LINK TO ORIGINAL ARTICLE
LINK TO AUTHORS' REPLY

A test is not a test

Christian Hoppe

In their Analysis article (Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013))[1], Button *et al.* state that insufficient statistical power owing to insufficient sample size undermines the reliability of findings in neuroscience. The authors rely on an earlier publication[2] for this statement. Although it is obvious that a small sample size increases the risk of missing an existing effect, these authors question the reliability of significant findings in cases of insufficient power.

In both papers, concepts from diagnostic testing are applied to statistical testing in basic research. In this different context, a test's positive predictive value (PPV) turns into the reliability of a significant statistical effect. However, the size of the critical effect of test sensitivity (or statistical power) on the PPV (or the reliability of significant findings) depends on prevalence (or the odds of effects among tested effects) and becomes negligible if prevalence is high (BOX 1).

Unfortunately, the authors do not discuss the role of odds of effects among all tested effects on the questionable correlation between power and the reliability of significant findings. Using odds instead of probabilities may cover the fact that the paper relies on the implicit assumption of low 'prevalence' of those effects. Button *et al.* state that, "in an exploratory research field such as much of neuroscience, the pre-study odds are often low" (REF. 1), and the diagram in figure 4 of their article uses maximum odds of 1, equalling a maximum prevalence of only 0.50. However, this key assumption is not explained in the article.

In the context of diagnostic testing, disease prevalence may be estimated to a reasonable extent. But I doubt that it makes sense to estimate the odds of true effects among tested effects in a research field. First, the total of tested effects may refer to all possible, reasonable or actually tested effects. Second, researchers usually have good reasons to 'believe' in effects under examination, which might substantially increase the odds of true effects. Third, particularly in neuroscience, one might argue that the 'prevalence' of effects approaches 100% (including small effects), as all neurophysiological phenomena tend to affect each other in some way. Last, estimating the odds of effects from the literature runs into self-contradiction if one claims that most studies have insufficient statistical power to detect existent effects. As we cannot know the odds of true effects among all tested effects, we do not know whether the reliability of significant findings is substantially affected by insufficient statistical power. Thus, it is possible, but not mathematically proven, that insufficient statistical power reduces the reliability of significant findings in biomedicine and neuroscience.

To put it formally, the authors inappropriately mix the rationale of Bayesian statistics with the rationale of statistical hypothesis testing by Neyman and Pearson. However, the paper[1] reminds us that a test of statistical significance never exempts researchers from defining what they consider to be a valuable effect and that it is only meant to ensure that an empirical finding is unlikely to be a mere random result. Pre-set standards for when an effect is accepted as conceptually relevant are needed in each field of research.

Box 1 | **Prevalence, sensitivity and positive predictive value**

In the table below, columns indicate the presence of a disease and rows indicate possible outcomes of a diagnostic test for this condition. The prevalence of the disease is (a + c) / total sample size. The false alarm rate is defined as the probability $\alpha = b/(b+d)$ that a test result is positive despite the disease being absent; specificity is then defined by $(1-\alpha) = d/(b+d)$. The missing error rate is defined by the probability $\beta = c/(a+c)$ that a disease that is present is not picked up by the test; sensitivity is then defined by $(1-\beta) = a/(a+c)$. The probability for a positive test indicating a disease that is present — that is, the test's positive predictive value (PPV) — is calculated by $PPV = \text{sensitivity} \times \text{prevalence} / (\text{sensitivity} \times \text{prevalence} + \alpha \times (1 - \text{prevalence}))$. Thus, a positive diagnostic test becomes less reliable if the test produces many false alarms ($\alpha \to 1$) or if the disease is rare (prevalence → 0). In addition, there is an effect of sensitivity on PPV, but notably this effect depends on the disease prevalence and becomes marginal if the prevalence is high. For example, in case of a disease with a prevalence of 0.10, the PPV of a test (with a specificity of 0.95) equals 0.68 if the sensitivity of the test is 0.80, but the PPV decreases to 0.31 (that is, less than half) if the sensitivity is reduced to 0.20. However, if the prevalence is 0.90, the respective numbers for PPV are 0.994 and 0.973. Crucially, Button *et al.*[1] transferred the rationale and arithmetic from diagnostic testing (that is, Bayesian statistics) to statistical hypothesis testing in a research field. In this framework, the size of the effect of statistical power on the reliability of significant findings depends on the 'prevalence' of positive effects among all tested effects, but this 'prevalence' is by definition not known.

Diagnostic test result	Disease	
	Yes	*No*
Positive	a	b
Negative	c	d

Christian Hoppe is at the Department of Epileptology, University of Bonn Medical Centre, 25 53105 Bonn, Germany.
e-mail: christian.hoppe@ukb.uni-bonn.de

doi:10.1038/nrn3475-c5
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. Ioannidis, J. P. A. *et al.* Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).

Competing interests statement
The author declares no competing interests.