Multiple analyses in clinical trials: sound science or data dredging?

Sarah J Lord, Val J Gebski and Anthony C Keech

Clinical trials typically require the collection of many data to describe the participants and for measuring their response to an intervention. In addition to the primary analysis of treatment effect, investigators can use these data to perform multiple analyses, but there are important pitfalls with their use.[1,2] Here, we discuss three common types of secondary analyses: analyses of multiple outcome variables; analyses of trial outcomes that account for prognostic factors (adjusted analyses); and using trial data to answer secondary research questions (see definitions in Box 1). The use of trial data for population subgroup analyses has been discussed earlier in this series.[3,4]

What are the problems?

The two main problems introduced by multiple analyses are, firstly, the increased probability of detecting intervention effects where none exist ("false positives" owing to multiple comparisons — type I errors), and secondly, the limited capability ("power") of trials to detect a true treatment effect in secondary outcomes if not enough participants are enrolled to show a statistically significant difference in these outcomes ("false negatives" — type II errors). One study compared trial protocols with their subsequent publications, and provided empirical evidence of the selective reporting of positive trial results.[5] The use of multiple analyses is therefore of particular concern when these are conducted post-hoc as a "fishing expedition", and undue emphasis is given to positive findings. Item 18 of the CONSORT checklist (Box 2) recommends that investigators report on all multiple analyses and declare which were prespecified and which were conducted as exploratory activities after investigators were unblinded to the treatment allocation of participants.[6]

Analyses of multiple outcomes

Advantages

Investigators may choose multiple outcomes to measure the effect of treatment. This is an advantage when different parameters provide information about different aspects of the treatment response.[1] Secondary analyses may also assist the interpretation of the primary analysis. For example, in a recently reported trial comparing chemotherapy regimens in the treatment of patients with metastatic breast cancer, investigators selected two primary outcomes as the most important measures of treatment effect: the overall tumour response rate (measured as complete and partial response) and time to treatment failure. Secondary outcomes, including overall survival, toxicity and quality of life, provided

NHMRC Clinical Trials Centre, University of Sydney, Sydney, NSW.
Sarah J Lord, MB BS, MScEpid, Research Fellow; Val J Gebski, BA, MStat, Associate Professor and Principal Research Fellow; Anthony C Keech, MScEpid, FRACP, Deputy Director.
Reprints will not be available from the authors. Correspondence: Dr Sarah J Lord, NHMRC Clinical Trials Centre, University of Sydney, Locked Bag 77, Camperdown, NSW 1450. enquiry@ctc.usyd.edu.au

1 Definitions

Primary outcome: The health parameter measured in all study participants to detect a response to treatment. Conclusions about the effectiveness of treatment should focus on this measurement.

Primary analysis: The statistical test performed to determine whether there is a difference in the primary outcome between participants allocated to receive the treatment and those allocated to the control arm.

Secondary outcomes: Other parameters that are measured in all study participants to help describe the effect of treatment.

Baseline variables: The characteristics of each participant measured at the time of random allocation.

This information is documented to allow the trial results to be generalised to the appropriate population/s. Specific characteristics associated with the patient's response to treatment (such as age and sex) are known as *prognostic factors*.

Multiple analyses: Comparisons between the study groups for more than one outcome. They increase the likelihood of detecting a difference between the treatment and control group owing to chance alone (false positive).

Common examples of multiplicity in trials include the use of:
- multiple outcomes, including surrogate endpoints;
- multiple treatment comparisons (in a multiarm trial);
- subgroup analyses to detect differences in the treatment effect in one or more subsets of trial participants;
- adjusted analyses to control for imbalances in prognostic factors between the study groups;
- repeated measures over time of the same outcome; and
- interim analyses of the treatment effect at different stages in the trial.

Exploratory analyses: Analyses that were not specified before the trial or, for blinded studies, analyses planned after the investigators were unblinded to the treatment allocation of participants. These analyses may be driven by the results of the primary analysis.

additional information about the treatment effect.[7] Including a set of supplementary outcomes may also be a practical solution when different investigators value outcomes differently.

Multiple-outcomes analysis is particularly useful when a statistically significant benefit of treatment on the primary outcome can be confirmed or strengthened by a consistent effect on other relevant outcomes. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) trial evaluated the effectiveness of pravastatin for preventing cardiovascular events in patients with diabetes or impaired fasting glucose and a history of coronary heart disease.[8] The finding of a statistically significant reduction in the risk of a major coronary event was supported by a similar reduction in the risk of a revascularisation procedure or stroke. Such findings may also advance the understanding of the relationships between outcomes.

Pitfalls

The type and number of analyses performed should be reported so that readers can assess the probability of detecting a treatment

2 CONSORT checklist of items to include when reporting a trial[6]

Selection and topic	Item no.	Descriptor
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory.

effect by chance alone. This is often poorly documented in trial reports.[5] Additionally, major discrepancies have been observed between the primary outcome specified in the trial protocol and that reported in the published article.[5] Chan et al reported that of 76 trials that prespecified a primary outcome in the trial protocol, 20 (26%) did not report on this outcome in the published article, and of 63 trials that specified a primary outcome in the published article, in 11 (17%) it was not mentioned in the trial protocol.[5] An example of the latter was a study reporting on the percentage of patients with graft occlusion as the primary outcome, even though the study was not originally designed to measure a difference in this outcome.[5]

Caution is needed when unexpected results from multiple analyses are interpreted. Inconsistent results are more credible if the outcome variables are restricted to those that were prespecified in the trial protocol, are clinically relevant, and are based on plausible biological mechanisms. A variety of statistical corrections can be performed to take into account the increased probability of a chance finding with multiple testing.

A statistically significant treatment effect for one outcome and not other clinically related outcomes may also indicate that the sample is too small — that is, the study lacks power. Interpreting the results of analyses that are underpowered is difficult. This is a common problem, for example, in the mandatory reporting of adverse events in drug trials. A trial may report on a large set of adverse events, but it will commonly not have been powered to detect a statistically significant difference in these outcomes between the trial's study groups.[1]

Composite endpoints

To overcome the problem of insufficient power, investigators may combine data from clinically related outcomes to form one or more composite endpoints. This approach reduces the number of analyses required while retaining all the potentially valuable information. The TAXUS IV trial was a double-blind randomised controlled trial to determine the safety and effectiveness in coronary artery disease of paclitaxel-eluting stents compared with bare metal stents.[9] The primary outcome of the trial was the incidence of revascularisation procedures due to reocclusion of the target vessel at 9 months. A composite endpoint, "major adverse cardiac events" (defined as death from cardiac causes, myocardial infarction, or revascularisation procedures), was a secondary outcome. At one year after the procedure, the rates of cardiac death and myocardial infarction were similar between the study groups, while the rate of target-vessel revascularisation was 62% lower (P < 0.0001) in the patients receiving paclitaxel-eluting stents than in those receiving bare metal stents.[9] The treatment effect on

revascularisation rates appeared to drive the results for the composite endpoint, resulting in a reported 49% reduction in major adverse cardiac events (P < 0.0001) at 12 months. Combining disparate events can lead to an overestimate of the clinical importance if a positive finding is largely driven by the less important events.

Overall, as a result of the potential to overinterpret or misinterpret the results of analyses of multiple outcomes, readers should seek information in the methods section of the report about the primary purpose and outcomes that the trial was designed to address and interpret any additional findings in this context.

Adjusted analyses

Clinical trials use a concealed randomisation process, with or without stratification by key prognostic factors, such as age and sex, to help to ensure the baseline similarity of the study groups.[10] However, even well-conducted random allocation may still result in chance imbalances.[11] If an imbalance in an important prognostic factor occurs, statistical methods can control for this imbalance by including the factor as a "covariate". This is referred to as adjusted analysis, or a multivariate analysis if more than one covariate is included. While adjusted analyses can statistically accommodate imbalances between study groups in non-randomised studies, in randomised studies they should usually be considered supplementary to the unadjusted analysis of the primary outcome. If the adjusted effect estimate differs from the unadjusted estimate, interpretation may be a problem. For example, if some covariate data are missing and these participants are excluded from the adjusted analysis, it will not be clear whether observed differences result from controlling for this factor or another, unknown effect of these exclusions. Adjusted analysis may be indicated when a factor is known to strongly predict the outcome (for example, age and survival), even when the imbalance observed between study groups does not reach statistical significance.[12] In general, adjusted analyses frequently improve the precision of the estimate of treatment effect, even when the correlation of the covariates with the study outcome is not strong.[13]

Another recent review of 50 consecutive published trials showed that the methods and reporting of adjusted analyses vary widely in clinical trials.[14] Of the 36 trials with an adjusted analysis of the primary outcome, 42% did not report on the methods used to select the covariates.[14] Using inappropriate methods for adjusted analyses may cause inaccurate and misleading results. Ideally, investigators should prespecify any prognostic factors that, if unbalanced, may affect the study outcomes and should plan for adjusted analyses accordingly. However, some strong predictors of the outcome may only become apparent at data analysis on formal testing (so-called exploratory analysis). In this situation, investigators should clearly describe when and how covariates were selected for the adjusted analysis. In any case, the primary emphasis should be on the unadjusted results, because investigators are able to conduct multiple adjusted analyses using different sets of covariates, which may lead to overinterpretation or selective reporting of significant findings. The findings of the primary unadjusted analysis are strengthened if the results of the adjusted analysis are consistent with them.

3 Checklist for multiple analyses

Design and methods

- Were the primary and secondary outcomes for the detection of treatment response prespecified?
- Was the trial designed to have adequate power for the analyses of all outcomes?
- Were the covariates for the adjusted analyses and/or the method used to select these covariates prespecified?
- Were the substudies based on an existing trial or biological data?
- Were the substudies planned prior to unblinding of data?

Analysis

- Have corrections for multiple-significance testing been performed?
- Was the combination of data into a composite outcome appropriate?
- Was the interpretation of the composite endpoint results appropriate?

Reporting

- Are the total number of analyses performed reported?
- Was the power calculation reported for the primary outcome? Secondary outcomes?
- Are the rationale and methods of any adjusted analyses reported?
- Are the number and type of covariates in the adjusted analyses reported?
- Are the unadjusted and adjusted results reported?
- Are the prespecified analyses clearly distinguished from the exploratory analyses?

Interpretation

- Is appropriate emphasis given to the primary outcome?
- Have the relationships between interrelated outcomes been explored with equal interest?
- Are the findings of the multiple analyses discussed in the context of current biological knowledge and current research?

Other ancillary analyses

A clinical trial may seek to address ancillary questions unrelated to the primary question so as to optimise the use of resources required for a large clinical trial. Ancillary questions may relate to the treatment effect on other conditions of interest, such as the association between hormone replacement therapy and dementia in women recruited to a large trial investigating hormone replacement therapy and cardiovascular disease.[15] Substudies may also use trial data to investigate epidemiological questions about the natural history of disease, the biological mechanisms of the disease[16] or the treatment response.[17]

Ideally, these ancillary studies should be designed before the trial starts. However, important new information or scientific debate may arise during or after the trial to justify the use of trial data to investigate new hypotheses. Their results are more convincing if the decision to conduct the analysis has been made before unblinding. The same potential for overinterpretation and selective reporting of the results of multiple comparisons and reduced power apply to exploratory analysis, and any new findings should be regarded as new hypotheses for validation in future studies.

The principles of planning, reporting, analysing and interpreting multiple analyses are shown in Box 3. These are not intended to discourage investigators from conducting potentially important exploratory analyses of plausible new hypotheses. Rather, they encourage the balanced reporting of all analyses to prevent unsound manipulation of data or undue emphasis on particular findings that may misdirect future research or compromise the interpretation of results for clinical practice.

Competing interests

None identified.

Acknowledgements

We thank Rhana Pike for expert assistance in preparation of this manuscript.

References

1 Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 1997; 18: 530-545.
2 Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977; 198: 679-684.
3 Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. *Med J Aust* 2004; 180: 289-291. www.mja.com.au/public/issues/180_06_150304/coo10086_fm.html
4 Simes RJ, Gebski VJ, Keech AC. Subgroup analysis: application to individual patient decisions. *Med J Aust* 2004; 180: 467-469. www.mja.com.au/public/issues/180_09_030504/sim10218_fm.html
5 Chan AW, Hrobjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; 291: 2457-2465.
6 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; 357: 1191-1194.
7 Sledge GW, Neuberg D, Bernardo P, et al. Phase III trial of doxorubicin, paclitaxel, and the combination of doxorubicin and paclitaxel as front-line chemotherapy for metastatic breast cancer: an intergroup trial. *J Clin Oncol* 2003; 21: 588-592.
8 Keech A, Colquhoun D, Best J, et al. Secondary prevention of cardiovascular events with long-term pravastatin in patients with diabetes or impaired fasting glucose: results from the LIPID trial. *Diabetes Care* 2003; 26: 2713-2721.
9 Stone GWM, Ellis SGM, Cox DAM, et al. One-year clinical results with the slow-release, polymer-based, paclitaxel-eluting TAXUS stent: The TAXUS-IV trial. *Circulation* 2004; 109: 1942-1947.
10 Beller EM, Gebski V, Keech AC. Randomisation in clinical trials. *Med J Aust* 2002; 177: 565-567. www.mja.com.au/public/issues/177_10_181102/bel10697_fm.html
11 White H. and Hirulog and Early Reperfusion or Occlusion (HERO) Trial Investigators. Thrombin-specific anticoagulation with bivalirudin versus heparin in patients receiving fibrinolytic therapy for acute myocardial infarction: the HERO-2 randomised trial. *Lancet* 2001; 358: 1855-1863.
12 Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *Am Heart J* 2000; 139: 745-751.
13 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; 21: 2917-2930.
14 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; 355: 1064-1069.
15 Shumaker SA, Reboussin BA, Espeland MA, et al. The Women's Health Initiative Memory Study (WHIMS): a trial of the effect of estrogen therapy in preventing and slowing the progression of dementia. *Control Clin Trials* 1998; 19: 604-621.
16 Barron HV, Cannon CP, Murphy SA, et al. Association between white blood cell count, epicardial blood flow, myocardial perfusion, and clinical outcomes in the setting of acute myocardial infarction: a thrombolysis in myocardial infarction 10 substudy. *Circulation* 2000; 102: 2329-2334.
17 O'Connor FF, Shields DC, Fitzgerald A, et al. Genetic variation in glycoprotein IIb/IIIa (GPIIb/IIIa) as a determinant of the responses to an oral GPIIb/IIIa antagonist in patients with unstable coronary syndromes. *Blood* 2001; 98: 3256-3260.