What Evidence in Evidence-Based Medicine?

John Worrall[†][‡]
London School of Economics

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 exploring in particular the EBM position on RCTs (randomized controlled trials).

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?” Telling clinicians that they ought to operate in accordance with EBM sounds about as controversial as telling people that they ought to operate in accordance with virtue.

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. The idea that clinicians ought to base practice on best evidence surely ought to win widespread acceptance, but what exactly counts as best evidence? How

[†]Send requests for reprints to John Worrall, Department of Philosophy, Logic and Scientific Method, London School of Economics, London WC2A 2AE, U.K.; e-mail: J.Worrall@lse.ac.uk.

[‡]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. I am also greatly indebted to my former colleague, Dr. Peter Urbach, who first brought me to question the orthodoxy on the value of randomization (see Urbach 1985, 1993 and Howson and Urbach 1993—as noted later, Urbach's work forms the basis for some parts of my reexamination of the issue). I am also indebted to Brian Haynes and Ken Schaffner.

Philosophy of Science, 69 (September 2002) pp. S316–S330. 0031-8248/2002/69supp-0027$10.00

persuasive are different kinds of evidence (or rather, how persuasive ought they to be)? What evidential role, if any, is played by 'clinical experience' or 'clinical expertise'? EBM needs, but I shall argue does not yet possess, 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. 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 has, of course, long been a central one in the philosophy of science.

There are 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 concerns 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 truly scientifically "valid" evidence to be obtained from clinical trials is that obtained from trials employing randomized controls. This view derives from the frequentist statisticians. Tukey (1977, 684) for example asserts 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 (1981, 1558) writes: "Randomized trials remain *the* reliable method for making specific comparisons between treatments."

While it is often supposed that EBM endorses this view, 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 with which to answer our clinical questions. To find out about the accuracy of a diagnostic

[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 *randomised controlled trial* is the most reliable method of conducting clinical research" (1983, 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 unbiassed data so essential in making therapeutic decisions" (24).

> 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, 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 toward 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,* **1,** 1, 2)

Finally, in a 1995 article titled "Clinical Practice is Evidence-Based," randomized 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. (Sackett et al. 1995, 408–409)

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 toward clinical experience.)

No one, of course, disputes that EBM needed 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 just quoted is qualified, but that the qualifications are not explained. Several justificatory questions emerge once an attempt is made to think through the various claims and concessions. These include:

(1) 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 headaches, that penicillin is effective for pneumonia, that appendectomy may be beneficial in the case of acute appendicitis and so on and so on—

[2] See, for example, Doll and Peto 1980.

yet none of these therapies has ever been subjected to an RCT. There is, note, 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.)

(2) The effectiveness of no therapy is "self-evident." 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?

(3) 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?

(4) 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" 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 traditionally been three answers to this question—to which, as we will see, a fourth answer of a reliabilist kind was added later. (My account of the first two of the traditional answers follows the earlier treatments of Peter Urbach—from which I have also taken a number of quotations from other authors.[3])

3a. The Fisherian Argument from the Logic of Significance Testing. Fisher argued that the logic of the classical statistical significance test re-

[3] See Urbach 1985, 1993 and Howson and Urbach 1993, chap. 11. A fuller treatment than I can give here would respond to the criticisms of Urbach's views by David Papineau (1994).

quires randomization. Fisher wrote that it is only "[t]he full method of randomisation by which the validity of the test of significance can be guaranteed" (1947, 19), and Fisher's claim is repeatedly echoed by classical frequentist statisticians.[4]

An argument that some observed outcome of a trial was "statistically significant" at, say, the 95% level, can be valid, so this line of reasoning goes, only if the division between 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 5% chance or less of happening if there is no real difference in therapeutic effect between the two treatments (standard and experimental or placebo and experimental) involved in the trial.

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;[5] and *secondly* that there are, of course, many—not all of them convinced Bayesians—who regard the whole of classical signficance-testing as having no epistemic validity, and hence who would not be persuaded of the need for randomisation 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 the one that seems chiefly to have persuaded the medical community that RCTs supply the "gold standard."

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 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

[4] Byer et al. 1976 for example assert, explicitly in connection with RCTs, that "randomisation guarantees the statistical tests of significance that are used to compare the treatments." Or, in the same paper: "It is the process of randomisation that generates the significance test."

[5] See, for example, Lindley 1982 and Howson and Urbach 1993, chap. 11.

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 was considerably more robust 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 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 is, 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 might of course be significantly skewed with respect to one or more unknown factors. Thus the possibility is always open that any observed positive effect might be due, not to the treatment 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, 296). And so did Fisher (1947, 19):

> The full procedure of randomisation [is the method] by which the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated [that is, not deliberately controlled for ahead of the randomisation].

This claim is of course, if taken literally, trivially unsustainable. It is perfectly possible that a properly applied random process might "by chance" produce a division between control and experimental groups that is significantly skewed with respect to some uncontrolled prognostic factor that in fact plays a role in therapeutic outcome. 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 randomisation produces two groups which are alike *as possible* with respect to all their characteristics, both known and unknown. (1980, 7; emphasis supplied).

While Byer et al. in their highly influential paper claim that

> randomisation *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. (1976, 75; the emphases are mine)

And Sheila Gore (1981, 1558) talks of randomisation 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—the improbability growing 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, 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.[6])

[6] Peto et al. (1976), for instance, hold that stratification (a form of matching) "is an unnecessary elaboration of randomisation." Stuart Pocock (1983) holds that while, to

But, moreover, whatever may be the case with respect to *one* possible unknown "confounder," there is, as the Bayesian Dennis Lindley among others has pointed out (1982), 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 anyone knows 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 arm 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, there is a good deal of leeway for those clinicians to affect the outcome of the trial. (The motive here might be simply, and worthily, to protect the interests of their patients; it may also, more worryingly, be that, because of funding considerations, they have a vested interest in producing a 'positive' result.) Clinicians may 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 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 generally mean that those patients who are frailer will be overrepresented 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 open to interpretation, if a clinician is aware of the arm of trial that a given patient will go into (as she will be in unblinded studies), 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. (Remember that anyone declared ineligible for the trial will automatically receive "standard treatment.")

the contrary, in some circumstances "stratification would seem worthwhile," this is not true in the case of larger trials.

[7] For example, Doll and Peto (1980) 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."

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 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.

It is undeniable that selection bias may play a role. 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 favors 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 that a given patient is assigned to will then be determined by the toss of a coin (or equivalent) and not by any clinician.

This seems to me 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 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. It is blinding (of the clinician) that does the real methodological work—randomization is simply one method of achieving this.

3d. Observational Studies Are "Known" to Exaggerate Treatment Effects. A quite different argument for the superior evidential weight of results from randomized trials is based on the claim that, whatever the methodological rights and wrongs of the randomization debate, we just know

[8] Doll and Peto (1980) 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."

on an empirical basis that random allocation leads to more trustworthy results. This is because there are studies that "show" that non-randomized trials routinely exaggerate the real effect. For example, the 1996 clarification of the EBM position (Sackett et al. 1996, 72) cites the following reason for downgrading "observational studies" (at least when it comes to assessing therapy): "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).[9]

This claim stems from work in the 1970s and 1980s[10] which looked at cases where some single treatment had been assessed using *both* randomized *and* 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 I here simply make a series of brief points:

(1) The claim that these studies, even if correct, show that historically controlled trials exaggerate the true effect 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 Chalmers et al. compared to the RCTs were comparatively poorly done; and indeed they themselves argued 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 such maldistributions can be recognized, then it is difficult to see any reason why they should not be controlled for post hoc, by standard adjustment techniques.)

(3) 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 than those arrived at by Chalmers et al. Kunz and Oxman (1998, 1188) found that

[9] Or, as Brian Haynes suggested to me, the superiority of random allocation 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."

[10] See in particular Chalmers, Matta, Smith, and Kunzler 1977 and Chalmers, Celano, Sacks, and Smith 1983.

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

> 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.[12]

Benson and Hartz (2000), comparing RCTs to "observational" trials with concurrent but non-randomly selected control groups, found, still more significantly,

> 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 suggested 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 include a more sophisticated choice of data sets and better statistical methods. Newer methods may have eliminated some systematic bias" (Benson and Hartz 2000, 1878).

In the same issue of the *New England Journal of Medicine,* Concato, Shah, and Horwitz argue that "[t]he 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" (2000, 1887).

These authors emphasize 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

[12] Kunz and Oxman take themselves to be looking at the variety of "distortions" that can arise from not randomizing (or concealing). They explicitly concede, however, that "we have assumed that evidence from randomised trials is the reference standard to which estimates of non-randomised trials are compared." Their subsequent admission that "as with other gold standards, randomised trials are not without flaws and this assumption is not intended to imply that the true effect is known, or that estimates derived from randomised trials are always closer to the truth than estimates from non-randomised 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 randomised 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.")

the 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 tend to believe when talking of RCTs as the "gold standard," that

> the outcomes of . . . large randomised, controlled trials that we studied were not predicted accurately 35% of the time by the meta-analyses published previously on the same topics (Lelorier et al. 1997, 536).

(That is, 35% of the individual RCTs came to a different judgment as to whether the treatment at issue was effective or not than that delivered by the meta-analysis of all such trials on that treatment.)

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

4. Toward 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.

Of course clinical practice should be based on best evidence. Best evidence for the positive effect of a therapeutic intervention arises 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 why should randomization continue to be thought essential.

Notice that my analysis supplies no reason to take a more negative attitude toward the results of RCTs, but it does seem clearly to indicate a more positive attitude toward the results of carefully conducted (i.e. carefully controlled) non-randomized studies.[13] If something like the line that emerges from that critical survey is correct, then we do indeed move toward a more unified overall account of clinical evidence than that summarized in points 1 to 4 considered on pages 4–5 above.

REFERENCES

Benson, K. and A. J. Hartz (2000), "A Comparison of Observational Studies and Randomised, 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): 74–80.

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

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

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

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

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

Fisher, R. A. (1947), *The Design of Experiments,* 4th ed. Edinburgh: Oliver and Boyd.

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

Gore, S. M. (1981), "Assessing Clinical Trials—Why Randomise?", *British Medical Journal* 282: 1958–1960.

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

Howson, C. and P.M.Urbach (1993), *Scientific Reasoning — the Bayesian Approach,* 2nd ed. Chicago and La Salle: Open Court.

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

Lelorier, J. et al. (1997), "Discrepancies between Meta-Analyses and Subsequent Large Randomised Controlled Trials", *Journal of the American Medical Association* 337: 536–542.

Lindley, D. V. (1982), "The Role of Randomisation in Inference", *PSA 1982,* vol. 2. East Lansing, MI: Philosophy of Science Association, 431–446.

[13] Right from the beginning EBM-ers seemed to have been ready to move from consideration of some spectacularly 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 that we need an alternative in the form of the RCT). A notable example is provided by Cochrane (1972, chap. 4), who is very much a father figure to the movement.

Papineau, D. (1994). “The Virtues of Randomisation”, *British Journal for the Philosophy of Science* 45: 451–466.

Peto, R. et al. (1976), “Design and Analysis of Randomised Clincial Trials Requiring Prolonged Observation of Each Patient: I. Introduction and Design”, *British Journal of Cancer*34: 585–612.

Pocock, S. J. (1983), *Clinical Trials—A Practical Approach.* New York: Wiley.

Sackett, D. L. et al. (1995), “Clinical Practice is Evidence-Based”, *The Lancet* 350**:** 405–410.

——— (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: 679–684.

Urbach, P.M. (1985), “Randomisation and the Design of Experiments”, *Philosophy of Science* 52: 256–273.

——— (1993), “The Value of Randomisation and Control in Clinical Trials”, *Statistics in Medicine* 12: 1421–1431.