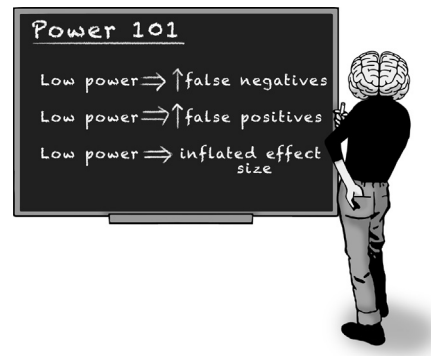
Power and design considerations in imaging research

12

Marcus R. Munafò, Henk R. Cremers, Tor D. Wager and Tal Yarkoni

Why should we care about statistical power? It turns out that many research findings may be false, and low power is one of the main culprits [1]. Low power, by definition, reduces the probability of discovering real effects. In other words, compared to well-powered studies, underpowered studies produce more false negatives—they conclude no effect exists when in reality one does. However, low statistical power also undermines the reliability of research findings in two less-appreciated ways. First, it reduces the probability that an observation passing the threshold for claiming discovery (i.e., statistical significance) actually reflects a real effect. Second, it can lead to an exaggerated estimate of the magnitude of an effect. This effect inflation is sometimes referred to as the "Winner's Curse," the analogy being an auction, where the winner typically pays an inflated price. It often occurs when researchers claim a discovery based on thresholds (e.g., statistical significance, or a Bayes factor of a given value; see additional readings for a more detailed description of these issues) [2,3]. In this chapter, we highlight not only the causes and consequences of low statistical power, but also how functional magnetic resonance imaging (fMRI) researchers are addressing these issues and why we can remain hopeful.



The concept of statistical power is intrinsically linked to the Null Hypothesis Significance Testing (NHST) framework that continues to dominate the biomedical sciences. We can, however, frame the problem in other ways. For instance, some researchers encourage a taxonomy that discusses errors of inference in terms of magnitude (Type M) and sign (Type S) [4], rather than the standard false positive (Type 1) and false negative (Type 2) used in the NHST framework. Smaller sample sizes, all other things being equal, will increase the risk of errors for

Casting Light on the Dark Side of Brain Imaging. DOI: https://doi.org/10.1016/B978-0-12-816179-1.00011-6

both magnitude and sign—in other words, estimates are more likely to deviate substantially from the true population effect, and also more likely to be in the opposite direction [4]. This is linked to the concept of "vibration of effects" [2]—the tendency of small, underpowered studies to be imprecise and therefore provide a wide range of estimates around the true effect size. This is particularly problematic when stringent significance thresholds and publication bias against "null" results conspire to select only the extremes of that range for publication.

Researchers have attempted to estimate the average statistical power of studies across the biomedical sciences. This endeavor remains challenging due to the difficulty in estimating the magnitude of "true" effects (because what is published is only a proportion of all the work conducted, and because of factors such as the Winner's Curse which means these published estimates will be imprecise and potentially inflated). Conventionally, scientists aim for at least 80% power (i.e., a 20% chance of accepting a false negative). Evidence suggests, however, that average power is considerably lower. Within the neurosciences, researchers revealed that average power ranges between $\sim$8% and $\sim$31% [2]—although the distribution may depend on the study type and methodology [5]. These numbers mean that somewhere between 69% and 92% of true effects go undetected. This pattern replicates across a wider range of biomedical sciences [6]. In the neuroimaging literature, estimating effect sizes is even more complex. Nonetheless, a summary of 1131 fMRI studies conducted over a span of more than 20 years suggests that sample sizes have increased only modestly in this time. As of 2015, the median fMRI study was only powered to detect large effect sizes ($d = 0.75$; [7,8]), whereas the typical effect size for the phenomena being tested is likely to be smaller ($d = 0.50$; [7]). Moreover, these estimates come from studies with relatively high-power compared to most fMRI studies [9]. In other words, most fMRI studies don't include enough participants to detect the effects they seek using the standard NHST approach.

Causes and consequence of low power in functional magnetic resonance imaging research

Low statistical power is a problem for any type of research, but certain aspects of fMRI research make the power problem more prominent, and the consequences more troublesome. Here, we elucidate at least three prominent causes with a simple example. Imagine we are conducting an fMRI study on working memory and want to compare patients with a major depressive disorder (MDD) to a group of healthy controls. First, the standard way of analyzing fMRI data divides the brain into about a hundred thousand small cubes, termed "voxels," and looks at the data from each voxel independently. This so-called *mass univariate* analysis requires adjusting the

significance threshold (for instance from $P < .05$ to $P < .0001$) in order to keep the probability of a false positive low, but this stringent threshold requirement necessarily reduces statistical power. Second, the sample size in fMRI studies has only risen modestly over the years [7] despite increasing awareness of the power problem for both functional and structural MRI [2]. In contrast to genetic research, where costs have fallen dramatically enough to allow for high-powered studies, fMRI research remains fairly expensive (around $500 per hour of scanner use) and these fees are unlikely to drop substantially anytime soon. This price tag limits the acquisition of large samples, and in addition, clinical samples, like MDD patients, are difficult to recruit. These first two causes of low statistical power—adjusted statistical thresholds and cost—would not be a major concern if the (expected) effect sizes were very large. However, it is becoming increasingly clear that effect sizes in fMRI are in the low to medium range, and this issue represents the third cause of low statistical power in fMRI research. Some of the large effect sizes in the fMRI literature may emerge due to selective publication of positive results (publication bias, [8]) and the selection of participants who are not representative of the wider population (sampling bias, [9]), or a combination of both. Of course, effect size varies by domain, research question, design, and other factors; however, researchers are realizing that extremely large effects in fMRI research—which would be required to achieve conventional statistical significance with current sample sizes—are rare. Novel approaches (described in the following section) may substantially increase power by increasing effect sizes.

As mentioned in the introduction, the consequences of low statistical power extend beyond its definition—a high chance of missing true effects. Some researchers may consider the increase in the false negative rate as an acceptable trade-off to control the false positive rate. However, the three causes of low power which we describe earlier (large number of dependent variables, small sample sizes, and small effect sizes) have at least three less-appreciated, but potentially far-reaching, consequences. First, as noted earlier, the combination of a small sample size and stringent significance threshold induces a large potential for inflation of statistically significant effects. Therefore effect sizes reported in fMRI studies with relatively small samples can be highly inflated, or even in the opposite direction of, the true effect [3]. In our example, we might find a difference in prefrontal activity between the MDD and control group. When we plot the extracted data of that region, the difference may look spectacularly large. However, this is most likely a case of the described winner's curse and the true effects are much smaller. Second, this potential effect size inflation, in combination with the availability of many dependent variables, can easily lead to misleading inferences about the neural architecture of cognition [9]. In particular, when true effects are small and diffusely distributed throughout the brain (a plausible model for many cognitive processes and differences between psychiatric patients and control groups), underpowered studies will tend to identify only a small subset of effects, but with substantially inflated effect sizes—often leading

researchers to incorrectly conclude that effects are strong and localized. When looking at the statistical map comparing the MDD patients and control group, we may observe just one or perhaps a few "spots" in the brain—yet the true difference in neural functioning between the two groups is much more likely to be distributed and small. Third, a consequence of low power is that different studies on the same psychological process and/or psychiatric disorder will report disparate results. One study might report a strong difference between MDD and controls in one region, another study an effect in a completely different region, so that the second doesn't replicate the first. This situation requires no explanation other than low power (although low power is rarely the first explanation a researcher will reach for) [2], and it is exacerbated by various forms of reporting bias [8], making it extremely difficult to achieve robust cross-study consensus.

Potential solutions and future directions

Fortunately, there are several solutions to the power problem in fMRI research. The most straightforward way to increase statistical power is to increase the sample size. This practice, of course, is easier said than done: fMRI scanning is expensive, and the recruitment of specific populations difficult (e.g., psychiatric patients). If we strive to maintain the conventional standard of a 5% chance of having at least one false positive among many analyses (termed full family-wise error rate correction), we would need hundreds of participants—particularly for between-groups comparisons (such as our MDD example) and analyses of individual differences [10]. Some scientists are tackling this problem head-on and initiating large-multicenter collaborations and building publicly available fMRI databases [7].

Another straightforward but controversial means to increase statistical power is to apply a more lenient statistical significance threshold. The argument here is that if true effects are small and distributed across the brain, we would need a lenient significance threshold to detect them. The controversy surrounding this approach stems from the parallel increase in false positives. This practice becomes hard to justify when chosing a threshold that increases statistical power just enough to detect at least one statistically significant effect [9].

A third way to increase statistical power is to apply a "region of interest" (ROI) approach, where researchers focus on a single brain region, chosen a priori, instead of all brain regions. Using an ROI approach, we can avoid the need to statistically correct for thousands of voxels, and in turn increase the statistical power. However, there is a great deal of flexibility in how one defines a region [10], and substantial uncertainty in whether a certain region was truly chosen a priori. This potential for "hypothesizing after results are known," or HARKing [7], limits the conclusions that we can draw. Preregistrating study protocols, and prespecifying regions of interest, can help address this limitation (see previous chapter for more details).

Finally, testing each voxel often yields low power and focusing only on ROIs might miss other relevant areas of the brain. An emerging family of measures test predefined patterns that involve multiple variables distributed across many brain regions and/or systems to address both these concern. For example, rather than testing 20 or so regions involved in working memory, you can define one a priori pattern across the images and test the "expression of" or response in that pattern. In the simplest terms, this would involve taking the average activity in the regions included in the pattern. One recent study, for example, did just this [11]. The researchers used neurosynth [12] to identify a working memory-related pattern, averaged over this pattern to develop a single measure of working memory-related activity, and then tested that single measure for effects of a psychosocial stressor [11]. Another recent study has extended this concept to test averages over predefined large-scale networks [13]. This example looked at seven predefined cortical networks [14] that span the cortex. This approach largely reduces bias and the potential for HARKing. Moreover, limiting the analysis to seven patterns reduces the problem of multiple comparisons.

Multivariate pattern-based approaches can also yield much greater effect sizes, and reduce the number of tests from many voxels to the expression of a single, predefined pattern [15,16]. For example, when researchers applied an established multivariate pattern—a pain-predictive model called the Neurologic Pain Signature [17]—to new individual participants, they found very large effect sizes for high versus low pain ($d = 1.2-3.50$) [17,18]. Similarly, a negative emotion-predictive model, the Picture Induced Negative Emotion Signature [19], differentiated emotionally negative images from neutral images with an effect size of $d = 4.69$. Effect sizes for a Vicarious Pain Signature [18], applied to comparisons of high versus low observed pain in independent samples, ranged from $d = 1.63-1.75$ [18,20]. These effect sizes are several times larger than those found in voxel-wise analyses [7], and do not require correction for multiple comparisons when testing the magnitude of the response in a pattern as a whole. These examples illustrate that novel analytic approaches can address statistical power concerns and provide biomarkers for cognitive and affective processes that can be validated and used across studies.

Conclusions

More and more researchers are beginning to appreciate the implications of low statistical power, from the need for a priori sample size calculations to inform study design, to the impact of low power on the robustness of a study's conclusions. In the context of fMRI, high costs, small effect sizes, small sample sizes, and multiple comparisons all exacerbate the problem of low statistical power. Fortunately, brain researchers are increasingly addressing the issue of low power using both solutions that apply to the wider scientific enterprise, as well as a number of fMRI-specific

advances, including novel analytical approaches. Taken together, we can remain cautiously optimistic that the robustness of the fMRI literature will improve.

Additional readings

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14(5):365−76. PubMed PMID: 23571845.

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafo MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 2017;18(2):115−26. PubMed PMID: 28053326.