Research Review
EDUCATIONAL SERIES

Identifying the Common Pitfalls in Randomised Controlled Trials – Is the Patient Unconscious or Dead?

About the Reviewer

Associate Professor Chris Frampton

Chris is an Associate Professor in the Department of Biostatistics at the University of Otago, Christchurch. He has 25 years of experience in statistical analysis and design. He has a particular interest in the statistics associated with drug development studies, with an emphasis on phase III RCTs.

Subscribing to Research Review

To subscribe or download previous editions of Research Review publications go to

www.researchreview.co.nz

About Research Review

Research Review is an independent medical publishing organisation producing electronic publications in a wide variety of specialist areas.

Research Review publications are intended for New Zealand medical professionals.

This review focuses on important and often flawed aspects of published randomised controlled trials (RCTs). Many of these failings occur within key features of the design of the study. These features are often fundamental to the principles of medical research as exemplified in RCTs, e.g. blinding and representativeness of participant samples. These weaknesses may be unavoidable in some contexts and in others may have no significant clinical ramifications. It is the implications of these non-ideal designs that provide the challenge to those reviewing the study, both from an academic and a clinical perspective. In any given situation, there may be no definitive conclusion as to the implications of a non-ideal design and for this reason the onus is on those designing and conducting studies to produce as rigorous a design as is possible in the specific clinical context.

Readers of this review should refer to "*Clinical Trials – An Overview*", a paper that explains and defines clinical trial terminology for a target audience of doctors, nurses and pharmacists reading clinical trial reports.

Blinding: to what extent has achievable blinding been undertaken; what are the implications for non- or partially-blinded studies?

Blinding within RCTs is undertaken to remove the possibility of bias contaminating the comparisons between the randomised treatments. Whether this bias may increase or decrease the differences between treatments is largely irrelevant.

Participants who are aware of which treatment they are receiving may believe that the treatment is offering an advantage (a new treatment) or if aware that they are receiving the control/placebo treatment may believe that they are receiving an inferior treatment. The bias introduced by a lack of participant blinding will therefore manifest most strongly when outcomes are self-reported or in circumstances where the participants' perceptions may influence the outcome. This will occur where quality of life (e.g. EORTC QLQ-C30, which assesses the quality of life of cancer patients), self-reported measures of functioning (e.g. SF12, which measures general functional health status) and symptoms (e.g. BDI [Beck depression inventory]), and where non-specific maladies not easily confirmable through objective diagnosis (e.g. headaches, fatigue) are reported. The bias is also likely to be more profound in placebo-controlled trials where participants may be aware that they are receiving 'no treatment'. Participants aware of their treatment allocation may also show better or poorer compliance to treatment as a consequence. Those on placebo/control may see less point in taking what they perceive to be an ineffective treatment.

Investigators/assessors who are not blinded to participant treatment also have the potential to introduce bias into the comparisons of treatments. While professionals can reasonably claim a degree of objectivity in terms of how they treat and assess patients, they are of course not free of prejudice. The motivation for undertaking or contributing to the trial may of itself indicate a preconception as to the benefit of a new treatment. If investigators are aware of the randomisation sequence for a RCT they have the potential to allocate a participant into the treatment they believe is the most appropriate for the participant, thereby introducing bias. If they are aware of which treatment a participant has been allocated to they have the potential to influence both concomitant treatment and the assessment of clinical status, both of which will introduce bias to the comparison of treatments. Rather than debating the integrity of investigators and the extent to which bias may be introduced into the comparison, it is better to improve the rigour of the RCT by blinding those involved in treating and assessing participants. For example, well-conducted later phase oncology trials, particularly those with progression-free survival or response rates as primary outcomes (e.g. **CRYSTAL**, a RCT of cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer) will have an independent blind assessment/review of participant progression and response data. Some studies go to extensive lengths (double-dummy and double-blind) to ensure that both participants and investigators are blind to treatment allocation (e.g. **ARISTOTLE**, a RCT of apixaban versus warfarin in patients with atrial fibrillation and **ADVANCE-3**, a RCT of apixaban versus enoxaparin for thromboprophylaxis after hip replacement).

There are many circumstances in which participant or investigator blinding is simply not possible (e.g. **SPARX**, a RCT comparing treatment as usual with a computerised self-help resource for adolescent depression and **ALCCaS**, a RCT comparing laparoscopic and conventional open surgery for colon cancer). When interventions differ profoundly (e.g. surgery vs no surgery), dosing forms are different (e.g. IV vs oral) or when one of the treatments has unavoidable side effects or clinical symptoms (e.g. changes in INR as occurred in the **ROCKET-AF** and **RE-LY** studies – where rivaroxaban and dabigatran respectively were compared with warfarin in non-valvular atrial fibrillation – warfarin alters INR but rivaroxaban and dabigatran do not), true blinding of participants or investigators may be impossible. In such situations, all efforts should be made to mitigate the potential bias introduced, by having objective outcomes (laboratory measures, survival) or an independent blinded assessment of outcomes. The assessment of compliance when relevant is also an important component to consider when studies are not blinded.

The onus in reporting a RCT is to show that bias introduced by non-optimal blinding could not have influenced the comparison and hence the clinical conclusions. It is not the responsibility of anyone critiquing the trial to show that the presence of bias has influenced results; the possibility of bias is enough to cast doubt on the conclusions. Studies should be designed so they reduce the possibility of bias as far as is practicable.

Representativeness: how typical is the participant population of the wider target clinical population?

For the results of a RCT to be generalisable to a wider clinical setting it is important that the participants enrolled in the trial are representative of the population presenting with the disorder. A clinician needs to evaluate the demographic and clinical features (notably co-morbidities, concomitant medications, disease severity) of the participant population to determine whether the results of the trial are likely to be relevant in their setting.

Frequently, RCTs are criticised because their participant populations are not, by design, representative of those with the targeted disorder. Inclusion/exclusion criteria are often very extensive, typically removing those with compromised kidney function, pregnant women, the very old and very young, those on medications potentially interacting with the study medication and those with serious co-morbidities. These are sensibly justified on safety grounds, usually because not enough is known about the new treatment at the point at which early phase studies are undertaken. In recent times, the common exclusion of participants with suicidal ideation from antidepressant studies has led to a protracted debate of the effects of antidepressants on those with severe, potentially suicidal depression.

Involvement in a RCT may involve a considerable burden for participants, such as additional visits and monitoring, restrictions on additional treatments or lifestyle options. The RCT is an experimental environment and by its very nature carries some degree of risk to the participant. Many participants eligible for the trial may not consent to be recruited into the trial, as they are not prepared to be part of an 'experiment' or may have an aversion to one or other of the treatments being assessed and are not prepared to risk being randomised to that treatment. These also have the potential to produce a trial population that is not representative of the wider clinical group. For example, females and retirees are generally more likely to consent to be included in trials.

The most important consideration in reviewing the repercussions of a non-representative study population is whether the relative effects of the treatments will translate to those not included or under-represented in the trial. The factors most likely to have a bearing in this context are age, gender, kidney function and, most importantly, disease severity. There are many studies across a broad spectrum of medical conditions which have shown an 'interaction' of treatment effects with disease severity. The treatment may work less effectively in absolute terms in the more or less severe group, which is why percentage change (effectively correcting for baseline differences in severity) is often chosen as an outcome measure. The treatment may also work more or less effectively in; younger or older patients (as was indicated by the *a-priori* stratification and analysis of the influence of age on the efficacy of a natriuretic peptide-guided treatment for heart failure in BATTLESCARRED), those with impaired renal function and male or female patients. These concerns may to some degree be allayed with a larger sample size if results are presented overall and separately for severity, age, renal function and gender groups.

A further repercussion of an inappropriate participant population occurs when results are combined or included within meta-analyses or systematic reviews. There are many reasons for conducting meta-analyses; however, it is rare for treatments to be introduced as a consequence of a single RCT, as a greater body of evidence is required. An assumption for pooling studies within a meta-analysis (less important for individual participant meta-analyses) is that the participant populations are comparable in terms of key demographic, disease severity and other clinical features. Therefore, studies with non-typical participant groups may not be included in these analyses.

The onus is again on the study to produce results which are as generalisable as possible.

A rigorously designed and conducted RCT may be of little clinical utility if the trial population is so manipulated that the results are not translatable to the wider target clinical group.

Participant management (treatment/assessment): is the participant management of the study group typical of that likely in the target patients?

The involvement of participants in a RCT means they are not subjected to usual clinical practice. As a rule, the participants in a trial will be seen more regularly and have potentially more extensive clinical management than would normally occur, particularly with regards to diagnostic and side effect monitoring. As a consequence, and putting to one side the specific effects of the treatments, it is often assumed that participants will do better if they are in RCTs. The generalisability of a new treatment into standard practice, therefore, does not only depend on the participant population; all study procedures undertaken should also be considered.

When translating RCT results into a clinical setting where monitoring is less intensive, are there any consequences for the treatment efficacy or risks? For example, will a lack of efficacy or the development of side effects be detected as quickly or accurately in a standard clinical setting? It may be that the negative aspects of a treatment are not so well managed in the usual clinical practice and that this counteracts any positive effects of treatment.

The RCT protocol for participant management needs to be considered when assessing whether a new treatment should be adopted.

A-priori form of study comparison: superiority, non-inferiority or equivalence

In broad terms, there are three statistical forms that the comparison of treatments from a RCT may take. Superiority trials hypothesise that a treatment is superior in terms of the primary outcome with regards to the comparison group. Non-inferiority trials hypothesise that a treatment has comparable or superior efficacy, so it may have an advantage in other regards, e.g. fewer adverse effects or cost and therefore, doesn't need to have superior efficacy to be considered a 'better' treatment. Equivalence trials, which frequently involve generic drugs, hypothesise that the treatments are comparable (not higher or lower) in terms of the primary outcome. The primary outcome for equivalence trials is more likely to be a pharmacokinetic measure (C_{max} or AUC) than an efficacy or safety measure, to ensure that drug levels remain within a therapeutic range.

How are the objective clinical definitions of superiority, non-inferiority and equivalence quantified and defended? It is important to understand the statistical null (default) scenario for each of these three designs and, therefore, to know how to interpret statistically significant and non-significant results.

Superiority trials are established and statistically powered to show superiority. In order to calculate statistical power the term 'superiority' with respect to the primary outcome needs to be defined. At what minimum point, e.g. with a 5%, 10%, 20%, or 50% advantage would we consider this new treatment in its entirety (considering cost, adverse effects) to be clinically superior? A larger advantage is more likely to be considered clinically significant and requires a smaller sample size for statistical power. However, if a smaller than anticipated difference is shown with the RCT, a difference which might perhaps be considered clinically relevant, then statistical significance may not be apparent, as the study was not powered for this smaller effect. Consequently, if the treatment is shown as 'not superior', the study is considered a 'failure' and the treatment may not be further researched. So there is a risk in overstating the superiority limit. In a superiority trial, if the primary result is not statistically significant, then the null hypothesis is not rejected and superiority is not shown (e.g. **ALCCaS**, the RCT comparing laparoscopic and conventional open surgery for colon cancer where the primary outcome, 5-year mortalities, were 77.7% and 76.0% respectively, p=0.94). The onus is on the study to establish superiority and if this is not shown then the null (non-superiority) scenario is maintained. From a confidence interval perspective, if the 95% confidence limits of the difference or ratio are not both above the superiority margin, then superiority has not been shown.

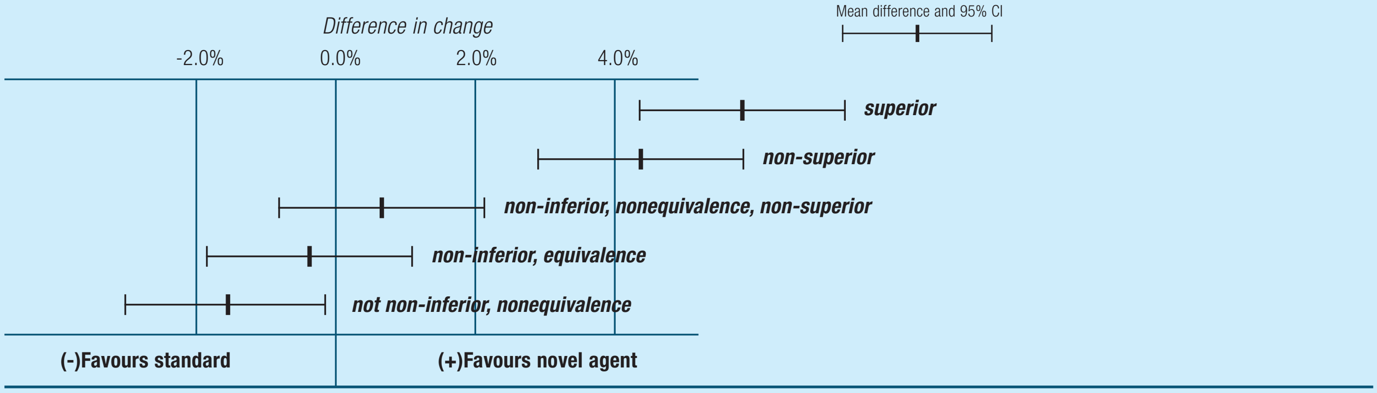
Non-inferiority trials are established and powered to show that the novel treatment is not inferior to the comparator (e.g. **ARISTOTLE** was designed to test the non-inferiority of apixaban compared with warfarin in the rates of ischaemic or haemorrhagic stroke or systemic embolism, and many of the recent HIV studies comparing LPV/r with e.g. DRV/r for the reduction in HIV-RNA copies). Non-inferiority trials are usually undertaken in circumstances whereby the novel treatment would be adopted if it were no worse than the accepted comparator treatment. In order to calculate statistical power, the term 'non-inferiority' with respect to the primary outcome needs to be defined. At what minimum point, e.g. with a 5%, 10%, or 15% disadvantage would we consider this new treatment in its entirety (considered in the context of reduced cost, fewer adverse effects) to be clinically inferior? A larger effect is more likely to be considered clinically inferior and requires a smaller sample size for statistical power. Non-inferiority limits are usually comparable to or slightly smaller in magnitude than those defined for superiority trials and as a consequence the sample sizes necessary for non-inferiority trials frequently exceed those for superiority trials. In a non-inferiority trial, if the primary result is not statistically significant then the null hypothesis is not rejected and non-inferiority is not shown. The onus is on the study to establish non-inferiority; if this is not shown then the null (not non-inferior) scenario is maintained. If non-inferiority is shown then superiority can be tested with no adjustment of the type I error, (alpha level) for the statistical comparison. From a confidence interval perspective, if the 95% confidence limit(s) of the difference or ratio are not above the inferiority margin, then non-inferiority has not been shown (e.g. **PHARE**, a non-inferiority trial comparing 6 and 12 months' trastuzumab treatment in adjuvant early breast cancer, with non-inferiority defined as a hazard ratio (HR) of 1.15 or less for disease-free survival. The null hypothesis of not non-inferior could not be rejected after an early interim analysis (HR=1.28, 95% CI 1.05 – 1.56; the upper limit of 1.56 being much higher than the non-inferiority margin of 1.15), so it is extremely unlikely from here that non-inferiority would be shown and therefore recruitment has been halted.

Equivalence trials are established and powered to show that the novel treatment is neither 'better' nor 'worse' than the comparator with regards to the primary outcome. These studies do not usually test efficacy or safety outcomes. Usually they involve pharmacokinetic or pharmacodynamic outcomes for which higher or lower values may be outside therapeutic limits and lead to inadequate efficacy or increased adverse events. Frequently these studies test generic products against the parent product. In order to calculate statistical power, the term 'equivalence' with respect to the primary outcome needs to be defined. Commonly with pharmacokinetic trials, the prescribed limits for AUC and C_{max} measures are 80% – 125% or 90% – 111%. In an equivalence trial, if the primary result is not statistically significant then the null hypothesis is not rejected and equivalence is not shown. The onus is on the study to establish equivalence; if this is not shown then the null (not equivalent) scenario is maintained. From a confidence interval perspective, if the 90% confidence limits of the difference are not both contained within the equivalence limits, then equivalence has not been shown.

Potential trial outcomes and decisions for novel agent

Defining difference in % change for 3 trial types

1.	**Superiority**	**≥4%**
2.	**Non-inferiority**	**≥-2%**
3.	**Equivalence**	**≥-2% and/or ≤2%**



Choice of outcomes and the length of follow up: to what extent are the outcomes surrogates for clinically meaningful outcomes?

There are many considerations determining the choice of the primary outcome for a randomised clinical trial. Ideally the outcome needs to be precise (reduces the necessary sample size), specific to the disorder, objective, have the potential to show a response in a timely manner and be widely considered as clinically indicative of the status of the disorder being studied. For many disorders, there may be a significant lag time before the full effects of a novel treatment manifest, (e.g. mortality benefit in oncology trials or fracture reductions from bisphosphonates). To complete RCTs as quickly as possible often the primary outcome chosen is one that will respond quickly but is believed to be strongly correlated with the ideal outcome. The choice of such surrogate endpoints is frequently problematic and most often restricted to phase II studies. Reduced lipid levels and lower blood pressure may serve as an appropriate surrogate for a reduction in cerebrovascular events in short-term phase II studies, but these reductions do not directly translate into a reduction in events and are not adequate for pivotal phase III studies. Even in phase III studies the choice of an appropriate timely outcome may not be straightforward. For example, some cancers show a strong link between disease-free survival and overall survival (e.g. metastatic colorectal cancer), whereas for others, the association is not strong (breast cancer) and therefore, a benefit with one treatment shown for disease-free survival may not be apparent in terms of overall survival.

The duration of follow-up is particularly important in the context of time-to-event outcomes (e.g. disease-free survival, overall survival). These studies need to explicitly stipulate in advance at what point the primary analysis of these outcomes will occur. This point might be defined on the basis of: (i) all participants having a minimum follow-up period or (ii) the total number of ‘events’ occurring in both groups. In studies with time-to-event outcomes, since data continues to accumulate after recruitment has ceased and the primary analyses have occurred, there is an opportunity for further unrestricted analyses of the outcomes. Such analyses can be considered *post-hoc* (similar to testing many outcomes looking to find one that shows statistical significance) and thereby inflate the true type I error rate, potentially leading to apparent statistical significance i.e. claiming statistical significance at a certain alpha level (e.g. 0.05) when in fact the level is much higher, when significance is really not achieved. However, on occasion, additional analyses are undertaken to determine whether significant effects identified at the primary analyses persist over a longer period of time (e.g. recent reporting of the median 8-year follow-up of the HERA trial confirmed the effect seen after 1 year of trastuzumab treatment compared with observation – the overall survival HR of 0.76 at 1 year persisted, with an 8-year HR estimate of 0.76).

When evaluating trials, it is important to consider the true clinical status of the outcome measures and in time-to-event studies ensure that the analyses you are assessing do not represent the ‘highlight result’ from many statistical comparisons of the same outcome.

Additional *post-hoc* analyses: are these data dredging undertaken in circumstances where key *a-priori* comparisons do not reveal the results investigators anticipated?

Post-hoc analyses are those statistical comparisons outside the key comparisons stipulated in the trial protocol. These may involve additional outcomes, sub-scales, or individual items and sub-groups of the participant population. How should we interpret such analyses? Analyses of outcomes or sub-groups not specified in the protocol should be treated as exploratory only and potentially justify further research if there is some biological plausibility to the association. Analyses of pre-specified secondary outcomes or sub-groups are frequently undertaken with corrections of the p-values limit used to define statistical significance so that the true alpha level (e.g. 0.05) is preserved. However, statistical significance evident in secondary outcomes even under these circumstances but not in the primary outcome is unlikely to be sufficient to change clinical practice. These results would need confirmation either in further clinical trials or by meta-analyses of other completed trials.

The possibility of sub-group analyses presents a dilemma for researchers. It is self-evident that not all participants will respond equally to any treatment. Disease severity, age, gender and existing comorbidities will all potentially impact the effectiveness or safety of a treatment. However, if there is the potential for significant differential effects of a treatment then these need to be considered and factored into the design of the trial (e.g. as age was considered in the BATTLESCARRED study). Clear inclusion and exclusion criteria will match the proposed indication for the treatment and thereby remove some of the extreme heterogeneity in the outcomes. The protocol may also indicate that all participants are expected to respond but some may respond better than others. In which case interactions between factors (e.g. disease severity) and treatment may be pre-specified for testing, so that this can be formally and appropriately tested on a restricted set of factors. This scenario contrasts with ‘someone must have responded to this treatment and if we look hard enough we will find what they are’.

The results of such data dredging are highly likely to find such a group but are unlikely to be robust or repeatable.

Disclaimer: *This publication is an independent review of clinical trial terminology and broadly summarises the process of clinical trials. Readers are asked to refer to the references provided for complete details of all the terms used.*

Roche

This publication has been created with an educational grant from Roche Products (NZ) Limited.
The content is entirely independent and based on the author’s opinions.