Brit. J. Phil. Sci. (2007), page 1 of 38

Why There's No Cause to Randomize

John Worrall

ABSTRACT

The evidence from randomized controlled trials (RCTs) is widely regarded as supplying the 'gold standard' in medicine—we may sometimes have to settle for other forms of evidence, but this is always epistemically second-best. But how well justified is the epistemic claim about the superiority of RCTs? This paper adds to my earlier (predominantly negative) analyses of the claims produced in favour of the idea that randomization plays a uniquely privileged epistemic role, by closely inspecting three related arguments from leading contributors to the burgeoning field of probabilistic causality—Papineau, Cartwright and Pearl. It concludes that none of these further arguments supplies any practical reason for thinking of randomization as having unique epistemic power.

1 Introduction

In a randomized controlled experiment (RCT) designed to test some new treatment—perhaps a new drug therapy or a new fertilizer—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 some 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 those in the control group receive instead may differ from case to case: in agricultural trials, the control plots may simply be

doi:10.1093/bjps/axm024

left unfertilized (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 of a proposed new drug treatment, 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 generally performed 'double blind': neither the subjects themselves nor those treating them know whether they are receiving the experimental or the control treatment.[1] (The first part of the condition can presumably be taken as read in the case of agricultural trials!)

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: the medical profession has been overwhelmingly convinced that RCTs represent the 'gold standard' by providing the only 'valid', unalloyed, genuinely scientific evidence about the effectiveness of any therapy.[2] Clinical science may occasionally have to rest content (perhaps for ethical or practical reasons) with evidence from other types of trial, for instance, so-called historically controlled trials (in which the control group is supplied by—arguably similar—patients who were treated for the same condition under the previous treatment regime), but this is always very much (at best) a case of epistemic second-best. For, it is widely believed, all nonrandomized 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).

Indeed, randomization has received *such* a favourable press that the educated layman could be forgiven for believing that its special scientific value is an entirely uncontroversial matter. However, this is far from true. Systematic treatment of the role of randomization in scientific inference began with R.A. Fisher in the early 1930s, and, as Ian Hacking points out in a fascinating historical article (Hacking [1988]), Fisher's reasoning was challenged right from the beginning by W.S. Gosset (aka 'Student'). Later, the necessity for randomization was challenged from a Bayesian point of view by Savage, and

[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 what is described above is the simplest RCT, and what most commentators have in mind in assessing the power of the methodology.

[2] Many clinicians will insist that alongside the objective evidence provided by trials, there is also evidence (sometimes misleadingly called 'subjective') garnered from clinical practice. They go on to insist that clinical experience or clinical intuition is an equally 'valid' source of evidence, often, again misleadingly, characterizing this view as that 'there is more to medicine than science' or ' there is an art to medicine as well as a science'. However, when it comes to the 'objective' 'scientific' evidence, even most of these clinicians would agree that RCTs provide the 'gold standard'.

while on this, as on every other substantive issue, there are major differences among those who regard themselves as Bayesians, the view that randomization has no essential (and certainly no *direct*) role remains the majority view within that influential school.[3]

In an earlier article (Worrall [2002]), and partly following the treatment of Peter Urbach (see Urbach [1985] and Howson and Urbach [1993]), I analyzed the main arguments that are routinely put forward to support the view that randomization is necessary for 'scientific validity' or at least carries very special epistemic weight. Urbach concentrated on two of these: first, Fisher's contention that randomization is necessary to underpin the logic of significance tests, and second, the idea that randomization somehow controls all at once, for not only known, but also *unknown* possible 'confounders'. It seemed to me that he succeeds in showing that neither of these arguments is at all compelling (though I shall need in the course of this paper to return more than once to the second argument, which is certainly seemingly the most persuasive).[4]

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' potential bias that is believed to have in fact operated to invalidate a number of trials—'selection bias'. If clinicians involved in trials are allowed to decide the arm of the trial to which a particular patient is assigned, then clearly it becomes possible that they will effect, perhaps subconsciously, a selection that distorts the result of the trial and hence gives an inaccurate view of the efficacy of the treatment. They might, for example—having a view on the likely effectiveness of the new drug and also its likely side effects—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, the pharmaceutical company) paymasters. This may well affect the outcome, especially when the effect of the new treatment, if it has one at all, is unlikely to be very dramatic. (Selection bias could, in principle, affect the result in either direction, though attention has predominantly been paid to cases where it has arguably produced a false *positive* result.) This does seem to me, as Urbach also concedes from a Bayesian perspective, a definite epistemological good that properly performed RCTs deliver. Notice, however, that there is of course nothing magical about the role of the coin-toss or random number table: selection bias is eliminated in an RCT because the procedure of random allocation means that the experimenter cannot affect the arm that

[3] For a survey of the complexities here, see (Kadane and Seidenfeld [1990]).

[4] An especially clear version of the argument that Fisher fails to show that randomization is necessary to underpin the logic of the significance test is developed in (Howson [2000], pp. 48–51).

particular patients are assigned to; if the trial was, and remained, double-blind then randomization could play no further role in this respect; moreover even in trials where the experimenters are not blind, any other means of taking the division into experimental and control groups out of their hands would eliminate this potential bias equally effectively; and finally, as we shall see, even outside formal experimental trials, in the case of 'observational studies' (aka 'historically controlled trials'), there may be solid evidential grounds for holding that selection bias can have played no (or at most, a negligible) role in the observed outcome.

In my ([2002]) paper, I analyzed a fourth argument (this one not mentioned by Peter Urbach) that has also been given a good deal of emphasis especially 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 fact* that the 'track-record' of RCTs is better than historically controlled trials (also sometimes known as 'observational studies') because the latter standardly give unduly optimistic estimates of treatment effects. This argument, so I suggest in my ([2002]) paper, is (i) circular (it depends on supposing that where an RCT and an 'observational study' have been performed on the same treatment it is the former—which after all provides the 'gold standard'!,—that reveals the true efficacy); (ii) based on comparing RCTs to particularly poorly performed observational studies that anyone would agree are obviously methodologically unsound; and (iii) to say the least, brought into 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 using the RCT protocol differ from one another much more markedly than do those using carefully performed and controlled observational studies.[5]

More recently, some different (or at least *seemingly* different) arguments to the effect that randomization has special epistemic virtues have, however, arisen from the burgeoning literature on 'probabilistic causality'. These arguments differ in detail, but all claim that randomization 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 good outcome are associated or correlated. This paper analyzes three such arguments from leading practitioners in the field: in the order in which I shall treat them here, David Papineau, Nancy Cartwright and Judea Pearl (Sections 3, 4 and 5). Before embarking on my analyses of these arguments, however, it will help briefly to outline (in Section 2) a case-study that *both* shows how immensely important are these apparently rather abstract arguments about the epistemic weight of different types of trial *and* will help me eventually to

[5] See, for example, (Benson and Hartz [2000]) and (Concato, Shah and Horwitz [2000]).

formulate (in Section 6) the rather nuanced view about the epistemic role of randomization that seems to me defensible.

2 Why the issue is of great practical importance—the ECMO case[6]

A persistent mortality rate of more than 80% had been observed historically in neonates experiencing a condition called persistent pulmonary hypertension (PPHS). A new method of treatment, extracorporeal membraneous oxygenation (ECMO), was introduced in the late 1970s, and Bartlett and colleagues at Michigan found, over a period of some years, mortality rates of around 20% in infants treated by ECMO (see Bartlett *et al*. [1982]). This new treatment could hardly be regarded as a mere stab in the dark. It was already known that the underlying cause of PPHS was immaturity of the lungs, leading to poor oxygenation of the blood, in an otherwise ordinarily developed baby. Those babies that survived were those that were somehow kept alive while their lungs were developing to maturity. ECMO, in effect, takes over the function of the lungs in a simple and relatively noninvasive way. Blood is extracted from one of the baby's veins before it reaches the lungs, is artificially oxygenated at a membrane, reheated to regular blood temperature and re-infused into the baby's carotid artery, thus bypassing the lungs altogether. Of course it does not follow from this that the treatment was *bound* to work. All interventions can have unexpected side-effects that sometimes outweigh direct benefits, but it does mean that it would hardly be a surprise if it *did* work.

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 randomized study'). This was in spite of the fact that their experience had already given them a high degree of confidence in ECMO as compared to the earlier treatment ('We anticipated that most ECMO patients would survive and most control patients would die ...').[7] They felt compelled to perform a trial because their claim that ECMO represented a significant improvement in treating PPHS would, they judged, carry little weight amongst their medical colleagues unless supported by a positive outcome in such a trial.[8] 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

[6] Peter Urbach first drew my attention to this case

[7] Both of these latter quotations are from (Bartlett *et al*. [1982]).

[8] This is another argument for RCTs that is frequently 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 them!

conventional medical treatment provided sufficiently rigorous controls; and hence, that the results of more than 80% survival that they had achieved with ECMO already showed that ECMO was a genuinely efficacious treatment for this dire condition. A comparison between the outcomes achieved using a new technique and the outcomes in an allegedly comparable earlier group of patients with the same condition but treated with the conventionally established treatment is, as already noted, called an 'observational study' or, more revealingly, a 'historically controlled trial'. Because such studies are generally considered to carry little or no weight compared to RCTs, these researchers in the ECMO case felt forced to go ahead and conduct the prospective trial.

They reported its outcome in 1985 (Bartlett *et al.* [1985]). Babies were allocated to ECMO treatment or to the 'control group' (which received the then conventional medical therapy, hereafter 'CT') in this particular trial using a modified protocol called 'randomized play the winner'. This 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 placed 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 clearly itself a compromise between what the researchers saw as the needs of a scientifically convincing trial and their own convictions about the benefits of ECMO.

As it turned out, the first baby in the trial was randomly assigned ECMO and survived, the second was assigned CT and died. This of course produced a biased urn, which became increasingly biased as the next eight babies all happened to be assigned ECMO and all survived. The protocol, decided in advance, declared ECMO the winner at this point, though two more babies were treated with ECMO (officially 'outside the trial') and survived. Thus 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 of around 80% for the disease.)

Ethics and epistemology are fully intertwined here. One's view of the ethics of undertaking the trial in the first place will depend, among other things, on what is taken to produce scientifically significant evidence of treatment efficacy. 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 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[9]

then you are likely to have a different view, perhaps even that

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

Many commentators in fact took this latter view and concluded that

> Further randomized clinical trials using concurrent controls and [. . .] randomization [. . .] will be difficult, but remain necessary. (Ibid.)

Those taking this second view held that it was entirely ethical to perform another trial, since neither the 'historically controlled' results nor the results from this initial 'randomized play the winner' trial had produced any truly reliable, scientifically telling information. The Michigan trial had not produced any real evidence because, in deference to the researchers' prior (and, according to these critics, unscientific) convictions, it had not been 'properly randomized'. Indeed, those taking this view even imply (see their 'will be difficult [to perform]' remark) that such trials and their 'historically controlled' antecedents, have, by encouraging the belief that the 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 doctors then 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 incompletely randomized trials should be actively discouraged. (Of course, since historical trials simply happen when some new treatment is tried rather than 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 (and reported in

[9] (Tukey [1977]; emphasis added) It is difficult to take this remark seriously, no matter how influential it, along with similar remarks from other classical statisticians, remains. First, as Jim Woodward reminded me, randomization is not at all required for the validation of a whole range of causal hypotheses (for example, in physics)—why should claims about the causal effectiveness of a medical therapy be necessarily different? And second, even within medicine, the vast majority of accepted treatments—right through from aspirin for mild headache, to diuretics in heart failure, to pretty well any surgical intervention (such as appendectomy for acute appendicitis)—have never been subjected to an RCT, and yet, no one seriously doubts their effectiveness.

O'Rourke *et al*. [1989]). This second trial involved a fixed experimental scheme requiring $p < 0.05$ with conventional randomization but with a stopping-rule that specified that the trial was to end once 4 deaths had occurred in either group (experimental or control). A total of 19 patients were, so it turned out, involved in this second study: 9 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 centers were still geared up to take trial-patients, a further 20 babies who arrived at the trial centres suffering from PPHS were then all assigned to ECMO (again officially 'outside the trial proper') and of these 20 extra patients, 19 survived.

Once again, views about the ethics of this further trial and in particular about the four 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 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 (together with 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.[10] 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 four infants treated by CT in it, would seem to be clearly unethical.

Some decisions, then, about which trials it is reasonable to perform have been motivated entirely by the claim that randomization has, at the very least, some highly privileged epistemological status. We should, therefore, be even more vigilant than usual in examining the credentials of any argument that this epistemological claim is correct, since it is carrying extraordinary ethical weight.

The conclusion of my ([2002]) paper was that, aside from the argument from selection bias (and, as I shall discuss later, it is difficult to believe that this particular bias *significantly* affected the original ECMO data from the 'historical trial'), it is in fact very difficult to see any cogent reason for thinking as highly of RCTs as the medical community does. The main purpose of the present paper is to see if the missing cogent argument can be found amongst

[10] Indeed some commentators, e.g. the statistician Stuart Pocock, have argued that because the stopping-rule meant that only 19 patients were involved in this second trial, even it could supply no solid scientific information, and hence, recommended still a third trial! But the point about the ethical significance of the strength of the arguments for randomization has already surely been made without going into this further twist in the story. (It led to a further UK trial that had to be stopped early because of too many deaths on the conventional treatment arm!)

those emanating from the recent literature on probabilistic causality, none of which I examined in that earlier paper. In the next three sections, I analyze such arguments developed by David Papineau, Nancy Cartwright and Judea Pearl.

3 Papineau on the 'virtues of randomization'

As indicated, my own investigation of the arguments for the special epistemological status of randomization took off from Peter Urbach's analysis. That analysis is criticized in (Papineau [1994]). Papineau argued that Urbach confuses issues about the role of random sampling with quite different issues about randomization; and that the role of the latter lies in underwriting genuinely *causal*, as opposed to merely correlational, conclusions from trial data.[11]

Papineau sees clinical researchers as making *two separate* inferences from trial data. The first inference proceeds from some observed difference in the frequencies of recovery (let us suppose that this is the outcome measure at issue) in the experimental and control groups to some conclusion about the 'objective probabilities' in the 'underlying population' of recovery (R) conditional on treatment or no treatment (T or $\neg$T). He takes it—I believe incorrectly—that this is where the main difference between the classical statisticians and the Bayesians lies, and that the difference is about whether random sampling is necessary in order for this first inferential step to be justified: classical statisticians insist that probabilistic conclusions can legitimately be drawn from sample frequencies if and only if the sample is random, while Bayesians deny this.

As Papineau notes, there is in fact no sense in which the initial set of patients involved in a typical clinical trial can be thought of as constituting a random sample from the 'target population' (even, I would add, where this is even remotely clearly defined): the members of the study-population are recruited *via* a mixture of deliberate design (all trials will have more or less explicit inclusion and exclusion criteria) and happenstance (which appropriate patients happen to walk at the appropriate time through the doors of the clinics

[11] Let me say at the outset that Papineau is definitely *not* one of those who believe that randomization is always necessary for any reasonable conclusion to be drawn about the effectiveness of some treatment. Indeed several remarks make it eminently clear that he would, in the ECMO case, vote with those who hold that the randomized trials were unethical. But he does definitely argue that randomized trials invariably produce results of greater epistemic weight, and hence that we are always settling for second best if we settle for a nonrandomized trial, even where ethical considerations strongly favour the latter sort of trial. I take it, therefore, that his argument, if accepted, would lend some support to the hard-line randomizer who would differ from David Papineau only on the issue of the possible price worth paying for the extra information derived from the trial and its extra security. (I should also add that Urbach himself responded in turn to Papineau's claim (Urbach [1994]). However, since I have a number of issues with Urbach's response I shall not take this up here and will proceed to criticise Papineau's account independently of that response.)

run by those clinicians who happen to be involved in the trial). Papineau sides with what he takes to be the Bayesian position here, agreeing that random sampling is neither necessary nor sufficient to legitimate the inference from finite frequency data to population probabilities. We can legitimately draw probabilistic conclusions if, and only if, we have no positive reason to think that the sample is in some systematic way misleading or unrepresentative. We may have no such reason even if the sample was not generated by a random process; and conversely, we may in fact be able to see that, even though the particular sample before us was drawn at random, it *is* biased in a clear way that we have reason to think may be significant, in which case it would be folly to follow the classical statistician in continuing to support the inference from observed frequencies to probabilities.

Papineau then proceeds to argue that random sampling and experimental randomization are very different things (as indeed they are) and that Urbach has confused the two (which I would dispute). Random sampling (selecting the group of patients involved in the overall study by some random selection from the 'set of all possible subjects') is not necessary to justify the step from observed relative frequencies to probabilities, but randomization (that is, the division of the overall study population, however assembled, into experimental and control groups *via* some random procedure) *is*, as Papineau sees it, necessary to justify the further inference from probabilities to *causes.* Having arrived at the conclusion that $p(R/T) > p(R/\neg T)$, the scientist conducting the clinical trial, on Papineau's account, is concerned to try to justify the further inference to the conclusion that the therapy T is a *cause* of recovery R.

Why is this a *further* step? Because, of course, of the problem of 'spurious correlation'—the possibility that the apparent connection between two variables (such as treatment and recovery in this case) may be 'accidental' rather than 'causal'. Suppose, for example, counterfactually so far as almost all real clinical trials are concerned, that the treatment under test was one already in use in the general population. And suppose that the trial was performed, not by dividing the study population by some random process, but simply by 'patient choice', that is, the experimenters simply recorded whether or not a particular person in the trial was taking the treatment at issue and then looked to see whether that person recovered in whatever time was specified by the terms of the trial. It could then easily turn out, of course, that choosing to take the treatment was itself 'correlated with' (that is, not probabilistically independent of) some other factor that plays a role in recovery. Suppose, for example, that more young people choose to take the treatment and so young people are over-represented in the 'experimental group' in the trial we are currently envisaging. It could easily be that young people (Y) are more likely to recover anyway (that is, independently of whether or not they receive T). Should it in fact be the case that $p(R/Y \& T) = p(R/Y \& \neg T)$ then Y 'screens

off' R from T—that is, R and T lose their probabilistic dependence when conditioned on being young. And in that case we would presumably want to infer that, despite the probabilistic dependence of R and T, T does *not* cause R.

I admit that it is tempting to think that experimental randomization can solve this problem of 'confounding' or of 'spurious correlation' (and we shall see that, albeit on the basis of rather different approaches, the same claim is also made by both Nancy Cartwright and Judea Pearl). But how *exactly* is this suggestion supposed to work?

Well, according to David Papineau's account, if you have already inferred that $p(R/T) > p(R/\neg T)$ (and remember that he admits that this is an undeniably fallible inference from observed frequencies) then it is a *guaranteed sure-fire* further inference to the conclusion that T causes R, *if* (but only if) the data that has more recoverers among the treated group than among the untreated is data from a randomized trial. He presents this sure-fire guarantee as itself the conclusion of the following argument.

The premise is the account of causality that he endorses, namely that specified in his principle **C** (*op. cit.,* pp. 439–440): 'A generic event like the treatment T is the cause of a generic event like the recovery R iff there are contexts (perhaps involving other unknown factors) in which T fixes an above average, single-case objective probability for R.'[12] He then argues that **C** entails that *if* $p(R/T) > p(R/\neg T)$ then *either* T causes R or T is correlated with one or more other factors that cause R; and finally that randomization entirely eliminates the possibility that the second disjunct holds:

> Since the treatment has been assigned at random—in the sense that all patients, whatever their other characteristics, have exactly the same objective probability of receiving the treatment T—we can now be *sure* that T is *not* itself objectively correlated with any other characteristic that influences R. (*op. cit.,* p. 440; my emphasis)

The claim that randomization is a sure-fire guarantee of a causal conclusion seems on the surface unsustainably strong. And indeed Papineau himself introduces perhaps the most obvious seeming objection:

[12] Notice that Papineau's notion of T 'fixing' a higher than average probability in some subpopulation is itself causally loaded. The best we can infer using standard statistical reasoning from what we observe is the probability of R within that subpopulation just *is* higher with T than without it. Cashing out what 'fixes' means might well drive Papineau towards the 'causal unanimity' account often ascribed to Nancy Cartwright (see Section 4). But my concern here is not to criticize any account of probabilistic causation directly, but to focus on the conditional question: if we were to accept such and such an account of probabilistic causality would it give us any reason to attach a higher value to evidence obtained from randomized trials?

> Suppose we notice, after conducting a randomized experiment, a relevant difference between the treatment and control samples[13]. For example, suppose that we notice that the experimental subjects who received treatment were on average much younger than those who did not. Common sense tells us that we shouldn't then take a difference in recovery rates to show that the treatment is efficacious. But advocates of randomized experiments, like myself, seem to be in danger of denying this obvious truth, since we claim that randomization is a sure-fire guide to causal conclusions. (*op. cit.,* pp. 446–7)

Quite so. But Papineau believes that this objection is easily dealt with:

> I agree that, if you think that age might matter to recovery, then you would be foolish to infer the efficacy of T solely from a difference in recovery rates between [these two particular randomized groups] [. . .] However, I don't think that this counts against my defence of experimental randomization. (*op. cit.,* p. 447)

We need, he suggests, constantly to bear in mind the difference between the two inferential steps: anyone (Papineau dubs him 'Quentin Quick') who makes the inference to cause from this particular set of data is, according to Papineau, automatically making a mistake at the *first* step—he has no right to the objective probabilities. *But*:

> *If* we *were* to grant Quentin his intermediate premise, that there is an underlying objective T-R correlation, then his inference to the efficacy of T would be quite impeccable. After all, if T did not cause R, how could there be such a correlation (an objective correlation in the underlying probability space, remember, which will show up, not just in this sample, but *in the long-run frequencies as the randomized experiment is done time and again*) given that the randomization will ensure that all other causes of R are probabilistically independent of T *in the long run*? (*Ibid*; my emphases)

Quentin Quick's first step is fallacious and his

> mistake is simply a variant of the case Urbach uses to argue against the classical theory [. . .] Assuming Quentin's sample was randomly generated (though remember that this is an extra assumption, over and above the random assignment of the treatment), then it was objectively unlikely that he would have found a statistically significant sample correlation[14], given the hypothesis that T and R are objectively uncorrelated. So

[13] In the clinical trials literature this is generally called a 'baseline imbalance'

[14] This is incorrectly formulated (it is important to keep in mind that we never observe correlations) but the point seems clear: many more patients in the sample are observed to have *either* both T and R *or* both ¬T and ¬R than would be expected on the assumption that $p(T/R) = p(T)$.

> classical theory advises Quentin to reject this hypothesis. But of course Quentin shouldn't reject this hypothesis on his evidence, for he can see that the freakishness of the sample is as good an explanation for the observed sample correlation as the alternative hypothesis that T and R are objectively correlated [. . .] Still, all this relates to Quentin's first inferential step [. . .] we shouldn't conclude from this that *randomization of the treatment* isn't needed for *causal inferences*, for randomization of treatment is crucial if we want to decide whether an objective correlation indicates a real causal connection (*op. cit.,* pp. 447–8).

There are a number of ways in which Papineau's analysis seems off-beam. For example, as he himself points out, no one thinks that the set of subjects in a trial can be thought of as a random sample from some specifiable population. So it is difficult to see how we can reconstruct Quentin Quick's alleged first inferential step in real cases. But even if we accept Papineau's argument on its own terms, does it really give any further *practical* reason to randomize? Peter Urbach was surely *not* conflating random sampling and randomization; and he and other Bayesians explicitly do find it difficult to see any direct role for the latter. A division of the study population into two groups may be achieved *via* an impeccably performed random experiment. But all agree that in any particular case this may produce a division which is, once we think about it, unbalanced with respect to some factor that plays a significant role in the outcome being measured. Those involved in clinical trials usually talk about this as 'checking for baseline imbalances.' Everyone agrees, as the Bayesian points out, that if there *is* a clear 'baseline imbalance' one should not proceed to draw any conclusion from the trial. The classical statistician (rather quixotically) insists that one should then re-randomize (if necessary again and again) until we see no reason to think the division unbalanced. For the Bayesian, however, the *only* issue is whether we have (essentially on the basis of 'background knowledge') some positive reason to think the control and experimental groups relevantly noncomparable. The Bayesian can, therefore, see no reason why one should not, in the first place, just match the groups with respect to factors that one does have some reason to think may be significant. If, having created groups matched with respect to those 'known' factors, one then goes on to decide which will be the experimental and which the control group by some random process—in the simplest case by tossing a fair coin—then one can do no epistemic harm, though one also does no further epistemic good.

In particular, randomizing cannot deliver us from the possibility that the two groups are—of course (by definition) unbeknown to us—relevantly different with respect to some factor that we have not yet thought about. Even when the two groups seem clearly comparable with respect to all 'known' other factors (possible alternative causes if you like), we may still

be in Quentin Quick's situation, but with respect to some factor that is not highlighted by background knowledge, that is, with respect to some 'unknown (better: unsuspected) confounder'. Indeed given that there are indefinitely many possible biases or confounding factors, it seems intuitively likely that we will be. Simply tossing the coin to decide which of the two matched groups is the experimental one, or not matching at all and it happening by serendipity that randomizing produces groups with no obvious baseline imbalance, *clearly* cannot remove this possibility. But if this possibility is actualized and there is an unknown common cause, then the conclusion that T causes R will be incorrect. Where, then, is David Papineau's 'sure-fire guarantee'?

Unsurprisingly—since Papineau's argument, after all, really amounts to the 'randomization controls for all possible confounders, known and unknown' claim newly-dressed in causalist clothing—the intuitive appeal of his argument seems to rest on the tempting but dangerous slip from consideration of what is justified on the basis of the sample we actually have (we have randomized only once!) and what might be justified if we re-randomized indefinitely (either on the same or 'equivalent' populations of subjects).

Indeed, rereading the passage from Papineau quoted above ("If we were ... in the long run?") with special emphasis on the portions I have italicized will reveal that he himself is aware of this issue. As we saw, he (incorrectly) refers to the observation of a 'correlation' between T and R in the randomized trial that has been performed (remember (again): what we observe are just relative frequencies in *that* trial); but then, in order to justify talk of a 'surefire guarantee' of a causal connection, slips into talking of 'an objective correlation in the underlying probability space', one that 'will show up, not just in this sample [really division], but *in the long-run frequencies as the randomized experiment is done time and again*.'[15] (Maybe talking of an observed 'correlation' was intended to encourage this slip, it certainly has that effect.) But, to repeat, we have only done the randomization once. The results supplied by the actually performed randomized trial can, on their own, tell us nothing about what would happen if we randomized repeatedly: the fact that T and R are 'correlated' in the trial that we perform (that is, there are (perhaps a lot) more recoverers in the experimental group than there are in the control group) can tell us nothing about what would happen if we re-randomized time and again. Once again, a view that is at least along Bayesian lines seems to underwrite what commonsense ought to tell us: that if we are to draw any general conclusion at all about the likely efficacy of T, we are forced to make an assumption about the typicality of the actual division we have before us;

[15] It is not clear to me whether Papineau means by this a repetition of the whole trial involving a fresh 'study population' or a repetition of the process of random division into the two groups of the same group of patients.

and this can only be done on the basis of whether or not we have some positive reason to think that the two groups are unrepresentative in some specific 'known' way; to think that the result of a coin toss can supply some further reason seems to be to indulge in 'probability magic'.

Papineau's argument, in other words, may show that if you randomized for ever, then the limiting-average effect could not yield more recoverers among those given the experimental therapy unless that therapy 'causes' recovery. But clinical researchers never do randomize for ever, they only do it once. There is no reason to think that any actual randomized trial reflects the 'limiting average'. Moreover, there is no sense in which we can ever know how close a particular randomized trial is to yielding this 'limiting average'. Any particular randomized trial may, therefore, (and *of course*) mislead about causes (even in Papineau's sense). It is entirely possible that any particular randomization may have produced a division into experimental and control groups that is unbalanced with respect to 'unknown' factor X, such that, although there are more recoverers in the experimental group, if we knew about factor X we would see that the frequency with which those in the placebo group recovered if they possessed X was no different (or 'not significantly' different) from the frequency of recovers in the experimental group who had factor X. Since X is by definition unknown, it seems obvious that we cannot possibly make any judgment as to whether or not this is indeed the case.

Papineau's argument therefore supplies absolutely no further *practical* reason to prefer the results of a randomized trial to, say, a historically controlled trial where there is no evidence that the newly treated group are importantly different from those forming the historical controls; and his own remarks about 'the long-run frequencies as the randomized experiment is done time and again' reveal this, even though it is at odds with what often seems to be the main thrust of his paper.

Another way of seeing that the argument supplies no practical reason at all to prefer the results of a randomized trial is through examination of the 'guarantee' that it is alleged by Papineau to supply. This is just the ineliminably conditional 'guarantee' that if, when you have randomized, you end up with a mistaken inference about causes, then you must already—though you will of course not know it—have made a mistake at the level of inferring underlying 'population' probabilities. I cannot see how a 'guarantee' of a conditional nature in which the antecedent is in principle undecidable can be thought of as amounting to anything very much or indeed to anything at all.

It is important to note here that I am not simply insisting on the trite point that it is logically possible for there to be no real causal connection between treatment T and recovery R, despite having found that there are more people with R among those given T in a randomized experiment. Instead, the point is that the results from the actual random allocation made in some particular trial

(as opposed to the results from an indefinite series of such random allocations) can give no extra reason at all for thinking that the division between an experimental and control group is not biased in some significant way. This is why no *practical* reason for randomizing is supplied by Papineau's invocation of causality. Once you have made sure that there is no positive reason to think the two groups are unbalanced (and this automatically means checking for imbalance in factors you know about), then whether or not the division was produced by following some table of random numbers or tossing a fair coin, or just by happenstance, can be of no epistemic account. This is what the Bayesian is saying, and it seems to me entirely convincing.

It should be remembered that there can be no doubt that—again as Bayesians allow—one reason (based on a 'known' possible 'confounder') for being suspicious of a particular division into experimental and control groups is if the experimenters have been allowed to make the division themselves. Standard methods of randomization do indeed rule out this possibility, which may result in 'selection bias'. But this is not a virtue of randomization as such, but rather an illustration of the fact that we should always seek to eliminate biases in 'known' factors. If, in some particular test, this bias has been eliminated by dint of randomizing, then that is obviously to the good; but if this bias can, in other particular circumstances, be eliminated by other means, then randomizing seems to be left with no further epistemic virtue.[16]

This is the main conclusion of this section of my paper. However it is useful, I think, further to explore the value (or rather lack of it) in David Papineau's 'conditional guarantee' by looking at a couple of cases taking the god's-eye-view: cases where, by supposition, 'we' know 'from the outside' what the relevant causal connections are, but where the investigators conducting the trials do *not* know these connections and are instead seeking to discover them.

Suppose, then, to take a simplified but not entirely unrealistic example, we *know* that disease D is caused by some bacterium and that antibiotic A kills the bacterium and hence cures D if, but only if, certain biochemical parameters within a particular person's body achieve certain values. Hence, taking A will cure any person P of D, exactly if P satisfies certain biochemical conditions $C_1, \ldots, C_n$. In anyone's book, in such a case, A certainly causes recovery from D for any particular person P who took A and satisfies conditions $C_1, \ldots, C_n$. Suppose however that these 'hidden variable' conditions are relatively rare in the population, while A causes nasty side effects in those who do not satisfy $C_1, \ldots, C_n$ to the extent that such people are actually made less likely to recover

[16] It should also be remembered that, as even the most noted of advocates of randomization, Doll and Peto, allow, it is intuitively unlikely that selection bias will produce a very large effect (see the discussion in Section 6 below).

from D by taking A. I do not myself believe that, in such a case, there is any answer to the question 'Does A cause recovery in the population as a whole?' (Nor need there be such an answer. The underlying causal facts as just outlined exhaust what sensibly can be said, and hence show the question to be ill-formed.) But certainly, according to Papineau's notion **C**, the answer we want is that A does cause recovery from D (because *there is* a subpopulation, consisting of those satisfying the conditions $C_1, \ldots, C_n$ within which A causes recovery). The researchers do not know about $C_1, \ldots, C_n$, but in fact, the randomization is (miraculously) 'perfect', which I suppose means (with an eye to 'external' as well as 'internal validity')[17] *both* that x% of the subjects in the trial overall have $C_1, \ldots, C_n$ (where x% is the frequency in the *target population*—which we can take to consist of those who will be treated with A should it prove to be effective), *and* exactly half of those x% go into each group in the trial.

Assuming that the investigators are not in the grip of a *false* 'background' theory that some different factor is a plausible confounder, it will follow from this, of course, that there is no overt reason for the investigators to worry about the randomization, that is, there is no 'known' factor that might plausibly be taken to play a causal role which is maldistributed in the two groups. Hence none of the investigators has motive to indulge in Quentin Quickery.

Since there are relatively few people who satisfy $C_1, \ldots, C_n$, the outcome will be that $f(R/A) < f(R/\neg A)$. So the 'conclusion' that the investigators draw will be that A causes $\neg R$. Hence randomization, allegedly a surefire guarantee of causal conclusions has, in this particular case, as a matter of practical fact, delivered what on Papineau's own account of causality is the *wrong* causal conclusion.

This might, however, plausibly be regarded as a criticism simply of his condition **C** rather than of his general claims about randomization. So now suppose we have a very similar case which again involves some disease D′, an antibiotic A′ and biochemical parameters $C'_1, \ldots, C'_m$; only now suppose that the conditions $C'_1, \ldots, C'_m$, rather than being rare are quite common in the population. I suppose that the judgment that most contributors to the literature on probabilistic causality would want to endorse in such a case is that A′ *does* cause recovery from D′ in the population at issue. Suppose, however, that considerably less than one half of the patients assembled in the trial satisfy

[17] A result is 'internally valid' if the observed result reflects the true situation with respect to the study population—that is, the set of people actually involved in the trial (many medics identify internally valid with randomized!); a result is 'externally valid' if the observed result reflects the true situation with respect to the 'target population' (roughly all those who would be considered for treatment with the treatment under test should it prove effective). It is important to note that, because the set of patients in a trial is in no sense a random sample from the 'target population' (itself only loosely characterized at best), there is not even the semblance of an argument that RCTs deliver a guarantee of 'external validity'. This is an important issue that I take up elsewhere (see, Worrall [forthcoming]).

$C'_1, \ldots, C'_m$; and moreover, by chance, almost all those that do satisfy $C'_1, \ldots, C'_m$ are assigned by the randomizer to the control group. Remember that the fact that conditions $C'_1, \ldots, C'_m$ are the vital ones is completely unknown to the experimenters in our story. Hence, they would have no reason to think in this regard that the random allocation was biased; and let's suppose that there is no factor that *is* known to those experimenters as a plausible confounder with regard to which the random allocation happens to be biased. In this case, of course, we will observe a lower rate of recovery in the experimental group, that is, in those given A′. (Remember we are supposing, as before, that A′ actually exacerbates D′ for those individuals who fail to satisfy the conditions $C'_1, \ldots, C'_m$.)

So again, since we know 'from the outside' the real causal situation, we see that the randomized trial has given the wrong answer. This is true despite the fact that there has been no 'sloppiness' or Quentin Quickery, *and* a genuine randomization. It is true, of course, that—again from the outside—we can see that a mistake has in fact already been made at the level of probabilities. But the investigators are bound to be in blissful ignorance of this, and hence this case fully emphasizes the nonpractical nature of Papineau's 'sure-fire guarantee'. Having initially withdrawn the drug A′ on the basis of this trial, only to cause, perhaps, large amounts of suffering from disease D′ that could in fact have been helped by treatment A′. These experimenters, once further research has revealed the true causal situation, will scarcely feel consoled at being told that at least the causal step in their initial two-step inference was flawless!

4 Cartwright on causality and the 'ideal' randomized experiment

Central to Nancy Cartwright's account of probabilistic causality, as to other accounts, is the problem of 'spurious correlation': A and B may be probabilistically dependent not because one causes the other but because they are both effects of a common cause C. As she points out, the 'conventional solution' to this problem is to hold C fixed:

> [T]hat is, to look in populations where C occurs in every case, or else in which it occurs in no case at all [. . .] [I]f A and B continue to be correlated in either[18] of those populations, there must be some further reason. If *all* the other causes and preventatives of B have been held fixed as well, the only remaining account of the correlation is that A itself is a cause of B (Cartwright [1989], p. 55).

[18] This should presumably read "both".

Hence, she arrives at a principle she calls 'CC' that is fundamental to her account[19]:

> 'C causes E iff
> $p(E/C \pm F_1 \pm \ldots \pm F_n) > p(E/\neg C \pm F_1 \pm \ldots \pm F_n)$, where $\{F_1, \ldots, F_n\}$ is a complete causal set for E.' (*op. cit.,* p. 56)

So this principle, in contrast to David Papineau's account, certainly *seems* to require that in order to count as a cause of E, C is required to raise E's probability in *every* cell of the partition of the population at issue into causally homogeneous subpopulations.[20] However, she emphasizes later in the book that her full account of causation makes it a *three*-place relation involving not just the putative cause C and putative effect E, but also some 'relevant population'. Hence principle CC is to be read as implicitly relativised to some underlying 'relevant' (sub)population: C causes E, relative to population S, if $p(E/C) > p(E/\neg C)$ in every cell of the partition of C into causally homogeneous subgroups, though C may well *not* thus cause E relative to some other part S' of what might be thought of as the 'overall population'. This would seem to take the steam out of what John Dupré, among others, saw as a debate between Cartwright and Papineau (and others such as Sober who favoured an 'average' impact account).

Relativized to a subpopulation or not, there are a number of more or less obvious difficulties with CC, several of which Cartwright addresses herself. I shall not, however, here attempt a detailed analysis of the principle, but, just as I did when analysing Papineau's argument, will concentrate on the purely conditional issue: supposing that this were the correct account of causality, would it give us any reason to randomize in clinical trials? As we shall see, Cartwright seems to claim that indeed it does, since principle CC and randomization 'dovetail'—a randomized experiment and principle CC are 'bound to agree', *provided*, that is, that the experiment is 'ideal' (and therein will lie much of the—unsurprising—rub).

The most obvious problem with principle CC is its apparent circularity. Cartwright explains that 'To be a complete causal set for E means, roughly, to include all of E's causes.'[21] But this seems circular since the notion of cause appears both on the left and on the right hand sides of CC. These

[19] She does, however, as we shall note later, eventually reject this principle in favour of a modified version. (I have changed her capital 'P's for probabilities to lower case 'p's in order to maintain consistency of notation within my paper.)

[20] It is true that no universal quantifier occurs explicitly in her statement of Principle, CC, but the prefatory remarks ('If *all* the other causes and preventatives of B have been fixed as well, the only remaining account is [. . .] that A [. . .] is a cause of B' (her emphasis)) strongly suggest it.

[21] As she makes clear later, she really of course means here 'all the other [potential] causes of E *aside* from C.' (*op. cit.,* p. 79)

days it is fashionable to claim that 'reductive definitions' (aka definitions!) are an impossible ideal and that there is nothing wrong with 'externalist' characterizations that tell us what it would be for something to count as, for example, a cause without giving us any method for deciding whether any particular putative cause is indeed a cause. (Just as Tarski's characterization of what it takes for a sentence to be true—it is true if it corresponds with the facts—gives us no indication as to how to go about deciding whether or not this correspondence holds.) But Nancy Cartwright as a self-avowed empiricist laudably wants her characterization of cause to have at least some methodological import, to supply at least some guidance as to how we might decide whether a particular putative cause is indeed a cause. And certainly, so far as RCTs are concerned, medics have eminently practical issues in mind. They would hardly be very impressed by a philosopher who told them 'I can tell you *what it would be* for, say, regular vitamin C intake to cause early recovery from colds; but of course I cannot tell you anything at all about whether or not regular vitamin C does indeed cause early recovery (or whether there is telling evidence that this is true)'.

Seeking, then, to base a *method* for identifying causes on principle CC, Cartwright squarely faces up to the apparent major problem:

> It seems that a method [for applying CC to 'discover', causes] that requires that you know all the other causes of a given effect before you can establish any one of them[22] is no method at all. (*op.cit.,* p. 62)[23]

But what should ride to the rescue here other than the randomized experiment?

> . . . the method does not literally require one to know all the other causes. Rather what you must know are some facts about what the probabilities are in populations that are homogeneous with respect to all these other causes, and that you can sometimes find out without first having to know

[22] This last phrase would more accurately read "before you can establish that some other possible cause is also in the complete set of actual causes."

[23] As my interpolations in this quotation indicates, it is important, even when attempting to forge a link between the metaphysical account of causation and the methodology of testing for it, to be aware of the inevitable differences between them. Suppose that it is correct that what it means for C to cause E, is that C raises E's probability whenever all the other *actual* causes of E are kept fixed at whatever values we choose. Still, when it comes to methodology, we would want to 'control' for factors that may not in fact be causes, but which 'background knowledge' makes it plausible *might be*. Suppose, for example, that as a matter of fact (forget how we come to 'know' it), age is irrelevant to recovery time from some disease whether or not you are given some therapy. Age, then, is not a cause of recovery time and it makes no difference so far as the objective issue is concerned of whether some therapy causes reduced recovery time, or what happens when we partition our groups into young and old (thinking for simplicity of age as a binary property). However, methodologically speaking, we would certainly want to eliminate this possibility and hence look at what happens when we effect this partition.

> what all those causes are. That is the point of the randomized experiment . . . (*ibid.*)

As Cartwright goes on to make clear, however, her claim is that a randomized experiment will perform its magic *only if* it is 'ideal'. An experiment involving treatment and control groups must satisfy 'two related conditions' if it is to count as 'ideal'. The first is that 'all the other causes that bear on the effect in question must have the same probability distribution in both groups' (p. 64). The second is that 'the assignment of individuals to either the treatment or the control group should be statistically independent of all other causally relevant features that an individual has or will come to have.' (*ibid.*)

Using a randomizing device to divide the subjects into the two groups 'is [. . .] a help to both ends', and then other 'clever and well-known devices'—such as double-blinding—also kick in to 'ensure that the results of the real experiment will be as close as possible to the results of an ideal experiment.' (*ibid.*)

There are, however, several difficulties here. Consider the first condition that an experiment is supposedly to satisfy if it is to count as 'ideal': that 'all the other causes that bear on the effect in question must have the same probability distribution in both groups'. To apply the notion of probability to the result of a *single* experiment—even an 'ideal' one!—is, at least on the frequency interpretation, a category mistake. Probabilities are defined not on single, but only on *repeatable*, events. It is not merely that, as Cartwright herself allows, experimental results are about relative frequencies, not probabilities (see, e.g., *op. cit.*, pp. 65–6), but moreover that probabilities are not defined at all relative to single experiments.

It seems reasonable to conjecture that what Nancy Cartwright really intends is that the 'ideal' experiment should involve a division between an experimental and control group that has the following characteristic: where the number of patients in the study group as a whole satisfying potentially causally relevant characteristic F_i is f_i, then, for all i, the number of those satisfying F_i in the experimental group *equals* the number of those satisfying F_i in the experimental group *equals* (exactly?) $f_i/2$. One might even, with a view to the issue of 'external validity'—that is, projecting the result from the study population to the 'target population'—like to have a 'super-ideal' case in which those frequencies in the study population exactly equal those in the target population. But given the way that the study group is standardly constructed—by a mixture of happenstance and design and not as in any sense a random sample from the target population—that would surely and palpably be asking for too much.

It is important to note, however, that the real instantiation of even the 'ordinary' 'ideal' experiment requires a miracle that casts the 'case' of the loaves and fishes into the shade. Given that we are talking, as Cartwright admits following Fisher, of 'innumerable' unknown (possible) causes, it would

clearly be a miracle if *all* of those factors just happened to be balanced in the two groups (even if we adopt a slacker notion of balance than outright equality) on a single random division.

As for Cartwright's talk of 'ensur[ing] that the results of the real experiment will be as close as possible to the results of an ideal experiment' (p. 64), this seems to amount to an empty tease. We can of course balance (or attempt to balance) *known* factors, as trialists do in the case of patients' and clinicians' expectations via double-blinding. It is, indeed, in this precise connection that Cartwright talks of making the result of the real experiment 'as close as possible' to the ideal result, but her remark sounds as if she holds that we can have at least some intuitive measure of how close a real experiment is to one that is ideal in respect of *all* 'other causes' known *or unknown.* But clearly there can be no estimate of how closely balanced a particular real trial is with respect to any unknown factor—this is so by definition, since the unknown factor is unknown!

Just as in the case of Papineau (indeed, despite starting from what seems to be different accounts of probabilistic causality, Cartwright's and Papineau's arguments—and hence the objections to them—are remarkably similar), the best that *might* be argued is that if we were to take the study population and divide it again and again by some randomizing device into control and experimental groups and keep a cumulative total of the relative outcomes in the two groups, then we would expect that in the indefinite long run, the innumerable other possible causal factors would balance out and the limiting cumulative relative outcome would reflect the true efficacy of the treatment.

Aside from the fact that we are trivially never in the limit, the idea of repeating a random division on the same group of patients even a large number of times is, of course, practically speaking, impossible. This is in part for ethical reasons, but there is also an epistemological issue about whether any repeated random trial would be comparable to the initial one. If a particular patient in the study receives, say, the 'active drug' on the first round, then, since this is expected to have some effect on his or her condition, the second randomization would not be rigorously a true repetition of the first. The second trial population, though consisting of the same individuals, would, in a possibly epistemically significant sense, not be the same population as took part in the initial trial. One might be ready to idealize this effect away in the case of palliative treatments for long-term, chronic conditions, though even then, due 'wash out times' would have to be allowed. But in any event, no one of course, seriously contemplates this possibility: randomized trials are performed once (though another trial may of course be performed later on the same treatment but invariably with a different group of subjects). Not only is there no reason to think that the results of a single randomized experiment, no matter how perfect the randomization (in that the coin was really fair, or, in practice, the table

of random numbers was used appropriately), is 'ideal', we have absolutely no access to a rational estimate of how far it differs from the 'ideal' with respect to unknown factors, and that was supposed to be exactly the point.

Nancy Cartwright's account of probabilistic causality, just like David Papineau's, supplies no valid reason to think of (actually performed, as opposed to 'ideal') randomized experiments as especially epistemically powerful or privileged.

Funnily enough (and again in striking similarity with David Papineau), I think that Nancy Cartwright may in the end agree with this. She ends the section of her (1989) book on randomized experiments as follows:

> The point of discussing [randomized experiments] here is to recall that the demand for total information that can seem to follow from CC is not necessarily fatal. Sometimes we can find out what would happen were all the other causes held fixed without even knowing what the factors are that should be held fixed. It is important to keep in mind, however, that it takes an ideal experiment to do this, and not a real one. For, as with Principle CC itself, the connection between causality and regularity is drawn already well above the level of real data and actual experiment. It is not frequencies that yield causes, but probabilities; and it is not results in real experiments, where subjects are assigned to groups by a table of random numbers, but rather in ideal experiments where randomization is actually achieved. (pp. 65–6)

I say only that in the end she *may* agree with me, because the first two sentences of this passage do not seem to me to cohere with the final three. Since no real experiment is ideal and since we have no way of telling how near to the ideal a real experiment is (except for known 'alternative causes' for which we anyway could deliberately control), how can we then take ourselves to 'find out what would happen were all the other causes held fixed without even knowing what the factors are that should be held fixed' by performing a randomized experiment?

In any event, it seems safe to conclude that *either* Nancy Cartwright agrees with me that her analysis of probabilistic causality gives no further practical reason to insist on randomized experiments as especially telling from an epistemic point of view *or* if she *is* claiming that her analysis does supply such a reason, then she has provided no sustainable justification for that claim.

I should add, that Cartwright later in her [1989] book argues that her principle CC is in fact in need of modification to deal with what she sees as cases of 'contrary capacities' and 'interactions'. There are again many issues here. In particular I think that the talk of 'contrary capacities' is based on a failure to understand the underlying physiology in cases like the often-discussed example (see Hesslow [1976]) of the contraceptive pill which, it is alleged, both causes thrombosis and prevents it (by inhibiting pregnancy which

is itself a cause of thrombosis). But, again, there is no need to go into these for present purposes. Her view is essentially that, because they 'dovetail', the 'ideal' randomized experiment yields the truth about causality just in cases where CC applies. In other cases, where 'contrary capacities' or 'interactions' are involved, even the ideal randomized experiment will not give the correct answers, although the modified version of principle CC (CC*) does give the correct answers. Hence, insofar as she argues that randomization plays a significant role at all, it is restricted to the simpler cases where CC as a matter of fact is, for her, true; and so the later modifications are not relevant to the issue discussed above.

5 Pearl on randomization, nets and causes[24]

The technically most sophisticated and formally elaborate account of causality currently available (along with the earlier, cognate account of Glymour, Scheines, and Spirtes)[25] is the one developed by Pearl ([2000]). Pearl explicitly argues that his account of causation 'provides a meaningful and formal rationale for the universally accepted procedure of randomized trials' (*op. cit.*, p. 348). However, when Pearl's views are analyzed more carefully, it once again becomes unclear just what this endorsement of the RCT methodology really amounts to. I first outline Pearl's account of causal nets, then report his argument for why this account is supposed to provide a rationale for the 'universally accepted procedure of randomized trials', and finally analyze and criticize that argument.

Pearl's account of causality developed out of his earlier work on Bayesian networks. Formally speaking, such a network consists of a finite number of nodes, occupied by variables X_n, some of which are connected by edges, some edges being directed (in which case they are arrows). If an arrow connects X_i to X_j then X_i is called a *parent* of X_j, while X_j is a *child* of X_i. (Naturally then X_i's children, children of children, *etc*., are called the *descendants* of X_i.) A joint probability distribution $p(X_1, X_2, \ldots, X_n)$ is also taken to be defined on all the variables in the network; and this probability distribution must cohere with the structure produced by the arrows in the sense that if, for example, node X_1, forms with X_2 and X_3 a 'conjunctive fork' in Reichenbach's sense (i.e., if there are arrows going from X_1 to both X_2 and X_3 but no other relevant arrows) then X_2 and X_3 must be probabilistically independent conditional on X_1: X_1 'screens off' X_2 from X_3. More generally, the probability distribution and network connections must together satisfy the 'Markov condition': that a

[24] I am especially indebted in this section to Jon Williamson for his patient help in increasing my understanding of some of the details of Pearl's position.

[25] See (Spirtes *et al*. [1993]) and subsequent literature from the group.

variable X is conditionally independent, given its parents, of any set of other variables that are not descendants of X.

One main project within Pearl's programme is the development of algorithms that will produce such a network (or more accurately, a class of such networks) from purely probabilistic 'data'. (Talk about 'data' here is of course, as Pearl admits *op. cit.*, (p. 45)[26], an idealization, and one that will be consequential when it comes to analysis of his approach to randomized trials just as it was in the case of Papineau and Cartwright. To be clear (if trite and repetitive), we observe finite relative frequencies *not* probability distributions.) However, although the networks that are produced by Pearl's algorithms certainly have a causal air about them, and although such a network is bound by construction to satisfy certain broad constraints of a generally causal nature—in particular the Markov condition—it is not yet *guaranteed* to be a *causal* net in Pearl's sense.

To see why, let us take Pearl's own initial example. Affluent southern Californians have sprinklers for their front lawns that they set to come on automatically for specified periods in the dry seasons but not during the wetter seasons. Whether or not a given southern Californian sprinkler is on, then, at any particular time, is probabilistically dependent on what the season currently is. Moreover, the chance of rain in California is also dependent on the season. If, in Pasadena or wherever, it either rains or the sprinkler is on, then usually the pavement will get wet (but not invariably; the rain may be light, the wind may unusually direct all the sprinkler water away from the pavement, there may unusally be a cover on the pavement, etc.); and finally, depending on how wet it gets and also on earlier conditions, the pavement may or may not become slippery.

Letting X_1 be the 'season variable' (this can of course by convention take on any one of four values: spring, summer, autumn, winter) and X_2, X_3, X_4 and X_5 binary variables representing the state of the rain (yes or no), the state of the sprinkler (on or off), the state of the pavement (wet or dry) and the slipperiness of the pavement (yes or no), then, on Pearl's account, background knowledge or, in his own words, 'causal intuition', sanctions a certain set of interconnections between these variables (see Pearl's figure 1.2).

A lot could be said about this. For example, it is not clear that it is sensible to talk about the seasons' 'causing' rain or even being *directly* causally connected with rain. But let this pass. Assuming further that it is sensible to model this whole setup probabilistically in the way Pearl suggests, then this network will be underwritten not just by 'causal intuition' but also by certain relationships between the *conditional probabilities* that each of the variables possess a

[26] Pearl there admits that his account of how to infer causal structure 'invokes several idealizations of the actual task of scientific discovery. It assumes, for example, that the scientist obtains the [probability] distribution directly, rather than events sampled from the distribution.'

particular value (spring, yes (rain), etc.), given values of other variables. For example, the absence in Pearl's graph of a direct connection between the season, X_1, and whether or not the pavement is slippery, X_5, is reflected in the fact that if we already know that the pavement is wet (or dry) we already know the probability that the slipperiness variable takes on the value it does, independently of any of the values of the other variables. That is, $p(x_5/x_4) = p(x_5/x_1, x_2, x_3, x_4)$, where lower case letters x_i stand for particular values of the variables X_i. In general, this network satisfies the Markov condition and might indeed have arisen as the output, given suitable inputs, of Pearl's algorithm.

However, even these sophisticated probabilistic constraints do not, on Pearl's account, exhaust the causal character of the connections portrayed—they do not *fully* flesh out the 'causal intuition' that underwrites the structure of the network. Where no such clear-cut 'causal intuition' operates and Pearl's algorithms are simply being used to *generate* networks from probabilistic 'data', the class of inferred networks will all be guaranteed to reflect the relevant probabilistic dependencies and independencies, but are *not* yet guaranteed to be causal.

The extra factor that is present when the network is genuinely causal, on Pearl's account, is centrally to do with *intervention*. Within his framework this is characterized as involving 'surgery' on one of his networks, which in turn means deleting some arrow, and fixing the value of the variable that had been at the tip of that arrow at some 'freely chosen' level. In the simple case we have been discussing, for example, we might intervene to set the sprinkler manually to 'on': this means that the 'normal' way in which the sprinkler variable is affected by the season is eliminated, hence the arrow from the season variable to the sprinkler is deleted, and the value of that sprinkler variable is set at some value of our choosing. Pearl calls the resulting network a 'mutilated' one—see his figure 1.4.

Finally and crucially, you know that you had a genuinely causal network on your hands (before the mutilation) if *the mutilated graph still, in some sense, continues to perform as previously advertised* and hence if you can still make predictions on its basis.

The sort of thing that this can mean is seen most clearly in the entirely deterministic, nonprobabilistic cases that Pearl discusses in the Epilogue of his book. Suppose that we have a machine that consists of two parts, a multiplier (it doubles any input) and an adder (it adds one to any input). 'Normally' some input X comes in from outside and is doubled by the multiplier and then passed (as variable Y) to the adder where one unit is added to it, producing the outcome Z. (Clearly then, in the 'normal' operation, if the outcome is z, then $z = y + 1 = 2x + 1$, where x was the input.) Moreover this is, at least in a stretched (!) sense, intuitively a causal setup: we might think of the above as

the abstract description of a two-component physical machine that takes any metal rod as input, and first stretches it to twice its initial length and then, in its separate second subcomponent stretches the rod further to add 1 meter to its length. If instead of putting a rod of length x into the 'front' of this machine (which would mean that its length y as it emerged from the first component of the machine was beyond our control, since y will then inevitably be 2x), we circumvent that first component and simply decide that the variable Y will have the value y (that is, of course, we decide to introduce a rod of length y directly into the second component), then that second component will operate exactly as it would have done before: it still adds one meter to the length of the introduced rod. The two subcomponents are, in other words, entirely autonomous. The hallmark, then, in Pearl's view, of a truly causal system such as this one is that, having performed an intervention (or surgery) on the system (in this simple example we have 'deleted' the connection between the 'normal' input X and the intermediate variable Y), we can still predict the outcome: having intervened to set the intermediate variable Y at the value y we can predict that the output will be $z = y + 1$.

A similar, but somewhat more complex notion, applies to the probabilistic case. Given the dependencies and independencies indicated by the structure of the graph, the joint probability distribution governing the 'natural' sprinkler set up must satisfy the following equation:

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2/x_1)p(x_3/x_1)p(x_4/x_2, x_3)p(x_5/x_4) \qquad (1)$$

The modified joint distribution of the remaining variables once X_3 has been deliberately set to 'on' (representing, remember, our 'autonomous decision' to turn the sprinkler on) will be

$$p_{X_3=\text{on}}(x_1, x_2, x_4, x_5) = p(x_1)p(x_2/x_1)p(x_4/x_2, x_3 = \text{on})p(x_5/x_4) \qquad (2)$$

Where the ps on the right hand side of this equation (2) are exactly the same as those in equation (1); and where, as Pearl puts it:[27]

> The deletion of the factor $p(x_3/x_1)$ represents the understanding that, whatever relationship existed between seasons and sprinklers prior to the action, that relationship is no longer in effect while we perform the action. Once we physically turn the sprinkler on and keep it on, a new mechanism (in which the season has no say) determines the state of the sprinkler. (*op. cit.,* p. 23)

Once again, in this probabilistic case, the hallmark of the network being truly causal is that we can predict what will happen when we make certain

[27] Again I have changed his notation for probabilities to ensure internal consistency within the present paper.

interventions on it. But what we predict in this case are not particular outcomes, given particular inputs, but rather, the new probabilistic structure, given the initial probabilistic structure. As Pearl strikingly (and also surely strictly incorrectly) puts it:

> . . . *intervention amounts to a surgery on equations* (guided by a diagram) and *causation means predicting the consequences of such a surgery*.[28] (*op. cit.,* p. 347)

Elsewhere (*op. cit.,* p. 345), he asserts that 'the very essence of causation' is 'the ability to predict the consequences of abnormal eventualities and new manipulations'.

Finally then X is (justifiably regarded as) a cause of, or more properly, as exerting a causal influence on, Y if there is a directed edge from X to Y in a network that has (i) been inferred from probabilistic 'data' in accord with Pearl's algorithm and (ii) is genuinely causal in the sense just explained. Though Pearl also seems to presume that, whenever the probabilistic 'data' is good and complete enough, it is *likely* that a network inferred in accordance with his algorithm will in fact be genuinely causal.

A number of criticisms could be (and have been) raised against Pearl's account of causation, but rather than launch into any general examination of its virtues and vices I want here to concentrate (as in the previous two sections) exclusively on the *conditional* issue of the strength of the case he makes, on the basis of his account, for the special epistemic power of randomization.

As noted, Pearl claims that his account of causation provides 'a meaningful and formal rationale' for performing trials according to the RCT protocol, a procedure he describes as 'universally accepted'. The only place where Pearl systematically develops this alleged rationale for RCTs is—tellingly—*not* in the main body of his book, but in its semi-informal 'Epilogue'. Let me quote the argument in full:

> Why do we prefer controlled experiment over uncontrolled studies? Assume we wish to study the effect of some drug treatment on recovery of patients suffering from a given disorder. The mechanism governing the behavior of each patient is similar in structure to the circuit diagram[29]. Recovery is a function of both the treatment and other factors, such as socioeconomic conditions, life style, diet, age, et cetera. [. . .][30] Under uncontrolled conditions, the choice of treatment is up to the patients and

[28] I take it that we'd all agree that what he really intends is not that this is what causation *means* but rather what the true hallmark of causality is (of course, on his view).

[29] This is a diagram that Pearl considers earlier in his 'Epilogue' and is essentially a description of the two-component deterministic machine that I discussed above.

[30] In his diagram 34, p. 347, he collapses all these extra factors into one, just called 'socio-economic conditions'.

> may depend on the patients' socioeconomic backgrounds. This creates a problem, because we can't tell if changes in recovery rates are due to treatment or to those background factors. What we wish to do is compare patients of like backgrounds, and that is precisely what Fisher's *randomized experiment* accomplishes. How? It actually consists of two parts, randomization and intervention.
>
> Intervention means that we change the natural behavior of the individual: we separate subjects into two groups, called treatment and control, and we convince the subjects to obey the experimental policy. We assign treatment to some patients who, under normal circumstances, will not seek treatment, and we give placebo to patients who otherwise would receive treatment. That, in our new vocabulary, means *surgery*—we are severing one functional link and replacing it with another. Fisher's great insight was that connecting the new link to a random coin flip *guarantees* that the link we wish to break is actually broken. The reason is that a random coin is assumed to be unaffected by anything we can measure on a macroscopic level—including of course, a patient's socioeconomic background (*op. cit.,* p. 348).

Let me say immediately in response that I am not, of course, against *controlled* studies! That is just scientific method. We are looking for evidence for our hypotheses from genuine tests of them and that means especially from evidence that if favourable, is entailed (or made highly likely) by the hypothesis, and at the same time *tells against* plausible rival hypotheses. Clearly, if we think that the recorded improvement in the health of particular patients who are taking drug D might in fact be due, in whole or in part, to their superior socioeconomic background (or rather to factors such as superior diet or living conditions generally associated with that background), then no one would deny that it is a good idea in studying the effect of the drug to control for socioeconomic background. The only point at issue is Pearl's immediate identification of 'controlled' with '*randomized* controlled'.

Pearl claims, remember, that his account of causation provides 'a formal and meaningful rationale' for RCTs. But surely Pearl cannot, consistently with his overall account, be claiming that having performed an RCT is *necessary* for a legitimately inferred conclusion of causation in his sense. After all, thinking back to his sprinkler example, he is entirely happy to infer that the net involved there is a causal one from the fact that we can predict the change in other probability distributions once *we decide* to turn the sprinkler on. We know that we have broken the initial causal link between the seasons and whether or not the sprinkler is on, because we deliberately set it to 'on'—no suggestion of course that in order to know for sure that the intervention or 'surgery' has been made we needed to toss a coin in order to decide whether or not to turn it on. Indeed, if we are thinking about repeating the sprinkler experiment very many times, as we need to in order to generate the probabilities, then simply

deciding to set the sprinkler to 'on' in all cases is an altogether surer way of 'severing the link between the season and the [state of the] sprinkler' than leaving it to the tosses of a coin. It could, of course, easily be the case that in any actual series of tosses the 'on' face (heads, say) as a matter of fact came up more in the drier periods of the year!

Indeed, Pearl devotes a good deal of effort in the main technical section of his book to showing how his method can be extended to deal in general with possible confounders, via introduction of intervening variables, in a way that sidesteps entirely any recourse to randomization (or indeed to intervention!). This is the way, for example, in which he seeks to differentiate the evidential bases of (i) Fisher's famous suggestion that there might be a common (genetic) cause for both cigarette smoking and cancer and (ii) the causal hypothesis that we all accept: that cigarette smoking itself causes cancer.

Even in Pearl's explicit argument only a *subsidiary* role is claimed for randomization. Pearl sees the RCT method as involving two parts: intervention and randomization. Now in fact, as my brief account indicates, for good or ill only intervention really matters according to Pearl's general account of causation, and randomization is at best simply a method of assuring ourselves that we have intervened effectively—that the path in the network that we wanted to sever really has been severed. As Pearl himself, remember, puts it:

> Fisher's great insight was that connecting the new link to a random coin flip *guarantees* that the link we wish to break is actually broken. The reason is that a random coin is assumed to be unaffected by anything we can measure on a macroscopic level—including of course, a patient's socioeconomic background. (*op. cit.,* p. 348)

But as we already saw, and as Pearl elsewhere happily allows, randomization cannot be necessary for such a 'guarantee': we can be guaranteed to have severed an erstwhile connection without having randomized. And indeed, in the case he considers here, in extolling the virtues of Fisher's idea, we again clearly could effectively be sure (or as sure as we are ever going to be) that we had broken any possible link between socioeconomic conditions and therapeutic outcome if, having identified those aspects of such conditions that background knowledge gave us reason to believe might play a causal role—such as diet, hygienic living conditions, likelihood of sticking to the treatment regime, *etc.*,—we deliberately matched for these factors in the experimental and control groups of our trial. (And indeed still better, and with an eye to 'external validity', deliberately matched at the same levels as are found in the general 'target' population.)

Ah! will come the cry, 'what about the *unknown* possible confounders?' We clearly do not know for sure, no matter how long our list, that we have ever

exhausted all the possible aspects of a person's socioeconomic situation that *might* play a role in outcome; and even if we had, there are other possible 'confounders' that are not linked to a socioeconomic situation. If we knew *all* the possible confounders, then in epistemological principle at least (though the practical difficulties might be immense), the most telling trial might arguably be the completely deliberately matched one. But once we acknowledge that there will always be possible confounders that we do not know about and therefore have not yet matched for, then it will just be a matter of happenstance if our deliberately matched groups are matched for this further factor: we certainly will not know if they are, and maybe in fact they are, or maybe in fact they are not. Does not randomization somehow or other guarantee (or perhaps, much more plausibly, provide the nearest thing that we can have to a guarantee) that *any possible* links to therapeutic outcome, aside from the link to treatment with the drug concerned, are broken?

Although he does not explicitly make this claim, and although there are issues about how well it sits with his own technical programme, this seems to me the only way in which Pearl could, in the end, ground his argument for randomizing. Notice, *first,* however, that even if the claim works then it would provide a justification, on the basis of his account of cause, only for randomizing *after* we have deliberately matched for known possible confounders. If what Pearl's causal inferences about the effect of treatment need is an assurance that we have severed the link with all other possible factors (possible 'confounders'), then we do this much more surely by deliberately matching with respect to known factors and then randomizing in the *hope* of dealing with the unknown factors. Once it is accepted that for any real randomized allocation known factors might be unbalanced—and more sensible defenders of randomization do accept this (though curiously, as we saw earlier, they recommend rerandomizing until the known factors *are* balanced rather than deliberately balancing them!)—then it seems difficult to deny that a properly matched experimental and control group is better, so far as preventing known confounders from producing a misleading outcome, than leaving it to the happenstance of the tosses. And *secondly*, the claim is equivalent to those I already analyzed and rejected when looking at Papineau's and Cartwright's defences of randomization.

Does this claim of control for known *and unknown* confounders fare any better within Pearl's framework? Well, let us stick firmly to practice here and not implicitly switch over—as, we noted earlier, it is tempting to do—to considerations of what might happen in the indefinite long run. Once we have admitted that a real single actually performed random allocation may well produce a division between experimental and control groups in which some known possible further causal factor is unbalanced, and hence in which, in Pearl's terms, the link between this other factor and outcome is not in fact

severed, then we cannot but admit that this *may* happen with *unknown* factors too. The random allocation *may* 'sever the link' with this unknown factor or it *may not* (since we are talking about an unknown factor, then, by definition, we will not and cannot know which). Pearl's claim that Fisher's method '*guarantees*' that the link with the possible confounder is broken is then, in *practical* terms, pure bluster.

But then, no one could seriously expect a literal guarantee, could they? Pearl surely should be taken as meaning that randomization somehow provides a '*probabilistic* guarantee' that the link with all factors, known and unknown, will be broken; and hence randomization will 'probabilistically guarantee' the inference that the treatment positively affects outcome (assuming more positive outcomes are observed in the experimental group). But what exactly could this mean? There are two main accounts of probability that might be applied here to give an answer: the frequentist and the Bayesian.

As always with Bayesianism, there are a variety of positions on offer (the phrase '*the* Bayesian account' always makes me smile), but the most straightforward one articulated, for example, by Savage (who later however, for reasons it seems difficult fully to understand, decided it was 'naïve') and Lindley, as we in effect noted earlier, sees no role for randomization here at all.[31] The basic argument is, I think, that the sensible person goes on the evidence that she has and can give no role to how that evidence was generated (or what other evidence she *might have* considered, but is not in fact considering).[32] If that person, in the case of a clinical trial, has no reason to think that the two groups are unbalanced with respect to a factor that she has reason to think might affect the outcome, then the fact that these groups were created by the toss of a coin rather than deliberately or by mere happenstance (or some combination thereof) can have no reasonable effect on the inference she makes about the relationship between treatment and outcome; and if she has done a systematic study of the ways in which the two groups might be significantly different ahead of receiving treatment or placebo (a study

[31] For the references to the different accounts of randomization in the Bayesian literature, and for the most sophisticated current Bayesian position, see again (Kadane and Seidenfeld [1990]).

[32] Except, of course, insofar as you have positive reason to think that the method of making the division resulted in unrepresentative groups in the particular case before you. Thus in trials using rats, for example, there is of course a perfectly clear Bayesian rationale for distrusting the outcome as a reflection of the true efficacy of the treatment if the experimental group is created by opening the door of a cage and closing it once half of the rats have got out of that cage. Not knowing that they were 'escaping' into an equally miserable cage, it seems reasonable to suspect that the escapees were predominantly the lean, mean macho rats, leading, in this case, to a likely overestimate of the effectiveness of treatment (or, if 'treatment' involves administering a potentially noxious agent to see if some awful illness is produced, to an underestimate of the agent's effectiveness). But in such a case again there is a reason (provided by the method of division) to suspect the *actual* evidence; potential evidence is again irrelevant.

effectively done for her by full and adequate matching) then she can have no better reason to believe that the groups are balanced.

This is one area in which Savage's claim that Bayesianism is just an articulation of commonsense—a claim that, as Colin Howson pointed out to me, mimics a much earlier one by Laplace—seems to me correct; and it is not immediately easy to see why Savage came to regard this attitude towards randomization as 'naïve'. The most sophisticated attempt to alleviate this interpretative difficulty and show why randomization might after all be given a Bayesian justification, that by Jay Kadane and Teddy Seidenfeld (*op. cit.*): (i) distinguishes between 'experiments to learn' (for yourself) and 'experiments to prove' (to someone else); (ii) concludes firmly (and convincingly) that the 'naïve' attitude—that is, that there is no Bayesian rationale for randomization—remains the justified view with respect to the former type of experiment; (iii) argues that there might be some reasons for randomizing when you are trying to convince someone else (basically you need to convince your readers that you have not rigged the experimental/control division in favour of the outcome you want); but finally (iv) argues that even then, there are nonrandomized designs that would do the job at least equally well (and also at smaller potential ethical cost).

Even from this more sophisticated Bayesian point of view, there is no serious sense to be made of the claim that randomization makes it probable that the trial has controlled for all unknown possibly confounding factors.

How does this alleged probabilistic guarantee look on a frequentist reading of probability? What sense can be made then of the claim that it is at least probable that any unknown possible confounder is balanced between experimental and control groups if, but only if, the division was made randomly? This is exactly the question that was raised by Papineau's and Cartwright's arguments and the answer, of course, remains the same: only so far as I can see that, *were we* to take the same population of subjects that we have randomly divided, and then randomized again, and then again . . . and so on, indefinitely, recording the cumulative mean responses in the experimental and control groups as we went along, then in the indefinite long run, the limiting frequency averages would reflect the real effect of the treatment, since in that limit that confounder must, with probability one, be balanced on average between the two groups, and balanced at the population frequency.

Even then, the 'population' in 'population frequency' here refers to the *experimental* (or study) population rather than the 'target population', that is, to the group of people who happened to have been recruited to the trial, rather than to the overall group of people it is thought might be treated with the therapy at issue. In other words, even in the indefinite long run we would have a guarantee only of 'internal' rather than 'external validity'. Moreover, as Lindley pointed out, even if this was convincing for the case of a single

confounder, it is not at all clear that the argument works even on its own terms when we take into account the fact that there are indefinitely many possible confounders. (Clearly 'it is probable that the groups are unbiased with respect to any particular possible confounder C' does not entail that 'it is probable that the groups are unbiased with respect to all possible confounders'.[33])

But let us concentrate on just the argument that this analysis makes good on the claim that in a randomized experiment, any particular possible confounder is probably balanced. Is this any source of justified consolation for the advocate of randomization? I cannot see it myself. The fact is that the subjects have been randomized between control and experimental group only once, and that division either is or is not balanced for the unknown factor at issue. Suppose it is unbalanced, and that this throws the conclusion about the efficacy of the treatment off, then it seems to me scant consolation to be told that—although you don't and can't know it,—you were 'unlucky', and if the randomization had been repeated indefinitely you would, in the indefinite long run, have inevitably realized your mistake. I know perfectly sensible people who do find consolation knowing that the *expectation* is that the groups will be balanced for a particular unknown confounder, but as I say, I just cannot see it.

It seems difficult, then, to avoid the conclusion that, like David Papineau, Judea Pearl has, *via* his argument for RCTs, provided no *practical* reason for randomizing or for automatically giving special weight to the results of trials that have been randomized.

6 Conclusion

So where then are we left with regard to the claim that RCTs possess special epistemic power?

No one should, of course, be against randomization in all circumstances. The idea that we should randomize is motivated by the patently laudable desire to be as scientific as possible in our approach to evidence. This means that we should attempt to 'control' the trial, essentially with the aim of attempting to rule out other explanations for any observed difference in outcome among those in the experimental group, alternative to the explanation that this is due to the effects of the therapy under test. Usually randomization can do no epistemological harm, *assuming* that 'known' factors (that is, ones that background knowledge tells us may well, where present, have a positive effect) have been balanced in the two trial groups, either deliberately ahead of time, or by checking for 'baseline imbalances' and rerandomizing whenever these have occurred.

[33] See Urbach's treatment in (Howson and Urbach [1993]) for references to, and development of, Lindley's argument.

Randomization *may* also do some uncontentious epistemological good, by controlling for a known possible confounder: 'selection bias'. Notice again, however, that it achieves this, not through any mystical power of the coin toss or random number table, but rather by preventing the experimenters from knowing in advance to which arm of the experiment a particular subject is assigned (if, for example, the experimenter were allowed to disqualify subjects from the trial *after* the toss had been made for them, then there would be strong reason to suspect selection bias even though it would still be true that each person who actually ended up in the experimental group did so because the coin landed, say, heads, while each person who ended up in the control group did so because the coin landed tails).

Just as there should be no automatic judgment against randomizing, so there should, equally obviously, be no automatic acceptance of the results of historically controlled trials (or observational studies). The issue, as always, from the more basic epistemological point of view, is the 'comparability' of the experimental and control groups; and clearly there is in some intuitive sense more leeway for noncomparability in the case of an observational study. There may, for example, have been some change in the natural history of the disease—the overall general health of the population could have changed, there may have been some marked change in the demographics of the catchment area of the hospital where the study was done—all might lead, in the event of a positive result, to a misplaced judgment of the effectiveness of the new therapy under investigation. Perhaps, above all, the possibility that selection bias might significantly infect the result is more marked in the case of an observational study. Whereas hospital records will reflect the results of treating *all* the patients suffering from some particular condition with the earlier 'conventional treatment' (given that that was the only available option), those testing out the proposed new treatment will, of course, know that all the patients they choose will be given that new treatment. If there is, whether consciously or subconsciously, selection of those patients to be given the new treatment on the basis of likely positive outcome (perhaps only those with milder forms of the disease are selected for treatment, or those patients that appear stronger) then a positive result in an 'historically controlled trial' *may* be misleading.

However, it is surely not true that there is no way of reducing the risks of biases of this sort except by randomizing. We might well have convincing evidence that no selection bias has occurred in some particular historically controlled trial. In the ECMO case, for example, if (as I believe to be true) hospital records show that, in the 'historical trial' phase, no baby diagnosed as suffering from PPHS and who would earlier have been given the then conventional treatment was excluded from being treated by ECMO, then there was just no selection of babies: one uniform method of treatment was replaced by a different method of treatment applied in both cases to *all* babies

suffering from the same condition. And since there was no selection, there was no selection bias.

Finally, and perhaps equally importantly, as even the most staunch of advocates of randomization Richard Doll and Richard Peto finally allow, selection bias, even if we have grounds for thinking that may have played a role, is hardly likely—at least in the vast majority of cases—to make a major difference. (In general there seems to be much too much attention given to the question of whether or not the result of a medical trial manages to make 'statistical significance' as opposed to trying to estimate the *size* of the effect.) Doll and Peto accept ([1980]) that selection bias is hardly likely to produce an outcome that is more than double the 'true effect'. But we should remember that at any rate, once staff had become adept at handling the ECMO procedure, around 80% of babies with PPHS were surviving compared to a traditional *mortality* rate of around 80%. It seems that unless someone can point to some other methodological defect here that might have meant that there is a significant difference between the conventionally treated and ECMO treated groups, then even Doll and Peto ought to have spoken out against the view that an RCT was 'necessary' in this case.

As this indicates, the problems arise—just as Peter Urbach's earlier treatment claimed—when randomization is treated as a *sine qua non* of scientific 'validity'. When pressed, even the most avid advocate of randomization admits that this is not true, and yet the medical profession is encouraged to act as if it *is* true. The claim of necessity, as we have seen, is generally made on the grounds that randomization controls for all factors, known *and unknown.*[34] By contrast, tests performed using any other protocol are inevitably at risk of bias by unknown factors. I have argued that none of the attempts to establish this—centrally, so far as the current paper is concerned, those attempts based on analyses of 'probabilistic causality'—achieves any success. All trial results are defeasible. We are always, quite trivially, at the mercy of the possibility that the two groups are, unbeknown to us, unbalanced in some significant way. And, whatever may be true in the theoretical indefinite long run of endlessly repeated random divisions, for real-world trials, randomization does exactly nothing to alleviate this worry (remember selection bias is a 'known' factor). The best we can do (as ever) is test our theories against rivals that seem plausible in the light of background knowledge. Once we have eliminated other explanations that we know are possible (by suitable deliberate, or *post hoc*, control) we have done as much as we can epistemologically. The unthinking pursuit of randomization in all circumstances seems simply to be bad epistemology. Given that in the ECMO

[34] Again, when pressed, all serious commentators accept that this is not really true, but again they act, and encourage clinical scientists to act, as if it were.

case the possibility of selection bias can be ruled out by other means, there does seem to be a case that the babies involved in those trials were—of course unknowingly—sacrificed to a false epistemological god.

Acknowledgements

I am indebted to the members of the 'Causality—Metaphysics and Methods' group (under the Arts and Humanities Research Council—funded project headed by Nancy Cartwright, Elliott Sober and myself) especially Nancy Cartwright, David Papineau and Jon Williamson for a number of enlightening discussions. I gave presentations of parts of the paper to a meeting of the British Society for the Philosophy of Science; to the Pittsburgh Centre Fellows Conference in Rytro, Poland; at a University of Ohio series in Philosophy of Science lecture; and to groups in Belfast, Crete, Dublin and London. I would like to thank all those who made comments or offered encouragement at those meetings—notably, Colin Howson, and John Norton. As I make clear at various points, I am indebted to Peter Urbach's earlier treatment of the issue of randomization. I received detailed, and enormously helpful, comments on an earlier draft from Peter Urbach, and from Nancy Cartwright and Jim Woodward. My knowledge of clinical trials (and their effects) has been greatly enhanced by countless conversations with Dr Jennifer Worrall. Finally, I am indebted to Michael Worrall and especially to Lefteris Farmakis, for research assistance and help in preparing the final version.

Dept of Philosophy, Logic and Scientific Method
London School of Economics
Houghton Street
London WC2A 2AE
J.Worrall@lse.ac.uk

References

Bartlett, R. H., Andrews, A. F., Toomasian, J. M., Haiduc, N. J., and Gazzaniga, A. B. [1982]: 'Extracorporeal membrane oxygenation for newborn respiratory failure: 45 Cases', *Surgery*, **92**, pp. 425–33.

Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W., and Zwischenberger, J. B. [1985]: 'Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study'. *Pediatrics*, **76**, pp. 479–87.

Benson, K. and Hartz, A. J. [2000]: 'A Comparison of observational studies and randomized, controlled trials', *New England Journal of Medicine*, **342**, pp. 1878–86.

Cartwright, N. [1989]: *Nature's Capacities and their Measurement*, Oxford: Oxford University Press.

Concato, J., Shah, N. and Horwitz R. I. [2000]: 'Randomized controlled trials, observational studies, and the hierarchy of research designs', *New England Journal of Medicine*, **342**, pp. 1887–92.

Hacking, I. [1988]: 'Telepathy: origins of randomization in experimental design', *Isis*, **79**, pp. 427–51.

Hesslow, G. [1976]: 'Two notes on the probabilistic approach to causality', *Philosophy of Science*, **43**, pp. 290–2.

Howson, C. [2000]: *Hume's Problem*, New York and Oxford: Oxford University Press.

Howson, C. and Urbach, P. M. [1993]: *Scientific Reasoning–the Bayesian Approach*, Second edition, Chicago and La Salle: Open Court.

Kadane, J. B. and Seidenfeld, T. [1990]: 'Randomization in a Bayesian perspective'. *Journal of Statistical Planning and Inference*, **25**, pp. 329–45.

O'Rourke, P. P., Crone, R. K., Vacanti, J. P., Ware, J. H., Lillehei, C. W., Parad, R. B., and Epstein, M. F. [1989]: 'Extracorporeal Membrane Oxygenation and Conventional Medical Therapy in Neonates with Persistent Pulmonary Hypertension of the New Born: a Prospective Randomized Study', *Pediatrics*, **84**, pp. 957–63.

Pearl, J. [2000]: *Causality–Models, Reasoning and Inference*, New York and Cambridge: Cambridge University Press.

Spirtes, P., Glymour, C. and Scheines, R. [1993]: *Causation, Prediction and Search*, New York: Springer-Verlag.

Tukey, J. W. [1977]: 'Some thoughts on clinical trials, especially problems of multiplicity', *Science*, **198**, pp. 679–84.

Urbach, P. M. [1985]: 'Randomization and the design of experiments', *Philosophy of Science*, **52**, pp. 256–73.

Urbach, P. M. [1994]: 'Reply to David Papineau', *British Journal for the Philosophy of Science*, **45**, pp. 712–5.

Ware, J. H. and Epstein, M. D. [1985]: 'Comments on "extracorporeal circulation in neonatal respiratory failure: A prospective randomized study" by R.H. Bartlett, *et al*', *Pediatrics*, **76**, pp. 849–51.

Worrall, J. [2002]: '*What* evidence in evidence-based medicine?', *Philosophy of Science*, **69**, pp. S316–30.

Worrall, J. [forthcoming]: 'Evidence in Medicine', in *Philosophy Compass*.