Rerum Causae, 6 (2014), 1-10

The Case Against the Special Epistemic Power of Randomised Control Trials

David Kinney*

Abstract

This paper evaluates and ultimately rejects the claim that Randomised Controlled Trials are a superior means of testing causal hypotheses. I argue that three common claims in favour of the epistemic power of RCTsthat they eliminate known nuisance variables, that they eliminate unknown nuisance variables, and that they avoid selection biasultimately fail to achieve their argumentative ends. Based on this argument, I conclude that rather than uncritically accept RCTs as the best test of causality, researchers should be prepared to use a variety of experimental methods.

Introduction

The literature surrounding evidence-based movements in policy and medicine often suggests a hierarchy of methods for testing causal hypotheses. At the top of this hierarchy are Randomised control trials (RCTs). In this paper, I argue against the idea that RCTs have any special epistemic power for testing causal hypotheses. In Sections I and II, I define the specific concepts relevant

* I would like to thank Katie Steele for all her help with this essay.

to this debate, as the notion of an epistemically powerful test of a causal hypothesis carries substantial conceptual baggage. In Sections III, IV and V, I argue that three common claims in favour of the epistemic power of RCTs – that they eliminate known nuisance variables, that they eliminate unknown nuisance variables, and that they avoid selection bias – ultimately fail to achieve their argumentative ends.

I. Randomised Control Trials

What is a Randomised control trial? An RCT is an experiment characterised by the following procedure: the experimenter randomly selects a test group of individuals from a larger target population and then randomly assigns each member of the test group to either the treatment group or the control group. The treatment group receives the intervention, and the control group does not. The effect on both groups is then measured. If a specific effect is observed in the treatment group with greater frequency than in the control group, then this is counted as evidence in favour of the causal hypothesis. If the effect is equally observed in both groups, then this counts as evidence against the causal hypothesis. Among practitioners of any given discipline, features such as double blindness, the use of a placebo on the control group, or specific randomisation methods may be considered a standard part of an RCT. For my purposes, however, I am content to define an RCT solely in terms of the basic features described above.

Determining whether RCTs have any superior epistemic power for testing causal hypotheses requires comparing them to other such methods. These methods include:

1. Observational and cohort studies, wherein the treatment and control groups are not randomly assigned. Instead, subjects who did or did not receive a given treatment are observed post-treatment in order to try and make an inference regarding that treatment's effect.

2. Case control studies, in which the histories of groups that do or do not have some symptom are traced back to see if there something common to the effected group as opposed to the non-effected group.

3. Theoretical research, e.g. a chemist showing that the molecular makeup of some medicine is likely to have a certain effect on the human body, based on foundational principles of chemistry and physics.

This list is not meant to be exhaustive of all methods of causal inference. Rather, it is a general survey of the "competition" that RCTs face.

The salient features of RCTs, as compared with these other methods of testing causal hypotheses, are their interventional and random character. By 'interventional' I mean just that in an RCT, the experimenter provides or withholds the intervention regardless of the intentions or preferences of each subject. By 'random' I mean specifically that RCTs select the test, treatment and control groups in a way such that each individual in the target population is assigned the same initial likelihood of being assigned to each group. For example, imagine that each person in a target population flips a fair coin. If it comes up heads, they enter the test group, if tails, they do not. Then the members of the test group each flip the same coin; heads results go to the treatment group, tails to the control group. This would mean that the groups were Randomised. We can compare this method to observational or case control studies, wherein each subject's individual decision, rather than chance, determines the makeup of each group.

II. Epistemic Power, Causal Hypotheses and Nuisance Variables

Having laid out the structure of an RCT, I can address two other relevant concepts: 'epistemic power' and 'causal hypothesis.' For my purposes, 'epistemic power' refers to the extent to which an experiment is capable of producing results that should raise the experimenter's beliefs about the truth or falsehood of

a given hypothesis. Note that I am not suggesting that any test could ever prove a hypothesis to be true. Empirical evidence can only effect the belief in the truth of the hypothesis.

As for the term 'causal hypothesis,' I follow Popper in defining 'hypothesis' as any statement that empirical evidence could show to be false.[1] A 'causal hypothesis' is just a hypothesis of the form 'C causes E.' This definition invites an obvious question: What does it mean for C to cause E? Although a thoroughgoing theory of causation is well beyond the scope of this paper, Nancy Cartwright's statistical definition of causation will suffice for my purposes. Her definition states that 'C causes E if and only if C increases the probability of E in every situation which is otherwise causally homogenous with regard to E.'[2] Two situations X and Y are causally homogenous with regard to E if and only if no factor other than a single cause C can explain a difference in the probability of E between X and Y.

Causal homogeneity must be included in Cartwright's definition of causation due to the potential for nuisance variables to influence statistical correlation. For an example of a nuisance variable, imagine that scientists want to test whether a new drug improves liver health. If they tested the drug on a group of heavy drinkers, the negative effect of heavy drinking on liver health may statistically outweigh the positive effect of the drug. As a result, the test would tell them nothing about the effect of the drug on liver health, even if such a causal relationship exists. Similarly, imagine that scientists are testing a drug not for its ability to improve liver health, but rather to see whether it causes liver damage as a side effect. In this case, too many heavy drinkers in the treatment or control groups would render the test useless, since scientists would be unable to attribute any observed decline in the treatment group solely to the drug.

In both of these cases heavy drinking is a nuisance variable. More generally, a nuisance variable is any variable other than the intervention that, if it is overly

[1] Thornton, 'Karl Popper'

[2] Cartwright, 'Causal Laws and Effective Strategies', 423.

present in any of the experimental groups, may effect the results of the experiment so as to inspire undue belief in the truth or falsity of the causal hypothesis. As they relate to Cartwright's definition of causation, nuisance variables are sufficient to undermine the ability of an experimental method to create a causally homogenous situation with respect to a given effect. So an experimental method that is able to eliminate all nuisance variables is a more epistemically powerful test of a causal hypothesis than any method that is unable to do so, all other things being equal.

III. Known Nuisance Variables

Having unpacked the necessary conceptual baggage, I am now in a position to address the claim that RCTs have a special epistemic power for testing causal hypotheses. One way of advancing this claim is to argue that the process of randomisation is uniquely effective in balancing the experimental group and the control group for nuisance variables, thereby creating something like a causally homogenous situation. By assigning each subject an equal probability of receiving the intervention, randomisation ensures that receipt of the intervention is dependent only on chance, and not on the presence of any nuisance variable. We can compare this with any non-Randomised test, wherein the non-interventional role of the experimenter makes it impossible in principle to remove any correlation between receipt of the treatment and some other causally relevant variable. Therefore, a Randomised experiment 'eliminate[s] the danger of any confounding factors which might be responsible for a spurious correlation.'[3]

John Worrall notes that this claim is, 'if taken literally, trivially unsustainable.'[4] A truly random assignment of each person into each group could easily yield an overabundance of a given nuisance in any one group; certainly nothing about randomisation precludes this from occurring. So a single instance of ran-

[3] Papineau, 'The Virtues of Randomisation', 439.

[4] Worrall, 'Evidence in Evidence-Based Medicine', 322.

domisation does not necessarily provide the kind of causal homogeneity that is required for testing causal hypotheses.

However, suppose that the same RCT is performed repeatedly over many trials. The mechanics of probability are such that over time it is unlikely that a large number of those RCTs would contain groups that were unbalanced with regard to any one confounding factor.[5] Further, scientists can be pragmatic and withhold treatment until a Randomised assignment of control and treatment groups is not unbalanced with respect to any known nuisance variables. This method seems like it works to create causal homogeneity, and therefore may count in favour of the epistemic power of RCTs.

Are randomisation and re-randomisation the most efficient ways of eliminating known nuisance variables? I do not think so. In the case given earlier, it seems obvious based on prior medical knowledge that scientists should not test a drug for its effect on liver health using any experimental group with an overabundance of heavy drinkers. An experimenter could achieve this end by randomising and re-randomising until such groups are formed, or she could just create the control and treatment groups herself and deliberately ensure that neither contains too many heavy drinkers. These two methods both eliminate the nuisance variable, and the latter does so without any potential need for inefficient multiple randomisations. So it does not seem that randomisation is superior to other methods in its ability to eliminate known confounders.

IV. Unknown Nuisance Variables

For many proponents of RCTs, the point of randomisation is not to eliminate known nuisance variables, but rather to eliminate these variables when they are unknown to the experimenter.[6] Returning to the example of the drug for liver health, suppose that the experimenter is doing an observational study, rather

[5] Worrall, 'Evidence in Evidence-Based Medicine', 323.

[6] Papineau, 'The Virtues of Randomisation', 441.

than an RCT, to study the drug's effect on liver health. However, she has no knowledge of the fact that people who seek treatment for liver health also tend to start exercising around the same time as they begin treatment with the drug. She might falsely conclude that the drug has the effect of improving liver health, when in reality it is exercise that is the cause of the improvement.

David Papineau argues that randomisation is the best method for giving the experimenter any chance of eliminating unknown nuisance variables. Since the experimenter does not know anything about the nature of the nuisance variables in question, she cannot engineer the experiment in any way so as to completely eliminate their influence. However, randomisation allows her to eliminate any dependency relation between a subject's receiving the intervention and a subject's possessing a given nuisance variable. This is because the only factor that affects the subject's receipt of the intervention is random chance. By contrast, in the observational study above, it may be impossible to separate the exercisers from the people who took the drug; perhaps the two actions may be are linked by a common desire on the part of some subjects to improve their health.

The problem with this argument is that it commits what Worrall calls a 'quantificational fallacy' regarding the nature of unknown variables.[7] In any given experiment, the number of unknown possible nuisance variables is indefinite. Therefore, any random assignment of subjects to different groups has a very high likelihood of being unbalanced with regard to any one member of the indefinitely large set of possible confounders and mutual enhancers.

I believe that Papineau's mistake here is to over-emphasise the practical importance of the distinction between those nuisance variables that cause experimental groups to be unbalanced solely as a matter of accident, and those nuisance variables that create unbalanced groups because they are somehow related to each subject's receipt of the intervention. Accidental imbalances can be just as pernicious as any other imbalance in terms of their ability to inspire undue belief in the truth or falsity of a causal hypothesis. Given an indefinitely

[7] Worrall, 'Evidence in Evidence-Based Medicine', 324.

large set of unknown possible nuisance variables, randomisation does nothing to remove these accidental imbalances from a given study. Therefore, randomisation does not justify a claim for the superior epistemic power of RCTs on the grounds that they eliminate unknown nuisance variables.

V. Selection Bias

One could argue that the superior epistemic power of RCTs is granted because randomisation is the best way to eliminate selection bias. Selection bias occurs where some prejudice of the experimenter effects how she assigns subjects to different experimental groups. The possibility of selection bias should lower the epistemic power of a test; this bias is a kind of nuisance variable that can prevent causal homogeneity with respect to a given effect. In a Randomised trial, it is argued, the experimenter does not influence the composition of the various experimental groups.[8] Therefore, RCTs avoid selection bias.

However, not all disinterested methods of selecting subjects are random.[9] For example, an experimenter could assign subjects to the experimental or control group based on whether their birthday fell on an even or odd day of the month. Here, the selection of the groups is not random because each subject already has probability 0 or 1 of being sorted into either group. This lack of randomness does not detract from the causal homogeneity of the situation; it would certainly strain any notion of scientific rationality to suggest that the numerical character of a person's date of birth had a causal relationship with liver failure. What is doing the work of eliminating selection bias is not randomisation but rather a disinterested method of assigning subjects to groups. Since disinterested selection of who receives treatment is a necessary but not sufficient condition of an RCT, the elimination of selection biases is not a reason to grant that RCTs have any special epistemic power for testing causal hypotheses.

[8] Worrall, 'Evidence in Evidence-Based Medicine', 325.

[9] La Caze, *Evidence-based Medicine: Evolution, Revolution or Illusion?*, 73.

Conclusion

Randomised control trials are not a bad way of testing causal hypotheses. Certainly, nothing here should be taken as an argument that RCTs are, in principle, worse for this purpose than other scientific methods. However, when the concept of epistemic power for testing causal hypotheses is taken seriously, I cannot find a philosophically satisfying justification for the superior status of RCTs. A fully-fledged program for how, specifically, scientists should test causal hypotheses would require significantly more work than I have space for here. However, my argument suggests that evidence-based movements would be well served by getting rid of hierarchies, and conceding that there are several justifiable scientific methods for testing causal hypotheses. The question of which method is best may need to be answered on a case-by-case basis, rather than as a matter of philosophical principle.

David Kinney

References

[1] Cartwright, N. 'Causal Laws and Effective Strategies', *Noûs*, 13, 1979: 419-437

[2] La Caze, A. 'Evidence-based Medicine Must Be…', in *Evidence-based Medicine: Evolution, Revolution or Illusion?*, PhD Thesis.

[3] Papineau, D., 'The Virtues of Randomisation', *The British Journal for the Philosophy of Science*, 45(2), 1994: 437–450.

[4] Thornton, S., 'Karl Popper', *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.)

[5] Worrall, J. 'What Evidence in Evidence-Based Medicine?', *Philosophy of Science*, 69, 2002: 316–330.

David Kinney is a MSc student in Philosophy and Public Policy at the London School of Economics and Political Science (2013-2014). He also holds a degree in Philosophy from Dartmouth College. His main areas of interest are the relationship between scientific methodology and public policy, and competing notions of wellbeing. You can contact him at [davidbkinney@gmail.com].