A call to action for publishing study designs and preliminary results in the *Archives of Clinical Psychiatry*

Andre R. Brunoni[1]

[1] *Service of Interdisciplinary Neuromodulation, Department and Institute of Psychiatry, University of São Paulo, São Paulo, SP, Brazil.*

Received: 10/10/2018 – **Accepted:** 10/10/2018
DOI: 10.1590/0101-60830000000175

Brunoni AR / Arch Clin Psychiatry. 2018;45(6):137-8

"*There is more than one way to skin a cat*", an old proverb says. Scientists know this well (metaphorically speaking, of course), as they often face complex challenges that require creative, out-of-the-box thinking, and are usually praised for their problem-solving skills.

The issue, however, is when these creative minds are faced with non-significant findings of their meticulously performed research; and, then, after ingenuity and art, obtain positive findings: in some cases, outliers are "rightfully" excluded and replaced by new data; in others, dozens of tests are run until, "out of serendipity", that association (never thought before) suddenly fits with the mainstream theory. Some data just needed to be handled more carefully, treated more softly, discussed more thoughtfully, to stand a chance in this wild world of peer reviewing and publishing. In fact, although "*God loves the [p=]0.06 as much as the 0.05*"[1], editors, reviewers and grant committees still love the latter much, much more[2].

The result of this mixed bag of controversial approaches to the data, driven by academic pressures and collectively described as "p-hacking"[2], has been detrimental to both basic and applied science, as it leads to an overinflation of false-positive – and, hence, non-replicable – findings. In psychology, a consortium of several research groups recently aimed to replicate 100 highly representative studies published in leading journals. The authors found that only one third of the replication studies had significant findings (*vs.* 97% of original studies) and the mean effect size of the replication studies was half the magnitude of the original studies[3]. Research that uses statistical methods to detect excess significance is also revealing. In a systematic review of meta-analyses investigating brain volume abnormalities, the mean effect size of each meta-analysis was employed to estimate the power to detect an alpha at 0.05 and then to estimate the number of expected positive datasets. The author found that there were too many studies with statistically significant results in the literature on brain volume abnormalities[4], which strongly suggests publication bias and/or p-hacking. A similar approach was used in psychotherapy studies, showing that the effect sizes of the psychotherapy interventions were overestimated, and that the literature had an unexpected high number (excessive) of positive findings according to the obtained evidence[5].

In another approach, Head *et al.*[2] used the p-curve to assess the reliability of published research. When the true effect of an investigated phenomenon is zero (true negative), each p value has an equal probability to occur (i.e., p = 0.04 is as likely as p = 0.03), whereas in true positive findings the p-curve right skews (i.e., has more smaller values) as the effect increases. In both cases, the p-curve shape will be changed if there is evidence of p-hacking. The typical pattern is an increased frequency of p-values just below 0.05, when researchers stop their efforts to obtain significant findings. The authors used text-mining techniques to extract p values from all open access papers available in PubMed, finding strong evidence for p-hacking across all disciplines. For instance, approximately 50% and 60% of studies in the medical and psychological sciences, respectively, had p-values between 0.045 and 0.05, "just in the limit" of significance. They concluded that p-hacking is widespread in scientific literature[2].

What can scientists do, therefore, to mitigate the p-hacking plague from our fields? Unfortunately, there is no single solution for this complex problem. There is still too much emphasis in the p-value, whereas more informative statistics, such as the effect size and its surrounding confidence interval, are usually neglected. In fact, the p-value only informs the probability of obtaining an equal or more extreme effect, *given the null hypothesis is true*. In fact, the pre-test probability (i.e., prior likelihood of the phenomenon) is the main determinant for rejecting a false negative finding[6]. Nonetheless, and in spite of having or not theoretical knowledge, editors, authors and policy-makers often simplify the p-value as being the probability of a true effect (i.e., of neglecting the null hypothesis). On the other hand, p-values > 0.05 are also informative in studies that were well-designed, powered, and robust to biases regarding sample selection, masking, performance and attrition[7].

Therefore, results should be contextually interpreted and, hence, a detailed description of the study design is crucial to critically assess its methodology. Interestingly, although the CONSORT statement emphasizes this need[8], studies are still insufficiently described in most fields[9]; possibly due to editorial restrictions regarding article size. Thus, publications of only study protocol and design are useful, which also allow assessment of study methodology separately and independently of the obtained results. Most importantly, though, is the *a priori* statement of the main research question, the primary and secondary hypotheses, and the planned statistical analyses. Authors are naturally not impeded to perform *post hoc* analyses, which also have its exploratory and hypothesis-generating value. In addition, new and important research questions might arise during an ongoing study that were not initially planned – for instance, a new neuroimaging analysis method that was not at first available or simply did not exist at study start. Researchers should and must explore novel research tracks – and also be transparent regarding *a priori* and *post hoc* hypotheses, a goal that study design publication can help to accomplish.

Considering these challenges, the *Archives of Clinical Psychiatry* issues a call to action for authors to publish the design, protocol and preliminary findings of their studies in the journal. We believe that this is one of the necessary steps to increase transparency, decrease p-hacking and tackle the reproducibility issues that ravage clinical neuroscience and psychiatry. The *Archives of Clinical Psychiatry*, an open access, peer review journal, welcomes authors to share their study protocols and methodology. Therefore, dear author, please let us know, in excruciating details, how you plan to skin your next cat.

Acknowledgements

The author is recipient of a Capes/Humboldt fellow for experienced researchers.

Address for correspondence: Andre R. Brunoni. University of São Paulo, Institute of Psychiatry. Av. Dr. Ovídio Pires de Campos, 785, 2º andar, Ala Sul – 05403-903 – São Paulo, SP, Brazil.

CC BY

References

1. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol. 1989;44(10):1276-84.
2. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015;13(3):e1002106.
3. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716.
4. Ioannidis JP. Excess significance bias in the literature on brain volume abnormalities. Arch Gen Psychiatry. 2011;68(8):773-80.
5. Flint J, Cuijpers P, Horder J, Koole SL, Munafò MR. Is there an excess of significant findings in published studies of psychotherapy for depression? Psychol Med. 2015;45(2):439-46.
6. Ioannidis J. Why most published research findings are false. PLos Med. 2005;2(8):e124.
7. Pocock S, Stone G. The primary outcome fails – what next? N Engl J Med. 2016;375(9):861-70.
8. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.
9. Aparício LVM, Guarienti F, Razza LB, Carvalho AF, Fregni F, Brunoni AR. A systematic review on the acceptability and tolerability of transcranial direct current stimulation treatment in neuropsychiatry trials. Brain Stimul. 2016;9(5):671-81.