LINK TO ORIGINAL ARTICLE
LINK TO AUTHOR'S REPLY

Small sample size is not the real problem

Peter Bacchetti

Widespread poor research practices raise difficult questions about how to bring about improvements. Unfortunately, I believe that the Analysis article by Button *et al.* (Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013))[1], along with previous similar discussion of sample size[2], misidentifies small sample size as a fundamental cause of problems in research and at the same time uncritically accepts a very harmful overemphasis on whether $p < 0.05$.

Much of their argument is undercut by the fact that the positive predictive value of $p < 0.05$ (PPV) is an unacceptably poor measure of the evidence that a study provides. PPV ignores distinctions between different p values below 0.05, such as $p = 0.049$ versus $p < 0.0001$, and therefore wastes information. Estimated effects, confidence intervals and exact p values should be considered when interpreting a study's results, and these make power irrelevant for interpreting completed studies[3–5]. In addition, any specific result (for example, $p = 0.040$) is not weaker evidence because of small sample size per se than the same p value would be with a larger sample size[6].

Button *et al.*[1] rightly distinguish the inherent consequences of small sample size from associated characteristics, but they do not acknowledge that questioning the validity of all small studies on the basis of associated factors is likely to be both ineffective and unfair. Associated problems should be addressed directly; trying to mitigate them by advocating larger sample sizes is distracting and confusing. Indeed, the concept of 'adequate' sample size promotes misinterpretation[7] of study results owing to the focus being only on whether $p < 0.05$. Importantly, the 'winner's curse' is caused by selection and is not an inherent problem for small studies if their results will be disseminated no matter what they turn out to be.

The discussion of ethics in the article[1] neglects a fundamental fact about power (and any other measure of a study's projected value): diminishing marginal returns[8,9]. Each additional subject produces a smaller increment in projected scientific or practical value than the previous one. This implies that efficiency defined by projected value per animal sacrificed will be worse with a larger planned sample size[8].

In addition, Button *et al.*[1] do not fully acknowledge the many conceptual and practical difficulties of power-based sample size planning. The fact of diminishing marginal returns precludes any meaningful definition of 'adequately powered' versus 'underpowered' (REF. 7); the goal of 80% power is only an arbitrary convention[10]. In addition, specifying the 'right' alternative effect size, along with other assumptions needed for calculations (such as the standard deviation), is often difficult; the true effect is not always a sensible choice for power calculations (for example, see the Xuan row in table 1 in the article[1]) and cannot be known with good accuracy in advance[7]. Power calculations therefore should not overrule cost–efficiency and feasibility[9], and this is impossible in real research practice anyway.

Manipulation of the design, conduct, analysis and interpretation of studies towards producing more 'interesting' results is a serious problem, as is selective dissemination of studies' results, but these are not caused by small sample size. In addition, it is counterproductive to analyse power and PPV, the very definitions of which contain the assumption that a study's results will be dichotomized. Trying to improve research while conceding that a study's information will be reduced to just one bit of information — whether $p < 0.05$ — is like starting an armistice negotiation by offering unconditional surrender.

Peter Bacchetti is at the Department of Epidemiology & Biostatistics, University of California, San Francisco, California 94143–0560, USA.
e-mail: PBacchetti@epi.ucsf.edu

doi:10.1038/nrn3475-c3
Published online 3 July 2013

1. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013).
2. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
3. Goodman, S. N. & Berlin, J. A. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann. Intern. Med.* **121**, 200–206 (1994).
4. Hoenig, J. M. & Heisey, D. M. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Stat.* **55**, 19–24 (2001).
5. Senn, S. J. Power is indeed irrelevant in interpreting completed studies. *BMJ* **325**, 1304 (2002).
6. Royall, R. M. The effect of sample-size on the meaning of significance tests. *Am. Stat.* **40**, 313–315 (1986).
7. Bacchetti, P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med.* **8**, 17 (2010).
8. Bacchetti, P., Wolf, L. E., Segal, M. R. & McCulloch, C. E. Ethics and sample size. *Am. J. Epidemiol.* **161**, 105–110 (2005).
9. Bacchetti, P., McCulloch, C. E. & Segal, M. R. Simple, defensible sample sizes based on cost efficiency. *Biometrics* **64**, 577–594 (2008).
10. Bacchetti, P., Deeks, S. G. & McCune, J. M. Breaking free of sample size dogma to perform innovative translational research. *Sci. Transl. Med.* **3**, 87ps24 (2011).

Competing interests statement
The author declares no competing financial interests.