LSE CPNSS

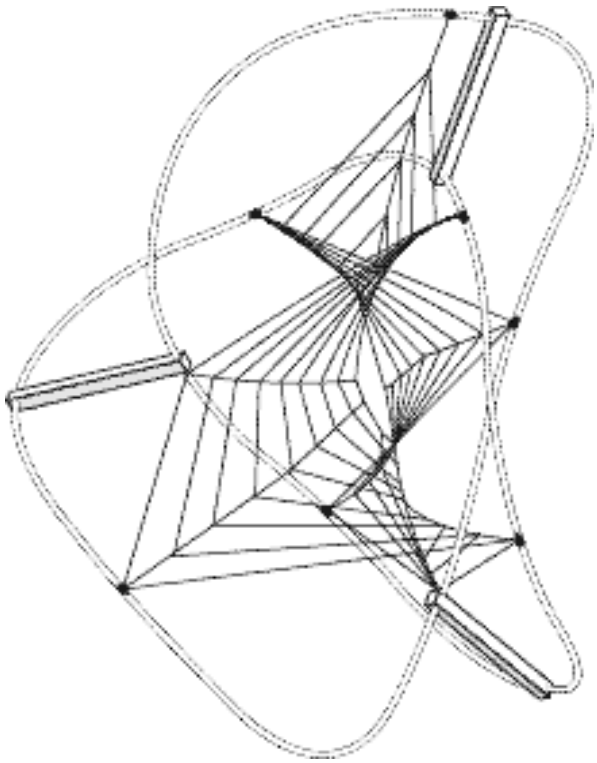
# Centre for Philosophy of Natural and Social Science

## Causality: Metaphysics and Methods

Technical Report 24/04

# *Why There's No Cause to Randomize*

John Worrall



Editor: Julian Reiss

# Why There's No Cause to Randomize

John Worrall*
Department of Philosophy, Logic and Scientific Method
London School of Economics
Houghton St
London
WC2A 2AE

*j.worrall@lse.ac.uk*

November 2004

* I am indebted for various discussions on various occasions to the members of the "Causality: Metaphysics and Methods" group (under the AHRB funded project headed by Nancy Cartwright, Elliott Sober and myself) and especially to David Papineau and Jon Williamson. I am also indebted to Julian Reiss for his own comments, organisation of discussions and editorial patience.

# 1 Introduction

In a randomized controlled experiment (henceforward an RCT) designed to test some new treatment – perhaps a new drug therapy or a new fertiliser – some experimental population (a set of patients suffering from some medical condition and recruited into the trial or a set of plots of land on which some crop is to be grown) is divided by 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 left unfertilised (all other obvious factors – such as frequency and extent of artificial watering, if any – being left the same as in the experimental regime), or, in clinical trials, the control group may be given either a "placebo" (generally a substance "known" to have no specific biochemical effect on the condition at issue) or the currently standard therapy for that condition. RCTs are almost invariably performed "double blind": neither the subjects themselves nor those treating them know whether they are receiving the experimental or the control treatment.[1] (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 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 – but this is always very much (at best) a case of (epistemic) second-best. For, it is widely believed, all non-randomized trials are inevitably subject to *bias*, while RCTs, on the contrary, are free from bias (or perhaps, and more plausibly, are as free from bias as any trial could possibly be).

Indeed, randomization has received *such* favourable press that the educated layman could be forgiven for believing that its scientific value is an entirely uncontroversial matter. However, this is far from the case. 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 while on this, as on every other substantive issue, there are major differences amongst those who regard themselves as Bayesians, the view that randomization has no justification remains the majority view within that (increasingly

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

[2] 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 misleadlingly, characterising 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".

influential) school.[3]

In an earlier article (2002), and partly following the treatment of Peter Urbach (see Urbach (1985) and Howson and Urbach (1993)), I tried to identify the main arguments that seem to underlie the view that randomization is necessary for 'scientific validity'. Urbach concentrated on two of these – Fisher's contention that randomization is necessary to underpin the logic of significance tests and 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 superficially the most persuasive).

A third argument is based on the idea that standard methods of randomizing control, not for some hitherto unconsidered possible bias, but for a known bias that is believed to have operated to invalidate a number of trials – 'selection bias'. If clinicians involved in trials are allowed to decide to which arm of the trial a particular patient is assigned to then there is the possibility that, perhaps subconsciously, they will effect a selection that distorts the result of the trial and gives an inaccurate view of the efficacy of the treatment. They might, for example, having a view on the effectiveness of the new drug and also 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 pharmaceutical company) paymasters. This does seem to me, as Urbach agrees from a Bayesian perspective, a definite epistemological good that properly performed RCTs deliver; though one should not of course infer that selection bias can never be eliminated by any means other than randomization.

In my (2002) paper I also analysed another 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 so-called observational studies ("historically controlled trials") because the latter standardly give unduly optimistic estimates of treatment effects. This argument, so I suggest in my (2002), is (a) circular (it depends on supposing that where an RCT and an 'observational study' have been performed on the same treatment it is the former, which after all provides the "gold standard"!, that reveals the true efficacy); (b) based on comparing RCTs to particularly poorly performed observational studies that anyone would agree are obviously methodologically unsound; and (c) is – to say the least – brought into severe question by more recent work that seems to show that, where a number of different trials have been performed on the same treatment, the results of those done according to the RCT protocol differ from one another much more markedly than to do carefully performed and controlled observational studies.[4]

More recently some different (or at least seemingly different) arguments to the effect that randomization has special epistemic virtues have, however, arisen from the currently burgeoning literature on 'probabilistic causality'. These arguments, though differing in detail, all claim that randomization plays an essential (or perhaps near-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. My current project is

[3] For a survey of the complexities here see Kadane and Seidenfeld (1990).
[4] See, for example, Benson and Hartz (2000) and Concato, Shah and Horwitz (2000).

to analyse three such arguments from leading practitioners in the field: Nancy Cartwright, David Papineau, and Judea Pearl. Only two of these – those from Papineau and from Pearl – will be examined in detail in the current paper. Before embarking on my analysis of these arguments, however, it will help briefly to outline a case which *both* shows how immensely important are the apparently rather abstract arguments about the epistemic weight of different trials *and* will help me formulate the rather nuanced view about RCTs that I would currently defend.

## 2. Why the Issue Is of Great Practical Import – The ECMO Case[5]

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 several years, mortality rates of less than 20% in infants treated by ECMO. (See Bartlett, Andrews *et al.* (1982).) I think it is important background here that this new treatment could hardly be regarded as a stab in the dark. It was already known that the underlying cause of this condition was immaturity of the lungs, 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. ECMO in effect takes over the function of the lungs in a simple and relatively non-invasive way. Blood is extracted from the body before it reaches the lungs, is artificially oxygenated, reheated to regular blood temperature and reinfused back into the baby – thus bypassing the lungs altogether.

Despite the appeal of the treatment, and despite this very sharp increase in survival from 20% to 80% the ECMO researchers felt forced to perform an RCT ("… we were compelled to conduct a prospective randomized study") despite the fact that their experience had already given them a high degree of confidence in ECMO ("We anticipated that most ECMO patients would survive and most control patients would die…")[6] They felt compelled to perform a trial because their claim that ECMO was significantly efficacious in treating PPHS would, they judged, carry little weight amongst their medical colleagues unless supported by a positive outcome in such a trial.[7] These researchers clearly believed that, in effect, the long established mortality rate of more than 80% on conventional treatment provided good enough controls – that babies treated earlier at their own and other centres with conventional medical treatment provided sufficiently rigorous controls; and hence that the results of more than 80% survival that they had achieved with ECMO showed that ECMO was a genuinely efficacious treatment for this dire condition. Given that there was an argument for thinking that there was no significant difference between the babies that Bartlett and colleagues had been treating using the earlier techniques and those that they had now been treating with ECMO, this counts as a (retrospective) "historically controlled trial". But, because historically controlled trials are generally considered to carry little or no weight compared to RCTs, these researchers felt forced to go ahead and conduct the prospective trial.

---

[5] Peter Urbach first drew my attention to this case.

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

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

They reported its outcome in 1985 (Bartlett, Roloff *et al.* (1985)). Babies suffering from PPHS were allocated to ECMO treatment or to the "control group" (which received the then conventional medical therapy – CT) using a modified protocol called "randomized play the winner". This protocol involves assigning the first baby to treatment group purely at random – say by selecting a ball from an urn which contains one red (ECMO) and one white (CT) ball; if the randomly selected treatment is a success (here: if the baby survives) then an extra ball corresponding to that treatment is put in the urn, if it fails then an extra ball corresponding to the alternative treatment is added. The fact that this protocol, rather than "pure" randomization, was used was 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 8 babies all happened to be assigned ECMO and all survived. The protocol, decided in advance, declared ECMO the winner at this point, though a further two babies were treated with ECMO (officially "outside the trial") and survived. So the 1985 study reported a total of 12 patients, 11 assigned to ECMO all of whom lived and 1 assigned to CT who died. (Recall that this is against the background of a historical mortality rate for the disease of around 80%.)

Ethics and methodology are fully intertwined here. How the ethics of undertaking the trial in the first place are viewed will depend, amongst perhaps 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 (Tukey (1977); emphasis supplied)

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

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

Many commentators in fact took this latter view and concluded that

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

Those taking this second view held that neither the "historically controlled" results (i.e. the comparison of the mortality rates achieved with ECMO with the historical mortality rate achieved with conventional treatment) nor the results from this initial "randomized play the winner" trial had produced any reliable, scientifically-telling information. The Michigan trial had not produced any real evidence because – in deference to the researchers' prior convictions – it had not been "properly randomized". Indeed, they even imply (see their "will be difficult" remark) that such trials and their "historically controlled" antecedents, have, by encouraging the belief that the new treatment was effective in the absence of proper scientific validation, proved pernicious by making it more difficult to perform a "proper" RCT: both patients and more

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

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

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

The question that surely needs to be asked about the initial 'historically controlled trial' is surely whether or not there is some plausible reason (on the basis of 'background knowledge') for the striking change in mortality rates, *aside from* the new treatment itself. So far as I can tell, no one takes seriously the idea that the natural history of the disease may have changed between the late 70s and the early 80s, or that the population from which the babies were drawn underwent some general change in robustness at the time ECMO was first introduced. Expectations of success either on the part of the patients or of the clinicians are hardly a plausible factor in the case of neonates. The outcome itself is as objective as you can get, so there is no question of the experimenters assessing outcome in a biased way. It is true that the evidence from the second ("properly randomized" study) suggests (weakly of course since the sample is small) that babies on *both* arms of the trial may have received better than average treatment (6 out of the 10 babies on the *conventional treatment* arm survived). There is indeed some reason to think that patients involved in trials in general receive more, and better informed, attention. (After all, the trials are often held in the big medical centres attracting first-rate staff and having superior facilities.) But this affects both arms of a trial equally. This leaves selection bias as a main contender – were, perhaps, babies directed to ECMO only if they were relatively strong compared to others suffering from PPHS, while the relatively weaker ones continued to be given the conventional treatment? Bartlett and colleagues in their initial, pre-RCT work seem clearly to have treated *all* babies in their care suffering from this condition with ECMO and

turned an 80% mortality rate into an 80% survival rate. There may of course be other plausible alternatives – but none has been suggested.

I can see no reason myself to rate the results of even the second trial (which involved, remember, only 19 babies in total between the two groups) as supplying any stronger evidence in favour of the new ECMO treatment than the initial experience of Bartlett and his colleagues (involving a good many more babies).

So what, then, seems to be the story so far concerning randomization and its alleged epistemological virtues? Well,

1. Randomization may certainly do some good by controlling for a known possible confounder – 'selection bias'; but
    a. it does this, not through any mystical power of the coin toss or random number table, but by preventing the experimenters from knowing in advance to which arm of the experiment a particular subject is to be 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); and
    b. there is no reason in some cases – such as the ECMO case – why the possibility of selection bias cannot be eliminated by other means (and indeed eliminated at least as firmly as it is by randomization, perhaps more so).
2. Randomization is certainly no *sine qua non* of scientific validity or guarantee of anything; and indeed it is difficult to see why anyone should think that, say, the second trial in the ECMO case gave one any stronger reason to believe that ECMO was a highly effective treatment for PPHS than that provided by the original, so-called historically controlled, trial. There seems to be a tendency to confuse the question of what can legitimately be inferred from one actual trial (here Savage and the Bayesians are surely correct that we must go on the evidence that we have from the actual outcomes for the people in the experimental and control groups and that this cannot at all rationally depend on how those actual people got into those groups) and what might – in (totally) unattainable – principle be inferred if we were to re-randomize indefinitely often. (As I shall argue later, this same tendency is exhibited by those defending randomization from the perspective of probabilistic causality.)

The ECMO case also points to a sort of moral lesson: given the things that have been done in pursuit of the idea that randomization has some privileged epistemological status, we had better be even more vigilant than usual in examining the credentials of any argument to that effect. This is what I proceed to do in the rest of this paper. As I indicated at the beginning, I did not in my earlier (2002) paper examine any of the arguments for randomization emanating from the currently burgeoning literature on probabilistic causality. In the next section I examine one such argument – from David Papineau, and in the final section I examine a similar argument from the

leading formal analyst of causality, Judea Pearl. (And I save analysis of another such argument from another leader in the probabilistic causality field – Nancy Cartwright – for a further paper.)

# 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. David Papineau criticized Urbach in a (1997) article – arguing 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.[8]

Papineau sees clinical researchers as making *two separate* inferences from trial data. The first inference takes them from some observed difference in the frequencies of recovery, say, in the experimental and control groups to some (alleged) conclusion about the "objective probabilities" in the "underlying population" of recovery (R) conditional on treatment or no treatment (T or ¬T). This idea of the population objective probabilities p(R/T) and p(R/¬T) is itself rather mysterious in the case of clinical trials. Almost invariably clinical trials are used to assess therapies that are not currently in use, and even if some such therapy were already being used to treat some patients, clinicians would be looking to the clinical trial to direct them to treat patients differently than they currently do – perhaps giving the therapy to patients currently not on it, or restricting its use by excluding currently treated types of patient. Clinical trials are first and foremost intended as guides to *intervention* – and interventions will standardly change the population frequencies.[9]

There is one further way in which even this initial suggestion by Papineau does not sit well with the reality of clinical trials. Even aside from issues that we shall need to raise almost immediately about the possibility of underlying factors ('common causes') engendering a 'spurious correlation', a clinician would not take the effectiveness of T for R as being established by a 'finding' that p(R/T) > p(R/¬T). This is because clinical trials never test a specific therapy against no therapy at all (indeed in the trial setting it is not clear what this would mean), instead they test at least against placebo (and in the trials most likely to be useful for clinical practice they test a proposed new treatment against a conventional treatment already known to be effective to some extent).[10] Given that the placebo effect is real, it may well be that in this particular case too p(R/placebo) > p(R). So in order to deliver a positive verdict on a new

[8] Let me say at the outset that Papineau is definitely *not* one of those who believe that randomization is necessary for any reasonably definite 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 produce results of higher epistemic weight and hence that we are always settling for second best if we settle for a non-randomized trial, even if 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 moral price worth paying for the extra, and extra security of, the information derived from the trial.

[9] Maybe the truly underlying probabilities should be thought of as fixed. Because, post trial, a different group of patients would be given T, then it is entirely conceivable (and indeed the trial would have been a failure if this were not true) that the number of people recovering on the treatment will be different (higher) than it was before, because of the more discerning selection of patients given the treatment. But presumably this will be because the trial has identified some general characteristic or characteristics G that were relevant so that p(new)(R/T) = p(old) (R/T&G).

[10] Papineau notices this but it is unclear from his article how he would accommodate this recognition within his general account.

treatment T, a clinical trial is demanding that T show itself not just to be effective, but to be *more* effective than placebo (at least). (You might want to get round this by separating out, so to speak, the 'real' 'specific' effect of T from its generic features, as just any sort of therapy – those effects it will share with placebo treatment (however this too is a tricky issue). Nonetheless the fact remains that placebo treatment is treatment and may be efficacious; it does not then follow from p(R/T) = p(R/¬T) – where ¬T is in fact placebo, that T is not efficacious.)

But let's lay these difficulties aside for the moment and return to the analysis of the main line of Papineau's argument. As noted, then, he holds that there is a first inferential step from the observed frequencies of recovery in the two groups in the trial to some sort of conclusion about some conditional probabilities in some general population. 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. (It is not clear whether this is supposed to mean that the sample of patients in the trial (the union of those in the experimental and control groups) should be random or if the *two* samples consisting separately of those in the experimental group and those in the control group should somehow or other be random samples. Neither suggestion seems coherent with standard practice. Certainly there is no sense in which the initial set of patients involved in the trial can be thought of as constituting a random sample: they are recruited in a more or less haphazard way (as Papineau records) and then, in all current clinical trials, are subjected to very non-haphazard selection criteria before being 'allowed' into the trial (they are also asked for their 'informed consent' to be included). But, in any event, Papineau sides with what he takes to be the Bayesian position here by agreeing that random sampling is not necessary – not necessary, that is, to justify the (alleged) inference from finite frequency data to the population probabilities he mentions. (Though he never explains how the Bayesian approach – involving of course subjective degrees of belief – is compatible with his repeated assumption that the population probabilities at issue are *objective*.)

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. Randomization, as Papineau sees it, is necessary to justify the further inference from probabilities to *causes*. Having arrived at the conclusion that p(R/T)>p(R/¬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? Well of course this is our old friend spurious correlation. It may well be, not only that there are more recoverers in the actually observed experimental group, but that if you kept on repeating the trial with different groups you would continue to find (on average) that the treated group did better than the 'untreated'. But this might not arise from anything specific to the treatment but rather because the treatment itself was 'correlated' with some other factor that was the real cause. Suppose for example that there are more young people in the experimental group than in the control group and that young people are more likely to recover anyway (that is, independently of T). This would be reflected presumably once the therapy had been made available in the general probabilistic truth about the population that, where Y is the property of being young, p(R/Y &T) > p(R/¬Y & T). Should it be the case that p(R/Y & T) = p(R/Y & ¬T) then Y 'screens' R from T – that is, R and T lose their probabilistic dependence, when conditioned on being young. And we would presumably want to say that in that case it may very well be that despite the probabilistic dependence of R and T, T does not cause R.

Once again, in order to make his two-stage story work, Papineau suggests that the treatment might be correlated with confounders *in the actual population of interest*: the idea presumably being that it might be that the therapy is preferentially offered to, say, younger people as a matter of course (that is, not just within the experimental population because of a quirk of the experimental/control division).

I admit that it is tempting to think that experimental randomization can solve the problem of spurious correlation (and we shall see that, albeit in a rather different approach, this claim is made by Judea Pearl too). But how *exactly* is this supposed to work?

Well, according to David Papineau, it is a *guaranteed sure-fire conclusion* from the premise that $p(R/T) > p(R/\neg T)$ (and remember that he accepts that, although a premise here, this itself is a conclusion of what is a very fallible inference) 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 as the conclusion of a tight logical argument, as follows:

First, the account of causality that he endorses is that specified in **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." He then argues (Claim 1) 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 (Claim 2) that randomization eliminates the possibility that the second disjunct holds (p. 440; my emphasis):

> 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.

So the conclusion of the argument is that, given the characterisation of cause **C**, then the fact that an experiment was randomized *entails* that if $P(R/T) > P(R/\neg T)$ then T causes R.

The claim that randomization is a sure-fire guarantee of a causal conclusion seems on the surface to be absurdly strong. And indeed Papineau himself introduces one obvious apparent objection (pp. 446-7):

> Suppose we notice, after conducting a randomized experiment, a relevant [??] difference between the treatment and control samples. 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.

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." (p. 447) We need, he suggests, constantly to keep in mind the differentiation 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 (*ibid*.; my emphasis):

> *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*.

Quentin Quick's first step is fallacious and his (*op. cit.*, pp. 447-8)

> 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, given the hypothesis that T and R are objectively uncorrelated. So 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.

However, the Bayesian does not use the point that no one regards randomization as infallible simply to point out that sometimes everyone (who is sensible) gives up on a particular randomly produced allocation and balances or "re-randomizes" (when they spot that the experimental and control groups happen to be biased in respect of some factor that background knowledge tells us might well play a significant role in recovery or whatever other end-point the trial is measuring). Instead the Bayesian is making the point that, *for all we know, we are always in this situation*. For all we know, we are always in Quentin Quick's situation, but with respect to some factor that is not so obvious. Indeed given that there are indefinitely many possible biases or confounding factors, it seems intuitively likely that we will be.

Again – unsurprisingly since this is, after all, just the 'randomization controls for all possible confounders, known and unknown' ploy dressed in causalist clothing – the intuitive appeal of the 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 only randomized once!) and what might be justified if we re-randomized indefinitely (either on the same or 'equivalent' population of subjects). Papineau himself claims (p. 447, though without my emphasis you might think of this as in the small print) only that "the randomization will ensure that all other causes of R are probabilistically independent of T *in the long run*". 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 gives the same results as would be got from the 'limiting average'. Moreover, there is no sense in which we can ever know how close a particular randomized trial is to the 'ideal'. Any particular randomized trial may therefore mislead about causes (even in Papineau's sense). This is not just a logical possibility: any *actual* randomized trial result cannot bring us any closer to any epistemic guarantee about causality. The 'guarantee' that Papineau's argument allegedly supplies is, remember, just the ineliminably conditional one 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 probabilities. This conditional guarantee with the undecidable antecedent reminds me a little of the famous Goldwynism : 'A verbal contract is not worth the paper it's written on.'

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 amongst 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 reason at all for thinking that the division between experimental and control group is not biased in some significant way.

Hence no further *practical* reason for randomizing is supplied by Papineau's invocation of causality.

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 anti-biotic 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, A in such a case 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 from D by taking A. I don't myself believe that, in such a case, there's 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, just 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 sub-population, consisting of those satisfying the conditions $C_1, \ldots, C_n$ within which A causes recovery). The researchers don't know about $C_1, \ldots, C_n$, but in fact the randomization is 'perfect' – which I suppose means (with an eye to 'external' as well as 'internal validity') *both* that x% of the subjects have $C_1, \ldots, C_n$ (where x% is the *population* frequency) and exactly half of those go into each group in the trial. Finally suppose that there is no overt reason to worry about the randomization – that is, there is no known factor which might plausibly be taken to play a causal role which is maldistributed in the two groups. Hence none of the investigators has scope for any 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 for Papineau 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 the antibiotic A and the disease D; only now suppose that the conditions $C_1, \ldots, C_n$, rather than being rare are quite common in the population. I suppose that the judgment that most people 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 satisfy $C_1, \ldots, C_n$; and moreover, by chance almost all the ones that do satisfy $C_1, \ldots, C_n$ are assigned by the randomizer to the control group. Remember the fact that conditions $C_1, \ldots, C_n$ 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, those given A. (Remember we are supposing that A actually exacerbates D for those individuals who fail to satisfy the conditions $C_1, \ldots, C_n$.)

So again since we know "from the outside" the real causal situation we see that the randomized trial has given the wrong answer. This, 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 inference to the probabilities. But the investigators are bound to be in blissful ignorance of this and hence this case emphasises again the non-practical 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 D that could in fact have been helped by A, these experimenters will scarcely feel consoled, once further research has revealed the true causal situation, to be told that the causal step in their initial two-step inference was flawless!

This second case was one where, despite impeccable randomization and lack of 'sloppiness', experimenters infer a lack of causal connection (or actually the wrong causal connection). It is equally possible, however, that despite randomization, experimenters will infer a causal connection where none actually exists. To see this, consider a third case.

Martians come down to Earth and notice that there seems to be a link between owning say more than 5 ashtrays and dying a particularly unpleasant kind of death – their background knowledge is so radically different from ours (that's why I'm playing around with Martians) that they have not the slightest inkling that the use to which ashtrays are generally put may have something to do with this death rather than the simple possession the ashtrays themselves. They decide to do a clinical trial . They take a group of patients and make them eliminate all ashtrays from their homes. Then they divide this group into two – those in the first group are to be forced to buy 10 ashtrays, while those in the 'control group' are forced to abstain from ashtray ownership. Suppose by happenstance that there are exactly x% of smokers in the overall set of subjects, which exactly reflects the population rate. However, it happens that, despite an impeccable randomization all of the smokers go into the experimental group and none into the control group. Again, exactly because they have no inkling of what we know to be the real causal factor, they have no reason to suspect that their randomization had produced a significantly biased division. The Martians wait for 10 years and obtain the data. They find many more cases of lung cancer in the ashtray group. So they conclude ashtray ownership causes cancer.

Exactly because all the smokers went into the ashtray group, and because, we may assume, there is a perfect – or near perfect – correlation in the general population between owning significant numbers of ashtrays and smoking tobacco, we can expect the data here to mirror fairly closely the correct population 'correlation' between ashtray ownership and lung cancer. So, once again, we have impeccable randomization, no 'sloppiness' in the statistical inference (the Martians thought that maybe hair colour was related to this sort of death, but had checked that the distribution of hair colour closely reflected that in the general population in both the experimental and control groups, they thought that owning green things might be toxic but had checked that there were just as many people with lots of green things in their house in the experimental as in the control group, etc, etc) and yet a causal conclusion that is entirely false.

So, again, what happened to David Papineau's 'sure fire' (albeit conditional) guarantee?

The Martians' inference went as follows. Their data was that there were many more cases of lung cancer in the ashtray-owning group than there were in the non-ashtray owing group. They inferred (first step) that in the population as a whole P(cancer/own ashtrays) > P(cancer/don't own ashtrays). And, because they had read David's paper, and knew that they had properly randomized, inferred (second step) that ashtray ownership causes cancer. Now, again from the outside, they were clearly lucky at the first stage: it was only because there is a pretty well perfect correlation between smoking and owning lots of ashtrays in the population, and because the randomization they performed gave them all the smokers in the ashtray group, that they arrived at the correct correlational view. Nonetheless the correlation they came up with is correct. So, so far as *outcome* of the inference is concerned, the first step was correct. In the second inferential step they were *un*lucky and of course 'we' can see (since we know the true causal set up) that had they re-randomized a number of times with the same or 'equivalent' experimental population then they would soon have had cause to doubt that P(cancer/own ashtrays) > P(cancer/don't own ashtrays). However, as always, they only randomized once and they made a causal mistake. It seems that Papineau's guarantee, for all its conditional character, is *not* in fact 'sure fire'. (Our Martians have, of course, made a mistake about a conditional probability – or, put more clearly, they have failed to discover a significant truth about conditional probabilities: namely, that P(cancer/smokes & owns ashtrays) = P(cancer/smoke & doesn't own ashtrays). But this is not anything to do with the first step in Papineau's treatment when applied to this case. They have been prevented from finding out that smoking "screens off" lung cancer from ashtray ownership by (a) this not being yielded as a plausible possibility by their background knowledge (that's why we needed Martians here) and (b) a quirk in the actual random division they made (all the smokers went in one group) but which because of (a) they were in no position to question.)

# 4 Pearl on Randomization and Cause[11]

## 4.1 Introduction

The technically most sophisticated and formally elaborate account of causality currently available – (along with the cognate account of Glymour, Scheines, and Spirtes)[12] – is the one developed by Judea Pearl and reported in his (2000) book. 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 analysed 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 analyse and criticise that argument.

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

[12] See Spirtes, Glymour and Scheines (1993) and subsequent literature from the group.

## 4.2 Pearl's Account of Causal Nets

As is well-known, 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 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 $X_i$.) There is also a joint probability distribution on all the variables in the network $P(X_1, X_2,\ldots,X_n)$ and this probability distribution must cohere with the structure produced by the arrows in the sense that, for example, if 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 variable X is conditionally independent, given its parents, of any set of other variables that aren't 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 (p. 45)[13], an idealization – and one that will be consequential when it comes to analysis of his approach to randomized trials: to be clear (if trite) 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's take Pearl's own initial example. It seems that 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 its raining in California is also dependent on the season. (Of course in England we have our own inbuilt – as good as automatic – lawn sprinkler system in the shape of the year-round weather.) If, in Pasadena or wherever, it either rains or if the sprinkler is on, then usually, but not invariably (the rain may be light, the wind may unusually direct all the sprinkler water away from the pavement, there may unusually be a cover on the pavement, *etc*.) the pavement (which I assume in English English means 'path') will get wet; 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 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 the following set of interconnections between these variables – Figure 1.

[13] 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 [then acknowledgedly conjectured] distribution.'

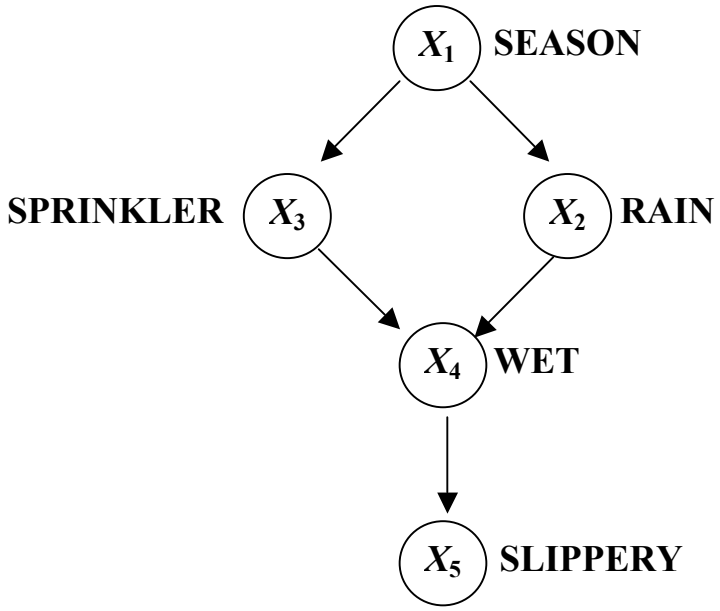


Figure 1: A simple Bayesian/causal net. (Source: Pearl (2000), p. 15)

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's let this pass. Assuming further that it is sensible to model this whole set up 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 particular value (spring, yes, *etc*.), given values of other variables. For example, the absence in the 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, has centrally to do with *intervention*. Within his framework this is characterised 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. In other words we have the following 'mutilated' network – Figure 2.

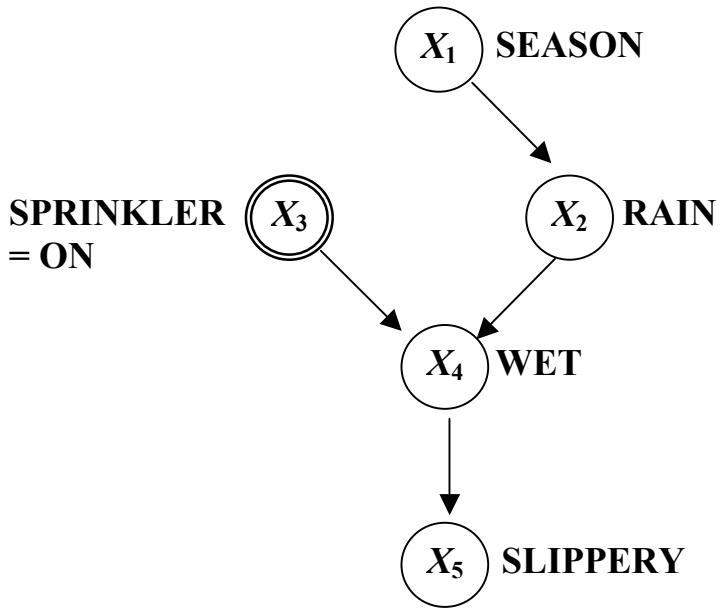



Figure 2: "Surgery" on a network (Source: Pearl (2000), p. 23)

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, and therefore non-probabilistic cases that Pearl discusses in the Epilogue of his book. Suppose to take a very simple case 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, say, doubled by the multiplier and then passed (as variable Y) to the adder where one 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 is the input.) Moreover this is – at least in a stretched sense – intuitively a causal set-up: perhaps we should think of the above as the abstract formal description of a two-component physical machine that takes any metal rod as input, and first stretches it to twice its length (whatever that initial length may have been) and then, in its separate second sub-component 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 sub-components 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 in figure 1 must satisfy the following equation:

$$P(x_1,x_2,x_3,x_4,x_5) = P(x_1)P(x_2/x_1)P(x_3/x_1)P(x_4/x_2,x_3)P(x_5/x_4) \quad (1)$$

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

$$P_{X_3=on}(x_1,x_2,x_4,x_5) = P(x_1)P(x_2/x_1)P(x_4/x_2, X_3=on)P(x_5/x_4) \quad (2)$$

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 (p. 23):

> 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.

Once again in this probabilistic case, the hallmark of the network's being truly causal is that we can predict what will happen when we make certain interventions in it. It is just that in this probabilistic case what we predict 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*." (p. 347; I take it that we'd all agree that what he really holds is not that this is what causation *means* but rather what the true hallmark of causality is (of course, on his view).) Elsewhere (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 line from X to Y in a network that has (a) been inferred from probabilistic 'data' in accord with Pearl's algorithm and (b) 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.

## 4.3 Pearl's Argument For RCTs

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 exclusively on the case he makes, on its basis, for RCTs; and will challenge aspects of his general account only as they become implicated in my response to that case.

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 which 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 (p. 348; all emphases in the original):

> 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 [that Pearl considers earlier – think about the two-component deterministic machine I discussed above]. Recovery is a function of both the treatment and other factors, such as socio-economic conditions, life style, diet, age, et cetera. … [In the diagram below – Figure 3 – he collapses all these extra factors into one, just called socioeconomic conditions.] Under uncontrolled conditions, the choice of treatment is up to the patients and 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.

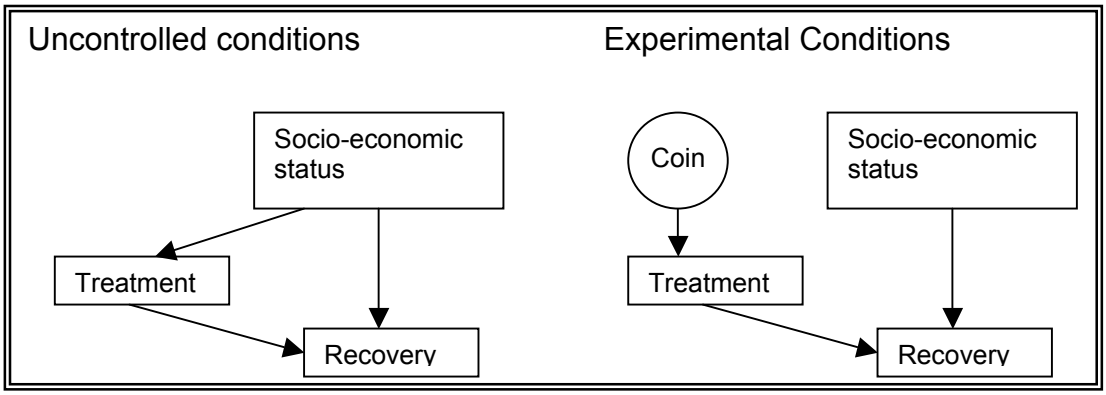



Figure 3. Controlled Experiments as "Surgery". (Source: Pearl (2000))

## 4.4 Analysis of Pearl's Argument For RCTs

In turning, then, from exposition of Pearl's defence of RCTs to criticism of it, let me say immediately that I am not, of course, against controlled studies! That's just science – we are looking for evidence for our hypotheses, whether causal or not, from genuine tests of them and that means from evidence that, as well as being compatible with the hypothesis, tells against plausible rival hypotheses. Clearly if we think that the apparent 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 socio-economic background (or rather to factors such as superior diet or living conditions generally associated with that background) then no one in her right mind would deny

that it is a good idea in studying the effect of the drug to control *in some sense* for socio-economic background. The only point at issue is Pearl's 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 celebrated 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 presumably 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 that in any actual series of tosses the sprinkler '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 Fisher's famous suggestion that there might be a common (genetic) cause for both cigarette smoking and cancer from 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 dividing into 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 ensuring 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.

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 be effectively sure (or as sure as we are going to be) that we had broken any possible link between socio-economic 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, *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 found in the general 'target' population.)

Ah! I hear you cry, as people are prone to do in this situation, 'what about the *unknown*

possible confounders?' We clearly don't know for sure, no matter how long our list, that we have ever exhausted all the possible aspects of a person's socio-economic situation that might play a role in outcome; and even if we had there are other possible 'confounders' that are not linked to socio-economic situation. (Pearl acknowledges, of course, that his diagram – here Figure 3 above – (deliberately) simplifies the real 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 would be the completely deliberately matched one. But once we acknowledge that there will always be possible confounders that we don't know about and *a fortiori* 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 won't know if they are, and maybe in fact they are, maybe in fact they are not. Doesn't randomization somehow or other guarantee (or perhaps, much more plausibly, provides the nearest thing that we can have to such a guarantee) that *any possible* links to therapeutic outcome, aside from the link to treatment with the drug concerned, are broken?

Although he doesn't 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, they recommend re-randomizing 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* it brings us back to exactly the same claim that I already analysed and rejected when looking at Papineau's defence of randomization.

Does this claim of control for known *and unknown* confounders fare any better within Pearl's framework? Well, let's stick firmly to practice here and not implicitly switch over, as people tend to, 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 won't and can't 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 recoverers 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, sees no role for randomization here at all.[14] The basic argument is, I think, that the sensible person goes on the evidence that he has and can give no role to how that evidence was generated (or what other evidence he *might have* considered, but is not in fact considering). 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 he 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 he makes about the relationship between treatment and outcome; and if he 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 effectively done for him by full and adequate matching) then he can have no better reason to believe the groups balanced.

This is one area in which Savage's claim that Bayesianism is just an articulation of commonsense seems to me correct; and it is not immediately easy to see why he 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 justification – that by Jay Kadane and Teddy Seidenfeld (*op. cit.*): (a) distinguishes between 'experiments to learn' (for yourself) and 'experiments to prove' (to someone else); (b) 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; (c) 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 (d) argues that even then, there are non-randomized designs that would do the job at least equally well (and also at less 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 about on a frequentist reading of this alleged probabilistic guarantee? What sense can be made on a frequentist reading 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? 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, then in the indefinite long run, the limiting frequency averages would reflect the real effect of the treatment, since in that limit that confounder would have, with probability one, to 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* population rather than the 'target population' – that is, to the group of people who happened to have been recruited

[14] 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).

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'.[15])

But let's 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 realised 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 can't see it.

In the end then it seems difficult to avoid the conclusion that, like Papineau (and, as I shall argue in a later paper, like Nancy Cartwright), 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. Nor has he therefore given any reason, returning to the ECMO case with which I began, that the evidence from even the second 'properly randomized trial' was in the slightest degree more compelling than the initial 'historically controlled trial' (which involved 45 babies rather than the 20 in the *official* second trial) and hence not the slightest justification for the apparent sacrifice of those babies in the two randomized trials we discussed.

Let me end by reiterating though, that the 'takehome message' of this article is NOT 'Don't randomize!' – it's a bit more nuanced than that. Here's what I think:

1. Randomization may, and often does, do some good – by controlling for a known possible confounder: selection bias; however
2. Randomization certainly cannot be trusted, as some of its extreme advocates suggest, to control for known confounders and should the trial suggest that the treatment at issue has a positive effect, this is clearly the more to be trusted the more plausible alternative explanations for the observed result other than that it was produced by the treatment itself have been eliminated by *deliberate* controls; and
3. If after that (and assuming that it's a good idea to perform a trial at all – see 4), then randomization at least can do no harm; however,
4. Where there is good reason to think that selection bias cannot have been a factor and no good reason to think that the newly treated group is in any significant way different from

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

> the historically treated group (as seems to have been true in the ECMO case), then, so far as I can tell, no one has provided any compelling, or even cogent, epistemological reason for insisting on a further, randomized trial.

If there is a simple takehome message it would be: don't believe the bad press that 'observational studies' or 'historically controlled trials' get – so long as they are properly done (that is, serious thought has gone in to the possibility of alternative explanations of the outcome), then there is no reason to think of them as any less compelling than an RCT, and, if the RCT has not been carefully matched for known possible confounders ahead of randomizing, there *is* reason to think that a properly conducted historically controlled trial may provide *more* compelling evidence.

# References


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

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

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

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

Hacking, I. (1988) "Telepathy: Origins of Randomization in Experimental Design", *Isis*, **79**, pp. 427-51

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

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

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

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

Spirtes, P., C. Glymour, and R. Scheines, (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. 679-684.

Urbach, P.M. (1985) "Randomization and the Design of Experiments", *Philosophy of Science*, **52**, pp 256-73.

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

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