LSE CPNSS

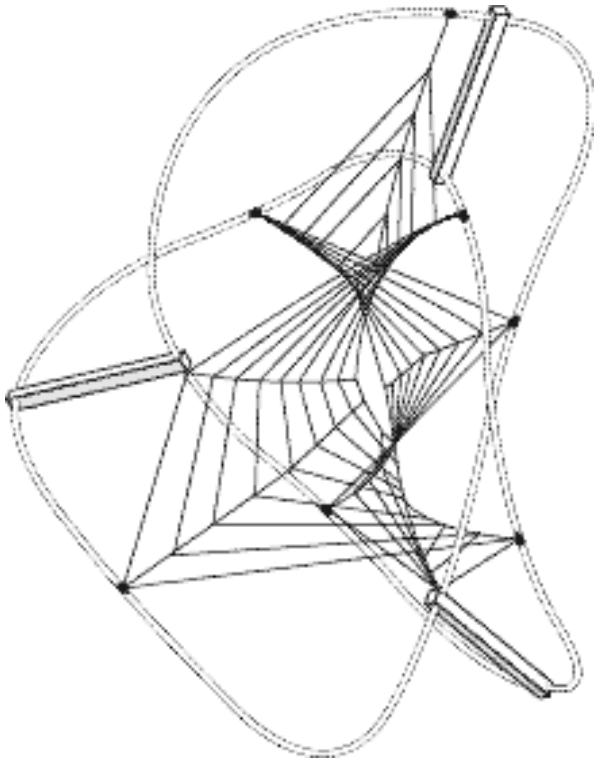
Centre for Philosophy of Natural and Social Science

Discussion Paper Series

DP 51/00

Controlling Clinical Trials: Some Reactions to Peter Urbach's Views on Randomisation

John Worrall
LSE



Editor: Max Steuer

Controlling Clinical Trials: Some Reactions to Peter Urbach's Views on Randomisation

by

John Worrall

comments To: j.worrall@lse.ac.uk

Controlling Clinical Trials: Some Reactions to Peter Urbach's Views on Randomization[1]

JOHN WORRALL

I have learned lots of things from having Peter Urbach as a colleague over more than twenty years, not the least of which is to be much more suspicious than I initially had been about orthodox classical statistics and its foundations. He has engaged statistical orthodoxy on a number of issues, including several invovled in the methodology of clinical trials.

Suppose a given group of patients who were all suffering from some disease D are given therapy T and that there is some overall improvement in their condition - say, to make it as simple as possible, that they all fully recovered from D . As everyone knows, it would be an error to infer without further ado that T was responsible for their cures and hence that T is in general an effective treatment for D.

If data alone is to tell us *anything* about treatment efficacy then it must include information about some sort of *control group* . Suppose, to take the hackneyed example, the condition at issue is the common cold and the treatment a course of vitamin C tablets taken for a week. Suppose further, to make the example as massively unrealistic as possible, that all the patients in the original experimental group had monozgygotic twins all of whom developed colds with exactly the same measured degree of severity of symptoms at exactly the same time as their twins and all of whom recovered in a week, even though none took any vitamin C (or any other relevant medication). Then, of course, far from data about the vitamin C group recovering within a week counting as evidence for the causal efficacy of vitamin C for colds, the data from the experimental and control groups *together* provides telling evidence for the *irrelevance* of vitamin C to recovery.

So control groups are important in judging treatment efficacy; if trials give evidence of such efficacy then it is only through observation in them of *differences* between experimental and control group in whatever is deemed the relevant outcome-measure. But how should control groups be formed?

It is not difficult to find in the literature - especially that part of it in which classical frequentist statisticians provide instruction in methodology for their clinical colleagues - very strong claims on behalf of the method of *randomization*. This is the requirement that some sort of random process be involved in dividing patients between experimental and control groups, so that "all possible assignments of treatments to patients are equally likely within the constraints of the experimental design." (Byar et al 1976, p.75).

Tukey (1977) 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) writes (p.1958) "Randomized trials remain *the* reliable method for making specific

[1] This paper was initially written for a one-day conference in honour of Peter Urbach to mark his leaving the Department of Philosophy, Logic and Scientific Method at the London School of Economics in 1996.

comparisons between treatments." Such remarks (and many others that could be cited) give the impression, as Peter Urbach has pointed out, that randomization is being claimed to be a *sine qua non* of a truly scientifically convincing trial. In fact, when it comes to detailed exposition and defence, all treatments make weaker, more guarded claims. But even those who allow that, at least under some conditions, some useful and "valid" information can sometimes be gleaned from studies that are not randomized, nonetheless continue to advertise randomized trials as undoubtedly epistemically *superior*. So, Stuart Pocock for example writes: "it is now generally accepted that the *randomized controlled trial* is the most reliable method of conducting clinical research." ((1983), p.5) Or Grage and Zelen (1972) assert (p.24) 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". And again similar quotations could be multiplied.

By contrast, the deliberate *matching* of the experimental and control groups in terms of factors that might be deemed relevant for the outcome (aside of course from the administration or non-administration of the therapy being tested) tends to be given a less central role. There is nothing approaching complete consistency in the statistical literature but Grage and Zelen again express a view that would, I think, meet widespread agreement amongst orthodox statisticians:

> An important component and refinement of most randomization trials is stratification by important known prognostic variables, which may influence the outcome of the trial. .... However, as every statistician knows, there are practical limits to the extent to which one can stratify for all known variables. When the number of patients in the trial is large enough, the randomization process will ensure that there will be an equal distribution of most of the important known and nuisance variables in both treatment arms. (*op. cit.,* pp25-6)

Some statisticians, eg Peto *et al* (1976), hold that stratification is an unnecessary elaboration of randomization. Some, eg Stuart Pocock (1983, p.81), hold that, while in some circumstances "stratification would seem worthwhile", this is not true for larger trials ("If the trial is very large, say several hundred patients, ... then stratification has little point" (*ibid*)) and that , even for smaller trials, the extra complexity involved may, practically speaking, introduce errors that outweigh any theoretical gain ("If the organizational resources for supervising randomization are somewhat limited then the increased complexity of stratification may carry a certain risk of errors creeping in, so that simpler methods may prove more reliable". (*ibid.*))

Peter Urbach has argued that, to the contrary, matching (*i.e*. stratification) is vital and the virtues of randomization have, to say the least, been overrated. He claims that neither of the two main arguments that have been traditionally regarded as demonstrating the necessity for randomization is cogent.

These are, *first*, the argument, due originally to Fisher, that the logic of statistical significance testing applies to a trial if and only if it is randomized. And *secondly* the argument that by randomizing the experimenter controls all at once for all

possible prognostic factors - either known *or unknown* - "at least in some probabilistic sense".

Urbach counters that Fisher's argument is unconvincing; and that the second argument does not bear examination (this despite the fact that it is undoubtedly the one that has won over a large section of the academic medical community to the necessity of randomization). He then goes on to provide a Bayesian analysis of clinical trials that, he contends, (a) shows that, although randomization may be useful, it is certainly no absolute requirement, and (b) underwrites the requirement that the two groups "*must* be adequately matched in respect of relevant prognostic factors" (Howson and Urbach (1993), pp.378-9).

There are a lot of issues involved in this more or less total disagreement between statistical orthodoxy and Peter Urbach, and it clearly is not feasible for me even to touch on them all. I shall not discuss *either* the cogency of his counter arguments concerning randomization *or* the merits of his alternative Bayesian analysis. Instead I shall concentrate on just one consequence that Peter Urbach draws from his arguments. This is the claim that the orthodox randomizers, as we might call them, have been unfairly and unjustifiably critical of trials that involve not randomized but *historical* controls.

Let me quote him at some length:[2]

> .... we shall argue that an intuitively satisfactory approach to the design of trials and the analysis of their results is furnished by Bayes's theorem, and that when looked at from that vantage point, randomization is seen no longer to be a sine qua non.
>
> It would not, of course, be correct to infer from this that randomization is necessarily harmful, nor that it is never useful - and we would not wish to draw such a conclusion - but removing the absolute requirement of randomization is a significant step which lifts some severe and, in our view, undesirable limitations on acceptable trials. For example, the requirement to randomize excludes, as illegitimate, retrospective trials using so-called historical controls. Historical controls suggest themselves when, for example, one wishes to find out whether a new therapy raises the chance of recovery from a particular disease compared with a well-established treatment. In such cases, since many patients would already have been observed under the old regime, *it seems unnecessary and extravagant to submit a new batch to that same treatment*. The control group could be formed from past records and the new treatment applied to a fresh set of patients who have been carefully matched with those in the artificially constructed control group (or historical control). But the theory of randomization prohibits this kind of experiment, since patients are not assigned with equal probabilities to the two groups;..." (*op. cit.*, pp. 279-280; emphasis supplied)

---

[2] I say 'him" - the passage is from Howson and Urbach (1993), but they each claim - or admit to! - separate responsibility for different chapters and this is from the Urbachian chapter 11 which involves a forward reference to an analysis to be provided in the equally Urbachian chapter 14.

However

> Freed from the absolute need to randomize, Bayesian principles .. imply that *decisive* information about a drug can be secured from properly controlled retrospective trials. (p378; emphasis supplied)

There is no theoretical reason why historical controls should not be regarded as epistemically entirely acceptable - so long of course as they involve "adequate matching in respect of relevant prognostic factors, on the basis of expert opinion" (pp. 378-9). There are also major practical and ethical advantages to historically controlled trials, at least in certain circumstances. Historically controlled trials seem *prima facie* to require fewer subjects (the control group pre-exists); and they do not require that physicians expose some patients to "ineffective placebos or what clinicians might judge to be inferior comparison treatments" (*ibid.*).

But there are practical *dis*advantages too, as Peter Urbach admits. In particular "historical controls that are adequately matched with current experimental groups can be compiled only with the aid of very thorough records of past patients, much more detailed than the records that are routinely kept". But this is, in his view, a *purely* practical difficulty that could be overcome were it not for "the widespread yet erroneous opinion that historical controls are intrinsically unacceptable or impossible in principle". This "widespread yet erroneous opinion" constitutes " a case where the mistaken principles of classical statistical inference are adversely affecting human welfare." (*op. cit.,* p.380)

Strong stuff. *Too* strong, I shall argue: there are genuine difficulties with historically controlled trials that Peter consistently either underplays or altogether ignores. My paper will consist of going through the arguments that have been raised against the reliability of historically controlled trials by orthodox statisticians and suggesting that at least some of them must be given more weight than he has been inclined to think. The question of what consequences this has for his underlying arguments against classical approaches and for a Bayesian analysis of clinical trials is one that must be left for some future occasion.

A preliminary terminological clarification

In order not to become further confused in an area not in any case known for its terminological exactness, I need to make one terminological point. Peter Urbach consistently equates "historically controlled trials " and "retrospective trials". In fact, although it is quite standard for commentators to use the term "retrospective" in connection with historically controlled trials, it is invariably used to qualify the term "controls": that is, historical controls and retrospective *controls* are treated as synonyms. "Retrospective *trials*", as standardly characterized, involve no intervention or experimentation on the part of the investigators: in such trials some group of patients is identified who *already suffer from some disease* and then the investigators look back to see how many of those patients were (earlier) affected by some conjectured cause C. This would then be compared with the relative frequency of earlier exposure to C within some control group of present non-sufferers. (So, for example, a retrospective trial of the old favourite claim that cigarette smoking causes lung cancer would involve taking a group of patients who already suffer from lung

cancer and noting how many of those had been smokers. This frequency would then be compared with the frequency of cigarette smoking in the population as a whole or amongst those not diagnosed as having lung cancer.)

There are a number of acknowledged problems with retrospective trials or studies. Non-retrospective studies are not unnaturally called "prospective". In prospective studies, patients in the experimental group are first identified as ones who have the putative cause present and are then followed, possibly over a number of years, to see how many of them develop the putative effect. Historically (or retrospectively) controlled trials are then in general *prospective* trials: some patients are given some new treatment and are then followed up to see how they react in respect of the relevant outcome measure; but instead of using a concurrent - and generally randomized - control group, the trial uses as controls "otherwise similar" patients who were earlier given the conventional treatment and whose subsequent progress has already been recorded. Peter Urbach is propounding the view that prospective, but historically controlled trials may be at least as compelling as randomized prospective trials.

Problems with historical controls

So what arguments do the statistically orthodox have for being suspicious of historically controlled trials and what cogency, if any, do those arguments have? I list the main arguments that I have found in trawling through the literature.

(i) Possible changes in the natural history of the disease and, more probably, of the population over time

There may be - possibly unquantifiable - differences in the historical population and the experimental population: given better levels of nutrition and general health, for example, the prognosis for certain diseases may be better now, even ahead of any effective treatment, than it was even a few years' ago. This consideration would, of course, invalidate any historically controlled trial in which the historical controls were very ancient. But usually such trials involve very recent historical controls - comparing, say, success rates on a new treatment with success rates in the same hospitals in some more or less immediately preceding set of patients.

It must of course be acknowledged as a possibility that some therpeutically significant change has occurred within the population in the shortest of interims. But in fact this consideration seems to be given relatively little weight by most opponents of historical controls. It seems likely that such historical changes would be relatively slow, so that, if the historical controls are from the recent past, they should be relatively unaffected by such changes.

(ii) Problems with the accuracy and completeness of the historical data

While the patients in the experimental group will all meet some clearly defined criteria for inclusion, since the historical controls on standard treatment were of course not involved in a trial, it will often not be known whether they satisfied these criteria. Moreover, record taking and record keeping differ not only between hospitals but between individual doctors within the same hospital.

Peter Urbach's reaction to this sort of problem is that it is a purely practical problem that could be overcome with sufficiently determined support for those advocating national or even international standardised data bases. Whatever the truth there, it does as a matter of fact mean that presently conducted historical studies can often be matched on many fewer variables (and, even then, only with considerably more conjecture) than would be optimal. It is also true of course that further research sometimes widens the range of variables that are considered potentially relevant - but then it will be highly unlikely that even systematic and consistently kept records will contain data about the newly considered variables.

(iii) Bias introduced by fortuitous changes in the samples

Byare *et al* (1976) discuss a major study - reported in 1967 - on the effect of oestrogen treatment on the survival of patients suffering from prostate cancer. 2313 patients were studied over a period of 7 years, using a stratified randomized design. If patients admitted in the last 2.5 years of study were looked at then there was no difference in survival rates amongst those treated with oestrogen and those treated with placebo. However there was a highly significant difference between the placebo patients in the first 2.5 years of the study and oestrogen patients in the last 2.5 years. The admission criteria had remained fixed, as of course had the outcome measure which was as objective as they get (length of survival). Moreover, since the experimental design involved randomization there was no question of selection bias (see below). These same data could have been involved in a historically controlled trial - using the placebo patients from the early part of the study as recent historical controls, whereupon the outcomes for the oestrogen treated patients in the later part of the study would have produced strong evidence for the effectiveness of oestrogen. In fact, the whole randomised study was taken to show that oestrogen was ineffective in increasing length of survival for prostate cancer sufferers.

Now in this case a finer-grained analysis of the data showed that the earlier patients were rather older and had poorer performance on a range of indicators that had been measured. The patients had seemed equivalent when a relatively coarse grained set of criteria were used but the data was available to correct this spurious result by finer-graining. But as Byar *et al* (1976) put it (p.76), while

> In this instance it was fortunate that data were available on variables that could explain the spurious significance. It is quite possible, however, that more subtle ... mechanisms could have been at work, producing biasses that could not be removed by adjustment technics, since the nature of the biasses could not have been identified.

More on post hoc adjustments shortly.

(iv) Selection bias

Unlike in randomized trials, all the patients actively involved in a historically controlled study are necessarily in the experimental group. The clinician therefore automatically knows what treatment anyone admitted to the trial will be given. This may lead to conscious, or perhaps more often subconscious, selection bias - ensuring that the experimental group is different from the historical control group in potentially important ways . Assuming the patient is fully informed then she too will

know that she is in the experimental group and again this may bias the result *via* the placebo effect. Finally the clinician or the assessor of outcome (if different) knows that all the patients were given the new treatment: this gives little room for bias in cases of unambiguous outcome such as death, but it may of course lead to quite substantial bias in the case of more ambiguous outcomes such as "improvement" of patients suffering from chronic disease.

Notice that the direct reasons for preferring randomized to historical controls in respect of possible selection bias have nothing to do with any special probabilistic considerations but simply with the failure of blinding (or "masking") in these cases.

There is no outright disagreement between the medical statistical establishment and Peter Urbach on this particular point, since Urbach acknowledges that

> This [that is,.in eliminating the possibility of selection bias] is where randomized allocation can play a role, for it eliminates those doctors' feelings and thoughts from the allocation process, and thereby makes it more probable than otherwise that the groups will be balanced on the factors mentioned. (*op. cit.*,p.376)

It does, however, seem to me that this admission of the virtues of random allocation is (i) somewhat *sotto voce* and (ii) over qualified - only showing that randomizing "might *sometimes* serve a useful purpose". (p. 377)

Peter Urbach's approach here, I suspect, reflects a presupposition that selection bias, though it must be acknowledged as a *possible* source of bias, is not in fact likely to play a very widespread or very substantial role. This presupposition if true would make it pretty astounding that selection bias is emphasised as the *main* criticism of historically controlled trials in pretty well every orthodox treatment: the orthodox might seem to making an Everest out of the smallest of molehills. Indeed it can seem that they are obsessed with the possibility of selection bias: to the extent, for example, that they are unhappy even with randomized designs if they involve *uncheckable* methods of randomization - such as coin flipping.

In fact, however, there is substantial evidence that this orthodox view is correct and that selection bias may be altogether stronger than might naively be thought possible. One type of evidence is from a different type of study again: the unblinded randomized study. Such concurrent studies involve dividing patients between experimental and control groups by some sort of open random process - whether their birth day of the month is even or odd, for example. The process is random but the investigator knows in which of the two groups a patient will end up , if she is declared to meet the admission requirements of the trial. A series of examples of such studies are cited in the literature in which the control and experimental groups ended up with substantially different numbers. For example an early study of anticoagulant therapy for myocardial infarction conducted on this unblinded randomized design had 589 treated (odd days) and 442 control patients (even days) ; while a study in the 60s of preoperative radiotherapy for rectal cancer based on the same design had 192 patients given the preoperative radiation (odd days) compared to 267 given surgical resection alone (even days). No one seems to have pointed out that there are in fact eight extra odd days in non-leap years and seven extra in leap years so one would not expect complete equality in divisions produced by this

method. Nonetheless, the imbalances in these two studies are unlikely to have occurred by chance and selection bias seems a likely explanation.

Even more telling evidence comes from some careful studies by Chalmers *et al.* (1983). These studies looked at imbalances between experimental and control groups in a large number of clinical trials on a number of prognostic variables *not explicitly taken into account by the trial protocol*. Three types of trial were represented in these studies - (i) "properly randomized", (ii) unblinded randomized and (iii) historically controlled. The imbalances were least in the case of fully randomized trials, greatest in the case of historically controlled trials and intermediate in the case of unblinded randomization. This, it is argued is exactly what you would expect if selection bias is operative: in both historically controlled and unblinded randomized (but not in fully randomized) trials, the clinician knows which arm of the trial a patient will be on, but whereas in historical trials, the clinician deliberately selects the control group, in unblinded randomized trials the clinician must find some (possibly) unconscious reason for exclusion of a patient who would otherwise be in the trial once his or her date of birth (or whatever) is known.

This study by the way was of trials all involving that most objective of outcomes, death. The authors drew the conclusion that

> the more frequently positive results found in studies in which assignment of controls is less blinded may be explained by bias in the selection or rejection of patients when the treatment to be given is known .. at the time of assignment (*op. cit.* p.1358).

(v) Treatment bias

One final bias often cited as a reason for being suspicious of historically controlled trials might be called "treatment bias" (though there is no generally accepted terminology): the possibility that patients involved in a trial receive more and better attention than 'standard' patients. There is, it is argued, likely to be greater than usual enthusiasm amongst physicians when a new treatment that may lead to major improvements is in the offing as well as higher levels of ancillary care. And remember it very often happens that patients involved in a trial are being treated at, or under the auspicies of, major research hospitals and institutes, while the historical controls may have been receiving conventional treatment at much less prestigious places. Since the patients in the historical control groups were generally not involved in any trial, this is again likely to lead to overestimates of the efficacy of the new treatment. (In the case of randomized trials, on the other hand, the physician will not know which treatment arm a particular patient belongs to and hence, although all the trial patients may receive "special treatment", there will be no reason to treat those on the treatment arm any more specially than those on the control arm.)

Peter Urbach, along with other Bayesians, has cited the ECMO trials[3] as a case where historically controlled investigations already made it extremely plausible that a newly

---

[3] "ECMO" (extra-corporeal membraneous oxygenation) was a method for treating pulmonary hypertension in neonates discovered in the 1980s. The treatment appeared to turn an 80% mortality rate in babies suffering from this condition into an 80% survival rate. However, rightly or wrongly,

introduced treatment was superior to existing treatment and where the further randomized trials that were in fact undertaken were unnecessary and, because unnecessary, ethically suspect to say the least.

But, whatever the overall truth in that case, there is one feature of it that can plausibly be explained by the sort of treatment bias that opponents of historical controls cite. The salient evidence in this case was a reported 80% survival rate of babies treated by the new technology ECMO compared to a reported 80% historical death rate on conventional treatment. This evidence convinced some commentators - and indeed some of the clinicians involved - that "*it seems unnecessary and extravagant to submit a new batch to*" conventional treatment. But when the "unnecessary" randomized trials were performed, while they confirmed a roughly 80% survival rate for ECMO, they also involved a 60% survival rate on conventional treatment. Now this might of course be explained as a sampling error (the sample of babies given conventional treatment in the trial was indeed - I personally believe thankfully - small) on a population frequency of 20% survival, but it might also - and more plausibly - be explained by what I've been calling treatment bias.

A solution of these problems: post hoc adjustment?

An obvious suggested solution of the problems cited against historical controls, and one of course strongly recommended by Peter Urbach's Bayesian treatment, is that if such biasses can be identified then they can be allowed for - if only after the event. If selection bias has led to an experimental group which is on average much younger or healthier, say, than the historical controls then this identifiable bias can be taken care of by statistical adjustment techniques. What counters do the defenders of randomized controlled trials as the near exclusive sources of real scientific evidence have to this obvious reply?

Well post hoc adjustments can in fact be made - using techniques that epidemiologists have always used. Nonetheless the defenders of RCTs argue that there are limitations to methods of post hoc adjustment and (I take it) that these limitations make it unlikely that even adjusted results from historical trials are as reliable as concurrent randomised trials. Amongst the limitations cited, the following seem to be fairly common:

1. Post hoc adjustment cannot solve the problem about the relatively poor historical data: since the historical controls will standardly not have been involved in a trial the data on them will often not be sufficiently precise or complete.

2. You cannot of course adjust for all possible factors that *might* have been important.

3. While you may be able to adjust for at least some of the biasses involved in patient selection, you cannot in general adjust for changes in what Pocock (*op. cit.*) calls "experimental environment" - in which he includes what I have called treatment bias ("it is very difficult to ensure that all aspects of managing the patient, other than the

---

statistical orthodoxy insisted on RCTs. For some of the controversy surrounding this fascinating and disturbing episode consult Ware (1989) and the associated commentaries.

treatment under study, remain constant" *op. cit.*, p.55) and a tendency to take more experimental patients off the trial, when for example they develop side-effects.

4. Perhaps the most telling point is that

> The validity of an adjustment technique depends on the correctness of the assumed mathematical model. Unfortunately one's ability to specify models correctly is to a large extent dependent upon the sample size. In samples of moderate size curvilinear relations and important interactions are often not detected. Consequently one is always uncertain whether adjustment does in fact eliminate the full effects of the confounding variables. (Byar et al (1976), p.76)

Certainly Urbach's particular Bayesian analysis seems - and this is an area that needs more development - to be motivated by a straightforwardly Millian account of eliminative induction. This account is known to presuppose for its validity an underlying metaphysical picture in which each factor has its own separate and simply additive effect on whatever aspect of the world we are investigating. This picture is certainly not universally accurate and may in fact apply rather seldom.

5. Relatedly, if the factor being adjusted for (usually called the confounder or covariate) is measured *after* the treatments have been administered, then clearly there is the possibility for some such covariates that they were affected by the therapy and therefore adjustment for it will remove some of the true treatment effect (whether positive or negative). Some factors that might indeed need to be taken into account can *only* be measured after treatments. One clear example is rate of adherence to a therapy. If say only 50% of experimental patients adhered to the whole course of therapy while 95% of the historical controls took the whole conventional therapy then one would want presumably to allow for this in judging how effective the new treatment was relative to the old. But if adherence was correlated to development of side-effects for example, then adjusting for this difference would favour the newer treatment.

Now there undoubtedly is some question of 'sauce for the gander' here - particularly concerning the adoption of mathematical assumptions that one cannot guarantee are correct. It is also true that these rejoinders seem at least implicitly to involve the supposition that randomization in effect eliminates all these problems. And Peter Urbach has argued that this supposition is, to sat the least, extremely suspect. Nonetheless I think these rejoinders do establish that his account gives a rather more rosy view of the power and trustworthiness of post hoc matching than is justified by the current state of the intellectual debate.

The track records of historical and of randomized trials

Suppose Peter Urbach is correct that there are difficulties in defending randomised trials at the general conceptual level; and suppose these defenders of RCTs are right that similar difficulties afflict attempts to justify historical trials. It might seem tempting to take the view that we should be looking - as good reliabilist, naturalised philosophers - rather to the track records of these two sorts of trial as our indicators of likely reliability. Certainly a crucial motivation for all orthodox critics of historically controlled trials is their conviction that such trials have, as a matter of fact, proved

substantially unreliable in the past. Thus Grage and Zelen (*op. cit.*, p.37) while going through a range of conceptual objections write:

> The major argument against the use of the historical control study are (sic) the large number of studies which have used historical controls and proclaimed an advantage for the new method over the [conventional treatment but] .. which have subsequently been debunked by a single well-conducted, well-controlled prospectively randomized clinical trial.
>
> The literature is replete with examples where a new treatment was highly touted by its proponents and enjoyed a good deal of popularity for a time, but did not withstand the test of a well-controlled randomized clinical trial.

Critics can certainly point to a long list of treatments that were indeed initially strongly recommended by historically controlled trials but were later dropped from accepted treatment as either worthless or, still worse, positively harmful. Examples include both interferon and laetrile as treatments for various cancers, intra-arterial infusion therapy for the treatment of metastatic colorectal carcinoma of the liver, hydrocortisone treatment after acute myocardial infarction, portacaval shunt operation for portal hypertension and freezing of the stomach for duodenal ulcer.

The list is, at any rate superficially, impressive. It could be extended almost indefinitely if, as Peter Urbach actually suggests, we accept as a form of historically controlled trial *any* comparison even by an individual physician of some new treatment with earlier experience. Several points should, however, be conceded to the Urbach line.

First the fact that the result of a historically controlled trial is contradicted by the result of a randomized controlled trial on the same treatment is only a criticism of the historical method if the randomized trial is taken as providing, as is generally said, the 'gold standard' - and that is exactly the issue here. After all, if it were decided that historically controlled trials were the gold standard then the fact that they contradict the result of randomized trials would of course be taken as indicating the *latter's* unreliability! It is true that there are many treatments that (a) were positively recommended in historically controlled trials but (b) are no longer acceptable medical treatments. But since the whole issue all along has been whether or not accepted medical treatments are really properly based on scientific evidence, and since it is just an empirical fact that what's an acceptable medical treatment has at least begun to be influenced by the results of randomized trials, this in itself is not convincing. (This is a general problem with reliabilism, but that's another paper.)

Secondly, critics of historical trials also point to a (substantial) number of cases in which different historically controlled trials on the same treatment have contradicted one another. However, the Bayesian Donald Berry has pointed out that there are also cases where the different randomized studies of the same treatment have come to opposite conclusions (see, e.g., his (1993)). Though the orthodox will no doubt respond that the fact that a study is randomized need not on its own mean that it was "well conducted" study. It would be interesting to have examples of trials that all the statistical experts agreed were well conducted and yet whose results were

subsequently overturned by later equally well conducted studies. Interesting, but of course far from conclusive.

There are two facts, for which defenders of RCTs claim there is already strong evidence and which would, if vindicated by a full history of clinical trials, surely be telling. These are that where there have been many studies of the same treatment, both historically controlled and randomised studies:

(i) the historically controlled trials have had much more variable results and hence have given much more variable estimates of the likely effectiveness of the treatment at issue and

(ii) that historically controlled trials have been consistently more favourable than randomized studies to the new treatment involved.

A fundamental difference in background motivation between Peter Urbach and orthodoxy

Several remarks in Peter Urbach's account suggest that he starts from the default position that any new treatment that has got to the stage of being tested in some clinical trial of whatever hue, or about which qualified doctors are enthusiastic, is more likely than not to be more efficacious than conventional treatments. So for example he remarks that an advantage of not insisting on randomized trials when historical trials have already pronounced in favour of some new treatment is that the physicians involved are thus "spared" from exposing those who would have been in the randomized control group to "ineffective placebos or what clinicians might judge to be inferior comparison treatments" (1993,p.1430). But this is clearly an advantage *only* if the new treatments are more beneficial to the patient than either placebo or conventional treatment. Lots of other remarks indicate the same underlying assumption.

On the other hand, the defenders of RCTs are mainly motivated by what they perceive to be rampant over-enthusiasm in the practice of medicine - by case after case of what they see as treatments that have been strongly endorsed by some doctors, and perhaps widely or even generally accepted within medicine, that in fact have no scientific backing (and which have subsequently been exposed as such). They see case after case where ill-founded judgments that some new treatment is superior to established ones have subsequently been debunked. They see case after case in which because of publicity for poorly controlled historical trials, it has become practically and ethically very difficult to perform "proper" "scientific" studies of the treatment concerned. For the defenders of RCTs, the historical record justifies an attitude towards proposed new treatments of guilty until proven innocent; while many remarks in Peter's treatment betray an attitude of innocent until proven guilty. For example, he is several times concerned about the ethical difficulties involved in a doctor's putting her patients in a randomized trial, "knowing" that half of them may receive what she takes to be an inferior treatment, that is, either the old one or placebo. The orthodox on the other hand are concerned with the ethics of having views about which is the superior treatment in the absence of any telling scientific evidence - about the ethics of elevating one's own inevitably highly biassed and subjectively interpreted experience into a standard of what is really effective.

Conclusion

So what conclusion is to be drawn from this wideranging and clearly preliminary survey? Well only that whether or not Peter Urbach's arguments against the attempted justifications of randomization are correct, whether or not he is correct about the central importance of matching, he himself underestimates the problems inherent in the method of historically controlled trials.

This is an area to which philosophers of science have given relatively little attention. Yet there are few areas to which they might contribute that have such enormous practical significance - involving, as it does, issues that are often literally matters of life and death.

REFERENCES

Berry, D.A. (1993) "A Case for Bayesianism in Clinical Trials" *Statistics in Medicine*, **12**.15/16, pp 1377-1393

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

Chalmers, T.C. *et al* (1983) "Bias in Treatment Assignment in Controlled Clinical Trials", *New England Journal of Medicine* **309**/22, pp 1358-1361

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

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

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

Pocock, S.J. (1983) *Clinical Trials - A Practical Approach*. Chichester and New York: John Wiley.

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

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

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