)

In the ECMO case, this line led to the recommendation of a further, and this time "properly randomised", trial which was duly performed. This second trial involved a fixed experimental scheme requiring $p<.05$ with conventional randomisation but with a stopping-rule that specified that the trial was to end once 4 deaths had occurred in either experimental or control group. A total of 19 patients were, so it turned out, involved in this second study: 9 of whom were assigned to ECMO (all of whom survived) and 10 to CT (of whom 6 survived, that is 4 died). Since the stopping-rule now specified an end to the trial but various centres were still geared up to take trial- patients, a further 20 babies who arrived at the trial centres suffering from PPHS were then all assigned to ECMO (again officially "outside the trial proper") and of these 20 extra patients 19 survived.

Once again, views about the ethics of this further trial and in particular about the 4 deaths in the CT group will depend on what epistemological view is taken about when it is or is not reasonable to see evidence as validating some claim. If it is held that the first trial was indeed methodologically flawed (because "improper" randomisation had resulted in only one patient being in the control group) and therefore that no real objective information could be gathered from it, then the conviction that the first trial result (let alone the historically controlled evidence) had already shown that ECMO was superior was merely a matter of subjective opinion. Hence this second trial was necessary to obtain proper scientific information. On the other hand, if the correct methodological judgment is that the evidence both from previous practice and from the initial trial was already reasonably compelling, then this second trial, and the deaths of 4 infants treated by CT in it, would seem to be clearly unethical.

The analysis provided above suggests that this evidence could perfectly well have been solid, provided other plausible alternatives can be discounted.

So far as I can tell, no one takes seriously the idea that the natural history of the disease may have changed or that the population from which the babies were drawn underwent some general change in robustness at the time ECMO was first introduced. Expectations of success either of the patients or the clinicians are hardly a plausible factor in the case of neonates. It is true that the evidence from the second ('properly randomized' study) suggests (weakly of course since the sample is small) that babies on *both* arms of the trial may have received better than average treatment (6 out of the 10 babies on the *conventional treatment* arm survived). There is indeed some reason to think that patients involved in trials in general receive more, and better informed, attention. (After all the trials are often held in the big medical centres attracting first-rate staff and having superior facilities.) But affects both arms of a trial equally. This leaves selection bias as a main contender– were, perhaps, babies directed to ECMO only if they were relatively strong compared to all those suffering from PPHS? Bartlett and colleagues in their initial, pre-trial work seem to have treated *all* babies in their care suffering from this condition with ECMO and turned an 80 per cent mortality rate into an 80 per cent survival rate. There may of course be other plausible alternatives, though it is difficult to see why they could not, in this relatively simple case at least, be assessed on the basis of hospital records. Once the idea of RCTs as epistemological miracle workers has been abandoned, as our earlier considerations suggest it should, then there *seem* to be grounds for holding that the evidence for the superiority of ECMO over conventional treatment were solid enough ahead of either trial.

The main practical conclusions of my analysis are, then

(a) No solid grounds seem to have been provided for the automatic downgrading of "observational studies"—the undoubted fact that there have been such studies that are significantly methodological flawed does not, of course, imply that all such studies are methodologically flawed (and, in any case, no solid reason has been given for thinking of RCTs as miracle methodological flaw-removers).

(b) Good science *is* an ethical issue; we may—of course unintentionally and unwittingly—make unethical decisions if our ideas of what counts as good science are awry.

References

Andersen, B. (1990) *Methodological Errors in Medicine.* Oxford:Blackwell.

Bartlett, R.H., Andrews, A.F. *et al.* (1982) "Extracorporeal Membrane Oxygenation for Newborn Respiratory Failure. 45 Cases" *Surgery*, **92**, 425-433.

Bartlett, R.H., Roloff, D.W. *et al.* (1985) "Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study" *Pediatrics*, **76**, 479-487.

Benson, K. and Hartz, A.J. (2000) "A Comparison of Observational Studies and Randomized, Controlled Trials" *New England Journal of Medicine*, **342**, 1878-1886.

Black, N (1996) "Why We Need Observational Studies to Evaluate the Effectiveness of Health Care" *British Medical Journal*, *312*, 1215-1218.

Byar, D.P. *et al.* (1976) "Randomized Clinical Trials: Perspectives on Some Recent Ideas", *New England Journal of Medicine* **295**/2, pp. 74-80

Cochrane, A.I. (1972) *Effectiveness and Efficiency. Random Reflections on the Health Service.* Oxford: The Nuffield Provincial Hospitals Trust.

Chalmers T.C., Matta R.J., Smith H, Jr, and Kunzler, A.M. (1977)"Evidence favoring the Use of Anticoagulants in the Hospital Phase of Acute Myocardial Infarction" *New England Journal of Medicine* **297**, pp 1091-1096

Chalmers, T.C., Celano P, Sacks H.S., Smith H Jr(1983) "Bias in Treatment Assignment in Controlled Clinical Trials", *New England Journal of Medicine* **309**/22, pp. 1358-1361

Concato,J., Shah. N., and Horwitz, R.I. (2000) " Randomized Controlled Trials, Observational Studies, and the Hierarchy of Research Designs", *New England Journal of Medicine* **342**, pp. 1887-1892

Doll, R. and Peto, R. (1980) "Randomized Controlled Trials and Retrospec-

tive Controls" *British Medical Journal*, **280**, p. 44

Fisher, R.A. (1947) *The Design of Experiments.* Fourth Edition. Edinburgh: Oliver and Boyd. (First edition 1926.)

Giere, R.N. (1979) *Understanding Scientific Reasoning.* New York: Holt, Rinehart and Winston

Gore, S. M. (1981) "Assessing Clinical Trials—Why Randomize?" *British Medical Journal*, **282**, pp. 1958-60.

Grage, T.B. and Zelen, M. (1982) "The Controlled Randomized Trial in the Evaluation of Cancer Treatment—the Dilemma and Alternative Designs" *UICC Tech. Rep. Ser.*, **70**, pp. 23-47

Howson, C. and Urbach, P.M. (1993) *Scientific Reasoning: The Bayesian Approach.* Second edition. Chicago and La Salle: Open Court.

Kempthorne, O. (1979) *The Design and Analysis of Experiments.* Huntington, NY: Krieger.

Kunz, R., and Oxman A.D. (1998) "The Unpredictability Paradox: Review of Empirical Comparisons of Randomised and Non-Randomised Clinical Trials" *British Medical Journal*, **317**, pp. 1185-1190

Lelorier, J. *et al.*, (1997) "Discrepancies between Meta-Analyses and Subsequent Large Randomized Controlled Trials" *Journal of the American Medical Association*, **337**, pp. 536-42

Lindley, D.V. (1982) "The Role of Randomization in Inference", *PSA 1982*, volume 2, pp. 431-446.

Peto *et al.* (1976) "Design and Analysis of Randomized Clincial Trials Requiring Prolonged Observation of Each Patient: I. Introduction and Design" *British Journal of Cancer*, **34**, pp. 585-612.

Pocock, S.J. (1983) *Clinical Trials—A Practical Approach.* Chichester and

New York: John Wiley.

Sackett D.L. *et al.* (1996) "Evidence-Based Medicine: What it is and What it isn't" *British Medical Journal*, **312**, 71-72

Schwarz, D *et al.* (1980) *Clinical Trials.* London: Academic Press

Tukey, J.W. (1977) "Some Thoughts on Clinical Trials, especially Problems of Multiplicity", *Science*, **198**, pp.679-684.

Urbach, P.M. (1993) "The Value of Randomization and Control in Clinical Trials", *Statistics in Medicine*, **12**.15/16, pp. 1421-1431.

Ware, J.H. "Investigating Therapies of Potentially Great Benefit: ECMO", *Statistical Science* , **4**/4, pp. 298-340

Ware, J.H. and Epstein, M.D. (1985) "Comments on "Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study" by R.H. Bartlett *et al.*" *Pediatrics*, **76**, 849-851.