BJPsych | The British Journal of Psychiatry (2025)
1–2. doi: 10.1192/bjp.2025.10383

Letter

Hypothesis testing miscommunication in psychiatric randomized controlled trials

Amin Sharifan

Keywords
statistical data interpretation; clinical relevance; scholarly communication; mental health; research design.

A *p*-value represents the probability of observing the study results, or more extreme results, assuming the null hypothesis is true. However, *p*-values do not provide evidence in support of the null hypothesis, irrespective of their value or the significance threshold. Consequently, statements suggesting that a non-significant *p*-value indicates 'no difference' or 'no effect' are inaccurate.[1] Bayesian statistics and Bayes factors[2,3] offer more appropriate approaches for quantifying evidence in favor of the null hypothesis, as they can directly measure support for 'no difference' versus 'a difference', unlike *p*-values that cannot support 'no difference' claims. On the other hand, significant *p*-values do not necessarily indicate the presence of a true effect, nor do small *p*-values represent large effect sizes.[4]

Furthermore, the interpretation of decision *p*-values is binary, and findings are either statistically significant (rejecting the null hypothesis) or not (failing to reject the null hypothesis).[5] Qualifying terms such as 'highly significant', 'trend toward significant', 'borderline significance', or 'near-significant' are discouraged, as statistical significance is not a spectrum. This concept is important in RCTs, where the *p*-value is often compared to a predefined threshold to assess an intervention's efficacy or safety.

Yet the reporting and interpretation of randomized controlled trial (RCT) findings, which are central to advancing and shaping evidence-based psychiatric care, often follow a frequentist approach that misinterprets these statistical principles. This approach tends to prioritize statistical significance, commonly defined by an alpha threshold of 0.05, over clinical relevance.[6] This practice risks conflating statistical significance with clinical relevance, potentially misleading clinicians, researchers, and policymakers about the clinical importance or magnitude of effect of interventions.[7] Misinterpretation of non-significant results, such as incorrectly accepting the null hypothesis by stating 'there was no difference between groups' or describing findings as 'trending toward significance', can distort perceptions of findings and influence decision makers.

When examining recent publications in leading psychiatry journals, the extent of this problem became evident. A review of RCTs using hypothesis testing published between January 2024 and May 2025 in six major psychiatry journals (American Journal of Psychiatry, Biological Psychiatry, BMJ Mental Health, JAMA Psychiatry, Molecular Psychiatry, and The Lancet Psychiatry) revealed widespread statistical misinterpretation.

Of 259 records, 70 were eligible for inclusion. Among the 189 excluded records, 149 (79%) were conference abstracts, and 3 (2%) had used Bayesian statistics. Of 70 included publications, 45 (64%; 95% CI, 53–75) contained null acceptance terminology, 13 (19%; 95% CI, 11–29) included trend terminology, and 6 (9%; 95% CI, 4–18) contained both types of terminology. Additionally, 3 (4%; 95% CI, 2–12) did not report effect sizes, and 13 (19%; 95% CI, 11–29) did not include confidence intervals for effect estimates. Table 1 contains the characteristics of publications with inaccurate statistical phrases.

Table 1 The characteristics of randomized controlled trials in six major psychiatry journals with null acceptance and/or trend phrases published between 2024 and 2025

	n/*N* (%)	95% CI
Study area		
Major depressive disorder	13/52 (25)	15–38
Mental health in general	8/52 (15)	8–28
Alcohol use disorder	5/52 (10)	4–21
Post-traumatic stress disorder	3/52 (6)	2–16
Other	23/52 (44)	32–58
Intervention type		
Pharmacological	14/52 (27)	17–40
Neuromodulation	10/52 (19)	11–32
Digital health	9/52 (17)	9–30
Psychotherapy	9/52 (17)	9–30
Other	10/52 (19)	11–32
Funding type[a]		
Government	44/77 (57)	46–68
Institution	14/77 (18)	11–28
Non-profit	13/77 (17)	10–27
Industry	6/77 (8)	4–16

CI, Confidence interval; *n*, Number of eligible entities; *N*, Total number.
a. Total exceeds the number of publications due to several sources per publication.

Null acceptance terminology had a median *p*-value of 0.52 (IQR, 0.30–0.74), with the most common examples being 'two groups did not differ', 'there were no differences between the groups', and 'there was no evidence of difference'. This terminology appeared in 12 abstracts (27%; 95% CI, 16–41) and 42 full texts (93%; 95% CI, 82–98).

Trend terminology had a median *p*-value of 0.0595 (IQR, 0.051–0.0805), with examples such as 'near-significant ($p = 0.06$),' 'marginally significant ($p = 0.067$),' and 'borderline significant ($p = 0.055$).' One exception involved a *p*-value less than 0.00001 interpreted as 'highly significant.' This terminology appeared in two abstracts (15%; 95% CI, 4–42) and 12 full texts (92%; 95% CI, 67–99).

The evidence based on this subset of the literature was striking: nearly two-thirds of publications contained inappropriate null acceptance claims, while almost one in five used misleading trend terminology. Nevertheless, the absence of assessing the presence of spin, a form of selective reporting, and the limitation to six journals may restrict their generalizability. Despite most publications reporting effect sizes and confidence intervals, problematic interpretive language remained common, indicating that the issue lies in how researchers interpret rather than report their findings, underscoring the need for greater clarity in statistical communication.

To summarize, the *p*-value alone provides limited information about the clinical relevance of a finding. As recommended by the American Psychological Association, researchers should accompany hypothesis testing results with effect estimates and measures of statistical precision, such as confidence intervals, to provide a more

comprehensive interpretation rather than relying solely on statistical significance. This approach is important in psychiatry research, where small sample sizes are common due to recruitment challenges in populations with psychiatric disorders such as schizophrenia. In these contexts, hypothesis testing may be less informative, and alternative methods such as Bayesian statistics may provide more intuitive interpretations by incorporating evidence-based prior knowledge and directly quantifying the likelihood that treatments differ, rather than relying solely on *p*-values. Researchers are also encouraged to predefine evidence-based significance thresholds that best fit their investigation and hypothesis, rather than focusing exclusively on an alpha level of 0.05. Additionally, researchers should determine and incorporate the minimal clinically important difference a priori to contextualize the clinical relevance of their findings. Collectively, these practices help avoid dichotomizing results solely based on statistical significance.

Amin Sharifan, PharmD, Researcher, Department for Evidence-based Medicine and Evaluation, University for Continuing Education Krems, Lower Austria, 3500 Krems an der Donau, Austria

Email: amin.sharifan@donau-uni.ac.at

First received 27 May 2025, final revision 5 Jul 2025, accepted 23 Jul 2025

Data availability

The data supporting this letter are available to bona fide researchers from the corresponding author, Amin Sharifan, upon reasonable request.

Acknowledgements

None to declare.

Author contributions

AS: the applicable CRediT Taxonomy roles.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of interest

The author declares no conflict of interest.

Ethics statement

Not applicable.

Consent statement

Not applicable.

Transparency declaration

Amin Sharifan affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted.

Analytic code availability

The R (version 4.4.2, 2024, Windows, The R Foundation, Vienna, Austria,) code used in this study is available to bona fide researchers from the corresponding author, Amin Sharifan, upon reasonable request.

Research material availability

Not applicable.

References

1 Ferr H. Misinterpretations of the p-value in psychological research: implications for mental health and psychological science. *PLOS Ment Health* 2025; **2**: e0000242.

2 Goligher EC, Heath A, Harhay MO. Bayesian statistics for clinical research. *Lancet* 2024; **404**: 1067–76.

3 Schmalz X, Biurrun Manresa J, Zhang L. What is a Bayes factor? *Psychol Methods* 2023; **28**: 705–18.

4 van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN. A new look at P values for randomized clinical trials. *NEJM Evid* 2023; **3**: EVIDoa2300003.

5 Greenland S. Divergence versus decision P-values: a distinction worth making in theory and keeping in practice: or, how divergence P-values measure evidence even when decision P-values do not. *Scand Stat Theory Appl* 2023; **50**: 54–88.

6 Thirthalli J, Rajkumar RP. Statistical versus clinical significance in psychiatric research—an overview for beginners. *Asian J Psychiatry* 2009; **2**: 74–9.

7 Citrome L. The Tyranny of the P-value: effect size matters. *Bull Clin Psychopharmacol* 2011; **21**: 91–2.