Kill the Statistical Significance Level (p-Value)

Egea-Guerrero JJ[1] and Vilches-Arenas A[2]*

[1]Neuro Critical Care Unit. Virgen del Rocío University Hospital, IBIS/CSIC University of Seville. Spain.

[2]Department of Preventive Medicine and Public Health, Faculty of Medicine, University of Seville, Spain

*Corresponding author: Prof. Vilches-Arenas MD. PhD., Department of Preventive Medicine and Public Health, Faculty of Medicine, University of Seville, Spain, Email: ava@us.es

Abstract

For some time, numerous articles and editorials in prestigious journals have warned about the misuse of the p-value in scientific research. Authors advocate for replacing statistical significance analysis with effect sizes and confidence intervals. Despite these recommendations, the emphasis on "significance level" persists in most biomedical journals. Ronald Fisher, who introduced the p-value, saw it as a measure of evidence strength against a null hypothesis, not a decisive tool for accepting or rejecting hypotheses. The common cut-off of $p<0.05$ is arbitrary and often misinterpreted, especially given its sensitivity to sample size. Large samples can yield statistically significant but clinically irrelevant results, while small samples might miss clinically important differences. The p-value's dichotomous nature (significant vs. non-significant) was never intended by early statisticians and has led to widespread misinterpretation. Calls to lower the threshold to 0.005 exacerbate this issue, potentially undermining the credibility of countless studies. P-values do not measure effect size, importance, hypothesis truth, or data randomness; they assess data-model incompatibility. The current practice encourages data manipulation to achieve desired p-values, leading to unreliable publications and misguided clinical decisions. To improve scientific rigor, impact journals should require effect sizes and confidence intervals, fostering more meaningful and reproducible research outcomes. It is time to replace the p-value's misuse and abandon "statistically significant" as a marker of true effect.

Keywords: P-Value; Statistically Significant; Research

Opinion Article

Dear Editor,

Numerous articles, opinions, and editorials in prestigious journals have long warned about the misuse of the p-value in scientific research. These authors advocate abandoning statistical significance analysis and instead endorsing the use of effect sizes and confidence intervals [1-4]. However, despite these warnings, the emphasis on "significance level" analysis persists in most biomedical journals.

Ronald Aylmer Fisher, who introduced the p-value, intended it to measure the strength of evidence against a null hypothesis, not as a definitive tool for rejecting or accepting ideas. The commonly accepted cut-off point, $p<0.05$, is arbitrary and often misinterpreted, especially given its sensitivity to sample size. Large samples can produce statistically significant but clinically irrelevant results, while small samples might miss clinically important differences. Unfortunately, there is not always a correlation between clinically relevant and statistically significant results. Scientists have often pursued differences solely because of their statistical significance, ignoring clinically relevant differences that were statistically insignificant. With the advent of Big Data, hypotheses can be rejected at any level of significance (p), complicating small group analyses (e.g., rare diseases).

Setting a p-value cut-off leads to dichotomous decisions, a practice not intended by the pioneers of statistical inference [5]. This division into "significant" and "non-significant" is problematic. It is troubling that the literature is filled with misinterpreted work based on p-value significance levels and that professional associations have recommended lowering the threshold from 0.05 to 0.005 [6-8]. This change could undermine the credibility of millions of scientific publications, while restricting medical research that does not achieve the established p-level. Addressing this issue would require a retrospective reconstruction of nearly all available scientific evidence, revealing that science may have been built on fallacies and that clinical decisions have often been based on convenience rather than evidence.

P-values and statistical significance do not measure the size of an effect or the importance of a result, nor do they determine the probability of a hypothesis being true. A p-value measures the incompatibility of data with a given statistical model. Scientific conclusions should not be based solely on whether a p-value exceeds a specific threshold. Many clinical journals still use this criterion for manuscript acceptance, encouraging researchers to manipulate data to achieve a desired p-value. This practice leads to misguided scientific processes and unreliable publications that do more harm than good. The misuse of p-values affects people with health problems and has economic repercussions due to non-reproducible statistically significant conclusions.

Beginning researchers unfamiliar with the p-value have an opportunity to avoid these misinterpretations and become better scientists. P-values often confuse rather than clarify, leading to flawed results in the literature. Fortunately, some professors disregard this institutionalized hindrance to medical research. It would be beneficial for impact journals to exclude scientific work that does not provide effect sizes and confidence intervals. This change would help researchers develop their inferential capabilities and offer scientifically significant results. This would be a welcome opportunity for intellectual development after years of decisions determined by significance level.

After nearly a century, we cannot continue to use a mathematical apparatus that no longer advances scientific progress, especially amid a crisis of non-reproducible results in many biomedical areas. It is time to end the debate on the use and interpretation of p-values, agree on a replacement, and abandon the term "statistically significant" as evidence of a real effect.

Conflict of Interest

The authors declare no competing financial interests. Each author has made a substantial contribution to qualify as such. No work resembling this manuscript has been published nor is under consideration for publication elsewhere.

References

1. Nuzzo R (2014) Scientific method: statistical errors. Nature 506(7487): 150-152.

2. Kyriacou DN (2016) The Enduring Evolution of the P Value. JAMA 315(11): 1113-1115.

3. Chavalarias D, Wallach JD, Li AH, Ioannidis JP (2016) Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA 315(11): 1141-1148.

4. Harrington D, D'Agostino RB Sr, Gatsonis C, Hogan JW, Hunter DJ, et al. (2019) New Guidelines for Statistical Reporting in the Journal. The New England Journal of Medicine 381(3): 285-286.

5. Sterne JA, Davey Smith G (2001) Sifting the evidence-what's wrong with significance tests?. BMJ 322(7280): 226-231.

6. Wasserstein RL, Lazar NA (2016) The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70 (2): 129-133.

7. Ioannidis JPA (2018) The Proposal to Lower P Value Thresholds to .005. JAMA 319(14): 1429-1430.

8. Waterstone RL, Umbrella AL, Lazar NA (2019) Moving to a world beyond "p < 0.05". On the stats 73(S1): 1-19.

Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats
  (Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission

https://juniperpublishers.com/online-submission.php