How necessary are randomized controlled trials?

Abstract
Randomized controlled trials (RCTs) are often deemed the gold standard for testing new treatments. This belief in turn justifies recruiting patients into such trials even when it is suspected that a new treatment is superior – although patients in the control group are thereby denied what might be the better treatment, we cannot know that the treatment actually *is* better, the thought runs, without conducting an RCT.

But Robert Northcott argues that RCTs are not always the best choice after all. Rather, like any other method, they can go wrong sometimes, in several different ways. The main alternative to them is historical studies, which try to assess a treatment's effectiveness from data not drawn from trials. These too can go wrong in several ways, and in the past have acquired a bad reputation. However, that prejudice has become outdated. The truer picture, Northcott argues, is that sometimes one method is preferable, sometimes the other. Things must be decided case by case. It follows that the ethical ramifications of conducting an RCT also must be examined case by case; there is no one-size-fits-all answer.

An especially striking and emotional example concerns ECMO, a treatment for newborn babies with life-threatening lung problems. Historical studies indicated that ECMO was a major breakthrough, offering hugely increased survival rates. But it was still insisted that it also be tested in RCTs, in the course of which many babies receiving the conventional treatment died. A properly nuanced view of RCTs suggests that these deaths were tragically unnecessary.

Historically, newborns with a form of respiratory failure called persistent pulmonary hypertension (PPHS) faced a mortality rate of more than 80 per cent. The main symptom is immaturity of the lungs, leading to poor oxygenation of the blood. Doctors seek desperately to keep the baby alive until the lungs mature. But recently it happened that suffering babies faced these terrible odds… when all along a new treatment that had recorded an 80 per cent *survival* rate was available but left unused. In particular, this occurred during several trials of the treatment known as extracorporeal membraneous oxygenation (ECMO).

Developed in the late 1970s, ECMO in effect takes over the function of the lungs by withdrawing blood, oxygenating and then reheating it artificially, before finally returning it to the baby. Nevertheless, despite immensely promising initial results, the researchers behind ECMO felt that, in order to prove their therapy, they needed to conduct a randomized controlled trial (RCT) of it. Otherwise, they worried, the treatment would never gain general acceptance (Bartlett et al 1982). Is this how it should be? Are RCTs really the unique 'gold standard' that should always trump other kinds of evidence? The

stakes are high. Of course, no one in this story *wanted* newborn babies to die unnecessarily. The dispute, rather, was a sincere one about how best to prove ECMO's effectiveness.

As is well known, RCTs work roughly as follows: they randomly divide a trial sample into two groups, one of which (the 'treatment group') receives the new treatment, the other of which (the 'control group') receives only a placebo or the old treatment. Outcomes in the two groups are then compared. If the first group does better this is strong evidence in favor of the new treatment, because randomization is designed to ensure that the only relevant difference between the two groups – and hence the only possible explanation for the different outcomes – is the different treatments received.

A common worry is that asking patients to participate in RCTs can be unethical, because often we have reason to believe that a new treatment is rather promising, in which case those in the control group are thereby being denied what is probably better care. Newborn babies being denied ECMO is an especially striking example. Are such unlucky patients being put at risk unjustifiably? The most powerful reply is that RCTs are the only way to establish secure knowledge of therapeutic effectiveness, and so in the long run it would be ethically disastrous to stymie medical progress by stopping them. Besides, advocates claim, until an RCT is carried out we cannot know for sure if a new treatment actually *is* better.

Reconciling these two positions is often a nuanced matter. The vital underlying issue is the extent to which, in the absence of RCT evidence, a physician is entitled to judge one treatment more effective than another. That is, matters turn on something more properly located in philosophy of science than in bioethics – namely, to what extent *are* RCTs the indispensable best method for judging a treatment's effectiveness?

Are RCTs special?

When it comes to establishing whether a particular treatment causes better outcomes, there is plenty of common ground. Everyone agrees the ideal situation is to compare the effects of a treatment to the effects of the alternative, *all else equal*. The problem, of course, is that usually many things affect an outcome besides the treatment itself: the patient's age, general health, sex, how early their condition has been diagnosed, and so on. The key is to balance treatment and control groups with respect to 'confounders', i.e. with respect to these other, potentially muddying factors.

Across the sciences, there exist many methods for doing this besides RCTs. Perhaps the most famous is the simple controlled experiment with no randomization element. Another is the historical or observational study. These latter studies use data that do not come from experiments. As a result, those data must be selected with extra care so as to make due allowance for potential confounders and thus avoid being distorted. In the context of medicine, it is these historical studies that are especially controversial. In the past, they were often conducted poorly. Common problems included inconsistent diagnostic procedures and interpretation, uneven rates of hospitalization across groups, and difficulty in blinding data collection. For example, if patients in the past were diagnosed

less quickly, then superior outcomes today might reflect simply earlier diagnosis rather than anything special about new treatments. Overall, historical studies had a tendency to overestimate treatment effects (Chalmers et al 1983). Compared to this, the rigor of testing for new treatments today is rightly lauded, and RCTs are central to that. It is this record that has motivated many to insist on RCTs – and only RCTs – as the gold standard.

However, recent work in philosophy of science casts doubt upon this insistence. To be sure, no one defends poorly conducted historical studies, let alone reliance on individual physicians' personal hunches. And no one denies that well run RCTs are often the most effective method, sometimes indeed perhaps the only available one. However, it is now argued, RCTs are not *always* best. First, techniques for running historical studies have improved enormously; great strides have been made in research design and in techniques for causal inference from statistics. Indeed, more recent analyses suggest that historical studies now perform equally as well as RCTs, and perhaps even *better* (Benson and Hartz 2000, Concato et al 2000). Much turns out to depend on the specifics of each case: sometimes one method may be the superior choice, sometimes another. This is especially true when one also takes into account 'external' considerations such as cost, timing or convenience.

Moreover, plenty of other sciences have prospered without RCTs. And even within medicine, some of our most reliable and important knowledge has come from historical studies, formal and informal. Examples include: most surgical techniques, that smoking causes cancer, and that aspirin relieves headaches.

Problems to be avoided

There are several well known difficulties that efficiently conducted RCTs have proved effective at guarding against:

-- Doctors might select only the most medically promising patients for the treatment group, which risks distorting the results. Or they might interpret a patient's symptoms differently depending on which group the patient is in. Avoiding these risks is the motivation for blinding a trial from *physicians*.

-- If patients know they are in the treatment rather than control group it can affect how they self-report their outcomes, often relevant for instance with regard to psychiatric conditions. Avoiding this risk is the motivation for blinding a trial from *patients*.

-- Generally, treatment and control groups can be imbalanced with respect to confounders both known and unknown. A *randomized* allocation seeks to head off this problem. (More on the issue of randomization shortly.)

However, it is helpful also to be aware of several other difficulties that can lead even an RCT astray. Examples include:

-- The patients recruited for a trial might be unrepresentative of the general population. For example, if a drug is especially effective only in one ethnic group but that ethnic group is under-represented in the trial, then this efficacy might well be missed. Observational studies are often *more* careful about matching the tested population with the wider eventual target population.

-- Not all patients signed up for a trial persevere with it. If the drop-outs are unrepresentative and are omitted from the final results, the trial can be skewed. For example, imagine testing a safe-sex program among sex workers: some of the subjects in the safe-sex treatment group will drop out rather than lose income. Those who drop out will likely not be typical, thus distorting the results. Again, such programs can often be assessed much better by using historical control groups rather than RCTs.
-- A treatment might have positive effects on one sub-group but negative effects on another. A trial's overall results might then simply reflect the net balance between these two sub-groups, missing the individual effects.
-- Mistaken units of analysis. For example, often an education program should allocate whole schools to treatment and control groups, not just classes or individuals. As a result, the effective sample size is number of schools, which is a much smaller number than number of classes or individuals. But analyses of such trial results might mis-state this.
-- False negatives, i.e. when the effectiveness of a treatment is missed, perhaps because of small numbers or insensitive outcome measures.
-- Problems with blinding and use of suitable placebo, meaning that bias is not controlled effectively.

Of course, well designed RCTs try their best to counter these difficulties. And poorly conducted historical studies will also be vulnerable to many of them. The point is only that conducting an RCT is no guarantee – they too can go wrong.

Some of these difficulties are especially likely to crop up when assessing social policy or public health interventions. As a result, in these domains observational studies are often not only more feasible than RCTs but actually give more accurate results too. It might be difficult to set up a useful RCT for a food policy program, for instance. But if we put weight only on RCTs, observational studies will inevitably be downgraded. As a result, the food program might be denied funding, even though it potentially has more impact on public health than most drug trials put together. That is, an over-emphasis on RCTs over all other methods risks unhealthily distorting which science is funded in the first place (Grossman and Mackenzie 2005).

ECMO again

Historical studies too, just like RCTs, need to avoid the various problems listed above. But sometimes they will. ECMO is a good example, for with that case there was good reason to trust the historical data. For instance, all newborns with PPHS were treated without exception; there was no ambiguity regarding the interpretation or reporting of outcomes; and there was no known muddying confounder, such as speed of diagnosis, correlated with the split of the historical sample between ECMO and conventional treatments. In sum, was not the established 80 per cent mortality rate for conventional treatment already a rigorous enough control group? And was not the 20 per cent mortality rate of the babies treated with ECMO already a rigorous enough treatment group? Moreover, the improvement in outcomes reported for ECMO was so huge that any confounding effect would have to have been correspondingly huge to nullify it.

The only element missing from the ECMO data was randomization. Yet this single omission was taken to be reason enough to demand a 'proper', i.e. randomized, trial. As we shall see in a moment, such an insistence was highly questionable. As it turned out, the complex procedure eventually adopted for ECMO's RCT, carried out in the early 1980s, resulted in 11 patients being assigned the ECMO treatment, who all survived, and one assigned the conventional treatment, who died (Bartlett et al 1985). Yet, despite this 100 per cent 11-for-11 survival rate for ECMO, as compared, remember, to historical survival rates of only 20 per cent, still critics insisted the evidence was insufficient to prove ECMO's efficacy, because only one baby had been assigned the conventional treatment.

Therefore another RCT was carried out, in the late 1980s. Nine babies received ECMO; all nine survived. Ten received conventional treatment, of whom six survived. At this stage, pre-specified criteria deemed the evidence sufficiently conclusive for the trial to be halted. A further 20 babies who arrived at trial centers suffering from PPHS were all assigned ECMO; 19 of those survived (O'Rourke et al 1989).

Finally, even after this second RCT some statisticians objected that the sample sizes were still too small for definitive conclusions. Amazingly, a *third* trial was begun in the UK. The result? It had to be stopped early because of too many deaths in the control group.

Is randomization really necessary?

The only virtue missing from the original ECMO data, recall, was randomization. But is randomization really so necessary? The case for it is well known. In particular, while we can arrange 'by hand' for a treatment and control group to be equally balanced with respect to known possible confounders such as age and health, we obviously cannot balance by hand with respect to *unknown* confounders, precisely because they are unknown. If we allocate randomly, however, that automatically ensures there will be no systematic bias with respect to *any* confounder at all, known or unknown.

However – return to that word 'systematic'. True enough, in the long run over many distributions, allocating patients randomly will ensure no systematic bias. But any *one* allocation might certainly turn out to be very biased indeed. For example, suppose age is important and that we randomly allocate 20 patients, ten young and ten old. Quite possibly, just by chance, we may end up with a treatment group of, say, eight young and two old, and a control group the other way round. (Such imbalances are especially likely with smaller sample sizes, and when there are many known confounds all needing to be balanced.) In such circumstances, the usual remedy is to re-randomize to get a better balance, a procedure known as 'baseline rebalancing'. Once we arrive at an allocation that is deemed sufficiently balanced with respect to all known factors, then a trial may go ahead. But if the only factors that are skewed are unknown ones, then we will have no way of recognizing the need for baseline rebalancing. In other words, in any *particular* case randomization provides no guarantee at all that we actually are balanced with respect to unknown factors, even though this is supposedly its unique advantage.

Compare balancing by hand instead. Known factors can obviously be allocated evenly that way; there is no need for randomization. With respect to unknown factors, meanwhile, there is no reason to expect a hand allocation to be any less balanced than a random one. If these unknown factors are correlated with any known ones, then the best we can do is balance with respect to the latter – and that is exactly what's done anyway.[1]

The conclusion is that randomization in itself adds nothing (Worrall 2002, 2007).[2] It might be objected that, be that as it may, in any case randomization does no harm either, assuming that any baseline rebalancing is performed thoroughly. But the real point is that the perception that randomization is necessary leads to the unjustified shunning of other forms of evidence. That was why, for instance, the dramatically heightened survival rates associated with the ECMO treatment were not thought decisive – even though there was no strong reason to think them the result of any of various known confounders. Only an unreasonable prejudice in favor of randomization for its own sake can explain why the existing ECMO evidence wasn't deemed sufficient.

Conclusion

To what extent, then, are RCTs necessary? Many times, they will indeed be the best evidence. But not always – other times, historical or observational trials may be equally or even more persuasive. Or, as with ECMO, the historical data should already be considered sufficient. Overall, there is no one-size-fits-all answer. Decisions must be made case by case. Whether it is indeed unethical to recruit participants for an RCT can therefore also only be assessed case by case.

References

-- Bartlett, R. H., A.F. Andrews, J.M. Toomasian, N.J. Haiduc, and A.B. Gazzaniga (1982). "Extracorporeal membrane oxygenation for newborn respiratory failure: 45 Cases", *Surgery* 92: 425-433.

-- Bartlett, R.H., D.W. Roloff, R.G. Cornell, A.F. Andrews, P.W. Dillon, and J.B. Zwischenberger (1985). "Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study", *Pediatrics* 76: 479-487.

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

-- Chalmers, T.C., P. Celano, H.S. Sacks, and H. Smith (1983). "Bias in Treatment Assignment in Controlled Clinical Trials", *New England Journal of Medicine* 309: 1358-1361.

[1] A large sample size is a red herring here. True, a large sample makes it more likely that a random assignment will be evenly balanced with respect to unknown factors – but equally, it makes it more likely a by-hand assignment will be too

[2] True, randomization can also help ensure the blinding of trials from physicians, something agreed by everyone to be desirable. But there are plenty of other ways to ensure such blinding.

-- Concato, J., N. Shah, and R.I. Horwitz (2000). "Randomised Controlled Trials, Observational Studies, and the Hierarchy of Research Designs", *New England Journal of Medicine* 342: 1887-1892.
-- Grossman, J., and F.J. Mackenzie (2005). "The Randomized Controlled Trial: gold standard, or merely standard?" *Perspectives in Biology and Medicine* 48: 516-534.
-- O'Rourke, P.P., R.K. Crone, J.P. Vacanti, J.H. Ware, C.W. Lillehei, R.B. Parad, and M.F. Epstein (1989). "Extracorporeal Membrane Oxygenation and Conventional Medical Therapy in Neonates with Persistent Pulmonary Hypertension of the New Born: a Prospective Randomized Study", *Pediatrics* 84: 957-963.
-- Worrall, J. (2002). "*What* Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69: S316-S330.
-- Worrall, J. (2007). "Why There's No Cause to Randomize" *British Journal for the Philosophy of Science* 58: 451-488.