EDITORIAL

Editor-in-Chief

Statistical testing for baseline differences between randomised groups is not meaningful

L. A. Harvey¹

© International Spinal Cord Society 2018

It is still common to see statistical testing of baseline data of clinical trials in an attempt to prove that two or more groups to whom patients have been randomised are comparable. For example, groups may be statistically compared on variables such as age, sex or type of injury that are measured before randomisation and before any intervention has been administered. p values less than 0.05 are interpreted as evidence that the groups are not comparable and hence do not provide a fair basis from which to compare the effects of the intervention. At one level this may seem like a reasonable approach. However, at another level, this practice defies the logic of hypothesis testing and encourages ongoing misuse of statistics.

This is not a new revelation. To the contrary, these issues have been talked about for nearly 30 years by leading biostatisticians [1]. Nonetheless, the practice persists. Here are some comments that should dampen enthusiasm for using p values in this way:

“...performing a significance test to compare baseline variables is to assess the probability of something having occurred by chance when we know that it did occur by chance. Such a procedure is clearly absurd.” p. 126 [2].

“ P -values for baseline differences do not serve a useful purpose, since they are not testing a useful scientific hypothesis.” p. 2928 [3].

“With few exceptions, the statistical literature is uniform in its agreement on the inappropriateness of using hypothesis testing to compare the distribution of

baseline covariates between treated and untreated subjects in RCTs.” p. 142 [4].

“Indeed the practice can accord neither with the logic of significance tests nor with that of hypothesis tests....I suspect that the practice has originated through confused and false analogies with significance and hypothesis tests in general.” p. 1716 [5].

Even if we ignore the criticisms of statistical testing for baseline differences, there is the added problem that an insignificant p value may merely reflect a small sample size. That is, there may be large differences that statistical testing fails to detect. And what if there is a significant difference on one baseline variable? It would be rather surprising if there was not given the high number of variables that are typically measured at baseline. A single p value less than 0.05 among many baseline statistical tests may just reflect a spurious finding.

Spinal Cord encourages authors to refrain from statistically testing for possible baseline imbalance in randomised studies.

References

1. Bland JM, Altman DG. Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials*. 2011;12:264.
2. Altman AR. Comparability of randomised groups. *Statistician*. 1985;34:125–36.
3. 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;21:2917–30.
4. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *J Clin Epidemiol*. 2010;63:142–53.
5. Senn S. Testing for baseline balance in clinical trials. *Stat Med*. 1994;13:1715–26.

✉ L. A. Harvey
spinalcord@iscos.org.uk

¹ University of Sydney, Sydney, Australia