BJA

British Journal of Anaesthesia, 134 (4): 909–913 (2025)

doi: 10.1016/j.bja.2025.01.008
Advance Access Publication Date: 4 February 2025
Editorial

EDITORIAL

The misunderstood *P*-value: why statistical significance is not enough in clinical practice

Ebadullah S. Ahmed and Mohsin N. Butt*

Department of Anaesthesiology, Aga Khan University, Karachi, Pakistan

*Corresponding author. E-mail: mohsin.nazir@aku.edu

Summary

P-values have traditionally guided clinical research, but over-reliance on them can lead to misinterpretation and poor decision-making. This article highlights common misconceptions about *P*-values and suggests incorporating the minimum clinically important difference (MCID) along with other metrics such as effect sizes and Bayesian methods. Evidence-based practice is essential in anaesthesiology, and research findings should be evaluated in the context of patient outcomes to guide clinical decisions.

Keywords: clinical relevance; confidence interval; machine learning; minimum clinically important difference; *P*-value; second-generation *P*-value; statistical significance

The *P*-value has been an essential component in assessing the statistical significance of a study in clinical research. Despite their widespread use, *P*-values are often misunderstood, leading to misinterpretations that can negatively impact clinical decision-making. In anaesthesia and critical care, choosing an intervention without considering prior probability (i.e. the likelihood of effectiveness based on existing evidence) can lead to suboptimal patient outcomes. Over-reliance on significance tests such as *P*-values without incorporating previous context can result in poor clinical decisions. This article discusses the importance of the *P*-value in research, its misconceptions, the consequences of over-reliance on this test, and the need to incorporate measures such as the minimum clinically important difference (MCID) in almost all fields of medicine, including anaesthesiology.

Misconceptions regarding *P*-values

One of the most common misconceptions about *P*-values is the belief that they represent the probability that the null hypothesis is correct. Many clinicians mistakenly believe that a *P*-value of 0.05 suggests a 5% chance that the outcome is a result of chance.[1] As emphasised by Staffa and Zurakowski,[2,3] *P*-values are often reported without accompanying effect estimates or confidence intervals (CIs), which leads to confusion or misinterpretation. As a result, this misunderstanding often leads to overconfidence in the results classified as significant. Another common misunderstanding is the belief that *P*-values alone determine the validity or importance of a finding. The common practice of labelling results as 'significant' if the *P*-value is <0.05 overlooks an important fact: *statistical* significance does not always mean the result is *clinically* relevant. Even if a study has a large sample size and produces a statistically significant *P*-value, the effect size might be so small that it does not make a real difference in practice.[4]

As Gelman and Stern[5] pointed out, the distinction between 'significant' and 'not significant' is not necessarily meaningful on its own, highlighting the limitations of relying solely on *P*-values for interpreting results. Goodman[6] identified a 'dirty dozen' of *P*-value misconceptions, noting that *P*-values do not measure the size of an effect or the probability that the data were produced by random chance alone. Consequently, this misunderstanding can lead to the incorrect assumption that a statistically significant result automatically has clinical importance, which is not always true.

Potential for type I errors in multiple comparisons

Significance testing is associated with some degree of error, such as type I and type II errors. A type I error, or 'false positive', occurs when the null hypothesis is incorrectly rejected, suggesting an effect exists when it does not. Similarly, a type II error, or 'false negative', occurs when the null hypothesis is incorrectly accepted suggesting no effect when one actually exists. In clinical trials, the concept of multiplicity arises where multiple comparisons are made in a single study.[7] This results in the amplification of type I error with an increase in the number of statistical tests made. Family-wise error rate, which is the probability of making at least one type I error across a set of hypothesis tests, increases with the number of tests in the set.

Multiple testing involves testing a set of hypotheses simultaneously, and multiple group comparisons pertain to investigations involving more than two study arms or the performance of subgroup analyses. Multiplicity can occur in various scenarios, including comparisons across multiple subgroups, evaluations of multiple treatment arms, assessments of multiple outcomes, or analyses of the same outcome at different time points.[8] Researchers have corrected this issue using *post hoc* analyses such as the Bonferroni adjustment. This method modifies the significance threshold by dividing the desired alpha level (typically 0.05) by the number of comparisons being made. For example, a study that has 10 comparisons would have an adjusted threshold of 0.005, which would help reduce the chances of type I error.[7] However, reducing the *P*-value to 0.005 can pose significant challenges, as discussed below. Even then, the possibility that results are attributable to random variability remains.

Consequences of over-reliance on *P*-values

In light of these challenges, an increase in dependence on *P*-values has led to several negative trends in clinical research. One major issue is publication bias, where studies with small *P*-values have a greater chance of getting published, whereas those with larger than standard cut-offs are often disregarded. Such bias can skew the evidence, potentially leading to the adoption of treatments that are either ineffective or harmful.[9] Specialities such as anaesthesiology, which focus heavily on patient outcomes, can be seriously affected by misinterpretations or manipulations of results. Moreover, the focus on achieving a *P*-value <0.05 has led some researchers to engage in unethical practices such as *P*-hacking, where data are manipulated or analysed in multiple ways to achieve significance.[10]

Greenland[11] has argued that although valid *P*-values function as intended, they are frequently misused and misunderstood, which can result in misleading criticism and improper applications in research. To counter these issues, he recommended incorporating *S*-values instead, which quantifies the strength of association in terms of bits, making the data less prone to misinterpretation. The *S*-value is the negative log base 2 transformation of the *P*-value, also known as 'Surprisal' or 'Shannon information', and is measured in bits. The *S*-value represents the amount of information in the data against the background assumptions, model, and test hypothesis. For example, a *P*-value of 0.05 corresponds to an *S*-value of 4.32, which is only slightly more surprising than getting all heads in four fair coin tosses. Thus, as the *P*-value approaches zero, the *S*-value increases in bits, making it a more user-friendly metric for assessing the strength of evidence.[12]

Understanding the *P*-value

The first step in addressing these problems is to understand what a *P*-value actually represents. A *P*-value indicates the probability of observing a result as extreme as that obtained, assuming the null hypothesis is true if the experiment were repeated a large (i.e. infinite) number of times.[19] It does not measure the probability that the null hypothesis is true, the size of an effect, or its clinical importance. A low *P*-value indicates that the observed results would be unlikely under the null hypothesis, suggesting that the null hypothesis may not be true. However, it is crucial to note that this does not necessarily confirm that the treatment is effective or that the observed effect has practical significance in a clinical context.[13]

Blume and colleagues[14] introduced the second-generation *P*-value (SGPV) to tackle some of the shortcomings of traditional *P*-values. Unlike the standard approach, the SGPV takes into account effect size and CIs, offering a more detailed view of statistical significance and highlighting the importance of effect sizes and their real-world relevance.[14] The SGPV builds on the traditional *P*-value by evaluating how uncertainty intervals (such as CIs) overlap with a null region representing trivial effect sizes. Rather than simply testing a single null hypothesis, it provides more nuanced conclusions: an SGPV of 0 indicates support for meaningful effects, 1 suggests only trivial effects, and values between 0 and 1 indicate inconclusive findings. This approach emphasises practical relevance over statistical significance, helping to lower false discovery rates, which represent the proportion of statistically significant results that are false positives in a set of hypothesis tests. SGPVs can be used with a variety of statistical methods, including Bayesian intervals, making them a more flexible and insightful alternative to conventional *P*-values.[15]

P-values in machine learning

In machine learning (ML), statistical significance testing is less commonly used. Instead, model performance is validated using metrics such as the area under the curve (AUC) and receiver operating characteristic (ROC) curves, which are better suited for assessing model accuracy. The commonly used evaluations in ML include accuracy, precision, recall, and F-1 score; however, examining these metrics individually has certain limitations. To assess the performance of binary, multiclass, and multilabel classifiers, additional metrics such as the area under the receiver operating characteristic (AUROC) and Kappa statistics should also be considered, as they provide deeper insights into model suitability.[16] A study in Sweden developed the ICURE model to predict 30-day mortality for ICU patients using nationwide data. The model outperformed traditional methods such as the SAPS 3 score. However, instead of relying solely on *P*-values, the model's performance was evaluated using the AUROC curves.[17] This reflects the shift in ML from focusing on statistical significance to predictive accuracy, reducing the emphasis on traditional *P*-value-based analysis.

Alternatives to *P*-values

Minimum clinically important difference and other clinically **meaningful measures**.

Given the limitations of *P*-values, there is a growing need for measures that better represent clinical significance. MCID is one such measure. It represents the smallest change in outcomes that can be considered beneficial for patients and would lead to a change in patient management.[18] In contrast to *P*-values, which focus only on statistical significance, MCID emphasises the practical importance of research findings from the patient's perspective. In anaesthesiology, where the primary goal is to improve patient outcomes, incorporating MCID into study design and interpretation is crucial. For example, instead of focusing solely on whether a new analgesic technique produces a statistically significant reduction in pain scores, researchers should consider whether the reduction is large enough to improve the patient's quality of life. According to Draak and colleagues,[19] MCID provides a framework for assessing changes in clinical outcomes, such as pain scores, which is crucial for patients.

In considering the relationship between sample size, power calculations, and clinical significance, it is crucial to acknowledge that traditional power calculations often assume statistical significance without regard for practical relevance. Using MCID in sample size calculations ensures that studies are powered to detect clinically meaningful differences that matter to patients. In addition, interpreting estimates and CIs in reference to the MCID highlights the practical importance of findings, helping clinicians decide if the findings are relevant for improving patient care.

Consider this hypothetical example of a study evaluating the effectiveness of a new analgesic in postoperative pain management. The study reports a statistically significant reduction in pain scores with a *P*-value of 0.03. However, the absolute reduction in pain score is only 1 point on a 10-point scale. Without considering the MCID, this result might be interpreted as evidence that the new analgesic is effective. However, if the MCID for pain relief is determined to be 2 points on the same scale, the reduction observed in the study would not meet the threshold for clinical significance. In this case, the treatment might not be deemed sufficiently effective despite the statistically significant *P*-value. This example illustrates the importance of integrating MCID with statistical analysis to provide a more accurate interpretation of the study's clinical implications.

An example from real life is a study that used both anchor-based and distribution-based methods to determine the MCID for pain. For the 100 mm visual analogue scale (VAS), an MCID of ~10 mm was identified as clinically significant across various postoperative procedures. Distribution methods included 0.3 SD, standard error of measurement (SEM), and 5% of the scale range. The study also reported pain reductions with 95% CIs to ensure that the changes were both statistically significant and clinically meaningful.[20] The MCID can hence be used to interpret whether observed differences exceed the threshold for clinical relevance. This way, MCID serves as a critical missing piece of a puzzle in interpreting results holistically.

Benjamin and Berger[21] made recommendations for improving the use of *P*-values, advocating for a more informed use and integrating *P*-values with other statistical measures. The authors recommend lowering the threshold to 0.005, which would increase the strength of evidence. They also recommend reporting *P*-values alongside their Bayes factor bound (BFB), which would help prevent overestimation of the significance of the results. They also emphasise the use of Bayesian analysis to provide a more complete picture of evidence, that is, combining Bayes factor with prior odds, which integrates evidence from other studies and provides a more comprehensive assessment of results.

In addition to the MCID, we recommend using CIs for effect estimates in inferential statistics. CIs and effect sizes should be used alongside *P*-values to provide a more comprehensive understanding of study results. CIs offer insight into the precision of an estimate and the range within which the true effect size is likely to lie, allowing for a clearer understanding of the uncertainty surrounding the effect size. Standard deviations and variance are useful for descriptive statistics to measure and describe the spread of data within a sample rather than making inferences about the population. Standard errors (SEs) play a dual role: they are descriptive in summarising variability and essential in inferential statistics as they are a key component in calculating CIs, hence aiding in making inferences about population parameters. CIs can be calculated as follows: **CI** = mean difference plus or minus 1.96 × **SE**, where *1.96 corresponds to 95% confidence*. Figure 1 summarises supporting methods that researchers can use in addition to *P*-values to improve the interpretation and clinical significance of their results.

Implications of alternative statistical approaches in anaesthesia research

Implementation of alternative statistical approaches will have significant implications in the field of anaesthesia research. Lowering the significance threshold from 0.05 to 0.005 would reduce false positives and improve reproducibility.[22] However, this would significantly increase the required sample size to maintain adequate statistical power for detecting meaningful effect sizes. This increase in sample size could pose challenges for study design and feasibility, particularly for those in low-resource settings. Hadjipavlou and colleagues[23] further illustrate this by adjusting the power and *P*-value while emphasising components of the Bradford Hill criteria to achieve greater positive and negative predictive values. This adjustment is crucial in trials evaluating anaesthetic drugs, where small effects, such as slight decreases in recovery time, must be interpreted carefully to avoid false conclusions. Furthermore, use of SGPVs also lowers false positive risk and provides researchers with results that are easy to interpret.[14] Adapting this approach will help researchers present results reflecting effect magnitude rather than relying solely on statistical significance. While implementing this as a standard will be challenging, it will emphasise the need to assess whether statistically significant results translate into patient benefits, ultimately highlighting the clinical relevance of findings.

For example, in a study comparing the effectiveness of regional *vs* general anaesthesia for a specific procedure, both might show similar statistically significant results, but using SGPVs might indicate better recovery times with regional anaesthesia, which is clinically more meaningful. The use of MCID ensures that observed changes in clinical outcomes are not only statistically significant but also meaningful from the patient perspective. A systematic review by Laaigard and colleagues[24] analysed 570 randomised controlled trials on postoperative pain management after total hip and total knee arthroplasty. Approximately 46% of trials with statistically significant primary outcomes did not meet the MCID criteria, indicating that although statistically significant, almost half of the randomised controlled trials might not have been clinically meaningful.

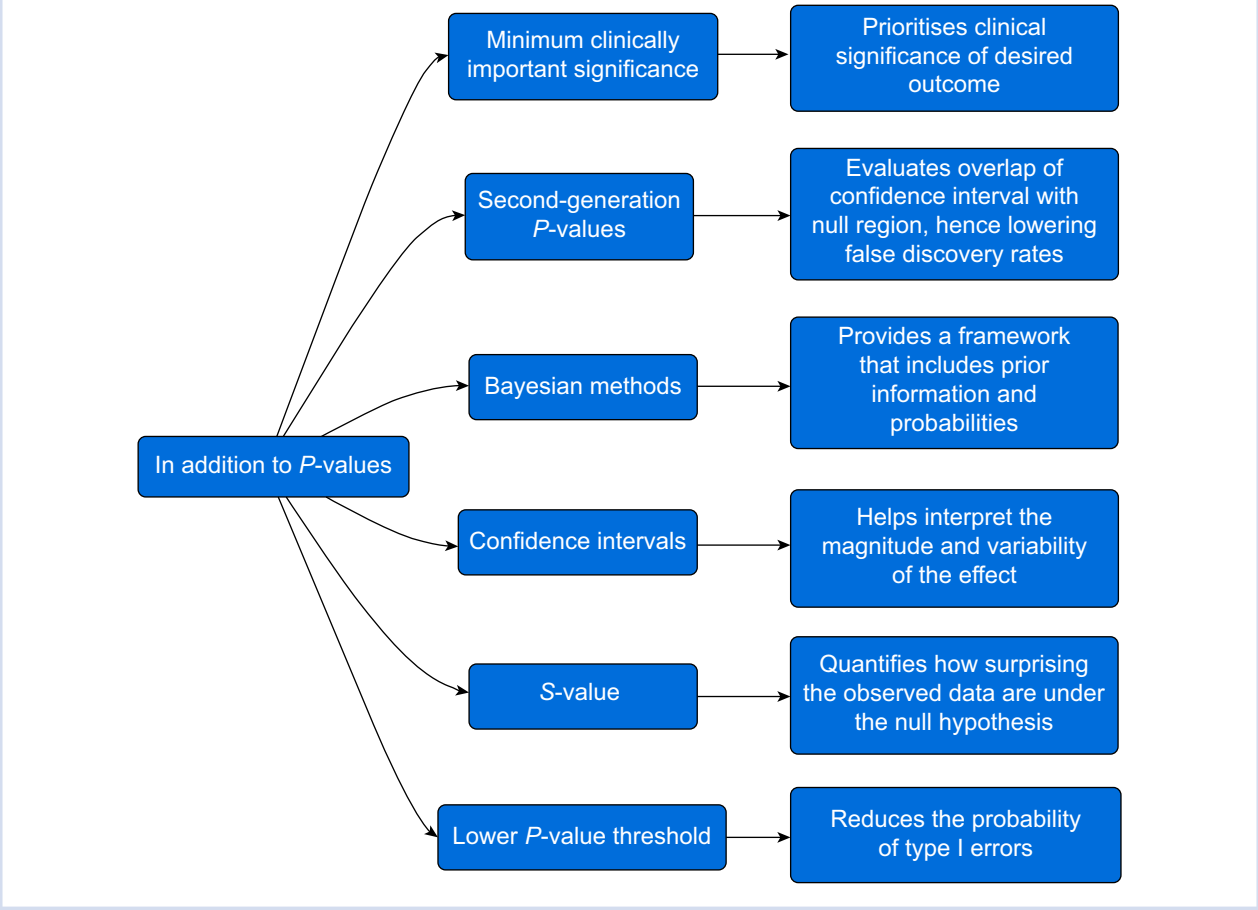

Fig 1. Summary of statistical methods in addition to *P*-values that further validate results.

Lastly, Bayesian methods provide a framework that includes prior information from relevant studies. A hypothetical example by Frost and colleagues[25] involves a sepsis trial and the effects of a new drug on improving 90-day mortality. The results reveal a *P*-value >0.05, indicating insufficient statistical significance for the new drug. However, Bayesian analysis of the hypothetical trial indicates that the treatment is likely beneficial. The posterior probability of an odds ratio (OR) being <1.0, indicating a potential benefit, is estimated at 95%. In addition, the probability of the OR being <0.9, which corresponds to a 10% reduction in mortality, is 80%.

Conclusions

Misinterpretation of *P*-values has led to widespread misconceptions in clinical research, which can result in inappropriate clinical decisions and potentially harmful patient outcomes. Although *P*-values are a useful statistical tool, they should not be the sole determinant of the significance or importance of study findings. Clinicians and researchers must understand the limitations of *P*-values and the importance of incorporating clinically meaningful measures, such as MCID, into their analyses. In anaesthesiology, where the goal is to improve patient outcomes through evidence-based practice, it is crucial to move beyond *P*-values and consider the practical significance of research findings. By adopting a more holistic approach that includes MCID, effect sizes, CIs, and Bayesian methods, clinicians can make better-informed decisions that truly benefit their patients.

Authors' contributions

Conception and design of the editorial, drafting of the initial manuscript, and critical revision for important intellectual content: ESA

Significant input in the development of the editorial and revision of the manuscript: MNB

Approved the final version for publication and agree to be accountable for all aspects of the work, ensuring that questions related to its accuracy and integrity are appropriately investigated and resolved: both authors

Declaration of interest

The authors declare no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT for improving language and enhancing the clarity of English expression. The substantive content, ideas, and arguments

presented remain the original work of the authors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016; **31**: 337–50
2. Zurakowski D, Staffa SJ. Commentary: confidence intervals are only one piece of the puzzle. *J Thorac Cardiovasc Surg* 2022; **164**: e37–8
3. Staffa SJ, Zurakowski D. Guidelines for improving the use and presentation of P values. *J Thorac Cardiovasc Surg* 2021; **161**: 1367–72
4. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat* 2016; **70**: 129–33
5. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat* 2006; **60**: 328–31
6. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol* 2008; **45**: 135–40
7. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int J Epidemiol* 2017; **46**: 746–55
8. Khan MS, Khan MS, Ansari ZN, et al. Prevalence of multiplicity and appropriate adjustments among cardiovascular randomized clinical trials published in major medical journals. *JAMA Netw Open* 2020; **3**, E203082
9. Sterne JAC, Smith GD, Cox DR. Sifting the evidence—what's wrong with significance tests? *BMJ* 2001; **322**: 226–31
10. Verhulst B. In defense of P values. *AANA J* 2016; **84**: 305–8
11. Greenland S. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. *Am Stat* 2019; **73**: 106–14
12. Mansournia MA, Nazemipour M, Etminan M. P-value, compatibility, and S-value. *Glob Epidemiol* 2022; **4**, 100085
13. Jiménez-Paneque R. The questioned p value: clinical, practical and statistical significance. *Medwave* 2016; **16**: e6534
14. Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. An introduction to second-generation *p* -values. *Am Stat* 2019; **73**: 157–67
15. Stewart TG, Blume JD. Second-generation P-values, shrinkage, and regularized models. *Front Ecol Evol* 2019; **7**, 460437
16. Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. In: Silhavy R, Silhavy P, editors. *Artificial intelligence application in networks and systems. CSOC 2023. Lecture notes in networks and systems*vol. 724. Cham: Springer; 2023. p. 15–25
17. Siöland T, Rawshani A, Nellgård B, et al. ICURE: intensive care unit (ICU) risk evaluation for 30-day mortality. Developing and evaluating a multivariable machine learning prediction model for patients admitted to the general ICU in Sweden. *Acta Anaesthesiol Scand* 2024; **68**: 1379–89
18. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trial.* 1989; **10**: 407–15
19. Draak THP, de Greef BTA, Faber CG, Merkies ISJ. The minimum clinically important difference: which direction to take. *Eur J Neurol* 2019; **26**: 850–5
20. Myles PS, Myles DB, Galagher W, et al. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. *Br J Anaesth* 2017; **118**: 424–9
21. Benjamin DJ, Berger JO. Three recommendations for improving the use of p-values. *Am Stat* 2019; **73**: 186–91
22. Chuang Z, Martin J, Shapiro J, et al. Minimum false-positive risk of primary outcomes and impact of reducing nominal P-value threshold from 0.05 to 0.005 in anaesthesiology randomised clinical trials: a cross-sectional study. *Br J Anaesth* 2023; **130**: 412–20
23. Hadjipavlou G, Siviter R, Feix B. What is the true worth of a P-value? Time for a change. *Br J Anaesth* 2021; **126**: 564–7
24. Laigaard J, Pedersen C, Rønsbo TN, Mathiesen O, Karlsen APH. Minimal clinically important differences in randomised clinical trials on pain management after total hip and knee arthroplasty: a systematic review. *Br J Anaesth* 2021; **126**: 1029–37
25. Frost SA, Alexandrou E, Schulz L, Aneman A. Interpreting the results of clinical trials, embracing uncertainty: a Bayesian approach. *Acta Anaesthesiol Scand* 2021; **65**: 146–50