



Access this article online

Quick Response Code:



Website:

<https://www.ovid.com/jnls/tjem>

DOI:

10.4103/tjem.tjem_350_26

Interpreting *P* values in clinical research: A practical guide for clinicians

Alp Şener^{1,2*}

¹Department of Emergency Medicine, Faculty of Medicine, Yıldırım Beyazıt University, Ankara, Türkiye,

²Department of Emergency Medicine, Ankara Bilkent City Hospital, Ankara, Türkiye

*Corresponding author

Abstract:

Accurate interpretation of medical literature is essential for evidence-based clinical practice. Among statistical measures, the *P* value remains one of the most frequently used and most frequently misunderstood concepts in clinical research. Conventionally, a $P < 0.05$ was considered the primary criterion for determining statistical significance. However, reliance on this threshold alone may lead to overestimation of clinically unimportant findings or failure to recognize clinically meaningful effects in underpowered studies. Recent recommendations emphasize the importance of interpreting *P* values together with confidence intervals, effect sizes, and study design characteristics. This review discusses the basic principles of *P* values, common misconceptions, statistical errors, practices that threaten scientific validity, and complementary statistical approaches that can improve clinical interpretation. The aim is to provide clinicians and researchers with a practical framework for evaluating statistical findings in medical studies.

Keywords:

Clinical significance, confidence interval, effect size, evidence-based medicine, *P*-value, statistical analysis

Introduction

The *P* value is one of the most commonly reported statistical measures in medical research. Despite its widespread use, it remains one of the most frequently misunderstood statistical concepts. For many years, a $P < 0.05$ was considered the primary indicator of a successful study outcome. However, this binary approach may lead to inappropriate conclusions regarding both statistical and clinical significance.

The threshold of $P < 0.05$ was originally introduced by Ronald Fisher as a practical guideline rather than a strict decision rule.^[1] Over time, this threshold

became widely accepted as a standard criterion for determining statistical significance in biomedical research. This practice has contributed to important methodological concerns and has led several scientific organizations, including the American Statistical Association, to publish recommendations regarding the interpretation and reporting of *P* values.^[2,3]

Current evidence-based medicine principles emphasize that statistical significance alone should not determine the interpretation of research findings. Clinical significance, effect size, confidence intervals, study design, and statistical power should all be considered when evaluating evidence. Therefore, clinicians should understand both the strengths and limitations of *P* values when interpreting medical literature.

This review addresses several common questions regarding the interpretation of

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License (CC BY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Şener A. Interpreting *P* values in clinical research: A practical guide for clinicians. Turk J Emerg Med 2026;26:281-7.

Submitted: 04-07-2026

Revised: 08-08-2026

Accepted: 29-08-2026

Published: 10-10-2026

ORCID:

AS: 0000-0002-0583-2936

Address for correspondence:

Dr. Alp Şener,
Department of Emergency
Medicine, Faculty of
Medicine, Yıldırım Beyazıt
University, Ankara Bilkent
City Hospital, Ministry of
Health, Bilkent, Ankara,
Türkiye.
E-mail: alpsener@gmail.com

P values in clinical research. It also discusses common misconceptions, statistical pitfalls, and alternative approaches that may improve evidence-based clinical decision-making.

Question 1: What Is a *P* Value and What Does It Not Mean?

Statistical hypothesis testing is based on two competing hypotheses: the null hypothesis that assumes there is no difference between groups or no effect of the intervention being studied, and the alternative hypothesis assumes that a difference or effect exists.

Within this context, the *P* value represents a conditional probability; it measures the likelihood of observing the current study data – or findings even more divergent – under the strict assumption that the null hypothesis is absolutely correct.^[4,5] In practical terms, it estimates how compatible the observed data are with the assumption that no true difference exists between the groups being compared.

Because results vary from one random sample to another, hypothesis testing should also be understood in terms of repeated sampling. If the null hypothesis is true and a significance level of $\alpha = 0.05$ is prespecified, repeated use of the same testing procedure would lead to rejection of the true null hypothesis in approximately 5% of studies in the long run; this is the Type I error rate. Importantly, the *P* value obtained in a single study is not the probability that the null hypothesis is true. Instead, it quantifies how compatible the observed data are with the null hypothesis.^[2,4]

The *P* value ranges from 0 to 1. Smaller *P* values indicate that the observed findings are less compatible with the null hypothesis. However, the *P* value does not directly answer many of the clinical questions that physicians seek to address. Therefore, several common misconceptions about *P* values persist in medical literature. The following statements are incorrect interpretations of *P* values:

If $P = 0.03$, there is a 97% probability that the treatment is effective

This interpretation is incorrect. *P* value does not estimate the probability that a treatment is effective or ineffective.

If $P = 0.09$, there is no difference between the two treatments

This interpretation is also incorrect. A nonsignificant *P* value does not prove that two treatments are equivalent. It only indicates that the available data do not provide sufficient evidence to reject the null hypothesis.

A *P* value of 0.001 indicates a larger clinical effect than a *P* value of 0.03

This statement is incorrect because the *P* value does not measure the magnitude of an effect. It only reflects the compatibility of the observed data with the null hypothesis.^[4,6]

Consider a study comparing a new thrombolytic protocol with standard treatment for reducing 30-day mortality in patients with acute pulmonary embolism. Suppose the analysis yields $P = 0.020$. This result indicates that if there were truly no difference between the treatments, the probability of observing the reported difference, or a more extreme result, would be 2%.

Because this probability is relatively low ($P < 0.05$), the null hypothesis may be rejected according to conventional statistical criteria. However, this result does not prove that the new treatment is definitively superior. It also does not indicate the magnitude or clinical importance of the observed effect. Additional measures, such as confidence intervals and effect sizes, are required to determine the clinical relevance of the findings.

Question 2: Does Statistical Significance Always Imply Clinical Significance?

No, it certainly does not. Statistical significance and clinical significance are related but fundamentally different concepts. Statistical significance tells us how unlikely the observed difference would be if there were truly no differences between the groups. Clinical significance, however, refers to whether the magnitude of the observed effect is meaningful for patients and healthcare providers.

One of the most important factors influencing statistical significance is sample size.^[7] As the sample size increases, even very small differences between groups may become statistically significant. Conversely, studies with small sample sizes may fail to detect clinically important differences. Studies with very large datasets are particularly susceptible to identifying statistically significant findings that have little or no clinical relevance. This occurs because increasing the sample size reduces the standard error and increases statistical power, making small differences more likely to reach conventional significance thresholds.

A practical example can be considered in emergency medicine. Suppose a multicenter study evaluates two analgesic treatments in 10,000 patients presenting with acute mechanical low back pain. Pain severity is assessed using the Visual Analog Scale (VAS), which ranges from 0 to 100 mm. The results show that analgesic A

reduces pain by an average of 42 mm, whereas analgesic B reduces pain by 44 mm. Although the difference between the groups is only 2 mm, the large sample size may produce a highly significant *P* value, such as *P* = 0.001. Despite its statistical significance, this finding may not be clinically meaningful. In pain research, the minimal clinically important difference generally accepted for the VAS is approximately 13 mm. Therefore, a difference of only 2 mm is unlikely to influence clinical decision-making or improve patient outcomes.

This example illustrates an important principle of evidence-based medicine: statistical significance does not necessarily indicate clinical importance. Interpretation of study findings should therefore include assessment of the effect size, confidence intervals, clinical relevance, and the potential impact on patient care rather than relying solely on the *P* value.

Question 3: What Is the Relationship between the *P* Value and Confidence Intervals?

A *P* value provides information about how compatible the study data are with the null hypothesis. However, it does not provide information about the magnitude, direction, or precision of an observed effect. Confidence intervals provide additional information and are therefore considered an essential component of statistical reporting.^[2,4]

Current recommendations emphasize that *P* values should be interpreted together with confidence intervals rather than reported alone. Confidence intervals provide an estimate of the range within which the true population effect is likely to lie. They also allow researchers and clinicians to assess the precision of the estimate and the potential clinical importance of the findings. The width of a confidence interval is influenced primarily by the variability of the data and the sample size. Larger sample sizes generally produce narrower confidence intervals, indicating greater precision. Conversely, smaller sample sizes often result in wider confidence intervals and greater uncertainty. For a sample mean, an approximate 95% confidence interval can be calculated as:

$$95\% \text{ confidence interval} = \text{Mean} \pm (1.96 \times [\text{standard deviation} / \sqrt{\text{sample size}}])$$

In this formula, the sample mean represents the center of the confidence interval. The standard deviation reflects the variability of the observations, whereas the sample size affects the precision of the estimate. The expression “standard deviation divided by the square root of the sample size” represents the standard error of the mean. The constant 1.96 represents the critical value from the

standard normal distribution that captures 95% of the data, meaning that 2.5% of the probability lies in each tail. As the sample size increases, the standard error decreases, resulting in narrower and more precise confidence intervals.

There is a direct mathematical relationship between *P* values and confidence intervals, but the relevant null value depends on the effect measure being reported. Importantly, this interpretation concerns the confidence interval for a between-group effect estimate, not the confidence interval for the raw values of a variable within a single group. For difference measures, such as a mean difference or risk difference, the null value is 0. Thus, if the 95% confidence interval for the between-group difference includes 0, the data are compatible with no difference between the groups at the corresponding two-sided significance level; if 0 is excluded, the result is generally statistically significant at *P* < 0.05. This does not mean that the measured variable itself is zero. For example, when comparing temperatures between two locations, the relevant quantity is the confidence interval for the difference in mean temperature, not whether the confidence interval for the temperature in either location contains 0°C. For ratio measures, such as odds ratios (OR) and relative risks (RR), the null value is 1. A 95% confidence interval that includes 1 is typically not statistically significant, whereas an interval that excludes 1 is generally statistically significant.^[4]

The importance of confidence intervals can be demonstrated using three hypothetical studies evaluating two vasopressor treatments in critically ill patients [Table 1].

Although Scenarios A and B have identical *P* values, their interpretations differ substantially. In Scenario A, with a small sample size, the wide confidence interval indicates considerable uncertainty, and a clinically important effect cannot be excluded. The lack of statistical significance may be related to insufficient statistical power. In contrast, Scenario B has a narrow confidence interval centered near the null value, suggesting that the two interventions likely have similar effects. In this case, the absence of statistical significance is more consistent with a true lack of clinically important difference. Scenario C with a huge sample size demonstrates another important limitation of relying solely on *P* values. Despite a highly significant *P* value, the confidence interval remains close

Table 1: The influence of sample size and confidence intervals on the interpretation of study findings

Clinical scenario	<i>n</i>	OR	95% CI	<i>P</i>
Scenario A	75	0.448	0.177–1.133	0.090
Scenario B	20,000	0.951	0.897–1.008	0.090
Scenario C	60,000	0.945	0.914–0.977	0.001

n: Sample size, OR: Odds ratio, CI: Confidence interval

to the null value, suggesting that the observed effect size may be small and of limited clinical importance.

These examples illustrate why confidence intervals should always be considered when interpreting study findings. Reporting only P values may lead to incomplete or misleading conclusions regarding the clinical relevance of research results.^[8]

Question 4: What Does “No Significant Difference” ($P \geq 0.05$) Mean?

One of the most common errors in interpreting medical research is to conclude that a nonsignificant result proves that no difference exists between interventions. However, a $P \geq 0.05$ does not demonstrate equivalence or lack of treatment effect. It only indicates that the available data do not provide sufficient evidence to reject the null hypothesis. In statistical hypothesis testing, the null hypothesis is never formally accepted. Instead, researchers conclude that there is insufficient evidence to reject it. Therefore, the statement “there is no difference between treatments” should be used cautiously when interpreting nonsignificant findings.

A nonsignificant result may reflect one of two situations. First, there may truly be no clinically meaningful difference between the groups being compared. Second, a real difference may exist, but the study may not have included enough participants to detect it statistically. This latter situation is known as a Type II error. A Type II error represents a methodological blind spot where a study overlooks a genuine treatment effect, typically because a restricted sample size underpowers the statistical analysis. Statistical power refers to the probability of detecting a clinically meaningful difference when such a difference actually exists. In clinical research, a statistical power of at least 80% is generally considered acceptable. Insufficient sample size is one of the most common causes of low statistical power. This issue is particularly important in studies involving rare diseases, uncommon complications, critically ill patients, or pilot investigations. In these settings, a clinically relevant treatment effect may remain undetected because the study population is too small.^[7]

For example, consider a study evaluating a new neuroprotective intervention after cardiac arrest. Suppose the intervention improves favorable neurological outcomes by 6%, but the study includes an insufficient number of patients. The statistical analysis may produce a $P \geq 0.05$. This result does not prove that the intervention is ineffective. Instead, it suggests that the available evidence is insufficient to establish a statistically significant benefit.

For this reason, clinicians should interpret nonsignificant findings together with information regarding sample size, confidence intervals, and statistical power. Studies reporting nonsignificant results should also be evaluated to determine whether an appropriate sample size calculation and power analysis were performed before patient enrollment.^[8]

In evidence-based medicine, the absence of statistical significance should not automatically be interpreted as the absence of a clinically meaningful effect.

Question 5: What Is p-Hacking?

P-hacking refers to the practice of performing multiple statistical analyses or modifying analytical strategies in an attempt to obtain statistically significant results. This practice can occur intentionally or unintentionally and represents a major threat to the validity and reproducibility of scientific research.

Several factors contribute to p-hacking in medical research. These include “Publish or Perish” pressure, preference for statistically significant findings, and the misconception that a study without significant results has limited scientific value. As a result, researchers may perform additional analyses that were not specified before the study began.

One common form of p-hacking involves conducting multiple subgroup analyses after the primary analysis fails to demonstrate a statistically significant effect. Researchers may divide the study population according to age, sex, disease severity, or other characteristics and perform repeated statistical tests until a significant result is identified. However, statistically significant findings obtained through multiple unplanned analyses may reflect random variation rather than true biological or clinical effects.^[7]

Another form of p-hacking involves repeated examination of the data during patient recruitment. In this situation, investigators repeatedly analyze the results before the planned sample size has been reached and terminate the study when a statistically significant result is observed. This practice increases the probability of obtaining false-positive findings.

Additional examples of p-hacking include selecting cutoff values after reviewing the data, performing multiple alternative statistical analyses, and selectively excluding observations that do not support the desired outcome. Similarly, increasing the sample size beyond the originally planned number of participants solely to achieve statistical significance may lead to the

identification of statistically significant but clinically unimportant differences.^[8]

To reduce the risk of p-hacking, study protocols and statistical analysis plans should be defined before data collection begins. Registration of clinical studies in publicly accessible databases, such as ClinicalTrials.gov, improves transparency and reduces the possibility of selective reporting. In addition, researchers and clinicians should interpret unexpected subgroup findings cautiously, particularly when these analyses were not prespecified in the original study protocol.

Recognition of p-hacking is essential for the critical appraisal of medical literature. Statistical significance alone should not be considered sufficient evidence of a clinically meaningful effect.

Question 6: How Do Multiple Comparisons Flaw Statistical Inference?

In hypothesis testing, the probability of a Type I error, also known as a false-positive result, is typically set at 5% ($\alpha = 0.05$). When a single primary hypothesis is tested, this threshold indicates that there is a 5% probability of incorrectly concluding that a difference exists when no true difference is present. However, when multiple statistical tests are performed within the same study, the probability of obtaining at least one false-positive result increases. This issue is particularly important in studies involving multiple subgroup analyses, secondary outcomes, or exploratory *post hoc* analyses.^[9]

For example, consider a study evaluating a new treatment strategy in patients with acute respiratory failure. Suppose the primary analysis shows no statistically significant difference in mortality between the intervention and control groups ($P = 0.120$). Investigators may subsequently perform additional analyses based on 5 respiratory support subgroups. For instance, they might isolate and analyze the Non-invasive Mechanical Ventilation subgroup and declare that the drug is significantly effective within this specific cohort (e.g. $P = 0.035$). However, such findings often occur by chance because of repeated testing. As the number of statistical comparisons increases, the cumulative probability of obtaining a false-positive result also increases. Therefore, studies involving multiple comparisons should apply statistical methods that adjust for this increased risk.

The Bonferroni correction is one of the most commonly used approaches for controlling Type I error.^[9] This method adjusts the significance threshold by dividing the conventional alpha value (0.05) by the total number of statistical comparisons performed. For the previous

example, if five independent subgroup analyses are conducted, the corrected significance threshold becomes 0.01 (0.05/5). The Bonferroni correction reduces the probability of false-positive findings and improves the reliability of statistical inference. However, because it is a conservative approach, it may also increase the risk of false-negative results, particularly in studies with limited sample sizes.

When interpreting studies that report significant findings from multiple subgroup analyses, clinicians should determine whether appropriate correction methods were applied. Findings derived from unplanned analyses without adjustment for multiple testing should be interpreted with caution.

Question 7: What Is the Future of the *P* Value?

In contemporary medical research, *P* values remain widely used but are increasingly interpreted alongside other statistical measures. Current methodological approaches emphasize that *P* values alone are insufficient for evaluating the clinical importance of study findings. Consequently, greater attention has been directed toward Bayesian methods, effect size measures, and estimation-based statistical reporting.^[2,6]

Bayesian analysis differs from conventional hypothesis testing by incorporating prior knowledge into statistical inference. Instead of focusing solely on whether observed findings are compatible with the null hypothesis, Bayesian methods estimate the probability that a hypothesis is true after considering both prior evidence and newly obtained data. This approach more closely reflects the clinical reasoning process used by physicians in everyday practice.

One practical application of Bayesian reasoning is the assessment of diagnostic tests. In patients with suspected pulmonary embolism, for example, clinicians combine the pretest probability of disease with the diagnostic characteristics of laboratory or imaging tests to estimate the posttest probability.^[10] Tools such as the Fagan nomogram facilitate this process and allow clinicians to make more individualized decisions based on the available evidence.^[10,11] In clinical practice, Bayesian analysis delivers a concrete percentage that is directly applicable to bedside decisions.^[12,13]

Another important concept in modern statistical interpretation is effect size. Unlike the *P* value, which only reflects statistical significance, effect size measures the magnitude of an observed effect and its potential clinical importance. Several effect size measures are commonly used in clinical research.

Standardized mean difference (Cohen's *d*)

This measure is used when comparing continuous variables between independent groups. By general consensus, effect size benchmarks are categorized using Cohen's thresholds, where a value of 0.20 signifies a subtle impact, 0.50 reflects a medium shift, and 0.80 or above indicates a prominent clinical difference.^[14]

Odds ratio and relative risks

These measures quantify associations between an intervention or exposure and a categorical outcome, but they describe different quantities. RR compares event probabilities directly (risk in the intervention/exposed group divided by risk in the control/unexposed group), whereas the OR compares event odds, defined as $p/(1-p)$, between groups. RR is often more intuitive when risks can be estimated directly, such as in randomized trials and cohort studies, whereas OR is commonly used in case-control studies and logistic regression. When an outcome is rare, the OR approximates the RR; when the outcome is common, the OR may appear farther from the null value than the RR and should not be interpreted as if it were an RR. For example, a RR of 0.60 indicates a 40% lower risk of the outcome in one group compared with the other.^[12] Another clinically useful measure is the number needed to treat, which represents the number of patients who must receive an intervention to achieve one additional beneficial outcome.

Correlation coefficients

Correlation coefficients describe the strength and direction of the relationship between two continuous variables. Values closer to zero indicate weaker associations, whereas values approaching -1 or $+1$ indicate stronger relationships. Here, the *P* value alone is not informative. Although cutoff values vary across fields, an absolute coefficient below 0.30 generally indicates a "poor" correlation; 0.30–0.59, "fair;" 0.60–0.79, "moderately strong;" and 0.80 or higher, "very strong;" and 1.00, a "perfect" correlation.^[15]

Area under the receiver operating characteristic curve

This measure evaluates the diagnostic or prognostic performance of the ROC analysis. Values range from 0.50, indicating no discriminatory ability, to 1.00, indicating perfect discrimination. In clinical research, values above 0.80 are generally considered indicative of good diagnostic performance.^[13,16]

Hazard ratio

This measure is used in time-to-event analyses and describes the relative occurrence of an event over time between two groups. For example, a hazard ratio of 0.50 indicates that the event rate is reduced by approximately 50% in one group compared with the other.^[17]

Conclusion

Modern evidence-based medicine requires a comprehensive approach to statistical interpretation. Clinicians should avoid relying exclusively on *P* values and should consider confidence intervals, effect sizes, study design, statistical power, and clinical relevance when interpreting research findings. Integrating these measures provides a more accurate assessment of the strength and applicability of clinical evidence.

Acknowledgments

The authors state that no additional individuals to acknowledge.

Author contribution statement

Conceptualization: A.Ş., Literature Search and Data Curation: A.Ş., Methodology and Critical Analysis: A.Ş., Writing – Original Draft: A.Ş., Writing – Review and Editing: A.Ş.

Conflicts of interest

None Declared.

Funding

None.

References

1. Fisher RA. Statistical Methods for Research Workers. 5th ed. Oliver and Boyd; 1934. p. 45.
2. Wasserstein RL, Lazar NA. The ASA's statement on *P* values: Context, process, and purpose. *Am Stat* 2016;70:129-33.
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "*P*<0.05". *Am Stat* 2019;73 Suppl 1:1-19.
4. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, *et al.* Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol* 2016;31:337-50.
5. Deeks JJ, Higgins JP, Altman DG, McKenzie JE, Veroniki AA, editors. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, *et al.*, editors. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5. Cochrane, 2024. Available from: cochrane.org/handbook. [Last accessed on 2024 Nov].
6. Sullivan GM, Feinn R. Using effect size or why the *P* value is not enough. *J Grad Med Educ* 2012;4:279-82.
7. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, *et al.* Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013;14:365-76.
8. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305-7. [doi: 10.1038/d41586-019-00857-9].
9. Bender R, Lange S. Adjusting for multiple testing-when and how? *J Clin Epidemiol* 2001;54:343-9.
10. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994;271:703-7.

11. Fagan TJ. Letter: Nomogram for Bayes's theorem. N Engl J Med 1975;293:257.
12. Akobeng AK. Understanding measures of treatment effect in clinical trials. Arch Dis Child 2005;90:54-6.
13. Altman DG, Bland JM. Diagnostic tests 3: Receiver operating characteristic plots. BMJ 1994;309:188. [doi: 10.1136/bmj.309.6948.188].
14. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. New York: Lawrence Erlbaum Associates; 1988.
15. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018;18:91-3.
16. Çorbacioğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turk J Emerg Med 2023;23:195-8.
17. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother 2004;48:2787-92.