LSE CPNSS

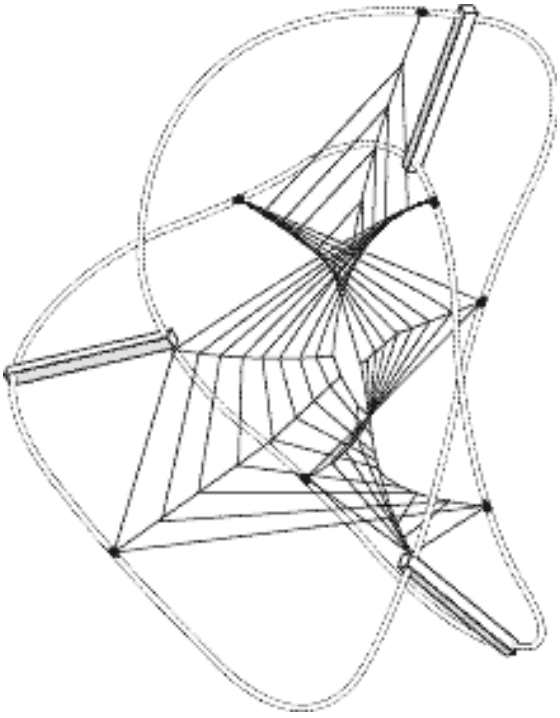
Centre for Philosophy of Natural and Social Science

Contingency and Dissent in Science

Technical Report 01/07

Are RCTs the Gold Standard?

Nancy Cartwright



Series Editor: Damien Fennell

The support of The Arts and Humanities Research Council (AHRC) is gratefully acknowledged. The work was part of the programme of the AHRC Contingency and Dissent in Science.

Published by the Contingency And Dissent in Science Project
Centre for Philosophy of Natural and Social Science
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

ISSN 1750-7952 (Print)
ISSN 1750-7960 (Online)

Are RCTs the gold standard?

Nancy Cartwright[1]

Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK

E-mail: n.l.cartwright@lse.ac.uk

Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0119

E-mail: ncartwright@ucsd.edu

Editor's Note

By exploring the conditions under which randomized controlled trials (RCTs) deductively imply their results, this paper makes explicit important assumptions on which RCTs depend. In this way, this paper contributes to the project's research on contingency in science and evidence, which focuses on the role and limits of evidence in reducing the contingency of scientific claims. This is particularly important for the RCT method, which is widely-used and treated as the 'gold standard' among experimental methods.

[1] I would like to thank participants of the BIOS 'Searching for Gold Standards Conference' June 2006 for their comments. This paper is forthcoming in the journal *BioSocieties*.

Abstract

The claims of RCTs to be the gold standard rest on the fact that the ideal RCT is a *deductive* method: if the assumptions of the test are met, a positive result *implies* the appropriate causal conclusion. This is a feature that RCTs share with a variety of other methods, which thus have equal claim to being a gold standard. This paper describes some of these other deductive methods and also some useful non-deductive methods, including the hypothetico-deductive method. It argues that with all deductive methods, the benefit that the conclusions follow deductively in the ideal case comes with a great cost: narrowness of scope. This is an instance of the familiar trade-off between internal and external validity. RCTs have high internal validity but the formal methodology puts severe constraints on the assumptions a target population must meet to justify exporting a conclusion from the test population to the target. The paper reviews one such set of assumptions to show the kind of knowledge required. The overall conclusion is that to draw causal inferences about a target population, which method is best depends case-by-case on what background knowledge we have or can come to obtain. There is no gold standard.

1. Introduction

The answer to the title question, I shall argue, is 'no'**.** There is no gold standard; no universally best method. Gold methods are whatever methods will provide a) the information you need, b) reliably, c) from what you can do and from what you can know on the occasion. Often Randomised Controlled Trials (RCTs) are very bad at this and other methods very good. What method best provides the information you

want reliably will differ from case to case, depending primarily on what you already know or can come to know.

Since I have no expertise in psychiatry, I shall discuss methods in general use in the human sciences without trying to approach special problems of psychiatry. The paper will have six parts:

I. Clinchers v Vouchers: A distinction and its implications

II. A Straddler: The hypothetico-deductive method

III. Examples of methods that clinch conclusions

IV. RCTs: Ideal RCTs, real RCTs and the scope of an RCT

V. The vanity of rigor in RCTs

VI. Closing remarks

Bits of part IV will rely on some formal results that I will present informally. I hope to convey a sense of the kind of information that is required to justify the claims of RCTs to be a gold standard as a basis for caution and for comparison with other methods that have an equal claim to this status (because they are what I shall call 'clinchers').

2. Clinchers v Vouchers: A distinction and its implications

Methods for warranting causal claims fall into two broad categories:

1. Those that *clinch* the conclusion but are *narrow* in their range of application, for example RCTs, derivation from theory or certain econometric methods.

2. Those that merely *vouch for* the conclusion but are *broad* in their range of application, for example qualitative comparative analysis, or looking for quantity and variety of evidence.

What is characteristic of methods in the first category is that they are deductive: *if* all the assumptions for their correct application are met, then if evidence claims of the appropriate form are true, so too will the conclusions be true. But these methods are concomitantly narrow in scope. The assumptions necessary for their successful application will have to be extremely restrictive and they can take only a very specialized type of evidence as input and special forms of conclusion as output. That is because it takes strong premises to deduce interesting conclusions and strong premises tend not to be widely true.

Methods in the second category are more wide-ranging but it cannot be proved that the conclusion is assured by the evidence, either because the method cannot be laid out in a way that lends itself to such a proof or because, by lights of the method itself, the evidence is symptomatic of the conclusion but not sufficient for it. What then is it to *vouch for?* That is hard to say since the relation between evidence and conclusion in these cases is not deductive and there are no general good practicable 'logics' of non-deductive confirmation, especially ones that make sense for the great variety of methods we use to provide warrant.

The fact that RCTs are a deductive method underwrites their claims to be the gold standard. But RCTs suffer, as do all deductive methods, from narrowness of scope. Their results are formally valid for the group enrolled in the study, but only for that

group. The method itself does not underwrite any strong claims for external validity, that is for extending whatever results are supposed to be established in the test population to other 'target' populations. This is important to keep in clear sight in comparing RCTs with other methods.

Compare then the costs and benefits of the two categories. Clinchers are deductive: *if* they are correctly applied *and* their assumptions are met, then *if* our evidence claims are true, so too will be our conclusions -- a huge benefit. But there is an equally huge cost. These methods are concomitantly narrow in scope. The assumptions necessary for their successful application a) tend to be extremely restrictive, b) can only take a very specialized type of evidence as input, and c) have only special forms of conclusion as output. In consequence we face a familiar kind of trade-off**:** We can ask for methods that clinch their conclusions but the conclusions are likely to be very limited in their range of application.

3. A Straddler: The hypothetico-deductive method

The hypothetico-deductive method is a straddler. Used one way – the way Karl Popper advocated – it is purely deductive and so is in the same category as the RCT. The method works, as all methods do, by presupposing a variety of auxiliary assumptions, otherwise nothing really follows from the hypothesis of interest.

Popper:

Hypothesis → outcome

¬outcome

Therefore, ¬hypothesis

This is a clincher.

Positivists:

Hypothesis → outcome

outcome

probability of the hypothesis increases (ceteris paribus)

This is a voucher.

Popper argued that the only correct use of the hypothetico-deductive method is as a clincher, to deduce that hypotheses are false. The argument accepted by the Positivists, he pointed out, is a deductive fallacy – the fallacy of affirming the consequent. And deductive logic, he maintained, is all the logic there is. This is borne out by centuries of failed efforts to establish some reasonable relatively uncontroversial theory of inductive confirmation. On the other hand, philosophers of physics maintain that the hypothetico-deductive method is the method by which physics theories are established. Nevertheless, medical science – and most of current evidence-based policy rhetoric – will not allow it.

Perhaps an example related to topics of interest to psychiatry will help. Consider the widespread correlation between low economic status and poor health and look at two opposing accounts of how it arises. (For a discussion and references see Cartwright

(forthcoming).) Epidemiologist Michael Marmot from University College London argues that the causal story looks like this:

Marmot:

Low status→ 'stress' → too much 'fight or flight' response → poor health

In contrast, Princeton University economist Angus Deaton suggests this:

Deaton:

Poor health → loss of work → low income → low status

Deaton confirms his hypothesis in the National Longitudinal Mortality Study (NLMS) data. He reasons: If the income-mortality correlation is due primarily to loss of income from poor health, then it should weaken dramatically in the retired population where health will not affect income. It should also be weaker among women than men, because the former have weaker attachment to the labour force over this period. In both cases these predictions are borne out by the data. Even more, split the data between diseases that something can be done about and those that nothing can be done about. Then income is correlated with mortality from both – just as it would be if causality runs from health to income. Also education is weaker or uncorrelated for the ones that nothing can be done about. Deaton argues that it is hard to see how this would follow if income and education are both markers for a single concept of socio-economic status that is causal for health.

Thus Deaton's hypothesis implies a number of specific results that are borne out in NLMS data and would not be expected on dominant alternative hypotheses. So the hypothesis seems to receive positive confirmation, at least if we share the Positivists' intuition. More carefully, it seems to receive some confirmation for the population sampled for the NLMS data. But what about other populations, i.e. what about *external validity*? The arguments I have just described that seem contra Popper to provide some evidence for Deaton's hypothesis in the population sampled do nothing as they stand to support any claims about alternative populations. More premises and more and different arguments are needed to do that. So here we are reminded how badly even a non-clinching method can suffer from problems of external validity.

4. Examples of methods that clinch conclusions

I list just a few other kinds of methods that work deductively.

1. Econometric methods
2. Galilean experiments
3. Probabilistic/Granger causality
4. Derivation from established theory
5. Tracing the causal process
6. Ideal RCTs

These are clinchers: It can be proved that if the auxiliary assumptions are true, the methods are applied correctly and the outcomes are true and have the right form,

then the hypothesis must be true. Even though I do not have the space to discuss them here, I mention them in order to stress that when it comes to clinchers – to methods from which the hypothesis can be rigorously derived from the evidence – RCTs are not the only game in town. There are lots of methods that can clinch conclusions.

It is important to keep in mind one caution, however. To buy the benefits of a clinching method we must be able to ensure that it is highly probable that *all* the requisite premises are obtained. That's because of the *weakest link* principle for deductive reasoning. The probability of the conclusion can be no higher than that of the weakest premise.

- Suppose you have 10 premises, 9 of them almost certain, one dicey. Your conclusion is highly insecure, not 90% probable.
- In a deductive argument P(conclusion) ≤ P(conjunction of premises)

I belabour this because of the benefits of clinching methods – clinchers are rigorous. It is transparent *why* the results are evidence: Given the background assumptions the hypothesis follows deductively from the results. And it is transparent *when* the results are evidence: When the background assumptions are met. This contrasts with ethnographic methods and expert judgment, for example. These can provide extremely reliable evidence. But there is no specific non-trivial list of assumptions that tell when they have done so. But if you want credit for this benefit of a clinching method, you must be able to show that the *conjunction* of your premises has high probability *in the case at hand*.

5. Randomised Controlled Trials

Ideal RCTs

I have claimed that ideal RCTs are clinchers. That of course depends on how they are defined. But there are perfectly natural definitions from which it can be proved that RCTs, as thus defined, allow causal claims about the population in the study to be deduced from probability differences between the treatment and control groups[2]. The one I have worked with extensively is the probabilistic theory of causality, formalized by Patrick Suppes (1970) but widely adopted throughout the human sciences, even if not consciously so under that title. Suppes's concept of probabilistic causality is similar to the concept of Granger causality (Granger, 1969) that is frequently used in econometrics.

The root idea of the probabilistic theory of causality is that if the probability of an 'outcome' O is greater with a putative cause T than without T once all 'confounders' are controlled for in some particular way, that is sufficient for the claim 'T causes O' in that particular setting of confounding factors. So, in a population where 'all other' causes of O are held fixed, any difference in probability of O with T present versus with T absent shows that T causes O in that population. The rationale supposes that differences in probability need a causal explanation and if all explanations relying on confounders are eliminated, then T causes O is the only explanation left. T must be causing O in at least some members of the population in order to account for the

[2] Cf. Cartwright (1989), Holland and Rubin (1988) and Heckman (2001).

difference in probability. I should note that whether one wishes to adopt the theory in exactly this form, some such assumption is necessary to connect causes and probabilities if we are to suppose that the probabilistic observations in RCTs can yield causal conclusions.

The definition so far only tells us when we can assert that T causes O for populations that have some fixed arrangement of 'all other' causal factors. To get a more general conclusion we may accept as well that if T causes O in a subpopulation of a given population φ, then T causes O in φ. This is consistent with my suggestion in the last paragraph that on the probabilistic theory of causality when we say T causes O in a population we mean that T causes O in at least some members of that population.

The proof that positive results in an ideal RCT deductively imply that the treatment causes the outcome would go something like this: To test 'T causes O' in φ via an RCT, we suppose that we study a test population φ all of whose members are governed by the same causal structure, CS, for O and which is described by a probability distribution P. P is defined over the event space $\{O,T,K_1,K_2,\ldots,K_n\}$, where each K_i is a state description over 'all other' causes of O except T.[3] The K_i are thus maximally causally homogeneous subpopulations of φ. Roughly,

- 'K_i is a state description over other causes' = K_i *holds fixed* all causes of O other than T.
- 'Causal structure' = the network of causal pathways by which O can be produced, with their related strengths of efficacy.

[3] This must include 'spontaneous generation'. More formally, K_i holds fixed one variable on each pathway that does not go through T, as judged by the causal structure CS.

Then assume

1. *Probabilistic theory of causality.* T causes O in φ if $P(O/T\&K_i) > P(O/\neg T\&K_i)$ for some subpopulation K_i with $P(K_i) > 0$.
2. *Idealization.* In an *ideal RCT* for 'T causes O in φ', the K_i are distributed identically between the treatment and control groups.

From 1 and 2 it follows that ideal RCTs are clinchers**.** If P(O) in treatment group > P(O) in the control group in an ideal RCT, then trivially by probability theory $P(O/T\&K_i) > P(O/\neg T\&K_i)$ for some K_i. Therefore: if P(O) in treatment group > P(O) in control group, T causes O in φ relative to CS,P.

What is going on here? We suppose that increase in probability of O with T does not show that T causes O in an arbitrary population. But it does in a maximally causally homogenous population. We of course are almost never in a position to identify what makes for a maximally homogeneous population, so how can we tell whether T increases the probability of O in some one of these? The RCT is a clever way to find out. The RCT tells us that in some one or another maximally causally homogeneous subpopulation of the population in the study, T does increase the probability of O. Given the probabilistic theory of causality that tells us that T causes O in that subpopulation. So, what is established in the ideal RCT according to the account based on probabilistic theory of causality is that T causes O in at least one maximally causally homogeneous subpopulation of φ. We may say we have established 'T

causes O in φ' and that is a fine way to talk, so long as we recall that this means that T causes O in some subpopulation of φ.

It is important to notice that on this account 'T causes O in φ' is consistent with 'T causes ¬O in φ'. This lines up with what we know of RCTs:

- RCTs deliver population-average results. A *positive* result shows that T causes O in at least one subpopulation. It could produce exactly opposite results in other subpopulations.
- Positive results are conclusive but negative are not: Equal probability for O in the treatment and control groups does not show that T does not cause O in φ. It shows that if T causes O in φ (because it does so in some $K_i \subseteq \varphi$) it must also cause ¬O (because it does so in some other $K_i \subseteq \varphi$).

Real RCTs

So, from positive results in an ideal RCT for 'T causes O in φ' we can deduce that the causal hypothesis is true. But we can be no more certain of our casual conclusion than we are of our premises, to wit, that the RCT is ideal and that the probability of O is indeed higher with T than without in the test population. What do we do to ensure the premises? Here are just some of the principal precautions we take: careful use of statistics to move from frequencies to probabilities, 'random' assignment to treatment and control groups, quadruple blinding, careful attention to drop-outs and non-compliance, and so on.

I mention them just to point out that the practical methodology must match and be matched with the kind of formal treatment I have outlined. RCT advocates claim that RCTs are extremely reliable if carried out properly. That claim can be justified by an account of the kind I have outlined. But then – what is justified is that positive results *as defined by the account*, in an ideal RCT *as defined by the account*, imply a causal conclusion *of the kind defined by the account.* The practical methodology then must be geared to ensuring that the premises required by the formal account are very likely to be true; and the conclusions drawn can only be of the kind admitted by the account. Of course the converse holds as well: A formal account that does not match well with our most careful, most well thought-out practical methodology should be viewed with at least a little suspicion.

The scope of an RCT

Starting as I have from the probabilistic theory of causality there are two kinds of causal conclusions we might naturally try to export from an RCT to some target population θ:

1. T causes O in θ. That is, T causes O in at least some members of θ.[4]
2. Some measure of 'average improvement' that holds in the experiment will hold in the target population. I shall consider the simple case of P(O/T) > P(O/¬T).

[4] In the 'long run' of course since all results are probabilistic.

Both conclusions need strong auxiliary assumptions to be warranted, well beyond those supported by the structure of the RCT. For the first, the RCT shows that T causes O in at least some members of some fixed causally homogeneous subpopulations. So to draw conclusions that T causes O in at least some members of θ, we need at least these kinds of assumptions:

- Auxiliary 1.a. At least one of the subpopulations (with its particular fixed arrangement of 'other' causal factors) in which T causes O in φ is a subpopulation of θ.
- Auxiliary 1.b. The causal structure and the probability measure is the same in that subpopulation of θ as it is in that subpopulation of φ.

For the second we need to show that the outcome is more probable with T than without in θ.[5] The simplest guarantee for this is

- Auxiliary 2. The causal structure (CS) and the probability (P) are the same in θ as in φ.

There are an indefinite number of other ways that guarantee $P(O/T)>P(O/\neg T)$ in θ given it holds in φ, depending on the exact strengths of efficacy and the exact probabilities involved. But this is the only rule that does not require explicit statement of the specific numbers, most (if not all) of which are unknown to us. To get a sense for this, just imagine a case where there are only two relevant

[5] Or as near enough as matters for our purposes. I shall here ignore these niceties and how to treat them in order to focus on the main point.

subpopulations, in one of which T is strongly positive for O and in the other it is equally strongly negative. The results will be positive in the RCT if the first subpopulation is more probable than the second, but will be reversed in targets where the second outweighs the first even if the new population has the same causal structure as the test population. Clearly if the causal structure differs, matters will depend on just how, just as the net result will depend on just what the probabilities are if the probabilities of the relevant subpopulations differ.

The central question for external validity then is, 'How do we come to be justified in the assumptions required for exporting a causal claim from the experimental to a target population?' Here rigor gives out. This is not to say that we do not have procedures or that we do not proceed in an intelligent way. We could aim to draw the test population 'randomly' from the target. We know that this is almost never possible. Moreover, we must not be deluded about sampling methods: You cannot sample randomly without any idea what factors are to be equally represented – which is just the issue that drives us to RCTs to begin with. One thing we certainly can do is to try to take into account all possible sources of difference between the test and target populations that we can identify. This is just what we do in matched observational studies. When it comes to internal validity, however, advocates of the exclusive use of RCTs do not take this to be good enough – matching studies are not allowed just because our judgements about possible sources of difference are fallible. Yet exactly the same kinds of 'non-rigorous' judgements are required if RCTs are to have any bearing outside the test population. For an RCT the reliability of the claims in the target population is only as good as our estimates that very demanding

auxiliaries like those above are met. The question then is about the trade-off between internal and external validity.

Lesson. We experiment on a population of individuals whom we take to have the same *fixed causal structure* (albeit unknown) and *fixed probability measure* (albeit unknown). Our deductive conclusions depend on that very causal structure and probability. How do we know what individuals beyond those in our experiment this applies to? We have seen some typical auxiliary assumptions about target populations that allow us to export conclusions from the experimental population to a target population and we have seen that these assumptions are very demanding, demanding of information that is not supplied by the RCT and that is hard to come by. But our conclusions about the target can be no more certain than these auxiliary assumptions. The RCT, with its vaunted rigor, takes us only a very small part of the way we need to go for practical knowledge. This is what disposes me to warn about the vanity of rigor in RCTs.

6. The vanity of rigor in RCTs

The title is borrowed from my paper 'The Vanity of Rigor in Economic Models' (Cartwright (forthcoming)). In both cases we see identical problems: that of internal versus external validity. Economists make a huge investment to achieve rigor *inside* their models, that is to achieve internal validity. But how do they decide what lessons to draw about target situations outside from conclusions rigorously derived inside the model? That is, how do they establish external validity? We find: thought, discussion, debate; relatively secure knowledge; past practice; good bets. But not

rules, check lists, detailed practicable procedures; nothing with the rigor demanded inside the models.

And RCTs? If we compare them with economic models on internal validity, economic models have the advantage: we can readily see when the results are internally valid in an economic model just by inspecting the derivation. This is clearly not so with RCTs. Consider the equal distribution of 'other' causal factors. Once we check the causes we know about, we have no further evidence that our precautions, our quadruple blinding and random assignment and so forth, indeed result in an equal enough distribution. And we know lots of things can go wrong. The best we can do is for people expert at what could go wrong to have a very close look at what actually happens in the experiment.

It is important though that these are not people like me (or independent experimental-design firms) who know only about methodology, but rather people with *subject-specific knowledge* who can spot relevant differences that come up. But this introduces *expert judgement* into the assessment of internal validity, which RCT advocates tend to despise. Without expert judgement, however, the claims that the requisite assumptions for the RCT to be internally valid are met depend on fallible mechanical procedures. Expert judgements are naturally fallible too, but to rely on mechanics without experts to watch for where failures occur makes the entire proceeding unnecessarily dicey.

This brief mention of economic models versus RCTs highlights the conventional trade-off I recalled at the start between internal and external validity. Despite the

claims of RCTs to be the gold standard, economic models have all the advantages when it comes to internal validity. As I remarked, we need just mathematics and logic to decide if the conclusions are internally valid, whereas RCTs need a number of demanding assumptions beyond valid reasoning. But it seems that RCTs have the advantage over economic models with respect to external validity. Surely no matter what the target population, people in experiments are more like people in the target population than people in models are. Even here there is a caution, however, for of course this claim depends on exactly what kind of knowledge about people in the target population we build into the construction of our experiments versus how much we build into our models, and how we do so.

7. Closing remarks

I close with some reminders for those who advocate RCTs as the gold standard:

The method of our most successful science – the h-d method – is not a clincher at all. (And we do have some biomedical theory!)

There are many other clinching methods. Which method provides the most secure conclusions in a given case depends entirely upon which kinds of premises we can be most secure about and the situation at hand.

An argument that certain procedures achieve a given result much of the time may not be a good argument that they do so on any one occasion.

External validity for RCTs is hard to justify. Other methods, less rigorous at the front end, on internal validity, can have far better warrant at the back end, on external validity. We must be careful about the trade-offs. There is no a priori reason to favour a method that is rigorous part of the way and very iffy thereafter over one that reverses the order or one that is less rigorous but fairly well reasoned throughout.

References

Cartwright, N. (forthcoming) *Hunting Causes and Using Them*, Cambridge; Cambridge University Press.

Granger, C. (1969) 'Investigating Causal Relations by Econometric Models and Cross-Special Methods', *Econometrica*, *37*, 424-438.

Heckman, J. (2001) 'Econometrics, Counterfactuals and Causal Models', Keynote Address, International Statistical Institute. Seoul, Korea.

Holland, P. W. and Rubin, D. B. (1988) 'Causal Inference in Retrospective Studies', *Evaluation Review*, 12, 203-231.

Suppes, P. (1970) *Probabilistic Theory of Causality*, Atlantic Highlands, N.J.: Humanities Press.