# Statistical methods in clinical trials

Val J Gebski and Anthony C Keech

APPROPRIATE STATISTICAL METHODS for analysing trial data are critical for the correct interpretation of the results. Item 12 of the CONSORT statement (Box 1) relates to the statistical methods used in the reporting of trials, together with scientific and statistical principles concerning analyses of subgroups, endpoints and appropriate statistical tests. These issues need to be carefully considered before beginning a study and should be outlined in a standard trial protocol, which may be supplemented by a more extensive statistical analysis plan.[1]

**1: CONSORT checklist of items to report when reporting a randomised trial[1]**

| Section and topic | Item no. | Descriptor |
|---|---|---|
| Statistical methods | 12 | Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analyses. |

## Primary outcomes

Both primary and secondary endpoints should be clearly described in the objectives sections of the trial protocol.[2] Statistical considerations appropriate to the design of the trial, including sample-size calculations, timelines for any interim analyses and a sketch of a proposed statistical plan for analysing these endpoints, should be detailed in the statistical section of the protocol and reported in subsequent publications.[3] The analysis principle for the primary outcome must be that of *intention-to-treat*, where the data are analysed according to the treatment group to which they were randomised.[4]

### *Statistical analysis plan*

Key components of the statistical analysis plan for the primary endpoint or endpoints include:

***Specifying how the outcome will be measured.*** Common measures are:

- ***Binary*** (whether or not an event has occurred) — for example, whether or not the subject has experienced a complete or partial response from cancer treatment at 12 months. Typical measures of the event are proportions (risk), rates or odds, and measures of treatment effect include *odds ratios* and *differences in the proportions (or rates)* between the intervention and control groups.
- ***Count*** (the frequency of an event in a set time period) — for example, the number of episodes of epilepsy experienced by patients in a 30-day period. A typical unit of measurement would be the rate (count per unit time), and measures of treatment effect include *incidence density ratios* (similar to odds ratios) or *differences between the rates* in the groups being compared.
- ***Time to event*** (how long it takes to observe the outcome of interest) — for example, the survival time of patients with advanced breast cancer. Endpoints of this type usually contain *censored* data (ie, the event of interest has not been observed by the end of the follow-up period), and analyses would involve comparing "averaged" *relative risks* or *hazard/risk ratios* (pooled across the time period of the study) between the groups.
- ***Measurement on a continuous scale***. Examples include blood pressure and temperature measurements, and analyses involve comparing the *difference between the means* of the intervention and control groups.
- ***Other measurements*** include *ordinal* scales (eg, quality-of-life ratings, 5-point trauma scales) and non-ordered scales (eg, patient preferences between oral, intravenous or combination treatment delivery). Outcomes measured on these scales require specialised statistical methods.

***Any transformations on the data likely to be required before analysis.*** This includes possible groupings or classifications of data (eg, into good, acceptable and poor quality of life), as well as mathematical transformations (logarithms, square root, etc) needed to "normalise" the outcome variables. Typically, these transformations are used if the distribution of the outcome exhibits skewdness, and where, after transformation, this distribution is symmetrical and thus satisfies the assumptions of the statistical method being used to make comparisons.[5] If statistical or graphical methods will be used to examine the distribution of the outcome, such as boxplots, histograms and scatterplots,[6] these should be detailed.

***Appropriate statistical tests which will be used to analyse the data.*** While the underlying assumptions of common statistical tests vary, underpinning all these tests is the assumption that either the outcome (or some transformation of the outcome) or other calculated measures (such as correlation coefficients, hazard or odds ratios) will be "normally" distributed. The normal distribution underlies most statistical inference for most continuous outcomes (and is the basis of $\chi^2$, F and *t*-tests, as well as comparison of odds, hazard and incidence density ratios). If the assumptions of proposed statistical tests do not apply (eg, the data are known to be bimodal), then alternative statistical

**NHMRC Clinical Trials Centre, University of Sydney, Camperdown, NSW.**
**Val J Gebski,** MStat, Principal Research Fellow; **Anthony C Keech,** FRACP, MSc(Epid), Deputy Director.
Reprints will not be available from the authors. Correspondence: Mr Val J Gebski, NHMRC Clinical Trials Centre, University of Sydney, Locked Bag 77, Camperdown, NSW 1450.
val@ctc.usyd.edu.au

approaches (eg, classifying the outcome into categories) for analysing such outcomes should be described.

***How missing data will be accounted for in the analyses (both scientifically and statistically).*** For example, missing data are sometimes omitted, assigned the baseline value or the group average, or imputed using statistical theory.[7]

***Whether statistical inference be will drawn using one-tailed or two-tailed tests (with appropriate justification) and if any statistical adjustments for multiple comparisons will be performed.***

In reporting the results of randomised trials, an unadjusted analysis for the primary outcome will provide a consistent, unbiased estimate of underlying treatment differences; this is guaranteed by the randomisation process. This analysis should usually be the primary comparison. However, if the randomisation was stratified, a primary analysis stratified by the stratification factors may be equally appropriate. Subsidiary analyses, which adjust for stratification factors, other potential confounders, or both, can further define the effect of treatment and may provide more efficient statistical comparisons.

*Parametric* tests are based on specific distributional assumptions such as the normal distribution.[8] Common misconceptions in analysing clinical data are that a *non-parametric* analysis (eg, Wilcoxon rank-sum test) is appropriate if the sample size is small (<30), the data appear skewed (ie, may not be normally distributed) or that the medians are being compared. Whether the distribution of the data departs significantly from the normal distribution may be formally tested; if no departure from normality is indicated, comparisons based on the normal distribution are usually still preferable. Tests based on the assumption of normally distributed data can also be statistically valid for small sample sizes (as low as three per arm). Of course, if there is clear evidence that the data are not normally distributed, the appropriate statistical tests (eg, "exact" tests or non-parametric tests) or appropriate data transformations are required. Finally, even non-parametric tests require some assumptions with respect to the underlying populations from which the samples are drawn.[8] If there is a choice of statistical method (ie, assumptions of a parametric test are satisfied), non-parametric methods are generally not as powerful (ie, do not have the same ability to detect a significant difference if it actually exists) as their parametric counterparts.

A checklist for a statistical analysis plan is provided in Box 2.

**2: Checklist for a statistical analysis plan for clinical trials**

- Provide a detailed description of the primary and secondary endpoints and how they are to be measured.
- Provide details of the statistical methods and tests that will be used to analyse the endpoints. The analysis of the primary outcome must follow the principle of intention-to-treat.
- Describe the strategy to be used (eg, alternative statistical procedures) if the distributional or test assumptions are not satisfied.
- Detail whether comparisons will be one-tailed or two-tailed (with appropriate justification if necessary) and specify the level of significance to be used.
- Identify whether any adjustment to the significance level or the final *P* values will be made to account for any planned or unplanned multiple testing or subgroup analyses.
- Specify potential adjusted analyses with a statement of which covariates or factors will be included.
- Identify any planned subgroup or subset analysis along with justification for the relevance of this analysis (eg, biological rationale) before commencement of the trial.
- Specify planned exploratory analyses, justifying their importance.
- Support claimed differential subgroup effects with biological rationale and supporting evidence from within and outside the study. Provide statistical evidence of interaction between the overall treatment effect and that observed in the subgroup(s) of interest.
- Remember that prespecified subgroups will have more interpretive value than those defined on an ad-hoc basis or as a result of multiple comparisons.

## Changing the primary outcome during the conduct of the study

Circumstances can arise where, after a trial commences, the primary outcome is deemed to be suboptimal. This most commonly occurs when the observed rate of the primary outcome is substantially lower than anticipated, reducing the ability (power) of the study to evaluate the effects of treatment on this outcome. This could be the result of a recent change in non-trial background therapy or to recruitment of a more healthy subset of the patients of interest. In these instances, it is possible to modify the primary outcome, provided the reason for so doing *is not based on knowledge of interim results of the effect of treatment in the study.* Thus, if study data indicate that the rate of myocardial infarction (the primary outcome) is much lower in the intervention or control arm than originally anticipated, it would be highly inappropriate to modify the primary outcome to, for example, include stroke, as this choice is potentially influenced by a knowledge of interim results of the effects of treatment in the study. However, using the overall event rate for myocardial infarction in the whole study cohort (blinded — not differentiated by treatment) could provide justification for endpoint modification in a valid way. Any change in primary outcome during the study requires careful thought, planning and documentation.

## Secondary outcomes

Analysis of the secondary outcomes needs to be described in the same way as that for the primary outcome, with sufficient documentation in the analysis plan as to how they will be analysed. Where possible, further exploratory analyses should be identified before the completion of the study, with a clear scientific rationale for the reason and value of such analyses.

## Subgroup analysis

It is essential that potential subgroup analyses are specified before the commencement of a study to guard against data "dredging" or "trawling". Applying many different statistical tests to the same data (eg, on subgroups or different

outcomes) has the effect of greatly increasing the chance that at least one of these comparisons will be declared statistically significant even if there is no real difference. This practice is often termed data dredging.[9] However, simply specifying a subgroup analysis in advance does not necessarily add scientific legitimacy to the interpretation. A number of strategies exist to ensure the credibility of subgroup analyses, and a checklist proposed by Simes (personal communication) suggests that the following criteria should be satisfied.

- That there is a *biological rationale* for considering the subgroup separately from the rest of the patients in the study. Lack of strong biological or clinical evidence for why the treatment should have different effects in a particular subgroup would detract from support of a true underlying differential effect, even if a conventionally significant difference were found.
- That there is *prior evidence or belief* that a differential treatment effect in a subgroup is plausible. Lack of prior evidence suggests that differential treatment effects observed in subgroups become hypothesis-generating observations rather than firm conclusions.
- That there is statistical evidence (ie, a *significant interaction*) of a difference in the effect of treatment for the subgroup in question compared with the other patients. For example, if there is an apparent advantage of treatment in younger compared with older patients, then careful (clinical and statistical) examination of this difference is required before it can be confidently concluded that a true differential treatment benefit exists in the subgroup of younger patients. Studies are frequently underpowered to detect such interaction effects; nevertheless, lack of statistical evidence of such interaction should prohibit firmly concluding any differential treatment effect in the subgroup.
- That there is *independent confirmation* from other factors in the study of the possible differential treatment effect in the subgroup. For example, if, in a trial examining the effect of chemotherapy in gastric cancer, it is observed that women survive longer after an intervention than men, supporting evidence could be to observe that the response rate to treatment was higher and time to disease progression was also longer in women compared with men.

Common pitfalls with subgroup analysis are focusing on the size of the *P* value and of the treatment effect in any subgroup, ignoring the play of chance. Other issues, such as the total number of subgroups examined, also play a major role in determining the credibility of any observed differential subgroup effect. Subgroups defined before initiating the study would be more credible in terms of true differences in effect on the study findings than those determined only at the time of analysis.

## Competing interests

None identified.

## References


1. Altman DG, Schulz KF, Moher D, et al, for the CONSORT group. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. *Ann Intern Med* 2001; 134: 663-694.
2. Gebski V, Marschner I, Keech AC. Specifying objectives and outcomes for clinical trials. *Med J Aust* 2002; 176: 491-492.
3. Kirby A, Gebski V, Keech A. Determining the sample size in a clinical trial. *Med J Aust* 2002; 176: 256-257.
4. Gebski V, Beller E, Keech AC. Randomised controlled trials, the elements of a good study. *Med J Aust* 2001; 175: 272-274.
5. Woodward M. Epidemiology: study design and data analysis. Boca Raton: Chapman and Hall/CRC Press, 1999.
6. Delaney G, Rus M, Gebski V, et al. An Australasian assessment of the basic treatment equivalent model derived from NSW data. *Australas Radiol* 1999; 43: 500-506.
7. Simes RJ, Greatorex V, Gebski VJ. Practical approaches to minimise problems with missing quality of life data. *Stat Med* 1998; 17: 725-737.
8. Hollander M, Wolf D. Nonparametric statistical methods. 2nd Ed. New York: John Wiley and Sons, 1999.
9. Martin G. Munchausen's statistical grid, which makes all trials significant. *Lancet* 1984; ii: 1457.


*(Received 10 Dec 2002, accepted 23 Dec 2002)*

# book review

## Australian guide to elder care

**Practical guide to geriatric medicine.** Ranjit N Ratnaike (editor). Sydney: McGraw-Hill, 2002 ($186.95, xxv + 958 pp).

ISBN 0 074 40801 5.

THE ELDERLY CONSUME an increasing amount of the health care dollar, hospitals are under pressure to discharge patients early and, increasingly, care is provided in the community. For these reasons *Practical guide to geriatric medicine* is a useful addition to the reference library of any general practitioner, general physician or geriatrician.

The book comprehensively covers a wide range of topics relevant to the clinical care of the elderly. It has a lot of practical tips on the management of common problems, such as psychiatric illness, neurological disorders and dementia, and includes lots of relevant validated screening tools. Whole-body systems are well covered, with an emphasis on screening and prevention.

The chapter authors are experts from various international backgrounds and almost half are Australian. This balance is a bonus, as it ensures that the information is relevant to clinical practice in Australia. For example, there is a chapter on the functional assessment of the over-75-year-old as part of the enhanced primary care package.

For the academically minded reader it would have been helpful to cross-reference the text with the bibliography and further reading references that appear at the end of each chapter.

This book is well written and easy to read, with a wealth of useful information. The presentation is excellent, with paragraph headings and useful tables, as well as figures highlighting the important points. It is excellent value for money and highly recommended for the busy clinician.

**George Szonyi**
Geriatrician
Balmain Hospital, NSW