

## EDITORIAL

# Statistical power calculations reflect our love affair with *P*-values and hypothesis testing: time for a fundamental change

*Spinal Cord* (2014) **52**, 2; doi:10.1038/sc.2013.117

Statistical power calculations are still considered essential when planning trials in spinal cord injury (SCI) and some may think it heresy to suggest otherwise.<sup>1</sup> Power calculations are reported in grant applications, ethic submissions and journal publications. The power of a trial refers to the probability (usually set at 80%) that the trial will reach statistical significance if there truly is the specified difference between two groups.<sup>2</sup> Power calculations rely on estimating variability and nominating an  $\alpha$ -level (usually, 0.05). As variability increases, then power decreases. Power is only a relevant concept during the planning stages of a trial. Once a trial is completed then power is meaningless.<sup>2,3</sup> To determine whether a completed trial has a sufficient sample size requires examination of the precision of treatment estimates (for example, the 95% confidence intervals of the mean between-group differences).

Power calculations are problematic for one fundamental reason; they perpetuate an out-dated focus on the hypothesis-testing approach to statistical analysis.<sup>2,4</sup> The hypothesis-testing approach dichotomises results as significant or not significant making interpretation misleadingly simplistic. Clinical trialists and 'statisticians have for a long time argued that this [approach] is inappropriate and that trials should be used to estimate treatment effects and not to test hypotheses' (p. 806).<sup>5</sup> This might go against everything some have been taught. However, the push away from *P*-values and hypothesis testing is not new, and journals are increasingly trying to encourage researchers and readers to focus on estimates of treatment effects<sup>6</sup> (the uncertainty of those estimates can be quantified with confidence intervals). One leading journals tried abolishing *P*-values altogether, and a number of others insist that estimates of treatment effects be reported in an effort to stop researchers and readers simply dichotomising results as significant or not.<sup>6</sup>

An alternative to statistical power calculations is called 'precision for planning' (p. 372).<sup>3</sup> This involves calculating the sample size required to ensure a pre-set level of precision in estimates. Precise estimates require larger sample sizes than imprecise estimates. This

approach still requires predicting variability that can be problematic but, importantly, it does not focus on *P*-values. Before we see a shift to 'precision for planning' there will need to be a fundamental change in attitudes and understanding. Trials will need to be valued for their ability to provide estimates of treatment effects rather than their ability to provide statistically significant findings. This shift in attitudes and understanding is not only important for how we determine sample size requirements for trials but also for how we conduct and interpret clinical trials, systematic reviews and clinical practice guidelines. We need this shift to advance evidence-based care for people with SCI.

## CONFLICT OF INTEREST

The author declares no conflict of interest.

## ACKNOWLEDGEMENTS

Thanks to Geoff Cumming, Joanne Glinsky and Ian Cameron for their valuable feedback.

LA Harvey

Sydney Medical School/Northern, University of Sydney, Sydney,  
New South Wales, Australia  
E-mail: l.harvey@usyd.edu.au

- 1 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; **365**: 1348–1353.
- 2 Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 1994; **121**: 200–206.
- 3 Cumming G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals and Meta-Analysis*. Routledge: London, UK, 2012.
- 4 Schmidt F, Hunter J. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: Harlow L, Mulaik S and Steiger J (eds) *What If There Were No Significance Tests?*. Erlbaum: Mahwah, NJ, 1997.
- 5 Edwards SJL, Lilford RJ, Braunholtz D, Jackson J. Why 'underpowered' trials are not necessarily unethical. *Lancet* 1997; **350**: 804–807.
- 6 Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychol Sci* 2004; **15**: 119–126.