Randomised controlled trials: important but overrated?

JF Boylan, BP Kavanagh, J Armitage

Practising physicians individualise treatments, hoping to achieve optimal outcomes by tackling relevant patient variables. The randomised controlled trial (RCT) is universally accepted as the best means of comparison. Yet doctors sometimes wonder if particular patients might benefit more from treatments that fared worse in the RCT comparisons. Such clinicians may even feel ostracised by their peers for stepping outside treatments based on RCTs and guidelines. Are RCTs the only acceptable evaluations of how patient care can be assessed and delivered?

In this controversy we explore the interpretation of RCT data for practising clinicians facing individualised patient choices. First, critical care anaesthetists John Boylan and Brian Kavanagh emphasise the dangers of bias and show how Bayesian approaches utilise prior probabilities to improve posterior (combined) probability estimates. Secondly, Jane Armitage, of the Clinical Trial Service Unit in Oxford, argues why RCTs remain essential and explores how the quality of randomisation can be improved through systematic reviews and by avoiding selective reporting.

**KEYWORDS** Randomised controlled trials, Bayesian analysis, subgroup analysis, clinical trials, bias, confounders

**DECLARATION OF INTERESTS** No conflict of interests declared.

Randomised controlled trials in critical care: caveat emptor

JF Boylan, BP Kavanagh

Several decades ago Yusuf, Collins and Peto made a strong case for large, simple randomised controlled trials (RCTs), and since then, several RCTs (e.g., thrombolysis in acute myocardial infarction) have produced results of great importance. However, there are concerns about RCTs, including high negativity rates, uncertainties in interpretation, and an increasing recognition that treatments, naturally, have varying impacts depending on individual patients and the severity of the disease. These concerns may be especially important in critical care, an area with high mortality, significant physiological variation, and a low anticipated yield in survival benefit. At the same time, some have suggested that trial results are best interpreted by ‘methodologists’ – as the best-trained and most objective assessors of evidence; this is, possibly, a paradox, given that the ‘large simple trial’ is supposed – by use of massive numbers – to keep interpretation simple. Here we focus on four elements: prior probability, trial bias, Bayesian analysis; and the application of trial results to diverse patients. We believe that appreciation of these issues could enhance the utility of RCTs in critical care.

**PRIOR PROBABILITY**

Browner and Newman have compared RCTs in populations with diagnostic tests in individuals. In both cases, we note the pretest probability, and, after the data are in, we then decide about our degree of belief. However, while we usually know the pretest probability (i.e. the prevalence) of disease before a diagnostic test, we almost never know the prior probability (π) of scientific hypotheses being true in an impending RCT. We are therefore, unable to place the results of the RCT in the context of our population of interest. If the lessons from diagnostics hold, usual RCT interpretation is likely to be flawed.

Attempts have been made to determine prior probability (in retrospect): the overall survival for experimental groups in a series of oncology trials was 0.96 (0.89–1.03; mean, 99% confidence intervals), suggesting that a prior probability (π) of scientific hypotheses being true in an impending RCT is also low there. A consistent negative trend in critical care trials suggests that it is also low there.

**What is the impact of π on interpretation of an RCT?**

Very low values mean that the chance of any given ‘significant’ result being a ‘true positive’ becomes very low, even with flawless study methodology. For 100,000 RCTs with π=0.01 (so 1,000 true hypotheses out of 100,000 tested) and with the usual α=0.05 and β=0.20, we could expect 5,000 false positives, 800 true positives, 200 false negatives and 94,000 true negatives. So where RCTs are ‘pragmatic’ (and especially where this equates...
to minimal reliance on physiological mechanisms) most positive results will be false positives.

Recognition of low \( \pi \) in RCTs may suggest that ‘most published research findings are false’.\(^6\) If a given \( \pi \) were high, a ‘positive’ trial would likely be a ‘true positive’; however, apart from trials designed to test a (truly) promising hypothesis, most \( \pi \) values would seem to be low.

**QUALITY AND BIAS**

Delivery of objective results hinges on eliminating experimental bias. The most widely accepted approach to confirm that bias is minimised is to grade the RCT methodology; quality scoring systems or methodological ratings might help the reader decide what to believe. There is some support for quality scoring; higher quality studies have lower effect-size estimates (perhaps suggesting less unrealistic findings), and non-randomised studies are more likely to be contradicted or have their effect estimate downgraded compared with randomised trials. But high ‘grades’ do not reduce false positive rates, and randomised designs are far from resistant to refutation. Thus no single aspect of study design allows us to draw robust conclusions.

Bias may be of most concern in ‘positive’ studies. There are several causes. Allocation bias (more ‘less sick’ patients in the treatment group improve outcome, independently of the treatment) involves manoeuvres, not all of which are detectable, which subvert randomisation and allocation concealment. Superior care in the treatment group could involve post-treatment co-interventions because of altruism in the context of a ‘risky’ therapy (e.g. increased vigilance); the corollary, inferior care in the control group, is also a possibility. Finally, biased analysis or presentation of data may be subtle, potentially reflecting prior beliefs or commercial pressure.

Among the most controversial questions in critical care in the last few years has been the role of intensive insulin therapy in reducing intensive care mortality. A single centre study suggested a major reduction in mortality,\(^7\) and after several failures to replicate this finding, a multicentre study suggested a significant worsening in mortality,\(^8\) the ‘quality’ scores in both studies were identical,\(^9\) suggesting that ‘quality’ does not help us decide which trial, if either, is ‘true’.

Randomised, concealed allocation, the cornerstone of trial validity, can be interfered with for benign motives. For example, in a trial investigating a therapy that cannot be blinded, sicker patients may be diverted into the treatment arm if caregivers ‘believe’ in the intervention. Of course, diverting sicker patients toward the control group would be much more harmful, since this results in a bias towards treatment even if the therapy is not efficacious. Thus even partial knowledge of a randomisation schedule could influence a patient’s group entry, and the traditional display of \( P \) values provides no protection against this.\(^10\) No one has modelled biased allocation of patients in trials, but simulation data suggest that even small-scale (2%) transfer of sicker patients out of an intensive care unit causes marked improvements in survival standardised mortality ratio.\(^11\)

**A BAYESIAN VIEW**

All clinicians are Bayesians in everyday practice, interpreting data in the light of what they already know (or believe). Reviewing test and RCT results, they think ‘how likely is the diagnosis/hypothesis to be correct, given the data?’. Trialists use \( P \) values – ‘how likely are these data if the null hypothesis is true?’ (not the same thing) – to persuade us of their conclusions. Many RCT results are not definitive, but it is hard to quantify this. Attempts to estimate ‘strength of evidence’ based on quality ratings and hierarchical scoring reflect this problem.

To estimate the posterior probability that a hypothesis is true, trialists need to combine a prior probability with experimental data. The prior probability may be based on data already available, or may reflect a belief (expressed as a probability that \( \pi \) lies within a ‘null range’). Because both the prior probability and the data can be expressed as distributions with means and standard deviations, we can generate a posterior distribution from them,\(^12\) giving us probabilities for any given treatment effect,\(^13\) including probabilities for a minimum treatment effect.

Bayesian interpretation of RCTs has not gained full acceptance, because of ‘subjectivity’ concerns, but looking at trial data with prior probability in mind gives us useful insights. The sensitivity of a trial result to the impact of different \( \pi \) values allows us to interpret it cautiously. Assuming no other pitfalls, how likely the hypothesis is to be true after the data are in depends on both the robustness of the data (sample size, event rates) and the prior probability of the hypothesis. Knowing that \( \pi \) in a research field is low may lead clinicians in that field to adopt at least a mildly sceptical stance when assessing a trial. In a Bayesian analysis of the most contentious recent critical care trials, Kalil showed that their interpretation was highly sensitive to the level of scepticism; a reluctance to implement RCT results may be driven by a (very reasonable) belief that the knowledge is provisional.\(^14\)

An important unresolved issue with RCTs is that negative results do not rule out clinically relevant benefit, and positive results do not mean that the reported benefit, an aggregate value, is the most likely benefit. Current trial analysis and reporting usually ignores this. In internal medicine, Wijesundera et al. have pointed out that the efficacy estimates in high-profile positive RCTs were very sensitive to the
assumed $\pi$, and many negative studies were compatible with moderate probability of at least some benefit.\textsuperscript{14} If we want to know the state of the investigators’ knowledge at the end of a positive RCT, they should be able to tell us, for example, ‘this treatment has an 80% chance of producing a relative risk reduction of 0.2 or better’, and to give us an idea of how this estimate varies, depending on assumptions about $\pi$.

**HOW MUCH BENEFIT IN THIS PATIENT?**

There is a widespread belief among methodologists that subgroups within RCTs are problematic, and that subdividing data for analysis does more harm than good, a stance backed up by statements, unsupported by data, that quantitative differences (different degrees of benefit) are rare and qualitative differences (benefit in some groups, harm in others) are very rare.\textsuperscript{15} Some authorities actively discourage collecting such data: ‘...less information may mean bigger numbers and hence better science: many trials still collect ten or a hundred times too much information per patient...we need to find ways of making trials much simpler and much larger...’.\textsuperscript{17}

If the correct approach to subgroup differences is to assume by default that they do not exist, or do not matter, then we can justify subgroup detection tests with a very ‘high bar’; such an approach was used in CRASH-2, an RCT suggesting that tranexamic acid reduces all-cause mortality in major trauma involving ‘significant haemorrhage’.\textsuperscript{18} The trial reported a net improvement in survival, but no difference in blood product usage (half of the patients were not transfused), a finding attributed – without supporting data – to non-survivors being transfused less than survivors. Univariate subgroup analyses (time since injury, blunt vs penetrating trauma, Glasgow coma score, admission systolic blood pressure) using a very stringent threshold ($p<0.001$) found no differences. This led the investigators to conclude that tranexamic acid conferred a general survival benefit; their analysis did not exclude benefit to patients with trivial bleeding, even though most patients (in the control group) did not die, a minority (12.5%) had large-volume bleeding, even though most patients (in the control group) did not die, a minority (12.5%) had large-volume bleeding, even though most patients (in the control group) did not die, a minority (12.5%) had large-volume bleeding.

By avoiding data on effect modifiers, or using analytical techniques that are underpowered, harm could result from generalising the conclusions to all patients. If experimental therapies are given to patients not ‘at risk’ in the first place, such trial participants will, at best, have been just making up the numbers. There is a strong case, articulated by Kent et al., that trialists should assume that summary estimates do not apply to individual patients, and should seek out heterogeneity of treatment effect using prospectively defined methods.\textsuperscript{24} In the real world, this concern is reinforced by a recent paper suggesting that compliance with a treatment guideline for healthcare associated pneumonia worsened survival.\textsuperscript{25} This finding surprised the authors, but not a commentator, who suggested that empirical therapy and failure to understand patient heterogeneity were the main reasons why compliance was harmful.\textsuperscript{26}

**GOING FORWARD**

Critical illness is a collective term for our lack of understanding of many acute disease states. It could be argued that some of the obvious, pragmatic (that is, non-mechanistic) questions in critical care (such as positive end-expiratory pressure, dialysis dosage, global oxygen delivery, gastrointestinal prophylaxis) have now been answered in the negative, suggesting a very low $\pi$ for the field. But this reflects a low $\pi$ for the hypotheses underlying ‘one size fits all’ therapies. Unless we accept the challenges inherent in studying seriously ill patients, with highly variable physiology and the likelihood of variable treatment-induced effects (ranging from benefit to harm), we can expect continuing difficulty. A trial – even a ‘simple’ one – of any intervention in complex
Randomised trials remain essential and are not overrated

J Armitage

INTRODUCTION

Randomised controlled trials (RCTs) lie at the heart of the practice of evidence-based clinical medicine and are accepted as the most reliable form of evidence. Results from RCTs have led to both the widespread use of effective treatments and a reduction in the use of worthless or harmful treatments. Examples include trials such as ISIS-2, which showed the benefits of aspirin and streptokinase in acute myocardial infarction, and more recently the large randomised statin trials, which first individually and then collectively, showed reductions in coronary events and strokes in a wide range of people.

These trials led to rapid changes in practice, widespread use of these effective treatments and undoubtedly avoided large numbers of premature deaths and morbidity and have produced significant health improvements.

However, despite a very large number of trials having been undertaken over the last 50 years, the trials that have delivered major health improvements represent only a minority and are generally large trials counting clinically relevant outcomes. Unfortunately, many trials are poorly designed and reported, often with too few patients to exclude chance effects or inadequately randomised and followed-up to be sure of minimising other biases. The

REFERENCES

3 Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987; 257:2459–63. doi:10.1001/jama.257.18.2459
12 Kahn JM, Kramer AA, Rubenfeld GD. Transferring critically ill patients out of hospital improves the standardized mortality ratio. Chest 2007; 131:68–75. doi:10.1378/chest.06-0741
challenge over the coming years is not to dismiss RCTs as overrated, but to improve the quality and particularly the size of randomised trials so that they continue to deliver valuable information to improve patient care. But how can we do this and are there alternative strategies?

**IMPROVING QUALITY**

Efforts are ongoing on many fronts to try to improve the quality of randomised evidence. Although randomisation in trials reduces systematic biases, only trials in populations which include large numbers of clinically relevant outcomes can produce reliable results. Systematic reviews of all the randomised evidence in particular disease or therapeutic areas is one way of increasing the number of endpoints available to be analysed. Provided there are a number of good quality trials in a particular area, such reviews (particularly if done in collaboration with the trialists) allow clear treatment effects to be demonstrated in a range of different types of patient. Examples include the collaborations reviewing anti-thrombotic treatments, blood pressure lowering therapy and the early treatment of breast cancer. Systematic reviews also illuminate areas where there is little evidence of value and so stimulate the initiation of new trials. But, frequently, such reviews reveal large numbers of inadequate trials sometimes also with clear evidence that not all the trials in that area have been published. In an effort to minimise this latter problem of selective reporting of trials with interesting or positive results and failure to report negative trials, major efforts have been made over many years to ensure that all RCTs are registered and the results made public. This should allow future reviews to be able to encompass all relevant randomised evidence and at least minimise publication bias.

Systematic reviews also highlight deficiencies in the way that trials are reported. The CONSORT group have been trying over many years to improve reporting standards but there is still room for improvement. It was hoped that improving reporting would feed back to improve trial design but there is little evidence to support this, particularly in terms of the size of trials. A recent review of 600 randomised trials reported in PubMed journals in December 2006, ten years after the first CONSORT paper, showed that the median number of patients randomised in these trials was still only 62 (10th–90th centile: 19–392) illustrating that the majority of trials are small and unlikely, therefore, to be having any material impact on clinical practice. So, despite some recent progress, there is still a way to go to make most trials informative.

Randomised trials of medicinal agents are currently subject to what many consider to be excessive regulation, partly as a result of the implementation of the European Union Directive on Clinical Trials and partly, in the UK, because of NHS Research Governance requirements. Although well intentioned, these regulations mean that significant amounts of time and effort are spent by those running trials ensuring compliance with the various regulations, thereby detracting from the efficient running the trial. This has led to higher costs and complexity of running trials which typically results in fewer patients being recruited and therefore less reliable studies. Again, active efforts are being made on several fronts to improve the situation.

**ARE THERE RELIABLE ALTERNATIVES TO RCTS?**

The aim of a randomised comparison is to achieve an unbiased assessment of the intervention or drug treatment being evaluated. Randomisation alone doesn’t guarantee this and, to minimise bias, blinding of outcome assessment, complete follow-up in each randomised group and intention-to-treat analyses are required.

Observational studies, either prospective or case-control can, in some circumstances, provide reliable results about effects of treatments, particularly if effect sizes are large, but confounding remains an issue if effect sizes are moderate. Routinely recorded information gathered into databases is increasingly providing an opportunity to readily investigate the quantitative effects of treatments. Such databases can be useful for looking at treatment patterns, however, as it is impossible to fully eliminate confounding by unmeasured (and unmeasurable) characteristics of both patients and prescribing doctors, such non-randomised analyses risk serious bias if used to try to measure treatment effects. Examples include claims based on routinely collected GP data that statins reduce the risk of dementia or increase the risk of cataracts, neither of which are borne out by data from large randomised studies and probably represent problems of residual confounding. So, observational studies generally do not allow for reliable assessments of treatment effects.

**CAN WE USE SURROGATE MARKERS RATHER THAN CLINICAL OUTCOMES?**

Epidemiologists spend much effort searching for risk factors for disease and in many well-established circumstances (such as blood pressure and risk of stroke) interventions which reduce the risk factor reduce the related outcome in line with expectations. However, there are also numerous examples in the literature of strong, consistent associations in observational studies between risk factors or biomarkers and disease outcomes which have not been shown to be causal in randomised trials. A recent example was the promising positive association between blood homocysteine levels and risk of cardiovascular disease which in observational studies remained clearly significant even after controlling for other cardiovascular risk factors. However, several large randomised trials of lowering blood homocysteine levels with B vitamins have failed to demonstrate any effect on cardiovascular events despite significant reductions in homocysteine levels.
In the past, blood pressure-lowering, lipid-lowering and many other drugs have been licensed based on small randomised trials showing reductions in risk factor or biomarker levels. But recent experience with the cholesterol ester transfer protein inhibitor, torcetrapib, which, despite reducing low-density lipoprotein cholesterol and increasing high-density lipoprotein cholesterol, caused harm not benefit, suggests that this strategy is not always a safe one.20 Torcetrapib was not licensed and appropriately was only being used in the context of randomised trials, but there are several past examples of treatments in common use which were later shown to be harmful in large randomised trials. To name a few: anti-arrhythmic drugs to suppress ventricular tachycardias,21 hormone replacement therapy22 and steroids in head injury.23 One wonders how many treatments are currently in widespread use without randomised evidence of effects on longer-term clinical outcomes which might be causing harm – perhaps non-steroidal anti-inflammatory drugs.24

**REFERENCES**


13 Stewart PM, Sears A, Tomlinson JW et al. Regulation – the real threat to clinical research. BMJ 2008; 337:a1732. doi:10.1136/bmj.a1732


25 Macintyre S. Good intentions and received wisdom are not good enough: the need for controlled trials in public health. J Epidemiol Community Health 2010; epub ahead of print 8 Dec. doi:10.1136/jech.2010.124198

**CONCLUSION AND FUTURE PERSPECTIVES**

Randomised trials are the cornerstone of the current practice of medicine and are a powerful lever for change. As illustrated above, experience indicates that there are no shortcuts and that large randomised trials are essential for reliably assessing the clinical effects of treatments. However, these large trials are increasingly expensive and complex to undertake and innovative methods, perhaps better exploiting electronic media, need to be found to allow the development of large cost-effective randomised trials of many more interventions. Steps are also urgently needed to simplify the regulation of clinical trials to ensure that they remain viable. Randomisation is now the accepted norm in healthcare evaluation and mandatory before most new interventions are introduced. This acceptance is a major advance which needs to be built on and strengthened rather than dismissed. Perhaps, wider use of randomisation in other areas of life might provide a more robust evidence base to inform policy in other areas.25

© 2011 RCPE