Evidence-based medicine. Part 2: Randomised controlled trials

Tim Holt

Article points

1. The three ingredients for a modern randomised controlled trial are: randomisation, allocation concealment, and blinding.
2. The value of unbiased randomisation is that all factors, even those we cannot currently imagine to be relevant, should be evenly distributed between the arms.
3. Ethical research requires transparency in the design, conduct and reporting of trials, data monitoring to detect conclusive benefits or harms, and equipoise to justify the involvement and randomisation of participants.

Key words

- Bias
- Equipiose
- Randomised controlled trial

Tim Holt is Academic Clinical Lecturer at the Department of Primary Care, University of Oxford, Oxford.

The modern randomised controlled trial is the most reliable inferential means of investigating treatment effects. Its development during the 20th century turned clinical medicine into a rigorous experimental science. The second in a series, this article discusses the practical challenges of conducting trials including randomisation, allocation concealment and blinding, and explains concepts such as intention to treat and equipoise. Large diabetes randomised controlled trials are used as examples.

The controlled trial has a long historical tradition dating back at least to Biblical times (Daniel 1:1–21, New International Version; Stolberg et al, 2004), but the modern approach is usually traced to the deck of the British frigate HMS Salisbury, navigating the Bay of Biscay in April 1747. It was here that James Lind, ship's surgeon and naval hygienist, compared the effect of citrus fruits with several control treatments (including untreated sea water) for sailors afflicted by scurvy (Lind, 1753). Improvements in those allocated to citrus fruits were dramatic, a discovery that changed maritime practice, improved public health, and led to the terms "limey" for the British and "ascorbic" acid for the responsible ingredient later identified: vitamin C. From this time such fruits were included in sailors' diets, to improve both health and productivity.

This was an early example of putting evidence into practice, but perhaps surprisingly, the fully developed method we now recognise had to wait a further 200 years. In a 1948 *British Medical Journal* editorial (Hill, 1948), Austin Bradford Hill described the Medical Research Council's (MRC) trials of streptomycin for pulmonary tuberculosis (TB). These represent a landmark in research methodology and his paper outlines three ingredients for a modern randomised controlled trial (RCT): randomisation, allocation concealment, and blinding.

Sealed envelopes

Bradford Hill described "an ingenious system of sealed envelopes" used to reduce various biases. Randomisation ensures that all participants are equally likely to be allocated to each study arm; allocation concealment means that the allocation for the next participant is not known to the recruiter. Blinding (unawareness of participants and observers over the treatment received) is very desirable but not always possible in practice. The outcomes of the MRC trials were based on interpretation of chest X-rays by assessors blinded to the treatment allocation.

Sixty years later, sealed envelopes remain a commonly used means of allocation

concealment, despite documented attempts to hold them up to a light bulb or to heat them open (Schulz, 1995). Large trials more typically remove such temptations by using automated electronic randomisation mechanisms that allocate participants using centralised clinical trials unit resources.

A large trial may have sufficient power to detect tiny differences in outcomes that are statistically significant but not clinically important. More often, power is less than desired due to the practicalities of recruitment, particularly in "hard to reach" populations, and this can lead to under-representation of such groups in published research (for example, treatment naïve or people with poor glycaemic control).

The larger the treatment effect, the fewer participants are required. A typical RCT that has 90% power at 5% statistical significance has a 10% risk of failing to demonstrate a true effect and a 5% risk of identifying a spurious effect arising by chance alone. The same trial will have lower power if it turns out that the effect under investigation is not as strong as assumed in the sample size calculation.

Sometimes factors likely to influence response to an intervention can be identified, such as age, sex, other treatments, or comorbidities. Trial reports typically confirm the success of randomisation by tabulating such characteristics in each arm at baseline, to show that differences are non-significant. Significant differences might be due to bias in the randomisation process, although inclusion of multiple factors makes it likely that at least one may be different purely by chance. The value of unbiased randomisation is that all factors, even those we cannot currently imagine to be relevant, should be evenly distributed between the arms. Following randomisation, it is essential that participants are treated, followed up and assessed in a way that maintains this same principle: the only difference is receipt of the intervention.

Intention to treat

Investigators later in the 20th century recognised sources of bias not originally evident in the 1940s. One of these involves "fidelity" to the treatment arm after randomisation. This is particularly evident in "pragmatic" trials, where we are interested not so much in whether an intervention works for those complying perfectly, but in whether it will also produce benefits despite lack of compliance, withdrawal due to toxicity, loss to follow-up, or departure from the protocol for whatever reason. This is the difference between efficacy and effectiveness.

The principle of "intention to treat" (ITT) ensures that outcomes are analysed on the basis of randomisation, not on who actually received the completed intervention. So a person allocated to surgical rather than medical treatment who dies before the operation, will be included as a death in the surgical arm in the ITT analysis. This principle is well established but still inadequately applied in many trial reports (Hollis and Campbell, 1999).

Equipoise

James Lind's citrus fruits were clearly effective despite a sample size of just two sailors in each arm. This raises the question of how necessary it is to actually trial an intervention when its benefit seems beyond doubt. In 2003 Smith and Pell questioned (with levity) whether parachutes might be subjected to a randomised controlled trial (RCT) (Smith and Pell, 2003). After all, many people have died despite using a parachute, while some have survived jumping out of an aeroplane without one. The number needed to supply with a parachute to save one life is assumed to be very low, but high-quality research evidence is lacking.

Only the bravest of the brave would take part in a parachute RCT*, and we should never expect trial participants to commit acts of bravery. Recruitment and randomisation is only ethical if it is not obvious that a participant will benefit from allocation to one arm over another. This is referred to as equipoise. While clearly an important ethical principle, it can be an obstacle to addressing questions that are probably but not definitely

* There are numerous apocryphal accounts of Gurkha recruits volunteering for an airdrop without parachutes.

Page points

1. Sometimes factors likely to influence response to an intervention can be identified, such as age, sex, other treatments, or comorbidities.
2. Following randomisation, it is essential that participants are treated, followed up and assessed in a way that maintains this same principle: the only difference is receipt of the intervention.
3. The principle of "intention to treat" ensures that outcomes are analysed on the basis of randomisation, not on who actually received the completed intervention.
4. Recruitment and randomisation is only ethical if it is not obvious that a participant will benefit from allocation to one arm over another. This is referred to as equipoise.
5. Diabetes research includes numerous examples that were controversial in their degree of equipoise.

Page points

1. In the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial, tight glycaemic control was defined through a target HbA_{1c} level of <6.0% (<42 mmol/mol), with 50% of the intervention arm population achieving 6.4% (46 mmol/mol) versus 7.5% (58 mmol/mol) in the control arm.
2. Glycaemic control in ACCORD was tighter than that achieved in the UK Prospective Diabetes Study, and the treatment strategies were also conspicuously different. The exact cause for the excess mortality is still unclear, and was not apparently due directly to hypoglycaemia.
3. The ACCORD study involved 10251 participants and was large enough to detect a difference of three deaths (14 versus 11) per 1000 per year between arms. Such a difference might have escaped notice in a smaller study.
4. In the past it has been relatively easy for investigators to use different outcome measures in the final analysis than those originally identified in the trial protocol.

answered. An intervention may be of proven benefit in an overall population, but unproven in a narrower subgroup. This subgroup may need to be trialled separately, and it is not surprising that discussions arise over whether such a trial is necessary. Diabetes research includes numerous examples that were controversial in their degree of equipoise.

The UKPDS (UK Prospective Diabetes Study; UKPDS Group, 1998a; b) clearly demonstrated the benefits of good glycaemic control in its population of people newly diagnosed with type 2 diabetes. Was it therefore necessary to recruit volunteers to more recent trials of tighter glycaemic control (each risking randomisation to the perhaps less popular comparator arm)? Control subjects were not expected to drink seawater instead of sucking oranges, let alone jump out of an aeroplane without a parachute. In fact participants were only risking allocation to a control regimen that was superior to background care operating outside the trial, but were nevertheless pitting their personal outcomes against an established consensus at the time: that good glycaemic control is generally beneficial in type 2 diabetes.

Without this commitment, we would not have discovered that people with type 2 diabetes of ≥8 years duration and at risk of cardiovascular events in fact are more likely to die through a strategy of tight glycaemic control (achieved if necessary through multiple medications) compared with a less intensive approach. This finding required the stopping of the ACCORD (Action to Control Cardiovascular Risk in Diabetes; ACCORD Study Group et al, 2008) trial. In this study, tight glycaemic control was defined through a target HbA_{1c} level of <6.0% (<42 mmol/mol), with 50% of the intervention arm population achieving 6.4% (46 mmol/mol) versus 7.5% (58 mmol/mol) in the control arm. This was tighter than the glycaemic control achieved in the UKPDS, and the treatment strategies were also conspicuously different. The exact cause for the excess mortality is still unclear, and was not apparently due directly to hypoglycaemia (ACCORD Study Group et al, 2011).

The current ASCEND (A Study of Cardiovascular Events in Diabetes) trial is testing the cardiovascular benefits of low dose aspirin (and omega-3 fatty acids) in people with type 1 or 2 diabetes aged >40 years (see www.ctsu.ox.ac.uk/ascend for further information). Recruitment began in 2008 and coincided with the publication of NICE clinical guideline 66 (National Collaborating Centre for Chronic Conditions, 2008), which recommended aspirin for most of this group, raising issues over equipoise. However, the evidence available to NICE was inconclusive, as stated in a subsequent Medicines and Healthcare products Regulatory Agency (2009) bulletin, and this important question needs to be answered through this trial.

Detecting harms

The ACCORD study involved 10 251 participants and was large enough to detect a difference of three deaths (14 versus 11) per 1000 per year between arms. Such a difference might have escaped notice in a smaller study. Conversely, a still larger trial might fail to detect a serious adverse event occurring with an excess of one person per 1000. Single RCTs are usually not large enough to exclude potentially important rare events, and many fail to demonstrate benefits due to inadequate sample sizes. To identify rare hazards, meta-analysis, post-marketing surveillance, or alternative study methods such as observational cohort or case-control studies are required.

Moving the goal posts

In the past it has been relatively easy for investigators to use different outcome measures in the final analysis than those originally identified in the trial protocol. In a study of this issue Chan and Altman (2005) reported that in 519 trials listed on PubMed as published during December 2000, over 20% of outcomes were incompletely reported, and these were more likely to be statistically non-significant than fully reported outcomes (Chan and Altman, 2005). This and other evidence

has prompted the move towards obligatory publication of RCT protocols. Registers such as ISRCTN (International Standard Randomised Controlled Trials Number) and ClinicalTrials.gov attempt to set the goal posts in concrete before outcome analysis begins. Those judging the quality of completed trial reports (or considering them for publication) can now consult these protocols, and check for signs of movement in the concrete.

CONSORT

Increasing reliance on RCT evidence created the need for standardisation in the reporting of trials. The CONSORT (Consolidated Standards of Reporting Trials) statement was last revised in 2010 (CONSORT, 2010), and has undergone numerous extensions. It sets a framework for the reporting of trials, allowing the progress of participants to be followed through the study. This promotes transparency and allows the reader to identify vulnerabilities in trial design and practical problems including loss to follow-up. Trials are difficult to conduct and almost inevitably include imperfections, but when clearly and openly reported the reader can gauge the importance of these issues and the resulting risk of bias where present.

The CARDS trial

Another example from the diabetes literature is the CARDS (Collaborative Atorvastatin Diabetes Study; Colhoun et al, 2004), which tested the effect of atorvastatin (versus placebo) for primary prevention of cardiovascular events in people with type 2 diabetes.

The protocol was registered on ClinicalTrials.gov (as NCT00327418) and published in *Diabetic Medicine* (Colhoun et al, 2002) before the outcome data were analysed. The trial report identifies that it is an RCT in the title. The abstract is structured and is followed by a scientific background stating the aims. This section raises the issue of equipoise. Could allocation to the placebo arm be justified given what was already known? Recruitment occurred between November 1997 and June 2001. For the primary prevention type 2 population, the Heart Protection Study showed benefit from lipid lowering although this trial was not published at this point in time (Collins et al, 2003).

Risk reduction in a subgroup of the ASCOT–LLA (Anglo-Scandinavian Cardiac Outcomes Trial – Lipid Lowering Arm) trial with diabetes but no cardiovascular disease (Sever et al, 2003) was not significant (despite a significant result for the trial as a whole). The CARDS investigators argued that the benefits were not proven conclusively and a definitive trial of this specific population was required. Participants were

excluded if their lipids were outside specified thresholds (LDL-cholesterol >4.14 mmol/L or fasting triglycerides >6.78 mmol/L) as the withholding of lipid-lowering drugs in such people was considered inappropriate.

The methods are described in detail including the randomisation mechanism and means to protect allocation concealment. Results include detailed characteristics in both arms at baseline and there is a flow diagram describing small numbers lost to follow-up. Of those initially randomised, 99% were available at termination of the trial for inclusion in the analysis. The study population had a mean age of 62 years, 68% were men and 94% were of white ethnic origin (suggesting possible under-representation of minority groups).

CARDS was stopped 2 years early (median duration 3.9 years) due to clear benefits in the atorvastatin arm in all major cardiovascular events (rate reduction 37% [95% confidence interval [CI], –52 to –17]). Acute coronary events were reduced by 36% [95% CI, –55 to –9], and stroke by 48% [95% CI, –69 to –11].

No lower limit of baseline LDL-cholesterol level was required for inclusion in CARDS, and this trial forms part of the basis for subsequent guidelines recommending statins for the majority of people with type 2 diabetes. In the final line of the published report, the authors return to the issue of equipoise for future research:

> *"The debate about whether all patients with type 2 diabetes warrant statin treatment should now focus on whether any patients can reliably be identified as being at sufficiently low risk for this safe and efficacious treatment to be withheld."*

The current NICE guideline (National Collaborating Centre for Chronic Conditions, 2008) recommends the identification of a low risk type 2 diabetes population not requiring such therapy and annual assessments of cardiovascular risk for these people. But the majority are at significant cardiovascular risk and should receive a statin unless contraindicated, irrespective of their baseline cholesterol level.

Conclusion

The modern RCT is the most reliable inferential means of investigating treatment effects. Its development during the 20th century turned clinical medicine into a rigorous experimental science.

However, the process of conducting trials is beset with practical challenges. Randomisation itself removes the influence of confounding variables, including hidden factors that may never be identified. This is an important source of bias affecting other methods. A single RCT may be insufficiently powered to measure the size of the treatment effect accurately, or to exclude important hazards. These require meta-analysis, post-marketing surveillance, and alternative study designs. Ethical research requires transparency in the design, conduct and reporting of trials, data monitoring to detect conclusive benefits or harms, and equipoise to justify the involvement and randomisation of participants. ■

"The modern randomised controlled trial is the most reliable inferential means of investigating treatment effects. Its development during the 20th century turned clinical medicine into a rigorous experimental science."

ACCORD Study Group, Gerstein HC, Miller ME et al (2008) *N Engl J Med* **358**: 2545–59
ACCORD Study Group, Gerstein HC, Miller ME et al (2011) *N Engl J Med* **364**: 818–28
Chan AW, Altman DG (2005) *BMJ* **330**: 753
Colhoun HM, Thomason MJ, Mackness MI et al (2002) *Diabet Med* **19**: 201–11
Colhoun HM, Betteridge DJ, Durrington PN et al (2004) *Lancet* **364**: 685–96
Collins R, Armitage J, Parish S et al (2003) MRC/BHF *Lancet* **361**: 2005–16
CONSORT (2010) The CONSORT Statement. Available at: http://www.consort-statement.org/consort-statement/ (accessed 01.03.11)
Hill AB (1948) *Br Med J* **2**: 791–2
Hollis S, Campbell F (1999) *BMJ* **319**: 670–4
Lind J (1753) A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Kincaid A, Donaldson A, Edinburgh
Medicines and Healthcare products Regulatory Agency (2009) *Drug Safety Update* **3**: 10–11
National Collaborating Centre for Chronic Conditions (2008) *Type 2 Diabetes: National Clinical Guideline for Management in Primary and Secondary Care (Update). Clinical Guideline 66*. NICE, London
Schulz KF (1995) *JAMA* **274**: 1456–8
Sever PS, Dahlöf B, Poulter NR et al (2003) *Lancet* **361**: 1149–58
Smith GC, Pell JP (2003) *BMJ* **327**: 1459–61
Stolberg HO, Norman G, Trop I (2004) *AJR Am J Roentgenol* **183**: 1539–44
UKPDS Group (1998a) *Lancet* **352**: 837–53
UKPDS Group (1998b) *Lancet* **352**: 854–65