European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-019-04467-5

EDITORIAL

The *p* value wars (again)

Ulrich Dirnagl[1,2]

More than 800 researchers, many of them prominent biostatisticians, called to "raise up against the *p*-value." [1] This recent battle cry was just the climax of a growing insurrection, which prominently surfaced in 2018 when another group of biostatisticians demanded that we should "redefine statistical significance," [2] and proposed to change the default *p* value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries. For many researchers and experts, this demand did not go far enough; in a follow-up statement, they demanded to "remove rather than redefine statistical significance." [3] This apparent upheaval even made it into the lay press. The *Financial Times*, for example, analyzed that "Scientists strike back against statistical tyranny." [4] What's all the fuzz about?

If you have not lived under a stone, you will be aware of a much wider discussion which started in earnest about a decade ago in Psychology, but then quickly percolated through the life sciences in general. According to a survey by *Nature* [5], the majority of researchers feel that a replication crisis has affected their discipline, as many experimental findings cannot be replicated and are likely to be false [6]. The search for underlying causes has spawned a whole new research field, meta-research [7]. It is currently widely believed that, among other issues (which include low internal validity [8] or publication bias [9]), low statistical power [10] and flawed statistics [11] are root causes of an exceedingly high false positive rate and hence difficulties to reproduce results. This is where the *p* value, or rather its interpretation, takes center stage.

In 2012, Craig Bennett and colleagues won the Ig Nobel Prize in Neuroscience [12] with a remarkable functional neuroimaging study. They positioned a dead salmon, purchased at a local supermarket, in an MR scanner and showed it a series of photographs depicting human individuals in social situations with a specified emotional valence. In a classical fMRI block design, the salmon was asked to determine what emotion the individual in the photo must have been experiencing. BOLD-imaging revealed a significant hemodynamic response indicating neural processing in the dead salmon's brain [13]. The reason why these authors found "neural correlates of interspecies perspective taking in the post-mortem atlantic salmon" is of course that they relied on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$), and did not appropriately control for multiple comparisons. This frivolous study is relevant because the authors also demonstrate that only 60–70% of published functional neuroimaging studies at that time controlled for multiple comparisons, questioning the results of a major portion of cognitive neuroscience studies. Other fields, in particular those based on gene expression [14] and association [15] studies, are also heavily affected by the testing burden and were initially drowning in a sea of false positives [16]. As a result, in functional imaging and genetics, techniques aimed at solving the multiple comparisons problem proliferated. Fortunately, it is nowadays unlikely that one can publish transcriptomic or functional imaging datasets without using some form of post hoc correction. It is good news that some research fields appear to have cleaned up their act. However, the bad news is that in many other fields, statistical problems, including insufficient correction for multiple testing, weak thresholds for type I errors, and low statistical power, are still rampant [17, 18].

But at the heart of the problem lie misconceptions about the *p* value. Many researchers believe that *p* is the probability that the null hypothesis is true, and that 1-*p* is the probability that the alternative hypothesis is true. Or more colloquially, *p* is confused with the false positive rate: "At an alpha of 5%, I am running a 5% chance that my hypothesis is a fluke, the drug is not effective, despite obtaining statistical significance." Other frequent misconceptions include the belief that the *p* value correlates with the theoretical or practical relevance of the finding, or that a small *p* is evidence that the results will be replicable. Another serious fallacy is the notion that failing to reject the null hypothesis ($p > 0.05$) is equivalent to

This article is part of the Topical Collection on Editorial.

✉ Ulrich Dirnagl
ulrich.dirnagl@charite.de

[1] Berlin Institute of Health, QUEST Center for Transforming Biomedical Research, Berlin, Germany

[2] Department of Experimental Neurology, Charité – Universitätsmedizin Berlin, Berlin, Germany

Published online: 03 August 2019

demonstrating it to be true. An excellent and comprehensive treatment of the most prevalent misconceptions about the *p* value can be found here [19].

So what then is the *p* value, and what can it tell us about our results? If we were to repeat the analysis many times, using new data each time, and if the null hypothesis were really true, at $p = 0.05$ on only 5% of those occasions would we (falsely) reject it. In other words: The *p* value represents the probability of getting data as extreme as our results if the null hypothesis is true. But do not these definitions sound compatible with interpreting the *p* value as a false positive rate? Let us look at this more closely: In the above definitions, probability is applied to the data, but the common fallacy is to apply it to the explanation (i.e., the hypothesis). In addition, we have no idea whether the null is true or not. And then there is the problem of the pre-study odds, and statistical power (i.e., the probability of detecting an effect given that there is one).

The question we all would like to answer is this: If we observe a "significant" *p* value after doing a single unbiased experiment, what is the probability that our result is a false positive? Unfortunately, the *p* value is only part of the equation we would need to solve, because the false positive rate depends on alpha, beta (power), as well as the prior probability of the hypothesis we are testing [10, 11]. The unlikelier our hypothesis is, and the lower statistical power, the more likely it is that we are looking at a false positive. Despite a significant *p* value. To put this into perspective: At an alpha of 0.05, power of 80%, and pre-study odds (prior probability that the alternative hypothesis is true) of 10%, almost 40% of the statistically significant results will be false positives. For a concise treatment of the positive predictive value as a function of the pre-study odds for different levels of statistical power, see [10, 20]. Importantly, in many fields of biomedicine, and in particular in preclinical research, statistical power often is much lower than 80%. In exploratory research, 10% pre-study odds may even be an overestimation. Together with a high prevalence of bias (and in the presence of low internal validity), this may well explain why "most published research findings are false." [6]

But wait a minute! Do not high energy physicists use statistical significance as a criterion to accept or refute new particles? And did not they just recently close in on the Higgs boson and reward a Noble prize using the "five-sigma" level of statistical significance, which corresponds to a *p* value of 3×10^{-7}? As The *Wall Street Journal* properly concluded [21]: "That is not the probability that the Higgs boson doesn't exist. It is, rather, the inverse: If the particle doesn't exist, one in 3.5 million is the chance an experiment just like the one announced this week would nevertheless come up with a result appearing to confirm it does exist."

Correct! The *p* value tells us something about the experiment, not the hypothesis. Another example from physics can help to clarify that even insanely low *p* values cannot prove a hypothesis. In 2011, a scientific revolution was announced. Neutrinos were found to travel faster than light. The *New York Times* [22] reported that "Tiny Neutrinos May Have Broken Cosmic Speed Limit" and asked "Einstein to roll over." Physicists working at the particle accelerator at CERN in Geneva had produced neutrinos and sent them on a 730-km-long trip. The arrival of the neutrinos was registered by a detector blasted through thousands of meters of rock in the Dolomites earlier than photons would have traveled such a distance. The physicists immediately realized that their finding would revolutionize physics. To be on the safe side, they raised the significance level from 5 sigma, the community standard for the discovery of new elementary particles, to 6 sigma. Moreover, they repeated the experiment several times. But the neutrinos did not bother about the speed of light, and the results remained significant at the incredible level of 6.2 sigma! [23] Unfortunately, this episode did not result in a Nobel prize, or the invention of time travel, but total embarrassment for the scientists involved. As it turned out later, a cable in the setup was loose, and the GPS system with which they had measured the distance was not properly calibrated.

Given the limitations of the *p* value, which are inherent to frequentist statistics ("Null hypothesis significance testing," NHST), and the serious misconceptions regarding its interpretation, shall we raise up against it and ban its use? I believe that this would be throwing out the baby with the bathwater. In a recent thoughtful commentary [24], John Ioannidis argued that "Significance (not just statistical) is essential both for science and for science based action, and some filtering process is useful to avoid drowning in noise." His point is that retiring significance testing would give bias a free pass, consequently "irrefutable nonsense would rule." [25] We are already drowning in a sea of false positive results, without any threshold for claiming an association this dire situation would almost certainly get worse. Instead, we should prespecify stringent rules before data collection and analysis. Albeit the norm in many fields, a 5% threshold of significance is insufficient for claiming genuine association [2] and, if anything, indicates that the results are "worth a look" and justify further study. Reporting a discovery based only on $p < 0.05$ is wrong. Without sufficient power, any *p* value is unreliable, while effect sizes (given a true effect) are overestimated. In an exploratory study mode, which is the norm in most of basic and preclinical biomedical research, we may want to emphasize power over alpha. This means that we should be less tolerant of false negatives than false positives: We surely do not want to miss an effect. But we need to be aware that many of our "statistically significant" results will be false positives. These we need to weed out through confirmatory experiments with stringent alpha levels. Almost certainly confirmation will require even larger sample sizes, and preregistering study protocols should be mandatory. Unfortunately, it is presently common practice to use exploratory studies to support confirmatory inferences [26].

In research, some connections must be established between chance and support. Criticism of NHST and in particular the *p* value as tools of hypothesis evaluation has been waxing and waning since its introduction [19, 27]. None of the arguments raised in the current debate regarding NHST and *p* value, or proposed alternatives, are novel. But the exponential growth of biomedical research over time has massively increased not only the use but also the misuse and misinterpretation of statistical methods to analyze data. Using significance levels (e.g., 5%) in hypothesis testing has become a "ritual." [28] This has grave consequences, as the *p* value underpins the majority of inferences in the biomedical literature. The declaration of statistical significance has become meaningless.

In conclusion, we should not depend on the *p* value when interpreting our results. Test statistics may guide our reasoning, but not determine it. As an editorial [29] in a recent special issue of *The American Statistician* on *p* values and statistical significance (containing 43 articles!) put it:

- Do not base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p* value passed some arbitrary threshold such as $p < 0.05$).
- Do not believe that an association or effect exists just because it was statistically significant.
- Do not believe that an association or effect is absent just because it was not statistically significant.
- Do not believe that your *p* value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Do not conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

We should be "Moving to a World Beyond '$p < 0.05$'" [29], where biological thinking rules, and sound data production is emphasized through careful planning, design, execution, and reporting of our studies. A world where methods and results are transparently described so that effects and inferences can be independently confirmed.

Compliance with ethical standards

Conflict of interest The author declares that there is no conflict of interest.

References

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305–7.
2. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:6–10.
3. Amrhein V, Greenland S. Remove, rather than redefine, statistical significance. Nat Hum Behav. 2018;2:4.
4. Anjana A. Scientists strike back against statistical tyranny | Financial Times. Financial Times; 2019.
5. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–4.
6. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:0696–701.
7. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: evaluation and improvement of research methods and practices. PLoS Biol. 2015;13:e1002264.
8. Henderson VC, Demko N, Hakala A, MacKinnon N, Federico CA, Fergusson D, et al. A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib. Elife. 2015;4:1–13.
9. Nissen SB, Magidson T, Gross K, Bergstrom CT. Publication bias and the canonization of false facts. 2016.
10. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14:365–76.
11. Colquhoun D. The reproducibility of research and the misinterpretation of *p*-values. R Soc Open Sci. 2017;4:171085.
12. Winners of the Ig® Nobel Prize. *Improbable research* Available at: https://www.improbable.com/ig/winners/#ig2012. (Accessed: 17th July 2019).
13. Bennett CM, Baird AA, Miller MB, Wolford GL. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for proper multiple comparisons correction. J Serendipitous Unexpected Results. 2011;1:1–5.
14. Mudge JF, Martyniuk CJ, Houlahan JE. Optimal alpha reduces error rates in gene expression studies: a meta-analysis approach. BMC Bioinformatics. 2017;18:312.
15. Pulit SL, De With SAJ, De Bakker PIW. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. 2016. https://doi.org/10.1002/gepi.22032.
16. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med. 2002;4:45–61.
17. Lew MJ. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. Br J Pharmacol. 2012;166:1559–67.
18. Diong J, Butler AA, Gandevia SC, Héroux ME. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. PLoS One. 2018;13:e0202121.
19. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000;5:241–301.
20. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of P values. R Soc Open Sci. 2014:1–15. https://doi.org/10.1098/rsos.140216.
21. Bialik C. How to be sure you've found a Higgs boson. Wall Street J. 2012.
22. Overbye D. Tiny neutrinos may have broken cosmic speed limit: The New York Times; 2011.
23. The OPERA Collaboration. Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. 2011. https://doi.org/10.1007/JHEP10(2012)093.
24. Ioannidis JP, The A. Importance of predefined rules and prespecified statistical analyses. JAMA. 2019;321:2067.
25. Ioannidis JPA. Retiring statistical significance would give bias a free pass. Nature. 2019;567:461–1.
26. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol. 2014;12:e1001863.
27. Harlow, L. L., Mulaik, S. A. & Steiger, J. H. What if there were no significance tests? (1997).
28. Gigerenzer G. Mindless statistics. J Socio-Econ. 2004;33:587–606.
29. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond 'p<0.05'. Am Stat. 2019;73:1–19.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.