Subgroup analysis in clinical trials

David I Cook, Val J Gebski and Anthony C Keech

CLINICAL TRIALS REPRESENT A MAJOR INVESTMENT by investigators, sponsors and participants, and it is reasonable to attempt to gain the maximum information from them. Practitioners and regulatory agencies are keen to know whether there are subgroups of trial participants who are more (or less) likely to be helped (or harmed) by the intervention under investigation, and a recent survey of trials published over 3 months in four leading journals found that 70% included subgroup analyses.[1,2] Furthermore, regulatory guidance documents (such as the Committee for Proprietary Medicinal Products September 2002 document *Points to consider on multiplicity issues in clinical trials*[3]) strongly encourage appropriate subgroup analyses. The results of subgroup analyses can also drive changes in practice guidelines. For example, the United States National Institutes of Health issued a clinical alert following the unexpected finding in the BARI (Bypass Angioplasty Revascularisation Investigation) trial that mortality after angioplasty in patients with diabetes was nearly double that after bypass-graft surgery ($P = 0.003$).[4]

Meaningful information from subgroup analyses within a randomised trial is restricted by multiplicity of testing and low statistical power. There is therefore a tension between our wish to identify heterogeneity in the responses of trial participants to trial interventions and our technical capacity for doing so. Surveys on the adequacy of the reporting of clinical trials consistently find the reporting of subgroup analysis to be characterised by poor practice.[2,5-7] Item 18 of the CONSORT checklist (Box 1) deals with the multiplicity issues that arise in subgroup analysis.[8]

1: CONSORT checklist of items to include when reporting a trial[8]

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 pre-specified and those exploratory.

Problems in subgroup analysis

The problem of multiple testing

Statistical investigation of large numbers of subgroups inevitably shows significant interactions with the effectiveness of the trial intervention. By definition, testing at the 5% level of significance will erroneously report a statistically significant difference between subgroup categories in about 5% of the tests performed (so-called false-positive results). Trials with multiple comparisons to assess the comparability of randomised groups at baseline confirm this prediction.[1,9]

Department of Physiology, University of Sydney, Sydney, NSW.
David I Cook, MD, FRACP, Professor of Cellular Physiology.
NHMRC Clinical Trials Centre, University of Sydney, Camperdown, NSW.
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: Professor David I Cook, Department of Physiology (F-13), University of Sydney, Sydney, NSW 2006. davidc@physiol.usyd.edu.au

In subgroup analysis, where a plethora of factors (eg, sex, age, race, centre, smoking status, stage of disease, and coexistent disorders) may influence outcome, the risk of false-positive results is high.[10] Overly enthusiastic analysis of subgroups can reveal statistically significant differences in outcome between subgroups even where neither arm of the study receives any intervention.[11] In some cases, such as in the ISIS-2 study, which found a slight adverse impact of aspirin therapy on patients born under the star signs Gemini and Libra, and that aspirin helped after the first, but not subsequent, infarctions,[12] the results of the subgroup analysis may be dismissed as contrary to current understanding of biological mechanisms. In other cases, such as the BARI trial,[4] whether the finding was valid could only be established by additional studies.[13,14]

The problem of statistical power

Most studies enrol just enough participants to ensure that the primary hypothesis can be adequately tested. Therefore, statistical tests on subgroups will have only power to detect substantially larger effects on the same endpoint. Loss of compliance, together with adjustments for multiple testing, will exacerbate this lack of power.[6] In consequence, when tested separately, many of the subgroups will fail to show the statistically significant treatment effect that was shown in the main population; at the same time, genuine differences in response to treatment (so-called heterogeneity) between study subpopulations may also go undetected.

Can the problems be overcome?

Despite subgroup analyses generally lacking statistical power, when used repeatedly to look for differences across many factors (eg, sex, age, smoking status, blood pressure) they have a proclivity to detect spurious effects. We are thus forced to reconcile our wish to find genuine differences between subgroups with the need to minimise the risk of accepting and publishing false positives.[2,6] One solution to this dilemma is to accept that the results of subgroup analysis are hypotheses. Guidelines such as those given in Box 2 are intended to help readers identify which hypotheses are strong and which are weak. How-

ever, even among experts, opinions range from only accepting pre-specified subgroup analyses supported by a very strong a priori biological rationale[15] to a more liberal view in which subgroup analyses, if properly carried out and carefully interpreted, are permitted to play a role in assisting doctors and their patients to choose between treatment options.[16]

Trial design

Are the subgroups appropriately defined?

Subgroups based on characteristics measured after randomisation, such as compliance, should be avoided, as allocation to the subgroup may be influenced by the intervention. Similarly, it is preferable to use the intention-to-treat population, as reasons for withdrawal may not be balanced between treatment arms. For example, adverse drug events may be a more important reason for withdrawals from an active treatment arm, whereas lack of efficacy may be more important in a placebo-controlled arm.[17]

Were the subgroup analyses planned before commencement of the study?

In general, subgroup analyses should be defined *a priori* and purposely on the basis of known biological mechanisms or in response to findings in previous studies. Ideally, the choice of the subgroups and the expected direction of the subgroup difference should be justified in the trial protocol. Where a particular subgroup analysis is of great interest, adequate power to show the results can be designed into the trial, for example by using an expanded endpoint for the subgroup analysis.

At the other extreme, subgroup analyses that are decided on once the dataset has been examined should be treated with scepticism. Intermediate between these two extremes are cases, such as occurred in the BARI trial, in which the subgroup analysis, although not originally planned, was decided on during the course of the trial in response to findings in other studies (with the investigators remaining blinded to the interim results of BARI).[4]

Reporting

The study report should include all the information required to assess the validity of subgroup analyses reported. In particular, the number of subgroup analyses should be declared, as this will enable readers to assess whether the issue of multiple testing is being dealt with. Analyses planned *a priori*, and the rationale for choosing them, should be clearly stated. Summary data, including event numbers and denominators for all the subgroup analyses, even the uninteresting ones, should be included, as this will facilitate future meta-analyses of the data and help prevent publication bias.[18] The impact of multiple tests on the chance of declaring as statistically significant at least one false-positive result is shown in Box 3.

2: Checklist for subgroup analyses

Design

- Are the subgroups based on pre-randomisation characteristics?
- What is the impact of patient misallocation on the subgroup analysis?
- Is the intention-to-treat population being used in the subgroup analysis?
- Were the subgroups planned *a priori*?
- Were they planned in response to existing trial or biological data?
- Was the expected direction of the subgroup effect stated *a priori*?
- Was the trial designed to have adequate power for the proposed subgroup analysis?

Reporting

- Is the total number of subgroup analyses undertaken declared?
- Are relevant summary data, including event numbers and denominators, tabulated?
- Are analyses decided on *a priori* clearly distinguished from those decided on *a posteriorl*?

Statistical analysis

- Are the statistical tests appropriate for the underlying hypotheses?
- Are tests for heterogeneity (ie, interaction) statistically significant?
- Are there appropriate adjustments for multiple testing?

Interpretation

- Is appropriate emphasis being placed on the primary outcome of the study?
- Is the validity of the findings of the subgroup analysis discussed in the light of current biological knowledge and the findings from similar trials?

3: Probability of at least one significant result at the 5% significance level given no true differences

Number of tests	Probability
1	0.05
2	0.10
3	0.14
5	0.23
10	0.40
20	0.64

Statistical analysis

Some investigators avoid the issue of multiplicity of testing by tabulating the observed outcomes for the subgroups of interest without undertaking any formal statistical analysis. The data become available for meta-analysis,[18] but there is the disadvantage that the investigator may fail to detect and draw attention to an important heterogeneity in the population.

The statistical methods used should be appropriate for the hypothesis being tested. The common practice of performing subgroup-specific tests of treatment effect is flawed in that it is testing the wrong hypothesis.[19] The hypothesis that should be tested is whether the treatment effect in a

subgroup is significantly different from that in the overall population.[19] Testing for a statistically significant treatment effect in a subgroup is hindered by a small sample size.

The appropriate tests to use when analysing heterogeneity of responses among subgroups are interaction tests,[2,10] for which worked examples are available.[19,20] One study found that these were used in only 43% of 35 trials which reported subgroup analyses in their sample.[2]

Finally, the article should state whether the statistical tests used included adjustments for multiplicity.

Interpretation

Because subgroup analyses have less power to detect a therapeutic effect than the main study, the trial report, especially in the Abstract or Conclusions, should emphasise the overall result. Given the risks of false-positive findings when multiple subgroup analyses are performed, it is not surprising if a subgroup-specific test shows a significant ($P < 0.05$) or suggestive ($P = 0.05$ to $P = 0.10$) effect of treatment, even when the trial failed to do so overall.[2,7]

Investigators are often tempted to highlight a particular subgroup analysis.[2,7] For example, in one trial the suggestion that a psychosocial nursing intervention following myocardial infarction was harmful for women ($P = 0.064$), but not men ($P = 0.94$), was highlighted, even though the intervention did not affect survival in the overall population[21] (and a test for interaction was not significant[2]).

A number of arguments may be used to support the validity of a claimed subgroup effect (see, for example, the BARI trial[4] and Rathore et al[22]):

- replication in another independent study;
- the presence of a dose–response relationship;
- reproducibilty of the observation in independent samples within the study, such as within individual sites; and
- the availability of a biological explanation.

Of these, the first is the strongest evidence. For example, even though the BARI study found no difference in survival following bypass surgery or angioplasty in the overall population, the validity of the subgroup findings was supported by other studies.[4] On the other hand, the report by Rathore et al that digoxin use is associated with a significantly increased risk of death among women ($P < 0.014$)[22] is weakened by the fact that it was a post-hoc analysis which was motivated by "biological suspicion" rather than by suggestive findings in earlier trials. Biological justifications for the findings of a posteriori (exploratory) analyses, on the other hand, carry little weight[6,23] — the reports that diabetes is more common in boys born in October,[24] and that lung cancer is more common in people born in March,[25] included (in)credible biological explanations after the findings had been revealed.

The strategies for overcoming some of these difficulties in interpreting subgroup analyses will be explored in a forthcoming article in this series.

Competing interests

None identified.

Acknowledgements

We thank Rhana Pike for expert assistance in preparation of this manuscript and Dr Jonathan Craig for helpful advice and comments.

References

1. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other mis(uses) of baseline data in clinical trials. *Lancet* 2000; 255: 1064-1069.
2. Pocock SJ, Assmann SE, Enos LE Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting. *Stat Med* 2000; 21: 2917-2930.
3. Committee for Proprietary Medicinal Products. Points to consider on multiplicity issues in clinical trials. September 2002. Available at: www.emea.eu.int\pdfs\human\ewp\090899en.pdf (accessed Feb 2004).
4. Bypass Angioplasty Revascularisation Investigation (BARI). Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. *N Engl J Med* 1996; 335: 217-225.
5. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in reporting of clinical trials: a survey of three medical journals. *N Engl J Med* 1987; 317: 426-432.
6. Yusuf S, Wittes J, ProbstfieldJ, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991; 266: 93-98.
7. Parker AB, Naylor CD. Subgroups, treatment effects and baseline risks: some lessons from major cardiovascular trials. *Am Heart J* 2000; 139: 952-961.
8. 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.
9. Burgess DC, Gebski VJ, Keech AC. Baseline data in clinical trials. *Med J Aust* 2002; 179: 105-107.
10. Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false positives and false negatives. *Health Technol Assess* 2001; 5: 1-56.
11. Lee KL. Clinical judgement and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980; 61: 508-515.
12. ISIS-2 Collaborative Group Randomized trial of IV streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction. *Lancet* 1988; 2: 349-360.
13. Hoffman SN, TenBrook JA, Wolf MP, et al. A meta-analysis of randomized controlled trials comparing coronary artery bypass graft with percutaneous transluminal coronary angioplasty: one- to eight-year outcomes. *J Am Coll Cardiol* 2003; 41: 1293-1304.
14. Niles NW, McGrath PD, Malenka D, et al. Survival of patients with diabetes and multivessel coronary artery disease after surgical or percutaneous coronary revascularization: results of a large regional prospective study. Northern New England Cardiovascular Disease Study Group. *J Am Coll Cardiol* 2001; 37: 1008-1015.
15. Peto R. Clinical trials. In: Price P, Sikara K, editors. Treatment of cancer. 3rd ed. London: Chapman and Hall, 1995: 1039-1044.
16. Coates AS, Goldhirsch A, Gelber RD. Overhauling the breast cancer overview: are subsets subversive? *Lancet Oncology* 2002; 3: 525-526.
17. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 3rd ed. New York: Springer, 1998: 289-293.
18. Hahn S. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Stat Med* 2000; 19: 3325-3336.
19. Matthews JNS, Altman DG. Interaction 2: compare effect sizes not P values. *BMJ* 1996; 313; 808.
20. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003; 326: 219.
21. Frasure-Smith N, Lesperance F, Prince RH, et al. Randomised trial of home-based psychosocial nursing intervention for patients recovering from myocardial infarction. *Lancet* 1997; 350: 473-479.
22. Rathore SS, Wang Y, Krumholz HM. Sex-based differences in the effect of digoxin for the treatment of heart failure. *N Engl J Med* 2002; 347: 1403-1411.
23. Altman DG. Within trial variation — a false trail? *J Clin Epidemiol* 1998; 51: 301-303.
24. Helgason T, Jonasson MR. Evidence for a food additive as a cause of ketosis-prone diabetes. *Lancet* 1981; 2: 716-720.
25. Dijkstra BKS. Origin of carcinoma of the bronchus. *J Natl Cancer Inst* 1963; 31: 511-519.

(Received 9 Feb 2004, accepted 9 Feb 2004) ❑