Revista de Investigación Clínica 78 (2026) 100042

Revista de Investigación Clínica

Clinical and Translational Investigation

journal homepage: www.elsevier.com/locate/ric

Brief Review

The meaning of "*p-value*": Pursuing clinical and statistical significance. A narrative review

Nancy Reynoso-Noverón [a], José F. Carrillo [b], Julissa Luvián-Morales [c], Francisco Javier Ochoa-Carrillo [b], Oscar Arrieta [d], Luis F. Oñate-Ocaña [b,*]

[a] *Unidad de Investigación en Epidemiología, Instituto Nacional de Cancerología (INCan), Mexico*
[b] *Subdirección de Cirugía, INCAN, Mexico*
[c] *Subdirección de Investigación Clínica, Programa MICAELA, INCan, Mexico*
[d] *Dirección General, INCan, Mexico*

ARTICLE INFO

Keywords:
Probability
Statistics
Clinical relevance
Confidence interval

ABSTRACT

The evolution of statistical thinking within the scientific community has developed over nearly three centuries. The concept of "probability," commonly represented by the *p-value*, has been applied in medical research for more than a century. Its integration into hypothesis testing has played an important role in advancing medical science. Nevertheless, the *p-value* has also become one of the most misunderstood and misapplied statistical measures. The common practice of rigidly categorizing the probability value as being above or below the 0.05 threshold has often taken on a dogmatic quality, at times resembling a doctrine rather than a purely scientific criterion. In reporting experimental results and conducting hypothesis testing, references to "statistical significance" are frequently equated with "truth", while findings deemed "not statistically significant" are mistakenly interpreted as "false". This is a conflation and it is incorrect. The objective of this article is to provide a concise review of the history of statistics in medicine, clarify the development of the "probability value" concept, discuss its role in hypothesis testing, and consider its interpretation alongside clinical significance.

For decades, researchers and statisticians have published numerous alarming statements about what the *p-value* is and is not. In recent years, much attention has been paid to this issue. However, the medical community has largely ignored these statements.[1]

These declarations recommend: don't base your conclusions solely on whether an association or effect was found to be "statistically significant"; don't believe that an association or effect exists just because it was statistically significant; don't believe that an association or effect is absent just because it was not statistically significant; don't 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; and don't conclude anything about scientific or practical importance based on statistical significance.[1,2]

This is a huge problem, and for this reason, many researchers developed a call to retire statistical significance from scientific publications and to use confidence intervals, stating that although it will eliminate many bad practices, it could well introduce new ones.[2,3] They suggest eradicating categorization of *p-value* to halt overconfident claims, unwarranted declarations of 'no difference' and absurd statements about 'replication failure' when the results from the original and replication studies are highly compatible.[3]

The objective of this narrative review is to elucidate the fundamental concepts that define the *p-value*, assess its function within hypothesis testing, and clarify its interpretation with respect to effect estimates, precision, and clinical significance. This review does not intend to provide a comprehensive overview of statistical theory; rather, it aims to present a conceptually sound and clinically relevant framework for professionals involved in medicine and health research.

History of statistics in medicine

The advancement of statistics represents a pivotal transformation in scientific research, replacing anecdotal evidence and subjective bias with methods grounded in empirical analysis. This evolution commenced in the seventeenth century, when quantitative methods were introduced to systematically study physical and biological phenomena and assess medical interventions.

* Corresponding author.
E-mail address: lfonate@gmail.com (L.F. Oñate-Ocaña).

https://dx.doi.org/10.1016/j.ric.2026.100042
Received 9 January 2026; Accepted 6 May 2026
Available online 4 June 2026
0034-8376/© 2026 The Author(s). Published by Elsevier España, S.L.U. on behalf of Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubiran. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Christiaan Huygens (1629–1695), was the precursor of the probability concepts necessary to describe the bell-shaped curve, nevertheless Abraham de Moivre (1667–1754), likely studied Huygens' treatise on probability and was the first to formally articulate the symmetrical bell-shaped curve. By introducing the normal curve, de Moivre resolved important challenges that had hampered the work of mathematicians, particularly regarding the average distribution of events. At the time, probability studies mainly addressed questions arising from games involving dice, whose outcomes could be accurately modelled using the binomial distribution to calculate the likelihood of specific results over a fixed number of throws. Subsequently, the de Moivre–Laplace theorem demonstrated that the normal distribution could approximate the binomial distribution.

Expanding the connection between statistics and medicine, Pierre-Simon Laplace (1749–1827), advocated for the comprehensive application of probability theory across all scientific fields, including medicine. He posited that more dependable and precise decision-making in medicine could be achieved through the systematic comparison of outcomes, leading him to advocate strongly for applying probability calculus within medical practice. He asserted that clinical inferences would gain robustness with an increased number of empirical observations, thereby mitigating uncertainty.[4] While Laplace did not employ modern terms such as "sample size" or "detection of small effects" his foundational concepts serve as a historical cornerstone for contemporary medical statistics, directly shaping methodologies such as power analysis and appropriate sample size determination.[5–7]

This assertion generated significant debate within the medical community. Pierre Jean Georges Cabanis (1757–1808), argued that the unique characteristics of each patient necessitate clinical judgment from an experienced physician, as opposed to relying solely on quantitative analysis. He maintained that medical practice involves relating patient-specific features to established knowledge, thereby requiring individualized assessments rather than statistical evaluation.[8] Subsequently, Philippe Pinel (1745–1826), proposed that physicians could assess treatment effectiveness by enumerating favorable outcomes, indicating that probability calculations might elevate medicine to a scientific discipline.[9]

Pierre Charles Alexandre Louis (1787–1872) embraced Laplace's assertion that probability theory could enhance judgment and reduce errors in practice. He implemented systematic observation, meticulous record-keeping, comprehensive analysis across multiple cases, prudent generalization, and validation through autopsy studies. In his research on typhoid fever patients, he noted that survivors tended to be younger than those who succumbed to the disease and found an association between bloodletting and increased mortality.[10,11] Louis is credited as one of the earliest advocates for randomized control studies, initiating significant debates regarding the role of probability theory and statistics within 19th-century medicine.

Friedrich Martius (1850–1923) argued that scientific medicine cannot be established by imposing mathematical methodologies on individuals due to their inherent subjectivity and variability. He asserted that science should be grounded in knowledge acquired through laboratory research rather than solely through data collection and analysis. Similarly, Friedrich Oesterlen (1812–1877) maintained that genuine scientific understanding in medicine must derive from comprehending the causality of phenomena, rather than merely identifying associations or correlations.[12]

Joseph Lister (1827–1912) initiated a major transformation in the scientific understanding of medicine by introducing the principle of antiseptic surgery at the University of Edinburgh.[13] Utilizing statistical methods and frequency analysis, Lister attributed the underlying scientific rationale to germ theory and fundamental immunology concepts developed by Louis Pasteur (1822–1895), particularly Pasteur's research on vaccination.[14]

Auguste Comte (1798–1857), maintained that empirical knowledge alone was insufficient for the advancement of medical science. Claude Bernard (1813–1878) further argued that progress in medical science arises from individual observations generated by experimental physiology and the systematic synthesis of these findings, rather than from statistics.[15]

Bernard's research demonstrated that the identification of physiological mechanisms is essential for comprehending and treating diseases, establishing this as a foundational principle in biology and medicine.[15]

The founders of the Statistical Society of London stated the primary objective of statistics as the accumulation of data. During the 19th century, biologists engaged in extensive data collection realizing the intricate nature of research, leading them to believe that numerous findings could be influenced by multiple causalities.[16] At that time, it was widely accepted that probability theory and empirical data were distinct and should not be combined.

Sir Francis Galton (1822–1911) and Karl Pearson (1857–1936), both members of the Statistical Society of London, were instrumental in shifting perspectives on statistics within the scientific community.[17]

Galton, a half-cousin of Charles Darwin (1809–1882), studied medicine at the University of Cambridge and inspired by Darwin, applying novel statistical concepts to investigating the determinants of heredity.

Pearson, a *protégé* of Galton, further advanced Galton's statistical principles and contributed to the establishment of the English Biometric School focused on inheritance. He also formulated the Pearson's correlation coefficient.[18]

William Sealy Gosset (1876–1937) worked for Guinness brewery in Dublin. In 1904 he wrote an internal report for Guinness titled "The application of the law of error to work of the brewery". Conducting experiments with small sample-sizes he realized the difficulties and asked for help from mathematicians interested in probability theory. In 1906–1907, Guinness sent him to Karl Pearson's laboratory and they work together.

In 1911, Pearson was the first Galton Professor of Eugenics at University College London, a position that, possibly inadvertently, lent pseudoscientific credibility to ideologies later associated with the Nazi regime.[19]

Pearson observed a division within the medical profession regarding the value of statistical approaches. Certain clinicians continued to prioritize the "art" of medicine, believing that statistics contributed little beyond the insights gained from experience. Conversely, other practitioners supported the concept of a "clinical science", utilizing physiological instruments or bacteriological methods for diagnosis and considering statistics as a tool to enhance objectivity in observation.[20]

Neither Galton nor Pearson regarded statistics as constituting "scientific" evidence. Up until that period, probability theory and empirical data were generally considered distinct fields within the scientific community, and the use of probability in making inferences remained a controversial issue.

Statistics were employed primarily for descriptive purposes, with this distinction persisting into the early twentieth century. The paradigm shifted significantly following Pearson's seminal contributions to statistical inference, he is regarded as the founding father of modern statistics due to his development of statistical methodologies and their promotion worldwide, effectively transforming statistics from a descriptive to an inferential science.[17,18]

Pearson advanced a comprehensive statistical framework that facilitated data analysis, pattern recognition, and the investigation of causality. This methodological development was instrumental in hypothesis testing, as it provided a means to quantify the probability that observed results arose from random variation rather than specific underlying factors.[17,18]

Ronald Aylmer Fisher (1890–1962), revitalized Darwinism by integrating Mendelian genetics with natural selection through mathematical analysis, establishing himself as founder of neo-Darwinism. He made significant contributions to modern hypothesis testing, including the development of the Chi-squared test, the formulation of significance testing, and the introduction of *p-value*. Fisher proposed that scientific hypotheses should be evaluated by testing the null hypothesis, which posits no effect or difference, against empirical data.

His approach involves calculating the *p-value* to determine whether results are consistent with the null hypothesis or indicate a statistically significant effect. Fisher asserted that a *p-value* less than 0.05 justifies rejection of the null hypothesis and lends support to the alternative hypothesis. This framework has rendered statistics an essential role for scientific investigation.[20]

In the 1930s, Jerzy Splawa-Neyman (1884–1981), advanced Fisher's concept of the *p-value* by introducing the Confidence interval (CI) into statistical hypothesis testing. Together with Egon Sharpe Pearson (1895–1990), son of Karl Pearson, developed the Neyman–Pearson framework – a systematic statistical method for making optimal decisions between two competing hypotheses: the null and the alternative.[21]

By formulating concepts such as Type I error (false positive) and Type II error (false negative), Neyman and Pearson formalized the process of hypothesis testing. Their contributions emphasized structured statistical decision-making, offering a rigorous method to evaluate conflicting hypotheses. The Neyman–Pearson framework incorporated the *p-value* alongside criteria like test power and significance levels, thereby enhancing the reliability and reproducibility of hypothesis testing.[22]

Globally, the introduction of *p-value* and hypothesis testing have profoundly influenced science and medicine. By the mid-20th century, new statistical methods were widely employed by researchers and clinicians to evaluate theories and assess treatments, prompting ongoing. Physicians and clinicians became divided, again, between those who emphasized the "art" of medicine – drawing upon experience and intuition – and proponents of "clinical science," which prioritizes objective diagnosis through statistics, bacteriological studies, and physiological measurements.[23] Even though the birth of Evidence-based medicine (EBM) proposed that both positions can be reconciled.[23]

The field of EBM emerged in the 1980s at McMaster University in Canada and represents the natural evolution of medical science, moving from a traditional, authority-based approach to a rigorous knowledge system grounded in adequate research question-based designs, and suitable data and statistical analysis. It focuses on integrating the best available evidence from medical research with clinical expertise and patient values to make safer, more informed medical decisions.[24]

Although initially met with resistance, the adoption of *p-values* and hypothesis testing established the foundations for EBM as an essential element of modern healthcare.[25,26]

p-Hacking

Despite their continued use in contemporary biomedical research, hypothesis testing and *p-values* have been increasingly criticized because they are often interpreted through rigid significance thresholds rather than within a broader inferential context. Such dichotomous interpretations distort the scientific record and may lead to misleading conclusions, inefficient research decisions, and unnecessary waste of resources.[1–3,8,27]

One consequence of this threshold-based culture is *p*-hacking, broadly understood as *the selective use of analytic flexibility* to obtain statistically significant results. This may occur when researchers repeatedly explore, modify, or reanalyze data until results cross conventional significance thresholds, rather than interpreting the data according to a prespecified hypothesis and its corresponding analytic strategy. The problem, therefore, is not hypothesis testing itself, but the misuse of inferential tools under selective and nontransparent analytical practices.[27–30]

This issue has been documented across several research fields. Prior work has shown that undisclosed flexibility in data collection, outcome definition, subgroup analysis, covariate adjustment, exclusion of observations, and stopping rules can materially increase false-positive findings and compromise the credibility of published results. In this context, *p*-hacking should be understood less as an isolated technical error than as a symptom of a broader research culture that rewards statistical significance more than methodological transparency. Representative examples of such practices are summarized in Table 1.[28–33]

A more appropriate response is not to abandon inferential statistics, but to strengthen them through pre-specification of analyses, transparent reporting, careful distinction between confirmatory and exploratory analyses, and interpretation of findings considering effect estimates, confidence intervals and their clinical relevance.[34]

Ethical research culture

The *p*-hacking rate in data science provides evidence towards the unintended (or foreseen) side effects of statistical abuse; this reflects on how there is tension between the methodological integrity and publication of large results. The need for an ethical research culture is at the heart of the *p*-hacking controversy.[27,28] Researchers can become aware of best practices by incorporating thorough statistical training and research ethics into curricula. Institutions, in the other hand, must promote an ethical research culture that encourage openness and make unethical behavior – like *p*-hacking – unwanted. Regulators are becoming more conscious of how *p*-hacking affects public trust and policymaking. Reactions involve calling for researchers to reveal all analytical methods and exploratory analyses used in their studies. Additionally, conducting random audits of studies that have been published to ensure adherence to best-practice standards and putting in place reward systems based on the reproducibility and transparency

Table 1
Summary of common *p*-hacking strategies and their potential impact on research validity.

p-Hacking technique	Brief description	Specific clinical research example
Multiple statistical tests without adjustments	Conducting multiple statistical tests without correcting for multiple comparisons.	Evaluating multiple secondary endpoints in a clinical trial without statistical adjustment, reporting only those with $p < 0.05$.
Arbitrary subgroup selection	Performing analyses on subgroups not pre-specified to achieve statistical significance.	In a negative clinical trial, arbitrarily selecting patients under 50 years old where statistical significance appears, and reporting only these results.
Selective data collection	Continuing data collection until a significant p value is achieved.	Recruiting additional participants after an initial non-significant result until $p < 0.05$ is reached.
Unjustified exclusion of data (outliers)	Removing extreme data points without appropriate justification to reach statistical significance.	Excluding patients with extreme results that negatively impact study significance.
Post hoc transformation or recoding of variables	Modifying variables after initial results to achieve statistical significance.	Post hoc changing the cutoff point in diagnostic tests to obtain a significant diagnostic performance.

of research, and not just the novelty or significance of the results. Such practices are bound to become increasingly frequent as the scientific community, along with regulatory bodies, tries to address issues with *p*-hacking.[29] Also, there are remedies available in the form of pre-registration of research protocols, appropriate statistical practices, open data, and regulatory interventions that can counter *p*-hacking and provide transparency and reproducibility of research. Data science must prioritize integrity over statistical significance and transparency over selective reporting to ensure that the results of our analysis add up to scientific advance.[30] Scholars, practitioners, and regulators must preserve such standards to make valid contribution to scientific progress.[31]

p-Value in the sight of editors of medical journals

In recent years, statistical consultants' and medical journal editors' initiatives to reform statistical practices have gained momentum. They have become more attuned to the challenges posed by the overuse and misinterpretation of significance testing, the meaning of *p-values* in scientific research, and *p*-hacking practices. These initiatives have gained traction, with many journals emphasizing the importance of reporting absolute effect sizes, Confidence intervals (CI), and complementary metrics alongside *p-values*, or even reduce or delete unnecessary *p-values*. This nuanced approach recognizes the inherent strengths and weaknesses of *p-values*, as underscored in recent publications by the American Statistical Association (ASA).[1,35]

Prestigious journal editors suggest replacing *p-values* with the appropriate point estimates of effect (absolute risk reduction, mean differences, odds ratios, risk ratios, as corresponds to methodology) along with their 95% CI, especially when multiple hypotheses are used.[36–38]

It is important to define that point estimates and their corresponding CI are indispensable to be able to interpret the *clinical significance* of the study findings, while their corresponding *p-value* is indispensable to interpret the *statistical significance* of the study findings.[39]

The gold standard for intervention studies in clinical research is a randomized controlled trial. High-quality randomized controlled trials require precise design, conduct, methods, analysis, and reporting of results. To properly interpret a randomized clinical trial, readers need complete and transparent information, from conception to publication. Unfortunately, evidence indicates that most trials lack this thoroughness, leading to biases that complicate the interpretation of the results. To address this, the updated Consort Statement 2025 aims, as many years ago, to improve the quality of randomized clinical trials and other experimental and observational designs.[40] Readers should not have to infer what was probably done; they should be told explicitly.[41]

The problem of *p-value* correction

If a null hypothesis test is introduced with a significance level set at 0.05, this implies that there is a 5% probability of committing a Type I error. This means, 5% probability of rejecting the null hypothesis and concluding that a difference or association exists when in fact it does not. This false-positive result can have serious implications in clinical research, particularly in studies comparing diagnostic protocols or intervention strategies.[5,6] This scenario is common in numerous studies where a single, specific null hypothesis is put forward. Nonetheless, researchers often design studies that incorporate two or more null hypotheses. Even in exploratory research, multiple "general" hypotheses may be considered. In these cases, it is essential to calculate a single Type I error measure that includes the group of hypotheses posed.

For example, a randomized clinical trial (RCT) has two outcome measures of equivalent interest, when there are three parallel treatment groups, or when there are more than two independent risk factors associated with the outcome. In this type of design, we must pose a family of hypotheses, and in the case of a family of k independent null hypotheses that we test with a significance level of α, the probability that at least one of them is false positive is $1-(1-\alpha)^k$. The Bonferroni correction works by dividing the original significance level (α) by the number of comparisons being made and then using this adjusted α level as the new significance threshold. Simply described, the adjusted α value $= \alpha/k$; if four null hypotheses are posed, $\alpha_{\text{Bonferroni}} = 0.05/4 = 0.0125$.[42]

The number of pairwise comparisons can be calculated as $[k(k-1)/2]$, where k is the number of study groups. The Bonferroni correction then adjusts the significance level (α) by dividing it by this number of comparisons, yielding a more stringent threshold for statistical significance.

The Bonferroni correction method is probably the most widely used method in the medical literature; however, it is highly controversial.

The main criticism of Type I error correction methods is that they always have an impact, decreasing statistical power and increasing Type II errors (β). In the design phase of the research project, it should be considered that, if we wish to keep the probability of false negatives in hypothesis testing (Type II errors) low, we must consider increasing the sample size as we proceed for Type I error correction. This is often an economic and logistic disadvantage that must be considered in the design phase of clinical trials, and different factors to consider are described in Table 2.[43–47]

The relationship between sample size and "*p-value*"

The *p-value* is essentially influenced by the sample size of the data under examination; bigger samples will produce smaller *p-values*. These can form statistical evidence against the null hypothesis in contrast to substantive differences among groups.[47]

In general, the sample size for a research project proposal is estimated based on the estimated size of the main effect, proposal of Type I error (usually 5%), and statistical power (usually 80%).[5] Furthermore, an increase in sample size enhances the accuracy of statistical tests, thereby providing stronger evidence against or in favor of the null hypothesis.[5,6,47,48]

Table 2
Factors to consider when selecting the best method for adjusting the level of statistical significance in the case of multiple comparisons.

Confirmatory study	Exploratory study
Pre-specified hypothesis	Pre-specified hypothesis, or multiple "general" hypotheses
One hypothesis	Observational studies
Randomized control studies	Randomized control studies
Laboratory experiments	Laboratory experiments
A protocol-specified statistical analysis for one hypothesis	Sometimes, large number of hypotheses produced
If multiple tests are required: correction of Type I error is necessary	If multiple tests are required: correction of Type I error is necessary
Consider the Type II error	Type II error increases
Increase sample size	Increase sample size
Interpretation easier	Interpretation challenging

The concept of *clinical significance* often requires looking beyond statistical results and considering the tangible impact of the findings on patient care and outcomes. While *statistical significance*, demonstrated through hypothesis tests and the *p-value*, provides insight into the reliability of results, *clinical significance* addresses whether the magnitude of the observed change is meaningful in real-world applications.[27,28] These differing perspectives create a nuanced landscape in which the importance of effect size, context, and patient health outcomes must be balanced thoughtfully.[29,30] Nevertheless, results that are statistically significant do not always lead to conclusions that are biologically relevant. Even minor effects can achieve statistical significance when the sample size is large, or the variability is minimal. An example of how sample size impacts *p-value* is shown as Appendix A.

Confidence intervals: clinical significance and statistical significance

The mean of a dataset is calculated by summing all values and dividing by the number of observations. As summary measure, it represents the central tendency for normally distributed data. Within the framework of hypothesis testing, the question often shifts from simply determining whether a new treatment is beneficial compared to an existing one, to quantifying the extent of its benefit. This can be addressed with point estimates, such as risk ratios, odds ratios, and differences in means, differences in proportions, which provide specific numerical measurements for comparative analysis.[48,49] Furthermore, it is essential to report the precision of the point estimate, typically conveyed through the CI, which reflects the probability that the population value lies within a specified range. The 95% CI – which has endpoints referred to as the 95% confidence limits – is most used, with a 95% likelihood corresponding to a *p-value* of 0.05.[49,50] Thus, the CI formally quantifies the uncertainty associated with the point estimate. Occasionally, 90% or 99% CIs are utilized when greater permissiveness or stringency is required.[51]

In medical research, a fundamental challenge is distinguishing between *clinical significance* and *statistical significance*. While statistical methods such as CIs and hypothesis testing offer a structured approach to data analysis, they may not fully address nuanced consequences for clinical practice. A *statistically significant* finding can indicate a credible association, but its *clinical significance* often depends on factors like patient demographics, healthcare environments, risk–benefit, quality of life, and cost-effectiveness assessments.[51] Confidence intervals present a range of plausible effect sizes rather than relying solely on *"p-values"*, enhancing the interpretability of outcomes and allowing practitioners to make evidence-based decisions that prioritize patient care. The integrity of these conclusions is determined by robust study designs, adequate sample sizes, and transparent reporting practices. Therefore, *statistical significance* should be viewed not as a final outcome, but as one component within a continuum that synthesizes quantitative precision with qualitative understanding.[6,50,51] Two representative e examples of the literature are described in Appendix B.

There are significant efforts underway to promote new strategies to advance knowledge of the debate surrounding *p-values*, including their major advantages and their poorly understood drawbacks, as well as the development of new statistical procedures to complement *p-value*. The idea of second-generation *p-values* has recently gained popularity, aiming to address criticisms of the inappropriate use of *p-values*. They consider scientific relevance by defining a priori a null hypothesis that includes null point values and those with no clinical relevance. This second-generation approach significantly reduces Type I error.[52]

Newer approaches to the *p-value* problem include methods to complement or modify the *p-value*, such as the Bayesian factor, the risk of false positives, the probability of rejection and credibility analysis, the credibility index, and Gannon's modification of the *p-value* based on the Bayesian factor. However, the medical community is extremely accustomed to and has adopted the *p-value* as fundamental. The simplest alternative to abandoning the *p-value* is to do away with its fixed thresholds and adopt the actual *p-value*.[53]

Overall, the goal is to promote transparency, rigor, and scientific reproducibility.

Conclusion

The events outlined illustrate how *p-values* in hypothesis testing were developed and show how statistical theory has evolved from descriptive analysis to inferential reasoning. The methodological framework that underpins contemporary science and clinical research has benefited greatly from the gradual contributions of pioneers.

Although *p-value* is still essential to hypothesis testing, its use is constantly changing as researchers try to strike a balance between rigor and useful information.

Artificial intelligence is considered an advanced, computational extension of inferential statistics, shifting from purely analyzing sample data to automating predictions and generalizations about vast datasets or populations. While classical inferential statistics relies on testing hypotheses via specific formulas (like t-test) to infer population characteristics, AI (particularly machine learning) automates this process to recognize complex, high-dimensional patterns and forecast future outcomes.

To ensure that advancements in statistics (including AI) lead to improved diagnostic and therapeutic accuracy, as well as greater equity in healthcare delivery, it is essential to enhance statistical education in medicine, promote diverse population representation in data collection, rigorously validate algorithms, institute robust regulatory frameworks, and facilitate equitable access to technology. Without these foundational measures, issues such as bias, misrepresentation, and inequality may undermine the benefits that new technologies offer within the medical field.

Ethical considerations

None declared.

Funding

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

Declaration of competing interests

The authors report there are no competing interests to declare.

Acknowledgment

Many thanks to Cittim B. Palomares-Palomares for his English-Language editorial review.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version available at https://doi.org/10.1016/j.ric.2026.100042.

References

1. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". *Am Stat.* 2019;73(51):1–19. http://doi.org/10.1080/00031305.2019.1583913.
2. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol.* 2008;45 (3):135–140. http://doi.org/10.1053/j.seminhematol.2008.04.003.
3. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature.* 2019;567(7748):305–307. http://doi.org/10.1038/d41586-019-00857-9.
4. Tröhler U. Probabilistic Thinking and the Evaluation of Therapies, 1700–1900. James Lind Library; 2020. Available from: https://www.jameslindlibrary.org/articles/probabilistic-thinking-and-the-evaluation-of-therapies-1700-1900/. Accessed May 29, 2026.
5. Chow SC, Shao J, Wang H, et al. Sample size calculations in clinical research.*Chapman & Hall/CRC Biostatistics*. third edition. CRC Press, Taylor & Francis Group; 2020.
6. 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(4):337–350. http://doi.org/10.1007/s10654-016-0149-3.
7. Todhunter I. *A History of the Mathematical Theory of Probability*. London: Macmillan & Co; 1865. http://doi.org/10.1017/CBO9781139923576.
8. Matthews JR. *Quantification and the Quest for Medical Certainty*. Princeton, New Jersey: Princeton University Press; 1995.
9. Pinel P. *Traité médico-philosophique sur l'aliénation mentale ou la manie. Paris*. second edition. 1809.
10. Louis PCA. *Anatomical Pathological and Therapeutic Researches Upon the Disease Known Under the Name of Gastro-enterite putrid Adynamic Ataxic or Typhoid Fever etc Compared With the Most Common Acute Diseases.* vols. 1 and 2. Boston: Translation Bowditch & Butts; 1836.
11. Morabia A. Pierre-Charles-Alexandre Louis and the evaluation of bloodletting. *J R Soc Med.* 2006;99(3):158–160. http://doi.org/10.1177/014107680609900322.
12. Oesterlen F. *Handbuch der Heilmittellehre*. Tübingen, Germany: Tübingen und Laupp; 1856.
13. Lister J. On the antiseptic principle in the practice of surgery. *Br Med J.* 1867;2 (351):246–248. http://doi.org/10.1136/bmj.2.351.246.
14. Pasteur L. Summary report of the experiments conducted at Pouilly-le-Fort, near Melun, on the Anthrax vaccination. *C R Acad Sci.* 1881;92:1378–1383.
15. Bernard C. *Introduction à l'étude de la médecine expérimentale*. Paris: J.B. Baillière et fils; 1865.
16. Willcox WF. Note on the chronology of statistical societies. *J Am Stat Assoc.* 1934;29 (188):418–420. http://doi.org/10.2307/2278614.
17. Pearson K. *The Grammar of Science*. third edition. New York: Macmillan; 1911.
18. Pearson ES. *Karl Pearson*. London: Cambridge University Press; 1938.
19. University College London. UCL makes formal public apology for its history and legacy of eugenics. UCL News 2021. https://www.ucl.ac.uk/news. [Retrieved 29 May 2 2025].
20. Fisher RA. *Statistical Methods for Research Workers*. first edition. Edinburgh, UK: Oliver & Boyd; 1925.
21. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond Ser A.* 1933;231:289–337. http://doi.org/10.1098/rsta.1933.0009.
22. Lehmann EL. The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc.* 1993;88(424):1242–1249. http://doi.org/10.1080/01621459.1993.10476404.
23. Simons M, Rapport F, Zurynski Y, et al. Links between evidence-based medicine and shared decision-making in courses for doctors in training: a scoping review. *BMJ Open.* 2022;12(4). http://doi.org/10.1136/bmjopen-2021-057335e057335.
24. Sackett D. How to read clinical journals: I. Why to read them and how to start reading them critically. *CMAJ.* 1981;124(5):555–558.
25. Institute of Medicine. *Evidence-based Medicine and the Changing Nature of Healthcare: 2007 IOM Annual Meeting Summary. Institute of Medicine*. Washington, DC: National Academies Press; 2008.
26. Subbiah V. The next generation of evidence-based medicine. *Nat Med.* 2023;29(1):49–58. http://doi.org/10.1038/s41591-022-02160-z.
27. Head ML, Holman L, Lanfear R, et al. The extent and consequences of p-hacking in science. *PLoS Biol.* 2015;13(3). http://doi.org/10.1371/journal.pbio.1002106e1002106.
28. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci.* 2011;22(11):1359–1366. http://doi.org/10.1177/0956797611417632.
29. Wigboldus DHJ, Dotsch R. Encourage playing with data and discourage questionable reporting practices. *Psychometrika.* 2016;81(1):27–32. http://doi.org/10.1007/s11336-015-9445-1.
30. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front Psychol.* 2016;7:1832. http://doi.org/10.3389/fpsyg.2016.01832.
31. Stefan AM, Schönbrodt FD. Big little lies: a compendium and simulation of p-hacking strategies. *R Soc Open Sci.* 2023;10(2). http://doi.org/10.1098/rsos.220346220346.
32. Teixeira PM. P-hacking: a call for ethics. *Pulmonology.* 2018;24(3):207–208. http://doi.org/10.1016/j.pulmoe.2018.03.001.
33. Raj AT, Salameh Z, Patil S, Sarode S. P-hacking: a wake-up call for the scientific community. *Sci Eng Ethics.* 2018;24(6):1813–1814. http://doi.org/10.1007/s11948-017-9984-1.
34. Ioannidis JPA. Retiring statistical significance would give bias a free pass. *Nature.* 2019;567(7749):461. http://doi.org/10.1038/d41586-019-00969-2.
35. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat.* 2016;70(2):129–133. http://doi.org/10.1080/00031305.2016.1154108.
36. Harrington D, D'Agostino RB, Gatsonis C, et al. New guidelines for statistical reporting in the journal. *N Engl J Med.* 2019;381(3):285–286. http://doi.org/10.1056/NEJMe1906559.
37. Moher D, Collins G, Hoffmann T, Glasziou P, Ravaud P, Bian ZX. Reporting on data sharing: executive position of the EQUATOR Network. *BMJ.* 2024;386:. http://doi.org/10.1136/bmj-2024-079694e079694.
38. Mansournia MA, Nazemipour M. Recommendations for accurate reporting in medical research statistics. *Lancet.* 2024;403(10427):611–612. http://doi.org/10.1016/S0140-6736(24)00139-9.
39. Morgan PP. Confidence intervals: from statistical significance to clinical significance. *CMAJ.* 1989;141(9):881–883.
40. Hopewell S, Chan AW, Collins GS. CONSORT 2025 statement: updated guideline for reporting randomised trials. *Lancet.* 2025;405(10489):1622–1640. http://doi.org/10.1016/S0140-6736(25)00672-5.
41. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ.* 1996;313(7057):570–571. http://doi.org/10.1136/bmj.313.7057.570.
42. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubbl. D. R. Ist. Super. Di Sci. Econom.. *E Commerciali di Firenze.* 1936;8:1–62. Zbl 0016.41103.
43. Ranstam J. Multiple p-values and Bonferroni correction. *Osteoarthritis Cartilage.* 2016;24(5):763–764. http://doi.org/10.1016/j.joca.2016.01.008.
44. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? *Korean J Anesthesiol.* 2018;71(5):353–360. http://doi.org/10.4097/kja.d.18.00242.
45. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol.* 2014;67 (8):850–857. http://doi.org/10.1016/j.jclinepi.2014.03.012.
46. Gómez-de-Mariscal E, Guerrero V, Alexandra Sneider A, et al. Use of the p-values as a size-dependent function to address practical differences when analyzing large datasets. *Sci Rep.* 2021;11(1). http://doi.org/10.1038/s41598-021-00199-520942.
47. Altman N, Krzywinski M. P values and the search for significance. *Nat Methods.* 2017;14(1):3–4. http://doi.org/10.1038/nmeth.4120.
48. Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: what do p values and confidence intervals really represent? *Anesth Analg.* 2018;126(3):1068–1072. http://doi.org/10.1213/ANE.0000000000002798.
49. Armitage P, Berry G, Matthewe JNS. *Statistical Methods for Medical Research*. fourth edition. Oxford: Blackwell Science Ltd.; 2002.
50. Braitman LF. Confidence intervals assess both clinical significance and statistical significance. *Ann Intern Med.* 1991;114(6):515–517. http://doi.org/10.7326/0003-4819-114-6-515.
51. Hazra A. Using the confidence interval confidently. *J Thorac Dis.* 2017;9(10):4125–4130. http://doi.org/10.21037/jtd.2017.09.14.
52. Blume JD, Greevy RA, Welty VF, et al. An introduction to second-generation p-values. *Am Stat.* 2019;73:157–167. http://doi.org/10.1080/00031305.2019.1537893.
53. Quatto P, Ripamonti E, Marasini D. Beyond p < .05: a critical review of new Bayesian proposals for assessing the p-value. *J Biopharm Stat.* 2022;32(2):308–329. http://doi.org/10.1080/10543406.2021.2009497.