Why Randomize? Evidence and Ethics in Clinical Trials

John Worrall

1. Introduction

In a randomized controlled experiment (henceforward an RCT) designed to test some new treatment —perhaps a new drug therapy or a new fertiliser— some experimental population (a set of patients suffering from some medical condition and recruited into the trial or a set of plots of land on which some crop is to be grown) is divided by a random process into two exhaustive and mutually-exclusive subsets: the "experimental group" and the "control group". Those in the experimental group receive the new test treatment while those in the control group do not. What they receive instead may differ from case to case: in agricultural trials, the control plots may simply be left unfertilised (all other obvious factors —such as frequency and extent of artificial watering, if any— being left the same as in the experimental regime), or, in clinical trials, the control group may be given either a "placebo" (generally a substance "known" to have no specific biochemical effect on the condition at issue) or the currently standard therapy for that condition. RCTs are almost invariably performed "double blind": neither the subjects themselves nor those treating them know whether they are receiving the experimental or the control treatment.[1]

It is widely believed that RCTs carry special scientific weight —often indeed that they are essential for any truly scientific conclusion to be drawn from trial data about the effectiveness or otherwise of proposed new therapies or treatments. This is especially true in the case of clinical trials (on which this paper will exclusively focus): the medical profession has been overwhelmingly convinced that RCTs represent the "gold standard" by providing the only truly "valid," unalloyed scientific evidence of the effectiveness of any therapy. Thus the statistician Tukey writes: "the only source of reliable evidence … is that obtained from … carefully conducted randomized trials," while Sheila Gore avers "Randomized trials remain the reliable method for making specific comparisons between treatments."[2] Clinical science may occasionally have to rest content (perhaps for practical reasons) with evidence from other types of trial, but this is always very much (at best) a case of (epistemic) second-best. For, it is widely believed, all non-randomized trials are inevitably subject to bias, while RCTs, on the contrary, are free from bias (or perhaps, and more plausibly, are as free from bias as any trial could possibly be).

In this paper I analyse the claim that RCTs provide far and away the best, perhaps the only really valid, scientific evidence; and I will closely scrutinise the arguments that have been put forward for it (section 1). In section 2 I shall consider a famous trial (or really a series of

[1] Other more complicated (and arguably more effective) randomized designs are possible —for example randomized blocks (where the two groups are matched for known prognostic factors and then which becomes the experimental and which becomes the control group is decided randomly). But the above is the simplest RCT and what most commentators have in mind in assessing the power of the methodology.

[2] Tukey, J. W., "Some Thoughts on Clinical Trials, especially Problems of Multiplicity," *Science,* v. 198, (1977), p. 1958; and Gore, S. M., "Assessing Clinical Trials-Why Randomize?," *British Medical Journal,* v. 282, (1981), p. 684.

trials) that shows that decisions have been made about which trials are ethical to perform that depend entirely on this epistemological claim that RCTs provide much stronger evidence than any other trial (and in particular than historically-controlled trials or "observational studies"). I will conclude (section 3) that because the arguments for the superior epistemic power of randomization seem on analysis to be weaker than they are widely considered to be, the trials outlined in section 2 —and no doubt others like them— raise enormous ethical concerns. And I also in that final section outline a seemingly more defensible approach to the question of the weight of evidence from clinical trials.

2. The Alleged Epistemic Superiority of RCTs

Many in the medical profession just accept it as a given that randomized controls are the proper scientific way of proceeding when it comes to judging the effectiveness of a new treatment. But it is not, of course, a given —instead it must be argued for on the basis of more fundamental principles of correct scientific method. The first question to ask from this more basic perspective is "Why "control" at all?".

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

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

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

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

Suppose we have matched the two groups with respect to every factor that anyone can think of as plausibly relevant to recovery. And suppose we find a substantially higher proportion of recoverers within the group given vitamin C. Surely now we have telling evidence that vitamin C aids recovery from colds?

Well, still "maybe not" and this for two separate reasons. First it may be (though we can't of course know it —a fact that we shall need to reflect on at length later) that the two groups now are indeed "equal" with respect to all factors relevant to recovery from colds (though it would actually be something of a miracle). In that case, we would indeed be justified in concluding that the observed difference in outcome was produced by the treatment rather than being due to some difference between the two groups. But it wouldn't follow that it was the fact that the treatment involved giving vitamin C that caused the improved outcome. It is conceivable —and there is a lot of evidence (albeit disputed by some commentators) that it is in fact the case— that a patient's expectations, fired by being treated by an authority figure, play a role in recovery from at any rate relatively minor complaints. In the vitamin C case, as we are now envisaging it, those in the control group receive no treatment at all —they are just the "otherwise equal in all respects" comparators. But the very fact that they receive no treatment —any treatment— at all is a difference and one that might, as just remarked, be relevant.[3]

This is the reason why the control groups in medical trials invariably will be either placebo controls or conventional treatment controls (sometimes there are two control groups in a trial: one of each). That is, those in the control group will not be left untreated, but will instead be treated either with the currently accepted treatment for the condition at issue or with a substance known to have no specific biochemical effect on the condition but which is intended to be indistinguishable to the patient from the (allegedly) "active" drug under test. This feature of clinical trials permits a further feature with clear methodological merits —namely that the trials can, again at least in principle, be performed "double blind". The two treatments ("active" drug and placebo (or conventional treatment)) can be sorted into packets marked simply, say, "A" and "B" by someone not involved in seeing patients and delivered to the patient in ways that are indistinguishable:

[3] Notice though that the plausibility of this suggestion depends upon the particular case. If, for example, the patients involved are neonates, as in the case examined in section 2, then the idea that expectations may play a role in reaction to treatment can surely be dismissed.

hence neither the patient herself nor the clinician involved knows whether that particular patient has received the "active" drug or not.[4]

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

As we shall see, the strongest argument for the epistemic superiority of randomized trials (strongest in the sociological sense that it is the one that has convinced most people in medicine) is precisely that randomization is alleged to solve the problem of "unknown factors": a randomized trial is controlled for all factors known and unknown.

There are various versions of RCTs. In the most straightforward, the division of those involved in the trial into experimental and control groups is effected by some random process —individual patients being assigned to treatment "A" or treatment "B" according to the toss of a fair coin (or, in more approved manner, on the basis of a table of random numbers). In what seem to me (for reasons to be discussed) much more sensible trials —some stratification or "blocking" is carried out with respect to factors that might plausibly play a role in recovery (or whatever the outcome at issue in the trial is). Again the most straightforward version of this type of design (in principle, if not in practice) would be to take the whole population to be involved in the trial and divide it consciously into two groups that are matched with respect to "known" prognostic factors and then use a random process to decide which of these groups is given treatment A and which B.

So these are the main ideas behind RCTs. Why should such trials be thought to carry more epistemological weight than any other, non-randomized trial? There are, so far as I can tell, five different arguments to be found in the literature that claim to show why this is so.

[4] This is not as straightforward as it sounds. The active drug will invariably have noticeable side-effects while the traditional bread- or sugar-pill will have none (or at least none worth the name in most cases: you wouldn't want to be too liberal with your sugar pills if diabetics were involved in your trials, nor with bread-pills if it involved some patients with gluten-intolerance). It would in such cases be easy for the blind to be 'broken' certainly for the clinicians involved and also for the subjects themselves. In recognition of this problem there are now attempts to perform trials involving 'active placebos' —substances that again are 'known' to have no characteristic effect on the condition but which mimic the side-effects of the (allegedly) active treatment under trial (See Moncrieff, J. et al., "Active placebos versus antidepressants for depression," *The Cochrane Database of Systematic Reviews* 2004, Issue 1, Art. No.: CD003012. DOI: 10.1002/14651858.CD003012.pub2.)

The first of these is Fisher's argument that randomization is necessary to underpin the logic of his famous significance test approach —universally applied (with some optional refinements) in medical trials.[5]

The second argument is the one we already noted as sociologically speaking the most persuasive —that by randomizing the trial is controlled not just for known factors but for all factors, known and unknown.

A third argument is based on the idea that standard methods of randomizing control, not for some hitherto unconsidered possible bias, but for a "known" bias that is believed to have operated to invalidate a number of trials —namely, "selection bias." If clinicians involved in trials are allowed to decide the arm of the trial that a particular patient is assigned to then there is the possibility that, perhaps subconsciously, they will make decisions that distort the result of the trial and thus produce an inaccurate view of the effectiveness of the treatment. They might, for example, have an opinion on the effectiveness of the new drug and also likely side effects, and therefore direct patients that they know to one arm or the other because of the perfectly proper desire to do their best for each individual patient, or because of the entirely questionable desire to achieve a positive result so as to further their careers or please their (often pharmaceutical company) paymasters. (I stress the "might" here.)

A fourth argument has also been given a good deal of emphasis of late within the Evidence-Based Medicine movement. This claims that, whatever the finer rights and wrongs of the epistemological issues, it is just a matter of historical fact that the "track-record" of RCTs is better than that of observational studies ("historically-controlled trials") because the latter standardly give unduly optimistic estimates of treatment effects.[6]

Fifthly and finally, an argument that randomization has special epistemic virtues has arisen from the currently burgeoning literature on "probabilistic causality". Several authors —notably Judea Pearl, David Papineau and Nancy Cartwright[7]— have argued that randomisation plays an essential role when we are seeking to draw genuinely causal conclusions about the efficacy of some treatment as opposed to merely establishing that treatment and positive outcome are associated or correlated.

Here I will concentrate on the second —and, as noted, most influential— argument, developing my analysis of it much further than in my earlier work.[8] And then I will briefly re-outline the analyses of the other four arguments that I have given in more detail elsewhere.

Mike Clarke, the Director of the Cochrane Centre in the UK, writes on their web-site: "In a randomized trial, the only difference between the two groups being compared is that of most interest: the intervention under investigation."[9]

[5] Fisher's argument can be found in his book *The Design of Experiments,* Oliver and Boyd, London, 1935.

[6] Historically-controlled trials (aka observational studies) are, as we shall see in more detail later, ones where the control group is considered to be supplied by (comparable) patients treated earlier with the previously accepted treatment. Hence in these trials, all patients actively involved are given the new treatment under test.

[7] See PEARL, J., *Causality-Models, Reasoning and Inference,* Cambridge University Press, New York and Cambridge, 2000, PAPINEAU, D., "The Virtues of Randomization," *The British Journal for the Philosophy of Science,* v. 45, n. 2, (1994), pp. 437-450, and CARTWRIGHT, N., *Nature's Capacities and their Measurement,* Oxford University Press, Oxford, 1989.

[8] I have analysed the first four of these arguments in WORRALL, J., "What Evidence in Evidence-Based Medicine?," *Philosophy of Science,* v. 69, n. S3, (2002), pp. S316-330, and the fifth in WORRALL, J. "Why There's no Cause to Randomize", *The British Journal for the Philosophy of Science,* forthcoming.

[9] See the website of the Cochrane Collaboration at *www.cochrane.org.*

That is, by randomizing, all other factors —both known and unknown— are (allegedly) equalised between the experimental and control groups; hence the only remaining difference is exactly that one group has been given the treatment under test, while the other has been given, say, a placebo; and hence any observed difference in outcome between the two groups in a randomized trial (but only in a randomized trial) can legitimately be inferred to be the effect of the treatment under test.

Mike Clarke's claim is admirably clear and sharp, but it is clearly unsustainable (as indeed he himself later implicitly allows). Clearly the claim as made is quite trivially false: the experimental group contains Mrs Brown and not Mr Smith, whereas the control group contains Mr Smith and not Mrs Brown, etc. An apparently more plausible interpretation would take it as stating that the two groups have the same means and distributions of all the [causally?] relevant factors. It is not clear to me that this claim even makes sense, though, and, even with respect to a given (finite) list of potentially relevant factors, no one can really believe that it automatically holds in the case of any particular randomized division of the subjects involved in the study. Although many commentators often seem to make the claim (and although many medical investigators blindly following the "approved' methodology may believe it) no one seriously thinking about the issues can hold that randomization is a sufficient condition for there to be no difference between the two groups that may turn out to be relevant.

Here is an amusing example that illustrates this point. A study by L. Leibovici and colleagues was published in the *British Medical Journal* in 2001 entitled "Effects of remote, retroactive, intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial."[10] The study looked at 3393 inpatients at the Rabin Medical Centre in Israel during 1990-1996 who had suffered from various bloodstream infections. In July 2000 (so, notice, between 4 and 10 years after they had suffered these infections), a random number generator was used to divide these patients into two groups; and which of the two became the treatment group was decided by a coin toss. 1691 patients were, so it turned out, randomized to the intervention group and 1702 to the control group. A careful check was made for "baseline imbalances" with regard to main risk factors for death and severity of illness. ("Baseline imbalances" are differences between the two groups in respect of known prognostic factors produced in a "purely" randomized trial —that is, one in which no deliberate attempt is made to make the two groups equal with respect to these known factors; and spotted after the random division has of course been made.) But no significant baseline imbalances were found. The names of those in the intervention group were then presented to a person "who said a short prayer for the well being and full recovery of the group as a whole.' Then, but only then, were the medical records of all the patients checked for those patients' mortality, for length of stay in hospital and for duration of the fevers they had suffered. The results were that mortality was 28.1% in the "intervention" group and 30.2% in the control group, a difference that orthodox statistical methodology declares "non-significant'; however both length of stay in hospital and duration of fever were significantly shorter in the

[10] LEIBOVICI, L. ET AL., "Effects of remote, retroactive, intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial," *British Medical Journal,* v. 323, n. 7327, (2001), pp. 1450-1451.

intervention group (p = 0.01 and p = 0.04).[11] Leibovici and colleagues drew the conclusion that "remote, retroactive intercessory prayer said for a group is associated with a shorter stay in hospital and shorter duration of fever in patients with bloodstream infection and should be considered for use in clinical practice."[12]

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

Leibovici himself fully agreed with this as he made clear in the subsequent discussion: "If the pre-trial probability [of the eventual 'result'] is infinitesimally low, the results of the trial will not really change it, and the trial should not be performed. This, to my mind, turns the article into a non-study, though the details provided (randomization done only once, statement of a prayer, analysis, etc) are correct."[13] The sentiment, entirely in line with Bayesian, as opposed to classical statistical, methodology, is that we need to take into account not only the "improbability" of a particular outcome occurring if some "null hypothesis" is correct (here the null hypothesis is that there is no difference between the two groups despite the remote intercessory prayer "intervention" and that any observed difference in outcome between the two groups is due to chance), but also the prior probability of the "negation" of the null (here that the prayer really did have a retroactive effect).

But although Leibovici may not have intended the study to be taken seriously as a basis for "treatment," it *is* to be taken seriously as a criticism of orthodox statistical methodology and in particular of the suggestion that a properly randomized study always produces real evidence of effectiveness. Leibovici insisted, note, that "the details provided (randomization done only once, statement of a prayer, analysis, etc) are correct." So the fact is that this was a properly randomized study (in fact an exceptionally and impressively large one) that happened to produce what we take ourselves to know must be the "wrong" result. Obviously what must have happened here is that although the division into experimental and control groups was done impeccably and although the double blinding was equally impeccable(!), "by chance" some unknown confounder/s were unbalanced and produced the difference in outcome. Not only is this not impossible, it ought to happen, according to orthodox statistical methodology, on average in one in every 20 or so such trials on treatments that in fact have no effect (assuming that the standard 5% "significance level" is invariably used)!

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

[12] LEIBOVICI, L. ET AL., "Effects of remote, retroactive, intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial," p. 1451.

[13] LEIBOVICI, L., "Author's Comments," *British Medical Journal,* v. 324, (2002), p. 1037.

The fact that —despite assertions like the one quoted from Mike Clarke— a particular random division may of course produce an importantly unbalanced division is indeed implicitly admitted even by the most orthodox advocates of randomization when they accept that if a particular ("pure") randomisation (involving no element of prior deliberate matching) has produced an imbalance in a "known" prognostic factor then one should not proceed to make any inferences. Though —surely rather quixotically— the orthodox then go on to assert that, in this situation, rather than deliberately match "known" factors, one should re-randomise until a random division is produced about which one has no concerns from the point of view of imbalances.

In sum, despite what is sometimes written, no one can seriously believe that having randomized is a sufficient condition for a trial result to reflect the true effect of some treatment. Is randomizing a necessary condition for this? That is, is it true that we cannot have real evidence that a treatment is genuinely effective unless it has shown itself to be so in a properly randomized trial? Again, some people in medicine sometimes talk as if this were the case, but again no one can seriously believe it. Indeed modern medicine would be in a terrible state if it were true. The overwhelming majority of all treatments regarded as unambiguously effective by modern medicine today —from aspirin for mild headache through diuretics in heart failure and on to pretty well any surgical procedure (appendicectomy, cholecystectomy, etc., etc.)— were never (and now, let us hope, never will be) validated in an RCT. Much of the impetus behind the "Evidence-Based Medicine" movement that emerged in the 1980s was the recognition that certain treatments that had been in regular use in medicine (such as grommets for glue ear, suppression of ventricular ectopic beats in heart failure and a *few* others) proved, when subject to systematic testing, in fact to be either ineffective or (worse) dangerous because adverse side-effects overwhelmed any positive effect on the target disorder. Although this is true, we must not let it blind us to the fact that the overwhelming majority of treatments in medicine that no one suggests for a moment are ineffective have never been subjected to an RCT.[14]

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

The problem is that for all its seeming plausibility and indeed for all its widespread acceptance and therefore immense practical significance, it seems difficult to make anything like full sense of this claim —especially on the basis of the orthodox approach to statistics. The latter (officially) refuses to deal in the probability of hypotheses at all,

[14] Nor should it blind us to the fact that it has also sometimes turned out to be true that treatments 'validated' in RCTs have later been withdrawn because of negative side-effects.

but only in the acceptance or rejection of hypotheses that attribute some probability to the values of some random variable. In order even to begin to make sense of the claim, we would need to be able to show that, for any particular (potentially) prognostic factor aside from whether a patient is given the treatment under test or, say, placebo, it is probable that that extra factor is identically (or near identically?) distributed in the two groups —treatment and control. Any plausibility that such a claim might appear to have depends, however, on confusing what can reasonably be asserted in the case of a single random division with what might reasonably be asserted about an indefinite number of repetitions of the random division.

What can it mean to claim that it is improbable that factor X is evenly distributed between the two groups? Assuming the classical non-Bayesian approach to probability (and there is no direct role for randomization according to the Bayesian approach),[15] it can only be a claim about an indefinite series of repetitions of the trial: that if you were to take a population and divide it at random into two lots and lots of times and record the cumulative relative frequency of positive values of X in the two groups (assume for simplicity that X is a two-valued random variable), then in the indefinite long run that frequency would be the same in the experimental and control groups and in fact the same as the actual frequency of positive values of X in the study population as a whole. But medical researchers involved in some particular trial do not make a random division indefinitely often, they do it once! In that one trial, factor X may be as substantially unbalanced between the two groups as you like, and there seems just to be no way to quantify what the "probability" of a substantial imbalance is: "single case probabilities" not being properly defined on this approach. Once you further take into account the fact that, by definition, the list of possible "unknown" factors is indefinitely long, then matters become even murkier. Even if one wanted to insist that despite the lack of any adequate formal analysis it was somehow "intuitively" clear that for any single factor X, it is "improbable" that it is significantly maldistributed between the two groups in a single randomisation, it would not of course follow even "intuitively" that it is improbable that there is no factor relative to which the single randomization is unbalanced —because of the lack of any real grasp of the list of potential other factors and of how large it is, this just seems to be, even intuitively, undefined.[16]

It is, then, difficult to see any objective weight in what is, as I mentioned earlier, sociologically speaking the most persuasive argument for the special epistemic power of randomized trials. What of the other arguments?

Let's start by accentuating the positive. There seems no doubt that the claim that randomization controls, not for all factors, but for the single "known" factor of "selection bias," carries some weight. This is accepted even by Bayesian critics of randomisation. If the experimenters are allowed consciously or unconsciously to influence the membership of the experimental and control groups, then this is undoubtedly a source of possible bias and hence an alternative (partial) explanation of any positive (or negative) result achieved. Notice however two things. First, this argument does not supply any reason to think that

[15] See for example KADANE, J. B. and SEIDENFELD, T., "Randomization in a Bayesian Perspective," *Journal of Statistical Planning and Inference,* v. 25, (1990), pp. 329-345.

[16] See for example LINDLEY, D. V., "The Role of Randomization in Inference," *PSA 1982,* volume 2, (1982), pp. 431-446.

the randomization itself has an effect —the procedure involved in randomized trials is just one way of ensuring that the clinicians are blinded to the group allocation of individual patients. If clinicians were prevented in some other way from influencing the allocation, or if it is clear for other reasons that they did not influence it, then randomization would, so far as this particular argument goes, become redundant. Secondly, as even Doll and Peto, the staunchest of RCT-advocates, allow,[17] selection bias, where operative (or possibly operative) at all, is unlikely to produce a large effect. It would therefore be a mistake to dismiss a non-randomized study that had revealed a large effect as not providing any "valid" evidence simply on the grounds that selection bias might have played a role.

Fisher's famous argument for randomization was that it is the only way in which the logic of his method of statistical significance testing can be underwritten. He argued in effect that when, but only when, you have randomized, is it legitimate to identify the "null hypothesis" (which would otherwise just be the very vague hypothesis that any positive result observed is not the result of the treatment being tested) with the "chance" hypothesis that each patient has the same probability of recovering (or whatever the outcome measure is) independently of whether they received the treatment under test (or the placebo/conventional treatment). Fisher's argument, based on his "Tea Lady" test (perhaps the most celebrated in the whole history of statistics), has appeared compelling to many and it does have a tenacious air of plausibility. However Bayesians have in my view conclusively shown that it holds no water. Since the considerations raised are rather technical, I shall not pursue this matter here but just refer to what seems to me the clearest Bayesian demolition —that by Colin Howson.[18] There is, in any case, an increasingly popular view (one that I share) that the whole classical statistical significance test methodology is itself fundamentally illogical and should be abandoned. If this view is correct then it would of course follow that, even were Fisher right that randomization is necessary to underpin the application of his methodology, this would still supply no cogent reason to randomize.

An argument that has had some impact in the recent Evidence-Based Medicine movement claims that whatever the finer rights and wrongs of the epistemological issues it is just a matter of fact that the "track-record" of RCTs is better than that of other types of study, and in particular better than the track record of so-called historically controlled studies. In this latter type of study, the control group is supplied by an allegedly comparable set of previous patients treated in the (recent) past with whatever is currently accepted treatment. Hence in such studies (also sometimes called "observational studies") all the patients actively involved in the trial are given the new treatment under test. The claim is that these studies standardly give unduly optimistic estimates of treatment effects. This argument, so I suggest in my paper "What Evidence in Evidence-Based Medicine?," is (a) circular (it depends on supposing that, where an RCT and an "observational study" have been performed on the same treatment, it is the former that reveals the true efficacy (after all randomized results provide the "gold standard"!) and this is precisely the question at issue; (b) based largely at least on comparing RCTs to particularly poorly performed observational studies that anyone would agree are obviously methodologically unsound;

[17] See Doll, R. and Peto, R., "Randomized Controlled Trials and Retrospective Controls," *British Medical Journal,* v. 280, (1980), p. 44.

[18] See Howson, C., *Hume's Problem,* Oxford University Press, Oxford, 2000, pp. 48-51.

and (c) is —to say the least— brought into severe question by more recent work that seems to show that, where a number of different trials have been performed on the same treatment, the results of those done according to the RCT protocol differ from one another much more markedly than to do carefully performed and controlled observational studies.[19]

The final argument for randomization is one that has emerged from the recent literature on "probabilistic causality". A number of authors —taking routes that are related though different in details— have claimed that it is only when you randomize that a trial can give evidence of a genuinely causal (as opposed to merely "associational") connection between the treatment and outcome. A central problem in this area of probabilistic causality is that of distinguishing between "real" (causal) and "spurious" correlations. Two variables may covary despite being causally unconnected with one another —they might, for example, be two independent effects of a "common cause". So, to take an obvious example, the probability that you will develop lung cancer is much higher if you own a reasonable number of ashtrays (say more than 3) than if you don't:

P (lung cancer / own more than 3 ashtrays) >> P (lung cancer).

Thus lung cancer and ashtray ownership are (strongly) probabilistically associated (or "correlated" as is often said in this literature —though this is not the usual statistical meaning of the term.) But we wouldn't say that owning ashtrays "increases the probability" of developing lung cancer, because there is, as we know on the basis of background knowledge, no causal connection between the two. The causal connections are instead between smoking cigarettes and developing lung cancer, and smoking cigarettes and "needing" ashtrays. In the jargon, smoking cigarettes is a common cause of both lung cancer and ashtray-ownership. The fact that this is so and hence that the "correlation" between cancer and ashtrays is "spurious" is revealed by the fact that smoking "screens off" cancer from ashtray-ownership. In other words, the conditional dependence between the latter two variables disappears when you further conditionalise on smoking:

P (lung cancer/ own more than 3 ashtrays and you smoke) = P (lung cancer / you smoke),

even though P (lung cancer / own more than 3 ashtrays) >> P (lung cancer).

The argument then in essence (and ignoring some important issues about the inference from (observed) relative frequencies to (theoretical) population probabilities) is that you are justified in taking an observed relationship between treatment and whatever your outcome measure is (recovery within some fixed period, say) when, but only when, this relationship is observed in a trial that was randomized. In effect, then, the claim is that randomization eliminates the possibility of a "common cause" of treatment and treatment outcome.

In a forthcoming paper,[20] I take the various versions of this argument —by Nancy Cartwright, David Papineau, and especially Judea Pearl— and show that they fail. I shall not repeat the details of my counterargument here. But, as the above brief outline will perhaps suggest, their claim is at root just a particular version of the "randomizing controls for all other factors" line and hence it falls to the same objection: that it trades on a confusion between what might be justified in the indefinite long run of reiterated randomizations on

[19] For detailed references, see Worrall, J., "What Evidence in Evidence-Based Medicine?," *Philosophy of Science,* v. 69, n. S3, (2002), pp. S316-330.

[20] Cf. Worrall, J., "Why There's no Cause to Randomize," *The British Journal for the Philosophy of Science,* forthcoming.

the same group and what is justified in the particular case where, by definition, the random division has been effected only once. It is of course possible in the single case that the two groups are unbalanced in respect of a factor that is in fact a common cause of treatment and treatment outcome.

No argument known to me, then, really establishes the almost universally held view that RCTs have a special epistemic status —except for the modest argument about controlling for "selection bias"(and that bias might be eliminable by other means). In the next section, I will look at some trials on a newly-introduced treatment that were motivated entirely by the view that only evidence for treatment efficacy from an RCT really counts scientifically. The lack of any substantial and cogent argument for the necessity of randomization makes the ethical acceptability of these trials extremely suspect. In the final section, I will make some suggestions about what seem to me the correct ethical and methodological judgements.

3. Why the Issue is of Great Practical and Ethical Significance-The ECMO Case[21]

A persistent mortality rate of more than 80% had been observed historically in neonates experiencing a condition called persistent pulmonary hypertension (PPHN). A new method of treatment —using a technique developed for other conditions and called "extracorporeal membranous oxygenation" (ECMO)— was introduced in the late 1970s, and Bartlett and colleagues at Michigan found, over a period of some years, mortality rates of less than 20% in these infants treated by ECMO.[22] I think it is important background information here that this new treatment could hardly be regarded as a stab in the dark. It was already known that the underlying cause of this condition was immaturity of the lungs in an otherwise ordinarily developed baby. The babies that survived were those that somehow managed to stay alive while their lungs were developing. ECMO in effect takes over the function of the lungs in a simple and relatively non-invasive way. Blood is extracted from the body before it reaches the lungs, is artificially oxygenated outside the body, reheated to regular blood temperature and reinfused back into the baby —thus bypassing the lungs altogether.

Despite the appeal of the treatment and despite this very sharp increase in survival from 20% to 80% the ECMO researchers felt forced to perform an RCT ("... we were compelled to conduct a prospective randomised study") even though their 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.[23] These researchers clearly believed that, in effect, the long established mortality rate of more than 80% on conventional treatment provided good enough controls that babies treated earlier at their own and other centres with conventional medical treatment

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

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

[23] This is another argument for RCTs that is not infrequently cited by medics and clinical scientists. It is however a very strange argument: if it were the case that randomizing was, in certain cases, neither necessary nor useful then it would seem better to try to convince the medical profession of this rather than turn their delusions into an argument for pandering to that delusion!

provided sufficiently rigorous controls; and hence that the results of around 80% survival that they had achieved with ECMO already showed that ECMO was a genuinely efficacious treatment for this dire condition. Given that there was an argument for thinking that there was no significant difference between the babies that Bartlett and colleagues had been treating using the earlier techniques and those that they had now been treating with ECMO (we will return to this point later), this counts as a (retrospective) historically controlled trial —one producing a very large positive result. But, because historically controlled trials are generally considered to carry little or no weight compared to RCTs, as we saw in the previous section, these researchers felt forced to go ahead and conduct the trial.

They reported its outcome in 1985[24]. Babies suffering from PPHN were allocated to ECMO treatment or to the control group (receiving 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 randomization, was used was no doubt itself a compromise between what the researchers saw as the needs of a scientifically (or is it sociologically?) 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 biased urn, which became increasingly biased 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%.)

Ethics and methodology are fully intertwined here. How the ethics of undertaking the trial in the first place are viewed will depend, amongst other things, on what is viewed as producing scientifically significant evidence of treatment efficacy: clearly a methodological/ epistemological issue. If it is assumed that the evidence from the "historical trial" (i.e. the comparison of the results using ECMO with the earlier results using CT) was already good enough to give a high degree of rational 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,"[25]

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

[25] Cf. Tukey, J. W., "Some Thoughts on Clinical Trials, especially Problems of Multiplicity," *Science,* v. 198, (1977), p. 1958. (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, ..."[26]

Many commentators in fact took this latter view and concluded that

> "Further randomized clinical trials using concurrent controls and ... randomisation ... will be difficult but remain necessary."[27]

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 "randomized 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 views— it had not been "properly randomised". Indeed, they even imply (note their "will be difficult" remark) that such trials and their "historically controlled" antecedents, have, by encouraging the belief that a new treatment is effective in the absence of proper scientific validation, proved pernicious by making 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 randomized trials should be actively discouraged. (Of course since historical trials in effect always 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 subjecting it to an RCT.)

In the ECMO case, this line led to the recommendation of a further, and this time "properly randomized," trial which was duly performed. This second trial involved a fixed experimental scheme requiring $p < .05$ with conventional randomization 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.[28]

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" randomization had

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

[27] Ware, J. H. and Epstein, M. D., "Comments on 'Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study' by R. H. Bartlett et al.," p. 851.

[28] O'Rourke, J. P. et al., "Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the new born: A prospective randomized study," *Pediatrics,* v. 84, (1989), pp. 957-963.

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.[29] On the other hand, if the correct methodological judgment is that the evidence both from previous practice and from the initial trial was already rationally compelling, then this second trial, and the deaths of 4 infants treated by CT in it, would seem to be clearly unethical.

Nor was this the end of the matter. Stopping rules of the sort employed in this second trial are anathema to orthodox statisticians, despite seeming entirely sensible from an intuitive point of view. (Surely another reason to be sceptical of that orthodoxy.)[30]

Hence many statisticians argued that even this second trial had not produced truly reliable scientific evidence of the effectiveness of ECMO for PPHN. Stuart Pocock, for example, wrote:

> "... a decision was taken to halt randomization when the data disclosed four deaths among ten infants receiving conventional medical treatment compared with none among nine infants having ECMO (p = 0.054) [R]andomization was stopped early on the basis of a fairly small amount of data, all subsequent patients being allocated to ECMO.
>
> The investigators were sensitive to the individual ethics of seeking parental consent and randomization for the next newborn infant ... However, with only 19 patients this does not represent strong evidence of the superiority of ECMO and provides little scope for making reliable judgments on the benefits of this treatment for universal use in such newborn infants in the future.
>
> Thus collective ethics may have been compromised by such early stopping…. [I]f ECMO really is effective the prolonged uncertainties maintained by lack of really substantial evidence may well have led to fewer newborn infants worldwide receiving it than would have been the case had the trial continued longer."[31]

But surely we would feel no such "ethical conflict" if we did not in fact already feel that we had sufficient evidence that ECMO is effective both from the "observational study" (that is, the striking comparison between the results of treating babies with PPHN using ECMO compared to what had historically been achieved using the earlier treatment) and from the earlier trials. Of course we need to distinguish carefully between merely subjective opinion and genuine evidence; and it is true that clinicians have occasionally

[29] There is still then of course the central, and in my view ultimately irresolvable ethical issue of the extent to which it is ethically acceptable to inconvenience (or worse) those patients involved in the trial for the promise of establishing a new treatment as effective for the benefit of future patients.

[30] For an account —and detailed criticism— of the reasons statisticians disapprove of stopping rules, see Howson, C. and Urbach, P., *Scientific Reasoning: the Bayesian Approach,* Second edition, Open Court, La Salle, IL, 1993, Chapter 9.

[31] See Pocock, S. J., "Statistical and Ethical Issues in Monitoring Clinical Trials," *Statistics in Medicine,* v. 12, (1993), pp. 1459-1469. For Pocock, "The basic ethical conflict in monitoring trial results is to balance the interests of patients within the trial —that is, the individual ethics of randomizing the next patient— and the longer term interest of obtaining reliable conclusions on sufficient data —that is, the collective ethics of making appropriate treatment policies for future patients," Pocock, S. J., "Statistical and Ethical Issues in Monitoring Clinical Trials," p. 1459.

firmly believed that they "knew" that certain treatments are effective that we now have seemingly-conclusive evidence are not; but, unless some reason can be given why RCTs should be so much superior in evidential value that nothing else really counts (and I have suggested that no such reason has so far been given), then we seem to be in a situation where proper scientific judgment is in conflict with the imposed statistical orthodoxy. The sensible solution would seem to be to reject the orthodoxy —there would then be no ethical conflict, but simply the "individual" ethical conviction that assigning any baby to conventional treatment in any of these trials was unethical!

One consequence of this view of Pocock's and of other statisticians was yet another "properly randomised" trial on ECMO for the treatment of PPHN —this one in the UK. It was stopped early by the "oversight committee" because of an excess of deaths in the conventional treatment arm. (Oversight committees are independent bodies who play no role in the treatment or assignment of patients in the trial, but who are allowed to keep a running total of the outcomes on the two arms and intervene where this seems to them the ethical thing to do. Needless to say such committees, like stopping rules, are anathema to classical statistical orthodoxy.)

4. Conclusion: Towards a More Defensible Approach to Evidence in Clinical Trials

We saw in Section 2 that there is —at least— some reason to be suspicious of the cogency of all of the arguments for the superior epistemic power of RCTs, except for the modest one that randomizing controls for "selection bias". Is it plausible that selection bias might invalidate the "observational" evidence for ECMO's effectiveness? Well, clearly if Bartlett and colleagues were achieving an 80% survival rate by dint of carefully selecting only some privileged subset of the babies that presented at their hospital with PPHN —those with relatively mild versions of the condition, those whose lungs were closest to normal development or whatever— then we would have good reason to be sceptical that their results were genuine evidence for the effectiveness of ECMO: perhaps CT would have achieved 80% survival in the same highly selected group (or even better!). There has never been any suggestion, however, so far as I am aware, that this was the case. Bartlett and colleagues seem just to have begun to treat all the babies that arrived in their hospital with PPHN, and who would earlier have been given CT, with ECMO. There is also no serious suggestion that the demographics of the catchment area of the Michigan University Hospital changed in any significant way or that the nature of the condition changed in some way or any other reason why the group against which the ECMO treated babies were compared was in any significant way different from it.

But these are surely the questions that should have been asked rather than the insisting on an RCT. What is operating, at root, in the methodology of trials is something like Mill's methods —the "controls" are in essence out to eliminate other potential explanations of any observed positive effect aside from the explanation that it is the effect of the treatment at issue. We cannot, I suggest, do any better than control for factors that background knowledge exhibits as plausible (partial) "causes" of response to treatment. I can see, as explained earlier, no basis for the belief that we can do better than this by randomizing —we are always at the mercy of the possibility that some other "unknown" factor happens to be significantly unbalanced between experimental and control groups and that it, rather than some specific effect of the treatment, is what explains (or perhaps chiefly explains) a positive outcome in a trial. It would be good to have grounds to eliminate that possibility but neither randomization nor anything else can supply them. We just do have to rely on plausibility assumptions grounded in "background knowledge."

This is not to say that randomization should always be avoided —it usually does no harm and may do some good. It does some good in terms of controlling for selection bias. Though I should again reiterate that, despite another well-entrenched myth, there seems to be no epistemological foundation for the view that randomization is the only way to eliminate selection bias: if Bartlett and colleagues, for example, simply treated every baby who would have been treated with the then conventional treatment with ECMO, then there could be no selection bias (and if they did not then there would be hospital records of PPHN babies being rejected for ECMO treatment). Randomization does no harm on two conditions. The first is that known prognostic factors are equally balanced —you surely do not want to leave it in the lap of the dice-playing gods whether the trials groups are balanced for obvious factors; I can see no epistemological virtue in the rigmarole of checking, after the event, for "baseline imbalances" rather than matching from the beginning, but so long as the groups do finish up reasonably well matched in terms such as age, general level of health and fitness and so on then it doesn't matter how exactly they get there. The second condition under which randomization does no harm is that some sort, any sort of trial, should be justified —this condition is not satisfied if there is evidence already available (as there arguably was in the ECMO case) that establishes the superiority of the new treatment.[32] (This will not in fact often be the case even if I am right that well-thought-through historically controlled trials are in principle just as telling as RCTs —the sort of very large effect produced by ECMO is unusual in current medicine and clearly the smaller the effect the more uncertain the evidence.[33]) We need the sort of systematic thought about plausible "confounders" that is suggested by Mill's methods and that forms the basic method of enquiry throughout epidemiology. The question to ask was not "has ECMO been 'validated' in an RCT?" but rather "Is there any plausible alternative explanation for such a large change in the mortality rate other than that it was produced by the change to treatment by ECMO?" It seems clear that the answer is "no" and that the evidence produced by the historical comparisons, because it was of the outcomes for so many more babies, in fact was weightier than that produced in either of the first two subsequent trials —"properly randomized" or not!

5. Bibliography

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

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

[32] Of course "establishes" here has the sense of "establishes defeasibly but as well as we possibly can at present." Equally obviously the neonates given ECMO should have been (and actually were) carefully monitored for possible side-effects of the treatment. But the side-effects issue gives no reason to prefer RCTs to historical trials.

[33] Not that there aren't troubling ethical issues about so-called mega-trials aimed at 'establishing' very small effects of treatments for very common conditions. See Penston, J., *Fiction and Fantasy in Medical Research: the Large-Scale Randomised Trial,* The London Press, London, 2003.

CARTWRIGHT, N., *Nature's Capacities and their Measurement,* Oxford University Press, Oxford, 1989.

DOLL, R. and PETO, R., "Randomized Controlled Trials and Retrospective Controls," *British Medical Journal,* v. 280, (1980), p. 44.

FISHER, R. A., *The Design of Experiments,* Oliver and Boyd, London, 1935.

GORE, S. M., "Assessing Clinical Trials-Why Randomize?," *British Medical Journal,* v. 282, (1981), pp. 679-684.

HOWSON, C., *Hume's Problem,* Oxford University Press, Oxford and New York, 2000.

KADANE, J. B. and SEIDENFELD, T., "Randomization in a Bayesian Perspective," *Journal of Statistical Planning and Inference,* v. 25, (1990), pp. 329-345.

LEIBOVICI, L., "Effects of Remote, Retroactive, Intercessory Prayer on Outcomes in Patients with Bloodstream Infection: Randomised Controlled Trial," *British Medical Journal,* v. 323, n. 7327, (2001), pp. 1450-1451.

LEIBOVICI, L., "Author's Comments," *British Medical Journal,* v. 324, (2002), p. 1037.

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

MONCRIEFF, J. ET AL., "Active Placebos versus Antidepressants for Depression," *The Cochrane Database of Systematic Reviews* 2004, Issue 1, Art. No.: CD003012. DOI: 10.1002/14651858.CD003012.pub2

O'ROURKE, J. P. ET AL., "Extracorporeal Membrane Oxygenation and Conventional Medical Therapy in Neonates with Persistent Pulmonary Hypertension of the New Born: A Prospective Randomised Study," *Pediatrics,* v. 84, (1989), pp. 957-963.

PAPINEAU, D., "The Virtues of Randomization," *The British Journal for the Philosophy of Science,* v. 45, n. 2, (1994), pp. 437-450.

PEARL, J., *Causality-Models, Reasoning and Inference,* Cambridge University Press, New York and Cambridge, 2000.

PENSTON, J., *Fiction and Fantasy in Medical Research: The Large-Scale Randomised Trial,* The London Press, London, 2003.

POCOCK, S. J., *Clinical Trials-A Practical Approach,* John Wiley, Chichester and New York, 1983.

POCOCK, S. J., "Statistical and Ethical Issues in Monitoring Clinical Trials," *Statistics in Medicine,* v. 12, (1993), pp. 1459-1469.

TUKEY, J. W., "Some Thoughts on Clinical Trials, especially Problems of Multiplicity," *Science,* v. 198, (1977), pp. 1958-1960.

WARE, J. H. and EPSTEIN, M. D., "Comments on 'Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study' by R. H. Bartlett et al.," *Pediatrics,* v. 76, (1985), pp. 849-851.

WORRALL, J., "What Evidence in Evidence-Based Medicine?," *Philosophy of Science,* v. 69, n. S3, (2002), pp. S316-330.

WORRALL, J., "Why There's no Cause to Randomize," *The British Journal for the Philosophy of Science,* forthcoming.