Trends
in Psychiatry and Psychotherapy

JOURNAL ARTICLE PRE-PROOF
(as accepted)

Trends

Medicine, artificial intelligence and uncertainty: Why is statistical thinking fundamental?

Claudio Córdova, Otavio Nóbrega

http://doi.org/10.47626/2237-6089-2025-1069

Original submitted Date: 19-Mar-2025

Accepted Date: 15-Jun-2025

Medicine, artificial intelligence and uncertainty: Why is statistical thinking fundamental?

Short title: Advancing Medical Education for AI Use

Claudio Córdova[1], Otavio Nóbrega[1]

1. Universidade de Brasília, Brasília, DF, Brazil.

Corresponding Author: Dr. Claudio Córdova
claudioucb@yahoo.com.br

Abstract

The medical field has historically resisted uncertainty, often delaying the integration of new scientific evidence into clinical practice—sometimes by nearly two decades. This inertia reflects deep-rooted cultural and epistemological barriers that also impede the adoption of innovations such as Artificial Intelligence (AI). Yet, the call for more rigorous decision-making in medicine is not new. In the 18th century, Pierre-Simon Laplace emphasized the value of probability theory in clinical reasoning, a view later echoed by William Osler, who famously described medicine as “the science of uncertainty and the art of probability.” These early insights gained traction through the work of Sir Austin Bradford Hill and Archibald Cochrane, whose contributions laid the groundwork for Evidence-Based Practice (EBP). In the 1990s, Gordon Guyatt formally introduced Evidence-Based Medicine (EBM), advocating for clinical decisions grounded in empirical data, professional expertise, and patient values. In this evolving landscape, basic statistical literacy is no longer sufficient. In this context, cultivating probabilistic reasoning and statistical thinking has become essential to support ethically sound and evidence-aligned decisions to guide a meaningful transformation in both clinical training and practice.

Keywords: Evidence-Based Medicine, Artificial Intelligence, Medical Education, Clinical Decision-Making, Probabilistic Reasoning.

Imagine a future where digital assistants analyze data in real time to support clinical decisions, and intelligent robots roam hospitals, optimizing workflows and improving care. Artificial Intelligence (AI) is no longer a concept confined to science fiction, but an advancing reality with the potential to profoundly transform the practice of medicine. However, despite this promising outlook, significant cultural and operational barriers still impede the full incorporation of these technologies into clinical practice. For example, it is estimated that it takes almost two decades for new scientific evidence to be integrated into medical routines [1]. This time lag largely reflects medicine's historical resistance to recognizing and managing uncertainties—whether diagnostic, prognostic, or therapeutic—which affects not only clinical decision-making but also the doctor–patient relationship and interprofessional collaboration. Discomfort with ambiguity can lead to an overreliance on apparent certainties, reinforcing a dichotomous view of reality that reduces complex clinical scenarios to simplistic black-and-white terms, neglecting the nuanced spectrum of gray.

However, cultural resistance to accepting uncertainty largely reflects the human brain's limited ability to engage in explicit probabilistic reasoning. Meanwhile, biological reality is inherently complex and characterized by high interindividual variability, especially in the early stages of diseases. In this context, probabilistic models [$Y \sim p(y|\theta)$]* which incorporate and quantify uncertainty become essential tools, contrasting with the deterministic view [$y = f(x)$]†, which assumes fixed and fully predictable relationships between variables. Aligning probabilistic approaches at both the clinical and statistical levels is a decisive step towards driving the evolution of health technologies and improving responsible information communication. This alignment strengthens the foundation of clinical decisions and contributes to better informed public health interventions adjusted to the complexities of the real world.

This cultural resistance to uncertainty in medical practice has deep historical roots. As early as the 18th century, Pierre-Simon Laplace and other

* Read the expression $Y \sim p(y|\theta)$ as "Real data, represented by Y, generated from a probability distribution $p(y|\theta)$, whose shape depends on the unknown parameter θ.
† Note that for each value of x, exactly one corresponding value of y is generated.

pioneers recognized the role of probability in improving conjectural sciences—a concept that would gradually influence medicine [2]. However, this perspective faced strong resistance and was slowly adopted in clinical settings. In the late 19th and early 20th centuries, William Osler emphasized that medicine is “the science of uncertainties and the art of probabilities” [3], but it took decades for this view to inspire tangible changes in clinical routines. It was only throughout the 20th century, with the contributions of researchers such as Sir Austin Bradford Hill [4] and Archibald Cochrane [5], that probability and statistical methods were more systematically integrated, laying the foundation for evidence-based practice (EBP).

The formalization of Evidence-Based Medicine (EBM) occurred in the 1990s, with Gordon Guyatt’s seminal article “Evidence-Based Medicine” (1991) [6], in which he highlighted the need for clinical decisions to be grounded in robust scientific evidence, integrating clinical expertise, empirical data, and patient values. Since then, EBM has become a cornerstone of modern medical practice, fostering more informed, personalized, and effective care.

This integration between scientific knowledge and clinical decision-making has continued to evolve, propelled by the emergence of new technologies**.** Artificial Intelligence (AI), particularly through Machine Learning and Deep Learning techniques, enables advanced analytical tools to uncover the mechanisms underlying health and disease. If traditional statistical methods helped incorporate uncertainty into diagnosis and prognosis, AI and big data analytics now offer opportunities to refine these uncertainties, making clinical decisions faster and more precise. However, as medicine’s historical trajectory reminds us, technological advancements must be accompanied by rigorous critical appraisal and a strong commitment to ethical and scientific standards to ensure that innovation genuinely enhances care quality and safety.

In this context, bridging the gap between theory and practice remains a central challenge—especially in training health professionals. Although EBM is widely discussed in academic curricula, institutional programs, and scientific forums, a significant gap persists between these principles and the realities of clinical care. This disconnection undermines the quality of care and underscores the urgency of translating scientific evidence into practical settings. To that end,

strategic investments in education and continuous professional development are essential.

However, the traditional approach to statistical literacy, often limited to interpreting numerical data and using analytical tools, has proven insufficient in preparing clinicians for evidence-based decision-making. The teaching of statistical concepts—such as measures of central tendency, confidence intervals, and p-values—is frequently included in medical curricula. **However, their interpretation and application in clinical practice are often overlooked, especially in institutions where statistics is not contextually integrated into medical decision-making.** Thus, it is essential to foster probabilistic reasoning and statistical thinking so that decisions are truly guided by evidence. These skills are fundamental to critically integrating scientific findings into clinical reasoning, while also honoring the biological complexity and individual variability of each patient.

In the management of psychiatric disorders, for instance, the decision to prescribe an antidepressant can be based on evidence that models treatment effectiveness as a conditional probability—that is, the likelihood of a patient exhibiting a significant therapeutic response given factors such as family history, psychiatric comorbidities, and genetic biomarkers. This probability, expressed as P (therapeutic response | patient profile), allows for the integration of multiple factors influencing treatment efficacy. Patients with major depressive disorder and a history of treatment resistance, for example, may exhibit differential responses to specific selective serotonin reuptake inhibitors (SSRIs), particularly when comorbidities such as generalized anxiety disorder or undiagnosed bipolar disorder are present. In this context, Bayes' theorem enables the continuous updating of these estimates as new clinical information and biomarkers become available, progressively refining therapeutic decision-making. Formally, this relationship can be expressed as:

$$\mathrm{P}\,(\mathit{therapeutic\ response} \mid \mathit{patient\ profile}) =$$

$$\frac{P(patient\ profile|\ therapeutic\ response\ )\ \times\ P(therapeutic\ response)}{P(patient\ profile)}$$

Complementary to the Bayesian approach, the Likelihood Ratio (LR+/-) plays a central role in individualizing clinical reasoning. In diagnostic testing, it enables the pre-test probability of a patient having a disease to be updated by incorporating the test's sensitivity and specificity, yielding a more accurate post-test probability. This allows new laboratory or imaging findings to be quantitatively integrated into clinical decisions, enhancing diagnostic precision. The approach illustrates how probabilistic reasoning moves beyond population-level inferences, tailoring decisions to the specific features of each patient.

On the other hand, advanced statistical thinking is essential for the critical appraisal of scientific evidence and its application to evidence-based practice (EBP) (Figure 1). Consider, for instance, a hypothetical study reporting that a new antidepressant is more effective than standard treatment for major depressive disorder, with a statistically significant result of P = 0.01. At first glance, this might be interpreted as evidence against the null hypothesis (H_0), suggesting the drug's efficacy. However, such interpretation requires caution, as a P-value < α does not provide information about the magnitude of the effect or confirm its clinical relevance.

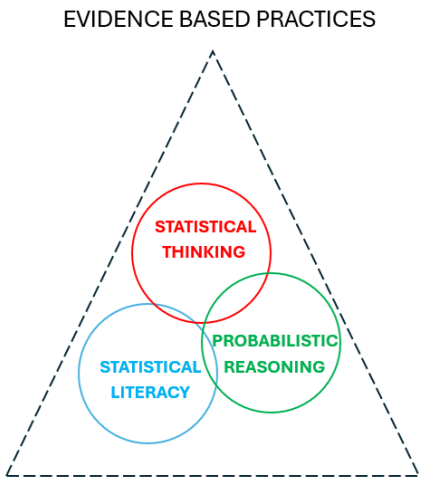

Figure 1. Diagram of the hierarchy of essential skills for evidence-based decision making. The three competence levels are interdependent and complement each other in clinical decision making.

In this sense, a more robust analysis requires formulating questions that guide the critical interpretation of results, such as: What is the magnitude of the observed effect, and is it clinically relevant? Do the confidence interval limits include effects of clinical importance? Are the findings consistent with previous studies? When assessing individual benefits, should absolute or relative measures be prioritized? Are the results applicable to patients with comorbidities or specific genetic characteristics, or are there limitations to their generalizability? These reflections are essential to ensure that clinical decisions are based on a contextualized assessment, respecting biological complexity and individual variability.

In addition to the critical appraisal of statistical outcomes, a key component of statistical thinking is evaluating the reliability of scientific evidence. This is particularly important in light of the high proportion of non-replicable findings in the literature, often due to methodological flaws or publication biases. In his influential article “Why Most Published Research Findings Are False” [7], John Ioannidis demonstrated that a significant proportion of studies in the medical field report irreproducible results, influenced by factors such as small sample sizes, publication bias, and indiscriminate use of multiple statistical comparisons. Moreover, questionable practices like p-hacking (repeated testing until statistical significance is achieved) and HARKing (Hypothesizing After the Results Are Known) compromise scientific credibility and hinder the translation of findings into clinical settings. These methodological issues not only weaken the evidentiary foundation of research but also affect the reliability of AI – based predictive models, which—despite their computational power—require rigorous validation to ensure clinical relevance.

The application of AI and predictive algorithms in healthcare raises technical and ethical challenges, particularly regarding the quality, integrity, and representativeness of the training data. Machine Learning models in biomedical contexts, including multi-omics approaches (integrating genomic, transcriptomic, proteomic, and metabolomic data) [8], have great potential to uncover complex biological patterns and improve diagnostic accuracy. However, poor data representativeness can compromise model generalizability, reducing

performance when applied to populations or scenarios that differ from the training set.

Beyond data representativeness, another major challenge in applying AI to medicine involves discrepancies between training data and real-world clinical data. In this context, two critical issues are **covariate shift and batch effects. Covariate shift** occurs when the distribution of predictive variables in the training dataset differs from that in real-world clinical data, impairing model accuracy. This phenomenon is particularly relevant in clinical studies, where population, environmental, and temporal variations may alter patient characteristics and reduce prediction reliability. Such discrepancies compromise a model's ability to generalize across diverse clinical contexts, limiting practical applicability **[9].** In addition **to covariate shift, batch effects** are a common issue in biomedical modeling. These non-biological variations stem from methodological inconsistencies between sample batches and can distort algorithmic pattern recognition. **Batch effects** are especially problematic in genetic and biomedical studies, where subtle differences in lab conditions may generate significant analytical discrepancies **[10].**

To mitigate these limitations, strategies such as statistical recalibration, data normalization, and tailored approaches for different types of omics data should be adopted [11]. These approaches help models adapt to real-world conditions, improving their robustness and clinical applicability. Additionally, it is critical to account for overfitting and data quality, as both directly influence model generalizability. **The use of flawed datasets or overtrained models can result in overfitting—a common issue in both Machine Learning and Deep Learning.** Overfitting occurs when a model memorizes spurious patterns from the training data, failing to generalize to new, unseen cases. Consequently, such models may perform well within the training context but break down in real clinical scenarios, producing unstable predictions and undermining their practical use. These impacts have been documented in multiple real-world applications.

An emblematic example of these challenges was highlighted in a recent systematic review [12], which emphasized the lack of representativeness in datasets used to train machine learning algorithms for healthcare prediction.

While these algorithms are effective at identifying latent patterns within available data, their clinical applicability is jeopardized when training data are biased or insufficiently diverse. This limitation is especially critical in primary and secondary prevention settings, where data collection biases—such as selection bias (when the sample does not reflect the target population) and spectrum bias (when cases do not reflect the actual clinical diversity)—can perpetuate inequalities in diagnosis and treatment.

Indeed, as demonstrated by Obermeyer et al. [13], a widely used AI algorithm in the U.S. healthcare system exhibited systemic racial bias when healthcare spending was used as a proxy for disease severity. This choice reduced the number of Black patients identified as needing additional care by over 50%, underscoring how flaws in data curation and processing can compromise equity and fairness in clinical decision-making. This case reinforces the importance of critically designing and validating predictive algorithms to ensure that their outputs reflect the complexity and diversity of the populations they aim to serve.

Therefore, moving forward requires ongoing training of physicians and healthcare professionals to rigorously and thoughtfully interpret the many variables involved in medical care. This effort goes beyond mastering the best available evidence; it entails embedding probabilistic reasoning and statistical thinking into clinical routines. Moreover, the adoption of new technologies must be inclusive, transparent, and responsible—not only enhancing diagnostic, prognostic, and therapeutic precision but also reducing health disparities and promoting more equitable access to care.

In this evolving landscape, balancing the transformative potential of technological innovations with the ethical, scientific, and humanistic principles that guide medical practice is essential. The growing role of AI in healthcare must be accompanied by strategic measures that ensure its ethical and equitable implementation, avoiding adverse effects that could undermine its effectiveness and acceptance.

One of the most pressing challenges is **data governance, particularly regarding the collection, storage, and use of medical information to train**

predictive models. The quality and representativeness of these data directly impact algorithmic accuracy, making it crucial to establish rigorous protocols that guarantee accessibility without compromising security and privacy. Regulatory frameworks such as the GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act) must be observed to protect sensitive information and ensure that clinical data are used ethically and transparently.

Beyond privacy concerns, legal accountability in AI use must be clearly defined. Currently, there is uncertainty about who should be held responsible for algorithm-based clinical decisions: the physician, the technology developer, or the healthcare institution? Developing regulatory standards that define responsibility and provide oversight is essential to mitigate risks and ensure patient safety. This issue becomes increasingly critical as intelligent systems influence diagnoses and prognoses, requiring continuous monitoring to prevent errors and distortions.

Another key aspect is the management of uncertainty in statistical and predictive models. Although AI offers powerful analytical capabilities, its outputs must be interpreted with caution, as data biases may compromise model applicability in real-world settings. The adoption of probabilistic methodologies—such as Bayesian approaches and Likelihood Ratios—enables ongoing refinement of estimates as new data are incorporated, reducing oversimplified interpretations and reinforcing the link between statistical analysis and medical decision-making.

In addition, ensuring equitable access to and application of AI should be a priority in implementing these technologies in healthcare. Models trained on biased datasets may perpetuate disparities, as demonstrated by Obermeyer et al. [13], who identified critical flaws in triage algorithms based on healthcare expenditures that led to the underdiagnosis of Black patients. Addressing these biases requires **continuous audits and a commitment to diversity in clinical datasets**, ensuring that technological advances benefit the population fairly and effectively.

Given these challenges, ongoing training of healthcare professionals is essential to strengthen their critical interpretation of statistical tools and AI models. In addition to improving diagnostic and prognostic accuracy, the adoption of these new technologies must contribute to reducing inequalities and building a more just and accessible healthcare system. In doing so, balancing innovation with accountability will allow AI to reshape medical practice without compromising its fundamental values.

Conclusion and Final Recommendations

The advancement of medicine requires a careful balance between technological innovation and ethical, scientific, and humanistic principles. To ensure that AI contributes meaningfully to clinical practice, it is essential to establish clear guidelines that promote its responsible, transparent, and equitable implementation. Key strategies include:

1. **Curricular reform: Integrating statistical and probabilistic thinking into medical education is crucial to empower professionals to critically interpret clinical data and evidence. This foundation enables physicians to assess predictive models and determine their applicability in real-world practice.**
2. **Rigorous AI validation: Standardized protocols must ensure data representativeness and minimize bias in predictive models, thereby preventing automated decisions from perpetuating inequalities or compromising patient safety.**
3. **Transparency and regulation: Clear guidelines are necessary to uphold fairness and accuracy in the application of AI in healthcare, including mechanisms for ongoing auditing and algorithmic adjustments as new evidence becomes available.**
4. **Adoption of probabilistic methodologies: Advanced statistical models should be encouraged in personalized medicine to support the continuous updating of clinical decisions and ensure better alignment with real-world patient data.**

5. **Equity in healthcare access: Technologies must be strategically used to reduce disparities, ensuring that medical advancements benefit all segments of the population in a fair and accessible manner.**

Improving medical education, validating algorithms rigorously, enforcing appropriate regulation, and promoting equitable access to technology are indispensable steps toward ensuring that technological progress enhances not only diagnostic and therapeutic precision but also fairness in healthcare delivery. Without these principles, the risks of bias, distortion, and inequality may undermine the potential benefits of AI in medicine.

Medicine is undergoing a period of rapid transformation, and it is the responsibility of the scientific community, healthcare professionals, and policymakers to guide this progress with accountability. The true challenge lies not only in developing technology, but in integrating it ethically and thoughtfully into clinical care, ensuring that innovation serves as a tool for more precise, safe, and equitable decisions.

If implemented ethically and equitably, AI has the potential to revolutionize medicine — making it more effective, accessible, and patient-centered. The key to this transformation lies in the adoption of responsible practices, continuous education, and the pursuit of solutions that foster a positive global health impact.

Handling Editor: Dr. Ives Passos

References

1. Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104*(12), 510–520. https://doi.org/10.1258/jrsm.2011.110180.
2. Laplace PS. Théorie Analytique des Probabilités. Paris: Courcier, 1814.
3. Osler W, Bean RB, Bean WB. Aphorisms from his bedside teachings and writings. Springfield, USA: Charles C. Thomas; 1961.

4. Hill, G., Forbes, W., Kozak, J., & MacNeill, I. (2000). Likelihood and clinical trials. *Journal of Clinical Epidemiology, 53*(3), 223-227. https://doi.org/10.1016/S0895-4356(99)00156-0.
5. Cochrane, A. L. (1972). Effectiveness and efficiency: Random reflections on health services. Nuffield Provincial Hospitals Trust.
6. Evidence-Based Medicine Working Group. (1991). Evidence-based medicine. *ACP Journal Club, 114*(A-16).
7. Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124.
8. Lee, S., & Sun, W. (2020). Using machine learning approaches for multi-omics data analysis: A review. *Computational and Structural Biotechnology Journal, 18*, 2419-2428. https://doi.org/10.1016/j.csbj.2020.05.002.
9. Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2), 227-244. https://doi.org/10.1016/S0378-3758(00)00115-4
10. Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics, 11*(10), 733–739. https://doi.org/10.1038/nrg2825.
11. Yu, Y., Mai, Y., Zheng, Y., & Shi, L. (2024). Assessing and mitigating batch effects in large-scale omics studies. *Genome Biology, 25*, 254. https://doi.org/10.1186/s13059-024-03401-9.
12. Nazer, L. H., Zatarah, R., Waldrip, S., Chen Ke, J. X., Moukheiber, M., Khanna, A. K., Moukheiber, D., Ma, H., Hicklen, R. S., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health, 2*(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278.
13. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of

populations. *Science, 366*(6464), 447-453. https://doi.org/10.1126/science.aax2342.