Reviewer's quick guide to common statistical errors in scientific papers

Design errors

Sample size for human subjects
Many studies are too small to detect even large effects (Table 1).

Table 1: Guide to sample size

Expected difference (p_1-p_2)	Total sample size required*
5%	1450-3200
10%	440-820
20%	140-210
30%	80-100
40%	50-60

* 5% significance level, 80% power. Smaller numbers may be justified for rare outcomes (p_1 <.1)

Look for:

- Clinical trials should always report sample size calculations
- Authors with 'negative' results (i.e. found no difference) should not report equivalence unless sufficiently powered -"absence of evidence is not evidence of absence"

Bias
Randomisation is the best way of avoiding bias but it is not always possible or appropriate.

Some biases affecting observational studies:

Treatment-by-indication bias: different treatments are given to different groups of patients because of differences in their clinical condition.

Historical controls: will tend to exaggerate treatment effect as recent patients benefit from improvements in health care over time and special attention as a study participant. Recent patients are also likely to be more restrictively selected.

Retrospective data collection: availability and recording of events and patient characteristics may be related to the groups being compared.

Ecological fallacy: an association observed between variables on an aggregate level does not necessarily represent the association that exists at the individual level.

Some biases affecting observational studies and clinical trials:

Selection bias: low response rate or high refusal rate – were patients that participated different to those that did not?

Informative dropout – was follow-up curtailed for reasons connected to the primary outcome? If so, imbalance in dropout rates between the groups being compared will introduce bias.

Bias in clinical trials:

No-one should know what the next random allocation is going to be as this may affect whether or when the patient is entered into the trial. Using date of birth, hospital number, or simply alternating between treatments is therefore inappropriate. Central randomisation is ideal.

Unblinded assessment of outcomes may be influenced by knowledge of the treatment group.

Look for:

- Appreciation and measures taken to reduce bias through study design
- Selection of patients, collection of data, definition and assessment of outcome and, for clinical trials, method of randomisation should be clearly described
- Number and reasons for withdrawal should be reported by treatment group
- Appropriate analytic methods such as multiple regression should be used to adjust for differences between groups in observational studies
- Authors should discuss likely biases and potential impact on their results

Method comparison studies
If different methods are evaluated by different observers then the method differences are confounded with observer differences. The study must be repeated with each observer using all methods.

Analysis errors

Failure to use a test for trend on ordered categories (e.g. age-group).

Dichotomizing continuous variables in the analysis (acceptable for descriptive purposes).

Using methods for independent samples on paired or repeated measures data. An example is using both arms or legs of the same patient as if they were two independent observations.

Using parametric methods (e.g. t-test, ANOVA or linear regression) when the outcome or residuals have not been verified as normally distributed.

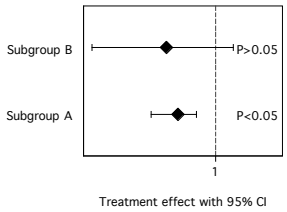
Over using hypothesis tests (P-values) in preference to confidence intervals.

One-tailed tests are very rarely appropriate.

Failing to analyse clinical trials by intention-to-treat.

Obscure statistical tests should be justified and referenced.

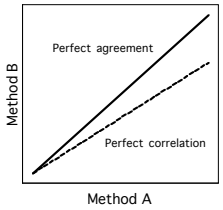
Comparing P-values between subgroups instead of carrying out tests of interaction is incorrect. Some may wrongly conclude from these results that:



the subgroup affects response to treatment, based on comparing P-values. A test of interaction would show no evidence of any effect of the grouping on response.

Correlating time series: any two variables that consistently rise, fall or remain constant over time will be correlated. 'Detrended' series should be compared instead.

Method comparison studies
Correlation ≠ agreement



Higher correlation can be induced by including patients with extreme measurements. Limits of agreement should be calculated according to method of Bland and Altman. Adequate agreement between methods is a clinical not a statistical judgement.

Multiple testing
Conclusions should only be drawn from appropriate analyses of a small number of clear, pre-defined hypotheses. Results from post-hoc subgroup or risk-factor analyses should be treated as speculative. If many such tests have been carried out adjustment for multiple testing should be considered.

Comparing groups at multiple time points should be avoided – a summary statistics approach or more complex statistical methods should be used instead.

Further reading:

CONSORT: http://www.consort-statement.org

Greenhalgh T. How to read a paper: Statistics for the non-statistician. I: Different types of data need different statistical tests. *BMJ* 1997;315:364-366

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-310. Available online at http://www-users.york.ac.uk/~mb55/meas/ba.htm

BMJ Statistics Notes: http://www-users.york.ac.uk/~mb55/pubs/pbstnote.htm

Produced by Tony Brady
Sealed Envelope Ltd
http://www.sealedenvelope.com