Brit. J. Phil. Sci. 45 (1994), 437–450 *Printed in Great Britain*

The Virtues of Randomization

DAVID PAPINEAU

ABSTRACT

Peter Urbach has argued, on Bayesian grounds, that experimental randomization serves no useful purpose in testing causal hypothesis. I maintain that he fails to distinguish general issues of statistical inference from specific problems involved in identifying causes. I concede the general Bayesian thesis that *random sampling* is inessential to sound statistical inference. But *experimental randomization* is a different matter, and often plays an essential role in our route to causal conclusions.

1 INTRODUCTION

Peter Urbach has argued that randomized experiments serve no useful purpose in testing causal hypotheses (Urbach [1985], Howson and Urbach [1989]).[1] In this paper I shall show that he misunderstands the role of randomization in this context, as a result of failing to separate issues of statistical inference sufficiently clearly from problems about identifying causes.

Urbach is a Bayesian, and in consequence thinks that *random sampling* is unimportant when inferring *objective probabilities* from *sample data* (Urbach [1989]). I am happy to concede this point to him. But I shall show that *experimental randomization* is a quite different matter from random sampling, and remains of central importance when we wish to infer *causes* from *objective probabilities*.

[1] Of the two authors of this book, Urbach is responsible for the sections on randomized experimentation.

This is a topic of some practical importance. Randomized experiments, of the kind Urbach thinks unhelpful, are currently extremely popular in medical research circles. I agree with Urbach that this medical enthusiasm for randomization is dangerous and needs to be dampened. But this is not because experimental randomization is worthless, which it is not, but rather because it is often unethical, and because the conclusions it helps us reach can usually be reached by alternative routes, albeit routes of greater economic cost and less epistemological security. It would be a pity if Urbach's spurious methodological objections to randomized experiments deflected attention from their real ethical deficiencies.[2]

I shall proceed by first giving a simple example of the kind of problem that a randomized experiment can solve. After that I shall consider Urbach's arguments.

2 AN EXAMPLE

Imagine that some new treatment (T) is introduced for some previously untreatable disease, and that it turns out that the probability of recovery (R) in the community at large is greater among those who receive T than among those who do not:

$$\text{Prob}(R/T) > \text{Prob}(R/-T). \tag{1}$$

Such a correlation[3] is a prima-facie reason to think T *causes* R. But perhaps this correlation is spurious: perhaps those who receive the treatment tend to be younger (Y), say, and so more likely to recover anyway, with the treatment itself being irrelevant to the cure. Still, we can easily check this: we can consider young and old people separately, and see whether recovery is still correlated with treatment *within* each group. Is

$$\text{Prob}(R/T \text{ and } Y) > \text{Prob}(R/-T \text{ and } Y), \text{ and} \tag{2}$$

$$\text{Prob}(R/T \text{ and } -Y) > \text{Prob}(R/-T \text{ and } -Y)?$$

If neither of these inequalities holds—if it turns out that T makes no probabilistic difference to R either among young people, or among old people—then we can conclude that T *doesn't* cause R, and that the original correlation

[2] For discussion of the ethics and methodology of randomized medical trials, see the symposia in the *Journal of Medical Ethics*, 9 [1983], pp. 59–93, and the *Journal of Medicine and Philosophy*, 11 [1986], pp. 297–404. For a historical account of the surprisingly recent origins of experimental randomization, see Hacking [1988].

[3] The standard technical definition of 'correlation' presupposes quantitative variables. However, equation (1) yields an obvious analogy for qualitative factors. In this paper I shall stick to qualitative factors, in the interests of simplicity, but the argument generalizes to the quantitative case.

(1) was due to the treatment being more probable among young people, who recover anyway, than among old.

On the other hand, if the inequalities (2) do hold, we can't immediately conclude that T *does* cause R. For perhaps some other confounding cause is responsible for the initial T–R correlation (1). Perhaps the people who get the treatment tend to have a higher level of general background health (H), whether or not they are young, and recover more often for that reason. Well, we can check this too: given some index of general background health, we can consider healthy and unhealthy people separately, and see whether the treatment makes a difference within each group. If it doesn't, then the initial T–R correlation is exposed as spurious, and we can conclude that T does not cause R. On the other hand, if the treatment does still make a difference within each group . . .

By now the problem should be clear. Checking through all the possible confounding factors that may be responsible for the initial T–R correlation will be a long business. Maybe those who get the treatment generally have some chemical in the drinking water; maybe their doctors tend to be more reassuring; maybe . . .

A randomized experiment solves the problem. You take a sample of people with the disease. You divide them into two groups at random. You give one group the treatment, withohold it from the other (that's where the ethical problems come in), and judge on this basis whether the probability of recovery in the former group is higher. If it is, then T *must* now cause R, for the randomization will have eliminated the danger of any confounding factors which might be responsible for a spurious correlation.

3 RANDOMIZED EXPERIMENTS HELP WITH CAUSES, NOT PROBABILITIES

In this section I want to explain in more detail exactly why experimental randomization is such a good guide to causation. But first a preliminary point. In the last section I ignored any problems which may be involved in discovering the kind of objective probabilities which are symbolized in equations (1) and (2). This was not because I think there aren't any such problems, but rather because I think experimental randomization has nothing to do with them. Experimental randomization does its work *after* we have formed judgements about objective probabilities, and at the stage when we want to say what those probabilities tell us about *causes*.

Let me now try to explain exactly why experimental randomization is such a good guide to causes, *if* we know about objective probabilities. The notion of causation is of course philosophically problematic in various respects. But here we need only the simple assumption that a generic event like the treatment T is a cause of a generic event like the recovery R if and only if there are contexts

(perhaps involving other unknown factors) in which T fixes an above-average, single-case objective probability for R. In effect, this is to say that the treatment causes the recovery just in case there are *some* kinds of patients in whom the treatment increases the chance of recovery.[4]

Once we have this assumption about causation, we can see why experimental randomization matters. For this assumption implies that if Prob(R/T) is greater than Prob(R/−T)—if the objective probability of recovery for treated people in the community at large is higher than that for untreated people—then *either* this is because T itself causes R, *or* it is because T is correlated with one or more other factors which cause R. In the first case the T–R correlation will be due at least in part to the fact that T itself fixes an above-average, single-case probability for R; in the second case the T–R correlation will be due to the fact that T is correlated with other causes which do this, even though T itself does not. The problem we faced in the last section was that a difference between Prob(R/T) and Prob(R/−T) in the community at large does not discriminate between these two possibilities: in particular it does not allow us to eliminate the second possibility in favour of the hypothesis that T does cause R.

But this is precisely what a randomized experiment does. Suppose that Prob(R/T) and Prob(R/−T) represent the probabilistic difference between the two groups of patients in a randomized experiment, rather than in the community at large. 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 we can rule out the possibility of a spurious correlation, and be sure that T does cause R.[5]

A useful way to think of experimental randomization is as a way of *switching* the underlying probability space. The point of experimental randomization is to ensure that the probability space from which we are sampling is a good guide to causes. Before the experiment, when we were simply getting our probabilities from survey statistics, we were sampling from a probability space in which the treatment might be correlated with other influences on recovery; in the randomized experiment, by contrast, we are sampling from a probability space in which the treatment cannot be so correlated.

[4] Some philosophers hold that T causses R iff T increases the chance of R in *all* contexts. (*Cf.* Eells and Sober [1983]; Eells [1987]; Humphreys [1989]; Cartwright [1989], Ch. 4.) While this assumption of 'causal unanimity' is certainly required by many familiar quantitative linear causal models, and is no doubt satisfied in some real-world cases, I see no reason for building it into our thinking about causation. Apart from anything else, this assumption would make it very difficult to explain why a non-spurious positive correlation is in general a valid sufficient (though not necessary) indicator of causation. (For further arguments against the assumption of causal unanimity, see Dupré [1984, 1990].)

[5] For a more detailed account of why causation follows from a T–R correlation, plus probabilistic independence of T from R's other causes, see Papineau [1985, 1989].

This way of viewing things makes it clear that experimental randomization is nothing to do with statistical inferences from finite sample data to objective probabilities. For we face the problem of statistical inference when performing a randomized experiment as much as when conducting a survey: what do the sample data tell us about population probabilities like Prob(R/T) and Prob(R/−T)? But only in a randomized experiment does a solution to this problem of statistical inference allow us to draw a secure conclusion about causes.

I shall return to these points below. But first it will be helpful to comment on one particular claim made by Urbach. For the moment we can continue to put issues of statistical inference to one side.

7 RANDOMIZED EXPERIMENTS HELP WITH *UNKNOWN* NUISANCE VARIABLES

One of Urbach's main reasons for denying that randomization lends superior objectivity to causal conclusions is that the decision about which factors to 'randomize' will inevitably depend on the experimenter's personal assumptions about which 'nuisance variables' might be affecting recovery (Urbach [1985], pp. 265, 271; Howson and Urbach [1989], pp. 150–2).

This seems to me to betray a misunderstanding of the point of randomized experiments.[6] Randomized experiments are important, not because they help with the nuisance variables we think we know about, but because they enable us to cope with all those we *don't* know about. If we can identify some specific variable N which might be affecting recovery, we can deal with the danger without conducting a randomized experiment. Instead, we can simply attend to the probabilities conditional on N in the community at large, as in (2) above, and see whether T still makes a difference to R among people who are alike in respect of N. It is precisely when we don't have any further ideas about which Ns to conditionalize on that randomized experiments come into their own. For when we assign the treatment to subjects at random, we ensure that all such influences, *whatever they may be*, are probabilistically independent of the treatment.

If we had a complete list of all the factors that matter to the probability of recovery, we could bypass the need for experiment, and use the probabilities we get from non-experimental surveys to tell us about causes. But, of course, we rarely have such a complete list, which is why randomized experiments are so useful.

Urbach writes (Urbach [1985] p. 271; Howson and Urbach [1989] pp. 153, 253) as if the alternative to a randomized experiment were a 'controlled' experiment, in which we explicitly ensure that the nuisance Ns are 'matched' across treatment and control group (for exmple, we might explicitly ensure

[6] A misunderstanding which is also present in some of the comments on Urbach in Mayo [1987].

that the two groups have the same distribution of ages). But this is a red herring. As I have said, if we know which Ns to match, then we don't need to do an experiment which matches them; indeed we don't need to do an experiment at all. Instead we can simply observe the relevant multiple conditional probabilities in the community at large. (If we like, we can think of this as using *nature's* division of the treatment and control groups into subgroups matched for the Ns.) It is only when we don't know which Ns to match that an experiment, with its potential for randomization, is called for.[7]

5 BAYESIANS VERSUS CLASSICISTS ON STATISTICAL INFERENCE

I suspect that Urbach has been led astray by failing to distinguish the specific question of experimental randomization from general issues of statistical inference. In his article on 'Randomization and the Design of Experiments' [1985] he explains that by 'the principle of randomization' he means (following Kendall and Stuart [1963], Vol. 3, p. 121):

> Whenever experimental units (e.g. plots of land, patients, etc.) are assigned to factor-combinations (e.g. seeds of different kinds, drugs, etc.) in an experiment, this should be done by a random experiment using equal probabilities.

This is the kind of experimental randomization we have been concerned with in this paper so far. But on the next page of the article Urbach says:

> The fundamental reason given by Fisher and his followers for randomizing is that it supposedly provides the justification for a significance test. ([1985], p. 259)

As a Bayesian about statistical inference, Urbach disagrees with Fisher and other classical statisticians on the importance of significance tests, for reasons I shall shortly explain. And on this basis he concludes that experimental randomization is unimportant. But the inference is invalid, for Urbach is wrong to suppose that the rationale for experimental randomization is to justify significance tests. What justifies significance tests are random samples. But these are a different matter from experimental radnomization.

In this section I shall briefly explain why Bayesians like Urbach disagree with classical statisticians like Fisher about the importance of significance tests and therefore of random sampling.[8] In the next section I shall explain why this dispute about random sampling is irrelevant to questions of experimental randomization.

[7] Urbach also observes that randomization provides no guarantee that the intuitively salient aspect of the treatment is in fact the casually efficacious aspect (Urbach [1985], p. 264; Urbach and Howson [1989], p. 149). On this point, which provides the rationale for double blinds and placebos, I fully agree. Still, randomized experiments do at least show that recovery is caused by something the experimenter *does* to the treated subjects, rather than something which merely happens to be correlated with the treatment in the community at large.

[8] My remarks in this section agree closely with Urbach [1989]. See also Johnstone [1989].

Suppose we want to evaluate some stochastic hypothesis H, about the probabilities of outcomes from some objective probability space, on the basis of sample statistic E. Classicists and Bayesians give different accounts of how this should be done.

The classical account is in terms of significance tests. Classicists say we should reject H if E falls in the 'rejection region', that is, if E falls in some chosen region with a low probability (normally 5 per cent) of containing E if H is true. The rationale for this strategy is that the objective probability of erroneously rejecting H when it is true (a 'type I error) is then itself only 5 per cent—since the rejection region is precisely designed to have a 5 per cent probability of containing E *if* H is true.

Bayesians, by contrast, invoke Bayes' theorem. They say that when we observe E, we should increase the personal probability we attach to H in proportion to our prior personal probability for E, given H, and in inverse proportion to our prior personal probability for E: that is, H should be favoured to the extent its acceptance would have increased our expectation of E.

Classicists dislike the Bayesian approach because it invokes personal probabilities. They claim that the classical account of significance tests (of type I errors, anyway) appeals only to the objective probability that H implies for E.

But in fact it's not quite that simple, and this is where the issue of random sampling comes in. A stochastic hypothesis H will only imply an objective probability for a sample statistic E if it is conjoined with some assumption about the probabilities that the sampling mechanism implies for different kinds of samples. For without such an extra assumption there is no way of getting from the objective probabilities specified by H to an objective probability for the sample statistic E. The normal form of the requisite assumption is that the sample displaying E is drawn from the probability space in an *objectively random* way, that is, all individuals, whatever their characteristics, have an equal objective probability of being sampled.

So classicists need an assumption of random sampling (or some alternative assumption about sampling probabilities) in order to derive objective probabilities for sample statistics. From the Bayesian point of view, by contrast, the requirement of random sampling is otiose. A Bayesian statistical inference depends only on your personal probability for E, given H, and not on whether this value rests entirely on objective probabilities. As a Bayesian you *might* assign a given personal probability to E, given H, because you believe an assumption of objective random sampling. But you might equally well have the same personal probability for E, given H, not because you assume random sampling, but simply because you lack any reason to think that the sample displaying E is unrepresentative in any particular way. And in both cases, argue the Bayesians, you will have an equally good basis for a statistical inference.

Urbach adds a specific criticism of the classical theory. He points out that it

has difficulty with cases where the sample displaying E is randomly generated, but we have reason to believe *post hoc* that it *is* unrepresentative in some particular way (Urbach [1989], pp. 146, 162–3). Suppose we are testing the hypothesis H that the average height of children in a school is 5′ or more. We sample at random, but by chance end up with a sample containing only children from the youngest class, giving a mean sample height of 3′ 2″. Common sense would say that this is a freak sample and so no basis for rejecting H. The classical theory, however, seems to imply that we ought to reject H, since the objective probability of getting such a low sample mean, given H and random sampling, was very small.[9] Bayesians, on the other hand, can reason that a sample mean in this range, given H *and* given that the sample is unrepresentatively young, is highly probable, and so can avoid rejecting H on the basis of E.

6 TWO STEPS TO CAUSAL CONCLUSIONS

As I said earlier, I am happy to concur with Urbach about statistical inference. I have no dispute with his thesis that classical insistence on using random samples for statistical inference is both unnecessary and misguided. My disagreement with Urbach is only that this is nothing to do with the use of randomized experiments to establish causal claims.

To see what is going on, let us once more divide the question of whether some treatment T causes R into two distinct stages. First, there is the question of whether the conditional probabilities Prob(R/T) and Prob(R/−T) in the underlying probability space are unequal. If they are, the second question then arises: is this because T causes R?

The first question is a question of statistical inference. We want to get from the frequencies with which R, T, and −T are found together in our sample to a conclusion about the underlying probabilities of R given T and not−T. This is the point at which Urbach takes issue with the classicists: they say the samples in question must be generated randomly, Urbach denies this.

But this leaves the second question untouched. Suppose that we conclude, either on Bayesian or on classicial grounds, that the underlying probability of R is different given T and not−T. Whichever way we reached this conclusion, we will still face the further question: is this difference due to T causing R? For, whether we are Bayesians or classicists, there remains the possibility that T is objectively correlated with R, but not because it causes R itself, but because it is objectively correlated with something else which does. And this is where I say

[9] Classicists can argue that a practice of 'stratified sampling', which builds up an overall sample by sampling separately from each class in the school, would be less likely to yield a type I error than simply sampling the whole school. But, as Urbach insists, this doesn't answer the objection. The problem with the *post hoc* unrepresentative non-stratified sample is not that it is an instance of a *general* practice which yields lower significance levels than stratified sampling, but that in this *particular* case it is telling us to reject when we manifestly shouldn't.

a randomized experiment can help. For, as I put it earlier, a randomized experiment will switch the underlying probability space to one in which T definitely *isn't* objectively correlated with any such other causes.

So there are two quite different ways in which *randomness* can enter into the kind of investigation at issue. First (random sampling), the overall *sample* can be *randomly drawn*—that is, every individual can have the same objective probability of entering the overall sample. Second (randomized experimentation), the *treatment* can be *randomly assigned*—however the sample is generated, all sampled individuals can have the same objective probability of getting the treatment. These two notions are two-way independent. Even if we insist, with the classicists, on *random samples* when surveying a non-experimental population to estimate the underlying probabilities, this does not mean that the treatment has been randomly assigned (and so yields no basis for concluding that a T–R correlation indicates causation). Conversely, even a Bayesian, like Urbach, who is prepared to draw statistical inferences from *non-random samples*, still has every reason to require that the treatment be randomly assigned (since even Bayesians need this further information to move from correlations to causes).

At first sight this latter notion, of a random assignment of a treatment within a non-random sample, might seem puzzling. But this is a familiar, indeed normal, situation in medical research. The sample of patients suffering from the disease is gathered relatively haphazardly. But once this sample has been gathered, then every care is taken to ensure that all its members have the same objective probability of getting the treatment. Urbach maintains, and I concur, that the initial non-randomness of the sample is no barrier to our using it to estimate objective conditional probabilities. But what he fails to recognize is that the subsequent randomized assignment of the treatment is crucial to our drawing causal conclusions from these conditional probabilities.

7 CAUSAL INFERENCES IN BAYESIAN TERMS

Bayesians may remain puzzled. I say experimental randomization of the treatment is relevant to the inference from probabilities to causes, rather than to the prior inference from statistics to probabilities. But still, this latter inference, from probabilities to causes, ought itself to be representable, like all scientific inferences, in Bayesian terms. And, once we do so represent it, then won't the requirement of objective randomization of the treatment be exposed as otiose, just as was the classicist's demand for objective random sampling?

No. We can indeed represent the inference from randomized-experimental probabilities to causes in Bayesian terms.[10] However, when we do so, it comes

[10] For two other attempts to justify randomization in Bayesian terms, neither of which, however, seems to me to distinguish sufficiently sharply between sampling and causal issues, see Rubin [1978] and Swijtink [1982].

out as the extreme case of a deductive inference, and indeed a deductive inference which goes through precisely because of the randomization of the treatment. Think of the inference in question as follows. The hypothesis H is that *T causes R*. The evidence E is that the *objective Prob(R/T) is greater than the objective Prob(R/−T) when the treatment is randomized*. Given this H and this E, the prior personal probability of H, given E, is *one*, since the randomization ensures that the probability of E is *zero* on any hyhpothesis other than H. So, if we knew E, then, by Bayesian conditionalization, we could be certain of H. (I am not forgetting the problem of getting *to* this E from finite sample data. But for the moment we are concerned specifically with the Bayesian logic of moving *from* this E to this H.)

This argument simply transposes a line of reasoning outlined in Section 3 above into Bayesian terms. In Section 3 I made a minimal assumption about causation, namely, that for T to cause R, there must be some contexts in which T fixes a higher than average single-case probability for R. It followed from this that there are only two possibilities consistent with an objective correlation between T and R: either T itself causes R, or T is objectively correlated with one or more other factors which are relevant to the single-case probability of R. So if we can be sure that T is *not* objectively correlated with any other possible causes of R, which is what experimental randomization tells us, then we can be sure that an objective T–R correlation means that T causes R.

One advantage of putting this argument in Bayesian terms is that it shows why randomized experimentation is *not* dispensable when inferring causes from probabilities, in the way that random samples arguably are when inferring probabilities from sample statistics. When we infer probabilities from sample statistics, a non-random sample and a random sample can yield just the same conditional personal probability for statistic E given the H under investigation, and so can underpin just the same statistical inference. But when we infer causes from probabilities, a non-experimental survey investigation certainly does not make it reasonable to give the same *zero* conditional personal probability to the claim (E) that T is correlated with R, on the hypothesis (H) that T does not cause R, that we can give it after a randomized experiment: even if you've got no special reason to suspect the presence of any confounding influences in the community at large, this is not the same as being certain that there aren't any, as you can be after the randomization of T in an experiment.

8 *POST HOC* UNREPRESENTATIVENESS IN A RANDOMIZED EXPERIMENT

Distinguishing the two inferential steps involved in inferences to causes enables us to deal with a difficult case raised by Urbach (Urbach [1985], p. 260; Urbach and Howson [1989], pp. 151–2). 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 the 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.

I agree that, if you think age might matter to recovery, then you would be foolish to infer the efficacy of T solely from a difference in recovery rates between a young group who get T and an old group who do not, however much the assignment of the individuals in the sample to the T and not−T groups was randomly arranged. However, I don't think that this counts against my defence of experimental randomization.

Once more, we need to distinguish two steps involved in inferring causes from finite sample data. Suppose a slapdash researcher—let us call him Quentin Quick, say—were to infer the efficacy of the treatment from the observed difference in recovery rates in the finite sample in question. On my analysis, Quentin has made a two-stage inference. First, Quentin has inferred objective population probabilities from sample statistics. Second, he has inferred causes from those objective probabilities. I still want to maintain, in line with my overall argument, that this second inference, from probabilities to causes, is quite infallible, in virtue of the randomization of the treatment in the experiment at hand. Quentin's error lies, rather, in his first step, from the sample data to objective probabilities, and it is this invalid first step that is responsible for his flawed eventual causal conclusion.

My point is simply that, *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?

However, as I said, Quentin's prior inference, from the sample data to probabilities, is fallacious. Indeed, we have already considered an entirely analogous inferential fallacy, in our earlier discussion of Bayesian versus classical accounts of statistical inference. Quentin's mistake is simply a variant of the case Urbach uses to argue against the classical theory. Urbach's argument, recall, was that classicists have trouble explaining what is wrong with significance tests based on random samples which we can see to be unrepresentative *post hoc*. The case at hand is an illustration. 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 the 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 of the observed sample correlation as the alternative hypothesis that T and R are objectively correlated. And this, as before, supports the Bayesian over the classical theory of statistical inference, since the Bayesian theory, unlike the classical theory, can happily accommodate this sensible reasoning.

Still, all this relates to Quentin's first inferential step, and is irrelevant to my claims about his second step. To repeat my central point, I agree with Urbach that the classical insistence of *random sampling* stems from a bad theory of *statistical inference*, and in particular may prevent us from explaining what is wrong with someone who infers an objective correlation from a manifestly unrepresentative sample. However, we shouldn't conclude from this that the *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.

9 DOING WITHOUT RANDOMIZATION

At the beginning of this paper I observed that there are often ethical objections to randomized experiments in medicine. The standard response by defenders of such experimentation is that the benefits of medical knowledge outweigh the ethical drawbacks of experimentation.

This response, however, assumes that randomized experiments are the *only* way to establish causal conclusions in medicine. So far in this paper I have been concerned to show that randomized experiments are a *good* way to find out about causes. But it is also a corollary of my analysis that randomized experiments are *not* the only way to find out about causes. It will be worth briefly explaining why, in the interests of making it clear that the ethical dangers of random experimentation are not always necessary for the good of medical knowledge.

As we have seen, the virtue of randomized experiments, as opposed to non-experimental surveys, is that they ensure that any further *unidentified* nuisance variables are *uncorrelated* with the treatment variable. It follows, then, that surveys will do just as well as randomized experiments whenever we are able to *identify* all those influences on recovery which *are correlated* with treatment.

In general this won't be an easy task. But note that it will certainly be a lot easier than identifying *all* influences on recovery *tout court*. There will

inevitably be a large number of Ns that affect the probability of recovery from any given medical ailment to some extent. But it is important that by no means all of these Ns will themselves be objectively correlated with the treatment T, and so not all of them will be capable of producing a spurious correlation. It is specifically Ns which *are* correlated with T which threaten this, and it is specifically these possibly confounding Ns that experimental randomization guards against. So if only we can identify those specific Ns which are correlated with T, we will be in a position to dispense with experimental randomization.

What is more, it is arguable that any given N, like age, will only be probabilistically associated with T in the general community if there is some explanation for this association (such as that doctors are more assiduous in treating younger people, or that younger people tend to have younger doctors who tend to know more about new drugs, or some such). So if only we can identify the limited number of influences on recovery which could conceivably have some connection with T, then we can ignore the rest. For then the unknown Ns will be uncorrelated with T, and our survey will be as good as a randomized experiment, namely, a sure-fire route to a causal conclusion.

Randomized experiments have the advantage of releasing us from the responsibility of identifying every N which is correlated with T. But this is not an impossible responsibility, and when there are ethical objections to randomized experiments we should try to shoulder it.[11]

[11] I would like to thank Peter Urbach for many helpful comments on this paper.

REFERENCES

CARTWRIGHT, N. [1989]: *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

DUPRÉ, J. [1984]: 'Probabilistic Causality Emancipated', in P. A. French, T. E. Uehling Jr., and H. K. Wettstein (*eds.*), *Midwest Studies in Philosophy IX: Causation and Causal Theories*. Minneapolis: University of Minnesota Press, pp. 169–75.

DUPRÉ, J. [1990]: 'Probabilistic Causality: A Rejoinder to Ellery Eells', *Philosophy of Science*, **57**, pp. 690–8.

EELLS, E. [1987]: 'Probabilistic Causality: Reply to John Dupré', *Philosophy of Science*, **54**, pp. 105–14.

EELLS, E. and SOBER, E. [1983]: 'Probabilistic Causality and the Question of Transitivity', *Philosophy of Science*, **50**, pp. 35–57.

HACKING, I. [1988]: 'Telepathy: Origins of Randomization in Experimental Design', *Isis*, **79**, pp. 427–51.

HOWSON, C. and URBACH, P. [1989]: *Scientific Reasoning: The Bayesian Approach*. La Salle, Ill.: Open Court.

HUMPHREYS, P. [1989]: 'Scientific Explanation: The Causes, Some of the Causes, and Nothing But the Causes', in P. Kitcher and W. Salmon (*eds.*), *Minnesota Studies in the Philosophy of Science XIII: Scientific Explanation*. Minneapolis: University of Minnesota Press, pp. 283–306.

JOHNSTONE, D. [1989]: 'On the Necessity for Random Sampling', *British Journal for the Philosophy of Science*, **40**, pp. 443–57.

KENDALL, M. G. and STUART, A. [1963]: *The Advanced Theory of Statistics*. London: Charles Griffin.

MAYO, O. [1987]: 'Comments on "Randomization and the Design of Experiments" by P. Urbach', *Philosophy of Science*, **54**, pp. 592–6.

PAPINEAU, D. [1985]: 'Probabilities and Causes', *Journal of Philosophy*, **82**, pp. 57–74.

PAPINEAU, D. [1989]: 'Pure, Mixed, and Spurious Probabilities and their Significance for a Reductionist Theory of Causation', in P. Kitcher and W. Salmon (*eds.*), *Minnesota Studies in the Philosophy of Science XIII: Scientific Explanation*. Minneapolis: University of Minnesota Press, pp. 307–48.

RUBIN, D. B. [1987]: 'Bayesian Inferences for Causal Effects: The Role of Randomization', *Annals of Statistics*, **6**, pp. 34–58.

SWIJTINK, Z. [1982]: 'A Bayesian Justification for Experimental Randomization', in P. Asquith and T. Nickles (*eds.*), *PSA 1982*, Vol. 1. East Lansing: Philosophy of Science Association, pp. 159–68.

URBACH, P. [1985] 'Randomization and the Design of Experiments', *Philosophy of Science*, **52**, pp. 256–73.

URBACH, P. [1989]: 'Random Sampling and Methods of Estimation', *Proceedings of the Aristotelian Society*, **89**, pp. 143–64.