BMJ Open

Power estimations for non-primary outcomes in randomised clinical trials

Janus Christian Jakobsen,[1,2] Christian Ovesen,[3] Per Winkel,[1] Jørgen Hilden,[4] Christian Gluud,[1] Jørn Wetterslev[1]

[1]Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[2]Department of Cardiology, Holbæk Hospital, Holbæk, Denmark
[3]Department of Neurology, Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark
[4]Department of Public Health Research, Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Correspondence to
Dr Janus Christian Jakobsen;
janusjakobsen@mac.com,
jcj@ctu.dk

ABSTRACT

Objective and methods: It is rare that trialists report power estimations of non-primary outcomes. In the present article, we will describe how to define a valid hierarchy of outcomes in a randomised clinical trial, to limit problems with Type I and Type II errors, using considerations on the clinical relevance of the outcomes and power estimations. Conclusion: Power estimations of non-primary outcomes may guide trialists in classifying non-primary outcomes as secondary or exploratory. The power estimations are simple and if they are used systematically, more appropriate outcome hierarchies can be defined, and trial results will become more interpretable.

To avoid problems with Type I errors (false rejection of a true null hypothesis) and Type II errors (false acceptance of a null hypothesis), and rash interpretations of the results of a randomised clinical trial, it is essential to (1) limit the number of outcomes;[1] (2) adjust CIs and thresholds for significance according to number of outcome comparisons;[1] and (3) define an outcome hierarchy (outcomes classified according to their type and how they ought to become interpreted).

Clinical success has many aspects and both beneficial and harmful effects ought to be interpreted, so selecting a single outcome variable is rarely feasible.[1] We have previously summarised how to adjust CIs and thresholds for significance if there are multiple outcome comparisons.[1] The European Medicines Agency has recently, conservatively and wisely, suggested using Bonferroni corrections.[2]

The present paper will describe how to define a valid hierarchy of outcomes in a randomised clinical trial, to limit problems with Type I and Type II errors, using power estimations of the non-primary outcomes. Our focus in the present paper is the overall outcome of a trial. Therefore, a Type I error will be defined as the case when the overall conclusion of a trial is that an intervention is effective—when it is not. Type II error will be defined as the case when the overall conclusion of a trial is that an intervention is *not* effective—when it is. In order to maintain simplicity, we will focus on dichotomous and continuous outcomes, but the described principles may be used for most other types of outcomes as well.[3]

SUMMARY OF FUNDAMENTAL CONSIDERATIONS WHEN DEFINING OUTCOME HIERARCHIES IN RANDOMISED CLINICAL TRIALS

Before considering power estimations of non-primary outcomes, we will briefly summarise what we believe are fundamental and essential considerations when defining outcome hierarchies in randomised clinical trials.

It is recommended to prespecify primary and secondary outcomes, including how and when they are assessed (http://www.consort-statement.org/checklists/view/32--consort-2010/80-outcomes).[3]

To limit problems with multiplicity and difficulties with interpreting the trial results, it is often optimal to use only one primary outcome and the sample size should be based on this outcome.[1] The primary outcome in a randomised clinical trial should be the outcome with the highest degree of clinical relevance for the patients, that is, patient centred outcomes. All primary and secondary outcomes in a randomised clinical trial should either be outcomes that are important for the decision to use the intervention or sufficiently validated surrogate outcomes for such important outcomes.[2 4 5] History has shown us that we cannot rely on surrogate outcomes, unless they are validated.[4] The most-often cited example is the Cardiac Arrhythmia Suppression Trial (known as CAST), in which two drugs that suppressed ventricular arrhythmias (a surrogate outcome correlated with a bad prognosis) were initially approved by the Food and Drug Administration, only to have the CAST demonstrate that, compared with placebo, individuals who had arrhythmias after myocardial infarctions and received antiarrhythmic drugs were 2.5 times as likely to die.[6] It is necessary to validate a surrogate

outcome before we can be confident that it can be used in clinical trials or practice.[4 7] Such validation requires randomised clinical trials that assess both the surrogate and clinical outcome and show that both are changed by the intervention in a comparable manner.[4 7 8] Moreover, a validated surrogate for one drug cannot guarantee that the surrogate outcome will not mislead when new drugs are being tested.[8] Non-validated surrogate outcomes should always be classified as 'exploratory outcomes', until formal validation has been proved and accepted by the scientific community.

When planning a randomised clinical trial, it is essential to estimate the required sample size.[1 9–11] However, the majority of randomised clinical trials have difficulties in obtaining the stipulated sample size,[12] and trials with too small sample sizes often suggest intervention effect sizes far from the 'true' effect sizes shown in subsequent larger trials and meta-analyses.[1 13] Even most Cochrane systematic reviews with meta-analyses do not have sufficient power.[14 15]

Power estimations of non-primary outcomes

Consider a single randomised clinical trial. If the estimated sample size has not been reached, the risks of Type I errors (false rejection of the null hypothesis) and Type II errors (false acceptance of the null hypothesis) should be estimated when interpreting the trial results.[1 16] The threshold for statistical significance (and consequently the CI) should be adjusted to the fraction of the preplanned number of participants randomised.[1 16] Clearly, there is no safeguard against all kinds of bias, but adjustment schemes in common use at the very least protect against the dangers of premature or repeated testing.[1] Such adjustments should, ideally, be common practice in all high quality trials.[1 16]

Analogous problems arise with non-primary outcomes when the information is deemed insufficient; that is, when statistical power is not known, the data cannot unreservedly be analysed as if based on a dataset large enough to draw conclusions about a minimal important difference (MID).[17] If MID effect estimates, as well as null effect, are included in the naïve 95% CI, then this indicates that more information may be needed. However, if MID effect estimates are *not* included in the naïve 95% CI, then it is unclear if more data are needed to uncover a worthwhile effect or if there is in fact no worthwhile difference between the groups.[1]

When null effect is *excluded* in the naïve 95% CI and it is unclear whether there is enough information, it will also be difficult to interpret the analysis results. Trial results tend to show spurious results of too beneficial or too harmful effect estimates if there is insufficient information.[1] Inspecting unadjusted naïve 95% CI when the sample size has not been reached will not suffice as such CIs would be inappropriately narrow, as stated above.[1 16]

In order to estimate the statistical power of an analysis, it is necessary to decide on an MID,[1 17] an incidence in the control group when assessing a dichotomised outcome or a SD when assessing a continuous outcome, and an acceptable risk of Type I error adjusted according to the number of outcome comparisons.[1 2] Alternatively, the sequence in which the secondary outcomes are tested may be prespecified and carried out without adjustment, but stopped when the first null hypothesis is not rejected after which the rest of the assessments will become exploratory.[1] Most statistical software can easily estimate both sample sizes and power estimations of non-primary outcomes.[18]

Power analysis should be part of standard trial methodology

For the reasons stated above, we recommend at the protocol stage to estimate the statistical power of all non-primary outcomes for confirming or rejecting a MID. If the power is less than 80% (or 90%), then this outcome should be classified as an 'exploratory outcome' together with the non-validated surrogate outcomes.[19] Alternatively, the CI and the thresholds for significance for the outcome in question may be adjusted due to sparse data,[1 16] or the sample size could be reconsidered and increased so the power of the non-primary outcome in questions becomes 80% (or 90%).[1 16]

We searched for all randomised clinical trials published in the British Medical Journal during 2017 and found 10. Only one randomised clinical trial briefly mentioned that 'A trial of this size will also give more than 80% power to detect important differences in secondary outcomes…'.[20] None of the remaining nine trials reported any considerations of power of non-primary outcomes, and it is generally rare that trialists report power estimations of non-primary outcomes. As we have described, trial results always ought to be interpreted in the light of the required sample size and the obtained sample size, and without power estimations it will be difficult to make valid conclusions based on non-primary outcome results. It is simple to estimate the power of outcome tests, so it is striking that this is not done regularly by trialists. Of course, MIDs (together with a measure of variance and an acceptable risk of Type I error) need to be estimated to estimate the power of an outcome comparison, which might seem troublesome. Nevertheless, MIDs need to be defined for all important outcomes regardless of the use of power estimations, otherwise it will be difficult to judge if statistically significant results are also clinically meaningful for patients.[1] All the necessary quantities (MIDs, estimations of proportion in the control group, SD) in the power estimations may possibly be estimated on the basis of a systematic review of studies, performed before the trial is conducted.

Considerations on the clinical relevance of outcomes and power estimations seem an important tool that may help defining appropriate outcome hierarchies. In addition to estimating a required sample size, we believe that future trialists when planning a randomised clinical trial, ought to estimate power of all non-primary outcomes and consider estimating power of subgroup comparisons. Power estimations of non-primary outcomes may guide trialists in classifying non-primary outcomes as secondary

or exploratory. The power estimations are simple and if they are used systematically, more appropriate outcome hierarchies can be defined, and trial results will become more interpretable.[21]

REFERENCES

1. Jakobsen JC, Gluud C, Winkel P, *et al*. The thresholds for statistical and clinical significance - a five-step procedure for evaluation of intervention effects in randomised clinical trials. *BMC Med Res Methodol* 2014;14:34.
2. Agency EM. Guideline on multiplicity issues in clinical trials. *Committee for Human Medicinal Products (CHMP) EMA/ CHMP/44762/2017* 2016.
3. Schulz KF, Altman DG, Moher D. CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010;152:726–32.
4. Garattini S, Jakobsen JC, Wetterslev J, *et al*. Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. *Eur J Intern Med* 2016;32:13–21.
5. Jakobsen JC, Gluud C. The necessity of randomized clinical trials. *Br J Med Med Res* 2013;3:1453–68.
6. Echt DS, Liebson PR, Mitchell LB, *et al*. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;324:781–8.
7. Jakobsen JC, Nielsen EE, Feinberg J, *et al*. Direct-acting antivirals for chronic hepatitis C. *Cochrane Database Syst Rev* 2017;9:CD012143. Art. No.: CD012143. DOI.
8. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Stat Med* 2012;31:2973–84.
9. Button KS, Ioannidis JP, Mokrysz C, *et al*. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013;14:365–76.
10. Chow S, Shao J, Wang H. *Sample size calculations in clinical research*. Boca Raton: Francis/CRC, 2003.
11. Fayers PM, Cuschieri A, Fielding J, *et al*. Sample size calculation for clinical trials: the impact of clinician beliefs. *Br J Cancer* 2000;82:213–9.
12. Sully BG, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. *Trials* 2013;14:166.
13. Levin GP, Emerson SC, Emerson SS. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Stat Med* 2013;32:1259–75.
14. Turner RM, Bird SM, Higgins JP. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PLoS One* 2013;8:e59202.
15. Jackson D, Turner R. Power analysis for random-effects meta-analysis. *Res Synth Methods* 2017;8:290–302.
16. Gkk L, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;70:659–63.
17. Zhang Y, Zhang S, Thabane L, *et al*. Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. *J Clin Epidemiol* 2015;68:888–94.
18. StataCorp. *Stata: Release 15. Statistical Software*. College Station, TX: StataCorp LP, 2017.
19. Winkel P, Bath PM, Gluud C, *et al*. Statistical analysis plan for the EuroHYP-1 trial: European multicentre, randomised, phase III clinical trial of the therapeutic hypothermia plus best medical treatment versus best medical treatment alone for acute ischaemic stroke. *Trials* 2017;18:573.
20. Epidural and Position Trial Collaborative Group. Upright versus lying down position in second stage of labour in nulliparous women with low dose epidural: BUMPES randomised controlled trial. *BMJ* 2017;359:j4471.
21. Kirkham JJ, Gorst S, Altman DG, *et al*. Core outcome set-standards for reporting: The COS-STAR Statement. *PLoS Med* 2016;13:e1002148.