

Study designs: Part 5 – Interventional studies (III)

Priya Ranganathan, Rakesh Aggarwal¹

Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, ¹Director, Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India

Abstract Several methodological and statistical aspects of clinical trials can affect the robustness of their results. We conclude the series of articles on “Interventional Studies” by discussing some of these features.

Keywords: Bias, clinical trials as topic, research design

Address for correspondence: Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, India.

E-mail: drpriyaranganathan@gmail.com

Received: 19-12-19, **Accepted:** 31-12-19, **Published:** 31-01-20.

In this last of the three pieces on interventional studies, we examine some additional aspects of clinical trials, which are crucial to ensure the validity of their results. These include:

- a. Choice of study outcomes
- b. Appropriate sample size
- c. Minimizing missing data
- d. Appropriate analysis technique
 - i. Intention-to-treat versus per-protocol analysis
 - ii. Choice of statistical test
 - iii. Adjustment for multiple testing.
- e. Complete and unbiased reporting.

CHOICE OF STUDY OUTCOME

The study outcomes are the variables that a research study sets out to measure. These should be chosen such that they capture the key effects of the study interventions. Study outcomes should be defined *a priori* (in the protocol; before the study commences), should be clinically relevant, should be amenable to quick and reliable measurement, should be sensitive to the effect of the study intervention, and should address the overall aim of the study. At times, a study may assess a few additional exploratory outcomes, which are

essentially hypothesis generating, and these hypotheses can then form the basis of future studies.

Most studies will have a single primary outcome (corresponding to the primary objective of the study) and a number of secondary outcomes (corresponding to the secondary objectives). For example, the DREAMS study compared the efficacy of dexamethasone versus standard therapy for postoperative nausea and vomiting in patients undergoing gastrointestinal surgery.^[1] The primary outcome was the occurrence of “any episode of vomiting within 24 h after surgery.” The study also assessed many secondary outcomes, including the number of episodes of vomiting, the need for anti-emetics, and severity of nausea and of vomiting.

Sometimes, a researcher may choose to study more than one (multiple) primary outcomes. Although this may provide a more comprehensive assessment of the effects of the experimental treatment, it carries an increased risk of false-positive results, as discussed in the section below on multiple testing. Hence, such studies need more careful planning and interpretation.

Access this article online	
Quick Response Code:	Website: www.picronline.org
	DOI: 10.4103/picr.PICR_209_19

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Ranganathan P, Aggarwal R. Study designs: Part 5 – Interventional studies (III). *Perspect Clin Res* 2020;11:47-50.

The sample size required for a study is calculated based on the expected difference in a primary outcome measure between the intervention and the control groups. Studies are often not sufficiently powered to definitively address the secondary outcomes.

Very often, in addition to the efficacy outcomes, some outcomes related to toxicity (e.g., the total number of adverse events or the number of individuals with specific adverse events, in each arm) are also included.

Outcomes can be of different types. Several considerations may influence the decision to choose some specific types of outcomes.

Surrogate outcomes

Researchers may choose to measure one or more biochemical or radiological parameters (which are often easier to measure and show a change over a shorter time frame) as substitutes for more direct outcomes - such as clinical improvement, improved survival, or reduced risk of disease recurrence. These are known as surrogate outcomes. For example, to assess the effect of a new treatment for diabetes, one may measure the change in glycosylated hemoglobin, although the real interest is the impact of experimental treatment on diabetic complications and end-organ damage. In prostate cancer, one could measure the changes in blood levels of prostate-specific antigen or tumor shrinkage after therapy; however, again, the real interest is in whether the treatment translates into a benefit in survival. Other examples include measurement of CD4 counts to assess the efficacy of antiretroviral therapy or of lipid levels for that of statins.

The use of surrogate outcomes is valid only if the changes in these correlate well with changes in clinical outcomes. Their use may sometimes lead to a misleading conclusion. Medical literature is replete with examples of drugs that were initially approved for marketing based on benefit in surrogate outcomes but were subsequently found to worsen clinical outcomes. For example, anti-arrhythmic drugs in myocardial infarction (MI) patients were found to suppress ventricular premature beats, which are known in this situation to be associated with increased mortality. Hence, these drugs were, for several years, recommended for post-MI patients.^[2] However, a subsequent trial showed that the use of these drugs, despite reducing the occurrence of premature beats (a surrogate outcome), was not associated with a reduction in more complex fatal arrhythmias (the desired clinical endpoint) and in fact led to increased mortality.^[2] Similarly, higher doses of erythropoietin in patients with renal failure improve hematocrit but lead to increased cardiovascular thrombotic events and death.^[3]

Composite outcomes

Researchers often combine many related outcomes into a single outcome measure known as a composite endpoint. For example, trials of cardiovascular diseases commonly use major adverse cardiovascular event (MACE) as a composite endpoint; this combines any myocardial infarction, cerebrovascular event (e.g., stroke), and cardiovascular death. Composite endpoints increase the total number of patients who have events of interest, improving the statistical power of the analysis of study results. However, one should be careful to combine only such outcomes that have the same biological pathway and are affected similarly by the study interventions.

Some considerations for integrating many outcomes into a composite endpoint include whether the components are of similar importance, whether they occur with somewhat similar frequency, and whether the intervention is likely to affect all the components similarly.^[4] A systematic review of studies with composite endpoints in cardiovascular medicine found that the largest treatment effects were seen in the components which were clinically less important, thus potentially misleading readers.^[5] Interestingly, in a trial of cariporide, a cardiovascular drug, the incidence of composite outcome (death or MI) showed a reduction from 20.3% in the placebo group to 16.6% in the treatment group; however, a closer look showed that though the incidence of MI had declined (from 18.9% to 14.4%), the mortality had in fact increased (from 1.5% to 2.2%).^[6]

Subjective versus objective outcomes

Objective or “hard” outcomes are those which are unambiguous and can be consistently measured by different assessors. On the other hand, subjective or “soft” outcomes are based on interpretation by the participant or assessor and can be associated with measurement bias. For example, in the DREAMS study, episodes of vomiting defined as projectile expulsion of gastric content would be a hard endpoint, whereas nausea (as experienced by the participant) is a subjective endpoint.^[1] Wherever possible, one should use objective endpoints, in order to minimize bias and improve the validity of study results. If subjective outcomes have to be used (since patient-reported outcomes are important though often subjective), all attempts must be made to reduce or eliminate bias, such as using blinding techniques (for patients and assessors) and standardized validated scales and scores. As an example, the DREAMS trial used standard validated scales to measure nausea, fatigue, and quality of life.^[1]

APPROPRIATE SAMPLE SIZE

Research studies begin with a statement of belief or a hypothesis. For conventional superiority studies, where the objective is to compare an experimental treatment (E) with standard treatment (S), we start with a null hypothesis – that there is no difference between the effects of treatment S and treatment E. The alternate hypothesis states that there is a difference between these effects.

Research studies are carried out in subsets (“samples”) from the entire universe (“population”) of individuals to whom the research question pertains. For example, to compare two drugs for the treatment of hypertension, ideally, we would randomly assign all the individuals with hypertension to receive either drug and compare the results. However, since this is not practical or feasible, we choose a sample of individuals with hypertension, compare the effects of the two anti-hypertensive drugs in them, and extrapolate the results to the rest of the population. In doing so, we run the risk of two types of errors.

1. Finding a difference between the effects of treatments when a true difference does not exist (i.e., there would be no difference if we could study the entire population). This is called a type 1 error or alpha error or a false-positive error. In terms of hypothesis testing, this means that we would falsely reject the null hypothesis and accept the alternate hypothesis
2. Not finding a difference between the effects of treatments when, in fact, a difference exists. This is known as a type 2 error or beta error or a false-negative error. This means that we falsely accept the null hypothesis and reject the alternate hypothesis.

Fortunately, statistical methods allow us to assess the likelihood of these errors. By convention, the upper limit of type 1 error is set at 5%. This means that if we observe a difference between the samples receiving new and the standard treatments, and the probability of this difference having occurred by chance is 5% or less, we conclude (with 95% or greater certainty) that the observed difference is a true difference.

In most studies, the type 2 error is set at 10% or 20%. This means that even if there is a true difference between the treatments in the population, there is a 10% (or 20%) probability that the study will fail to pick up this difference. The converse of beta error is the “power” of a study, which is defined as the ability of the study to detect a true difference in treatment effects (90% or 80%, in the above example).

These errors are more likely if the sample sizes are small. In particular, studies with a small sample size have low study power and a high risk of beta error. Thus, if a study with only a few subjects fails to find a difference between two treatments, this may reflect a failure to detect a difference even if one existed, rather than a true absence of difference. Hence, it is important to ensure that a study is designed to be sufficiently large to have a reasonable power, i.e., to have a reasonable likelihood of picking up a difference if one exists.

The formula for the calculation of the sample size required for a clinical trial is based on type 1 and type 2 errors that one is willing to accept and the expected difference between the treatment effects. The lower the type 1 and type 2 errors one permits, the larger is the required sample size. One may wish both these errors to be zero; however, this would mean an infinite sample size – an impossible task. Hence, as indicated above, we conventionally limit the acceptable type 1 error to 5% and the type 2 error to 10% or 20%. As for the treatment effect, if the expected difference in outcomes (or the difference that one wishes to detect) between the two groups is smaller or if the outcome measure (on a continuous scale) has a larger standard deviation, the required sample size is larger. The estimate of expected difference can be based on previous literature, a pilot study or the researcher’s assessment of what would be a clinically meaningful yet feasible difference between treatments. The calculated sample size is inflated by 10%–20% to account for protocol violations and losses to follow-up (please see the section on “Minimizing missing data” below), so that an adequate final number of observations is available for the analysis when the study ends.

Researchers are often tempted to use a large expected treatment difference to obtain a smaller estimate of the required sample size. However, if this is not a realistic difference, one would run a greater risk of negative study results.

All trial protocols (and reports) should include a detailed section on sample size calculation, allowing readers to assess whether the assumptions made are valid.

MINIMIZING MISSING DATA

During a trial, there are likely to be protocol deviations or violations, and participant losses to follow-up, resulting in missing data. This has a negative impact on the validity of the study results. Statisticians have developed methods to deal with missing data, such as multiple imputation

techniques, best- and worst-case scenarios, and the last-observation-carried-forward technique. However, the best way of ensuring the validity of results is to have as complete data as possible. There are no absolute cut-off points to define the acceptable level of missing data – these vary with the clinical condition being studied and the duration of follow-up required.

Some ways to improve completeness of data collection include training of the study personnel to minimize protocol violations, keeping the study protocol simple so that compliance is better and motivating participants to adhere to the protocol.

APPROPRIATE STATISTICAL ANALYSIS

Intention to treat versus per-protocol analysis

Intention-to-treat analysis refers to the analysis of participants in the group to which they were randomized, irrespective of what treatment they received. On the other hand, per-protocol analysis refers to the analysis of only those participants who adhered to the protocol. To minimize bias, as discussed in a previous article in the journal,^[7] intention-to-treat analysis should always be reported in superiority studies; per-protocol analysis may be reported in addition, if desired.

Choice of statistical test

The choice of statistical test used for the analysis depends on the type of data, the number of groups to be compared, the objective of the study, and the study design (paired versus unpaired). The use of an inappropriate test can give misleading results. Readers can refer to published articles for further details on the different types of tests and their application.^[8]

Adjustment for multiple testing

In a previous article, we had discussed how the comparison of several outcomes, interim analyses, or multiple subgroup comparisons increases the possibility of spuriously significant results.^[9] For such analyses, the validity of positive results without examining and correcting for multiple comparisons is questionable.

COMPLETE AND UNBIASED REPORTING

The CONSORT statement lists the elements which are mandatory for the reporting of randomized clinical trials.^[10]

This ensures that the readers can better assess the quality of a study and hence the validity and applicability of its results.

It is not uncommon for investigators to compare multiple outcomes or to use multiple statistical tests for a particular comparison and then cherry-pick the results that show a positive impact of a treatment. This is inappropriate. It is important to report the results of a trial in totality and without bias so that readers can assess the validity of the study findings. Mandatory registration of clinical trials, with the investigators being required to specify the primary and secondary outcomes before starting a trial, is aimed at promoting such behavior.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. DREAMS Trial Collaborators and West Midlands Research Collaborative. Dexamethasone versus standard treatment for postoperative nausea and vomiting in gastrointestinal surgery: Randomised controlled trial (DREAMS Trial). *BMJ* 2017;357:j1455.
2. Connolly SJ. Use and misuse of surrogate outcomes in arrhythmia trials. *Circulation* 2006;113:764-6.
3. Singh AK, Szczech L, Tang KL, Barnhart H, Sapp S, Wolfson M, *et al.* Correction of anemia with epoetin alfa in chronic kidney disease. *N Engl J Med* 2006;355:2085-98.
4. McCoy CE. Understanding the use of composite endpoints in clinical trials. *West J Emerg Med* 2018;19:631-4.
5. Ferreira-González I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM, *et al.* Problems with use of composite end points in cardiovascular trials: Systematic review of randomised controlled trials. *BMJ* 2007;334:786.
6. Mentzer RM Jr, Bartels C, Bolli R, Boyce S, Buckberg GD, Chaitman B, *et al.* Sodium-hydrogen exchange inhibition by cariporide to reduce the risk of ischemic cardiac events in patients undergoing coronary artery bypass grafting: Results of the EXPEDITION study. *Ann Thorac Surg* 2008;85:1261-70.
7. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. *Perspect Clin Res* 2016;7:144-6.
8. Bellolio MF, Serrano LA, Stead LG. Understanding statistical tests in the medical literature: Which test should I use? *Int J Emerg Med* 2008;1:197-9.
9. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: The perils of multiple testing. *Perspect Clin Res* 2016;7:106-7.
10. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c332.