LINK TO ORIGINAL ARTICLE
LINK TO INITIAL CORRESPONDENCE

Empirical evidence for low reproducibility indicates low pre-study odds

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò

In response to our Analysis article (Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013))[1], Hoppe (A test is not a test. *Nature Rev. Neurosci*. http://dx.doi.org/10.1038/nrn3475-c5 (2013))[2] correctly points out that the figures in our article cover only pre-study odds of an effect (R) up to 1 (that is, up to 50% prior prevalence of a non-null effect). In principle, of course, R can be higher — the prior prevalence of a non-null effect can vary from 0 to 100%. However, in the large majority of research studies, it will be 50% or lower.

When the prior probability of an effect is very high, such as that required to justify a large, confirmatory clinical trial, it will still only approach 50% (R = 1). Empirically, Djulbegovic *et al.*[3] have recently shown this to be the case: only slightly more than 50% of Phase 3 clinical trials show superiority of the intervention arm over the comparator arm. As large-scale clinical trials are arguably the end stage of a research pipeline that begins in the basic sciences, such trials should represent the case when R (on average) can be expected to achieve the highest values. As R increases above 1 (that is, prior prevalence >50%), the incremental value of further research decreases; when it is very high (for example, prior prevalence >90%), further research is probably not necessary, because there is already high confidence in the outcome. Most neuroscience research is far removed from this situation, as most neuroscience involves testing exploratory hypotheses and addressing measurements with high complexity and extreme multiplicity. In other words, there are many variables that can be explored for associations and effects, often with little prior insight regarding which of them may be important.

Although the true value of R will vary somewhat from one field to another, empirically we know that it is likely to be low in many (or even most) fields. Recent attempts to replicate key findings from the biomedical science literature have indicated that the proportion of studies that replicate effects is usually less than 50%, and even this may be optimistic[4–6]. If one assumes that the true prevalence of effects is approaching 100%, then one has to also assume that these ubiquitous effects are very small, otherwise they should be replicated routinely. However, when effect sizes are tiny, even very large studies will generate substantial type S errors[7] (that is, many statistically significant effects will be in the opposite direction of the true effect). In this situation, the credibility of individual, meticulously performed, extremely large studies would still be close to 50% at best.

Finally, although the extent of the impact of statistical power on positive predictive value will vary for different values of R, ultimately statistical power is important for the whole range of R. There is growing evidence for the poor reproducibility of reported findings, and there is therefore a need for greater focus on possible reasons for this problem and on the identification of solutions. In our opinion, low statistical power is an important part of this equation.

Katherine S. Button, Claire Mokrysz and Marcus R. Munafò are at the School of Experimental Psychology, University of Bristol, BS8 1TU, UK.

Katherine S. Button is also at the School of Social and Community Medicine, University of Bristol, BS8 2BN,UK.

John P. A. Ioannidis is at Stanford University School of Medicine, California 94305, USA.

Brian A. Nosek is at the Department of Psychology, University of Virginia, Charlottesville, Virginia 22903, USA.

Jonathan Flint is at the Wellcome Trust Centre for Human Genetics, University of Oxford, OX 7BN, UK.

Emma S. J. Robinson is at the School of Physiology and Pharmacology, University of Bristol, BS8 1TD, UK.

Correspondence to M.R.M.
e-mail: marcus.munafo@bristol.ac.uk

doi:10.1038/nrn3475-c6
Published online 23 October 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. Hoppe, C. A test is not a test. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c5 (2013).
3. Djulbegovic, B., Kumar, A., Glasziou, P., Miladinovic, B. & Chalmers, I. Medical research: trial unpredictability yields predictable therapy gains. *Nature* **500**, 395–396 (2013).
4. Begley, C. G. & Ellis, L. M. Drug development: raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
5. Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M. & Zwelling, L. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS ONE* **8**, e63221 (2013).
6. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Rev. Drug Discov.* **10**, 712 (2011).
7. Gelman, A. & Tuerlinckx, F. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Stat.* **15**, 373–390 (2000).

Competing interests statement
The authors declare no competing interests.

© 2013 Macmillan Publishers Limited. All rights reserved