The Art of Medicine
A philosopher's view of the long road from RCTs to effectiveness

For evidence-based practice and policy, randomised controlled trials (RCTs) are the current gold standard. But exactly why? We know that RCTs do not, without a series of strong assumptions, warrant predictions about what happens in practice. But just what are these assumptions? I maintain that, from a philosophical stance, answers to both questions are obscured because we don't attend to what causal claims say. Causal claims entering evidence-based medicine at different points say different things and, I would suggest, failure to attend to these differences makes much current guidance about evidence for medical and social policy misleading.

The first big question to be clear about is "what's so good about RCTs?" The canonical answer is that "RCTs control for unknown confounders". This answer jumps into the middle of a discussion long underway. Two special features of ideally conducted RCTs provide more fundamental grounding: first, ideal RCTs can clinch causal conclusions; second, ideal RCTs are self-validating.

Some methods merely vouch for their conclusions. Although it is problematic to say exactly what it takes for a finding to vouch for a hypothesis, it generally involves at least that the finding is surprising but not surprising given the hypothesis. Others methods also clinch their conclusion in the ideal: if the assumptions defining the method are met, positive results deductively imply the conclusion. The ideal RCT—ie, one where all the requisite premises are met—is a clincher. Roughly, RCT logic assumes a general metaphysical premise (premise 1) that probabilistic dependence calls for causal explanation. Experimental design acts to ensure premise 2: all features causally relevant to the outcome other than the treatment (and its downstream effects) are distributed identically between treatment and control groups. If the outcome is more probable in the treatment than the control group, which is premise 3, the only explanation possible is that the treatment caused the outcome in some members of that group.

Evidence-based medicine focuses on clinchers. Yet studies in philosophy of science suggest that physics claims are primarily warranted by large collections of varied results merely vouching for them. There are no checklists for handling vouching evidence, however; perhaps this is why evidence-based-medicine guidelines favour clinchers. Clinching is not unique to RCTs, however. Economists use the rigorous methods of econometric modelling to estimate the degree to which one factor predicts another in a given population. This could be mere correlation. But given the right assumptions, their results can deductively imply causal conclusions from non-experimental data. Deduction from accepted theory can also clinch causal conclusions; as can ideal case-control studies, since these have the same logic as RCTs. The difference between RCTs and these others is the grounds for accepting the requisite premises, which is the second special feature of RCTs: that they are self-validating.

All methods have assumptions that must be met before conclusions from them are warranted. For causal conclusions, some of these premises must be causal: "no causes in; no cause out". For most studies, such as the economic ones mentioned, the warrant for these assumptions comes from outside the study design. The metaphysical assumption aside, support—though no guarantee—for premises 2 and 3 is built right into RCT design: premise 2, by policing of treatment administration, blinding, random assignment, and the like; premise 3, by techniques—including large sample size—for reliably inferring probabilities from observed frequencies. RCTs are thus self-validating.

Self-validation is a virtue but not a necessity. We often have good reason to accept the premises necessary for other study designs, including case-control studies, which is where unknown confounders enter. By definition we do not know "unknown" causal factors. We may nevertheless know enough about underlying mechanisms and the study environment to assume no strong unknown causes obtain. Sometimes we even shield studies to prevent unknown sources of confounding. RCTs trust to procedure; other methods import information. Which strategy provides most support for a particular conclusion depends on how confident we can be that the procedures achieve their aim in the case at hand versus the strength of justification for the information imported.

All the studies I have discussed so far justify what I would term "efficacy claims'", where "efficacy" is what happens

South Coast Press/Rex Features

in ideal circumstances. But recall the logic of RCTs. The circumstances there are ideal for ensuring "the treatment caused the outcome in some members of the study"—ie, they are ideal for supporting "it-works-somewhere" claims. But they are in no way ideal for other purposes; in particular they provide no better base for extrapolating or generalising than knowledge that the treatment caused the outcome in any other individuals in any other circumstances.

For policy and practice we do not need to know "it works somewhere". We need evidence for "it-will-work-for-us" claims: the treatment will produce the desired outcome in our situation as implemented there. How can we get from it-works-somewhere to it-will-work-for-us? Perhaps by simple enumerative induction: swan 1 is white; swan 2 is white…so the next swan will be white. For this we need a large and varied inductive base—lots of swans from lots of places; lots of RCTs from different populations—plus reason to believe the observations are projectable, plus an account of the range across which they project. Electron charge is projectable everywhere—one good experiment is enough to generalise to all electrons; bird colour sometimes is; causality is dicey. Many causal connections depend on intimate, complex interactions among factors present so that no special role for the factor of interest can be prised out and projected to new situations.

Sometimes it can. Magnets are tested in ideal circumstances; their power to attract metal objects can be relied on widely. The Heimlich manoeuvre is good for removing airway obstructions in almost anyone. Knowledge like this involves a third kind of causal claim, a power or capacity claim: the treatment reliably promotes the outcome, or reliably contributes across a given range of circumstances. "Reliably promotes" means roughly that across a wide range of circumstances there will be more cases, or a higher level, of the outcome with the treatment than there would be without it. What the actual numbers are depends on what other factors are present.

Where available, knowledge of capacities is a powerful tool. To use RCT results as evidence for effectiveness we are generally told to look for populations or settings like those of the study. This is advice difficult to follow. We do RCTs because we do not know all the major relevant factors, so judging whether other situations are relevantly similar is hard. Moreover, similarity is rare. But then, similarity is not necessary if the treatment reliably promotes the outcome. Magnets attract metal objects almost everywhere. The Heimlich manoeuvre depends on almost universally shared structures in the human body, so it can be relied on to encourage removal of obstructions across a wide variety of settings and individuals. So capacity claims provide evidence for effectiveness even in situations very different from those of any study. And where no capacity claims obtain, there is seldom warrant for assuming that a

treatment that works somewhere will work anywhere else. (The exception is where there is warrant to believe that the study population is a representative sample of the target population—and cases like this are hard to come by.)

But there are problems for using capacity claims. First, although knowledge that a treatment reliably promotes an outcome is evidence that it will cause that outcome for us, it is only part of an evidential argument. We also need to know that our situation contains all requisite helping factors and that there are no overwhelming countering causes. Magnets lift objects only if the objects are metal and they will not lift even metal objects when gravity is too strong; nor will the Heimlich manoeuvre remove objects if the oesophagus is too swollen by disease.

I highlight these additional factors not because they are unfamiliar but because influential guidelines for evidence-based medical and social policy often do not mention them let alone discuss standards of evidence for claims about them, despite the fact that such information is necessary for any reasonable predictions about effectiveness. Furthermore, capacity claims are hard to warrant. Worse, there is no explicit methodology describing exactly what it takes to warrant them, even in physics, despite the fact that most of our successful interventions using physics depend on capacity claims. What is clear is that even a handful of RCTs by themselves will not do the job. In general to support a capacity claim, a general understanding is needed of why the treatment should have the power to produce the outcome. Happily this is often available although few guidelines direct us to look out for it, let alone provide advice about what counts as good evidence that the backup understanding is sound. Probably that's because we try to rely on procedures, as with RCTs, to avoid relying on claims of a general theoretical nature.

But an RCT supports only an "it-works-somewhere" claim. How can we put hard-won RCT results to use for predicting "it will work for us"? Similarity is problematic to judge and the kind of similarity necessary for warranting direct extrapolation from RCTs is rare. Capacities provide a conduit from RCTs to effectiveness, often the only one. But these are hard to warrant and even when warranted are only part of a good evidence base for predicting effectiveness. Effectiveness predictions are always dicey. Use of scientific evidence makes them far less so. But to use this evidence we need to tackle, not ignore, the messy issue of "theoretical" warrant for capacities in medical and social contexts.

Nancy Cartwright

Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London WC2A 2AE, UK; and Department of Philosophy, University of California, San Diego, La Jolla, CA, USA
R.Robinson1@lse.ac.uk

Further reading

Cartwright N. What is this thing called "efficacy"?. In: Mantzavinos C, ed. Philosophy of the social sciences: philosophical theory and scientific practice. Cambridge: Cambridge University Press, 2009

Cartwright N. Predicting "it will work for us": (way) beyond statistics. In: McKay Illari P, Russo F, Williamson J, eds. Causality in the sciences. Oxford: Oxford University Press, 2011

Reiss J. Empirical evidence: its nature and sources. In: Jarvie IC, Zamora-Bonilla J, eds. The Sage handbook of philosophy of social science. Thousand Oaks, CA: SAGE Publications, 2010

Worrall J. Evidence: philosophy of science meets medicine. *J Eval Clin Pract* 2010; **16:** 356–53

Worrall J. Evidence in medicine and evidence-based medicine *Philosophy Compass* 2007; **2:** 981–1022