Statistical Rigor and Reproducibility in the AI Era

Selçuk Korkmaz

Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye

In today's data-rich biomedical landscape, traditional statistical practices increasingly intersect with powerful artificial intelligence (AI) tools. This convergence presents both opportunities and challenges. AI and machine learning (ML) can reveal patterns that manual analysis might overlook, potentially accelerating discovery. However, their use also heightens longstanding concerns about validity and reproducibility. The inappropriate application of AI is leading to findings that are often unreproducible or clinically irrelevant. This editorial examines issues of reproducibility, the limitations of *p* -value-based inference, and the integration of AI and provides practical guidance for clinicians and researchers.

Reproducibility has long been a concern in medical science, with flawed statistical practices contributing to the problem. In the AI era, new challenges have emerged. Modern AI models-such as deep neural networks-are inherently nondeterministic and sensitive to variations in data and hardware.[1] Simply sharing code does not guarantee identical results, as random initialization, complex processing pipelines, and computational differences can alter outputs. For example, Ball[2] describes striking cases in which AI diagnosed disease based on image background artifacts rather than true pathology and highlights how issues such as data leakage (e.g., accidentally using test data during training) and poorly documented model changes have led to reproducibility failures across multiple fields.

To address these issues, the research community is increasingly emphasizing transparency and standards. AI-driven studies should thoroughly report data preprocessing steps, model architectures, and training conditions. Whenever feasible, sharing data and code facilitates independent verification. Journals have begun requiring adherence to reporting guidelines-such as CONSORT-AI, TRIPOD-AI, and CLAIM-which mandate clear documentation of methods and validation procedures. Moreover, replication-ideally by independent teams or using new datasets-is essential. Replication is the most reliable way to confirm findings, as statistical inference alone, including small *p* -values, can be misleading.[3] In practice, clinicians should interpret AI results from single studies with caution and seek confirmatory evidence or prospective validation before applying findings in clinical care.

Traditional hypothesis testing and *p* -values remain widely used, but their limitations are now better understood. Modern analyses often involve large datasets ("big data"), where even trivial differences can become statistically significant. Experts increasingly advocate moving beyond rigid thresholds. For instance, a recent review emphasizes that the *p* -value is a useful decision-making tool only when interpreted in context-taking into account study design, effect sizes, and prior knowledge.[4] In short, a *p-value* is just one piece of evidence, not a definitive marker of truth. The American Statistical Association's 2016 statement cautioned against overreliance on the "$p < 0.05$" threshold, discouraged the use of hard cutoffs, and recommended transparent, contextual reporting of effect sizes and confidence intervals rather than focusing on a single *p* -value.[5]

In practice, clinicians and researchers should prioritize clinical significance. For example, a drug that reduces systolic blood pressure by just 1 mmHg may yield a *p* -value of 0.049 in a large trial-yet the effect is clinically negligible. Conversely, a 10 mmHg reduction with a *p* -value of 0.051 could represent a meaningful benefit that warrants further investigation, rather than dismissal based on an arbitrary threshold. Some experts have proposed replacing binary significance labels with graded measures of evidence strength-such as likelihood ratios or Bayesian factors-to better reflect how data influence our belief in a hypothesis.[6,7] At a minimum, *p* -values should always be reported alongside effect sizes and confidence intervals (or Bayesian posterior estimates). Transparency demands making uncertainty explicit and avoiding overstatements driven by arbitrary statistical cut-offs.

AI and ML are increasingly integrated into research workflows. For example, algorithms now analyze imaging data[8,9] or genomic sequences[10] to detect patterns or risk factors that may be overlooked by traditional methods. In well-designed pipelines, AI can assist with tasks ranging from automated data cleaning to advance predictive modeling, thereby improving efficiency and standardization. Large language models (LLMs)-such as ChatGPT-and automated ML (AutoML) systems can already perform routine statistical tasks, such as conducting standard tests and proposing or tuning candidate models,

Corresponding author: Selçuk Korkmaz, Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye
e-mail: selcukorkmaz@gmail.com

Available Online Date: 01.09.2025 • **DOI:** 10.4274/balkanmedj.galenos.2025.2025.040825
Available at www.balkanmedicaljournal.org
ORCID iDs of the authors: S.K. 0000-0003-4632-6850.
Cite this article as: Korkmaz S. Statistical Rigor and Reproducibility in the AI Era. Balkan Med J.; 2025; 42(5):386-7.

effectively acting as statistical assistants.[11] When combined with AutoML and code-execution environments, LLMs can even transform a dataset and a research question into a first-draft analysis report. Nonetheless, experts must still verify model specifications, check assumptions, control for multiplicity, and ensure reproducibility.[12]

Despite these advances, AI-driven automation has limitations. While AI models often excel at interpolating within known datasets, they may fail on out-of-sample data or when biases are present.[13] Key concerns-such as lack of transparency ("black box" models), overfitting, and dependence on high-quality training data-highlight the need for human oversight. As automation expands across the analytical pipeline, integrity, transparency, and interpretability still require multidisciplinary input. Statisticians, clinicians, ethicists, and regulators must work together to ensure rigorous study design, model validation, bias assessment, and reproducible reporting.[14] In practice, multidisciplinary collaboration is critical. Clinicians should partner with statisticians from the outset of study design through analysis. Koçak et al.[15] recently warned that a "language barrier" between AI developers and clinicians can result in errors or low-quality studies. Engaging clinical experts early helps ensure that AI models address medically relevant questions, use appropriate endpoints, and are evaluated using clinically meaningful-not merely technical-metrics.

Emphasize transparency and reproducibility: Preregister studies when possible, share analysis code and data (with appropriate patient privacy safeguards), and document each step of the process. This enables others to verify findings and reuse analytical pipelines.

Move beyond *p* -values as sole evidence: Report effect sizes, confidence intervals, predictive accuracy metrics, and where feasible, Bayesian measures. Interpret findings in light of clinical relevance and existing knowledge-not just statistical thresholds.

Validate AI models rigorously: Use true hold-out datasets or external cohorts to assess generalizability. Prevent data leakage and overfitting through careful separation of training, validation, and test sets. Always report model performance on independent data, when available.

Encourage interdisciplinary expertise: Include clinicians, statisticians, and computer scientists as part of research teams. Promote cross-training: clinicians should acquire basic AI/ML literacy, while statisticians and data scientists should understand the biomedical context. Bridging these gaps reduces the risk of methodological errors.

Practice continuous education: Stay current with evolving guidelines for AI in medicine (e.g., CONSORT-AI, TRIPOD-AI, CLAIM), and monitor updates from statistical and regulatory authorities. Journal editors and peer reviewers should also improve their AI/statistics literacy to ensure critical and informed manuscript evaluation.

These measures will help ensure that AI enhances-rather than undermines-statistical rigor in biomedical research.

AI is transforming medical research by enabling analyses that were previously impractical[16], but this shift brings greater responsibility to uphold scientific rigor. Clinicians and researchers should embrace these new tools while remaining cautious of overly simplistic or unverified findings. Robust statistical thinking-emphasizing transparency, thorough validation, and interdisciplinary collaboration-remains as essential as ever. Ultimately, the goal is not to replace hypothesis-driven medicine but to enhance it. This requires using AI-driven insights that are trustworthy, reproducible, and truly beneficial to patients. By balancing innovation with rigor, the medical research community can ensure that the promise of AI is fulfilled in a reliable and clinically meaningful way.

REFERENCES

1. Han H. Challenges of reproducible AI in biomedical data science. *BMC Med Genomics.* 2025;18:8. [CrossRef]
2. Ball P. Is AI leading to a reproducibility crisis in science? *Nature.* 2023;624:22-25. [CrossRef]
3. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 2016;31:337-350. [CrossRef]
4. Chén OY, Bodelet JS, Saraiva RG, et al. The roles, challenges, and merits of the p value. *Patterns (N Y).* 2023;4:100878. [CrossRef]
5. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *The American Statistician.* 2016;70:129-133. [CrossRef]
6. Perneger TV. How to use likelihood ratios to interpret evidence from randomized trials. *J Clin Epidemiol*. 2021;136:235-242. [CrossRef]
7. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med*. 1999;130:1005-1013. [CrossRef]
8. Usta U, Taştekin E. Present and future of artificial intelligence in pathology. *Balkan Med J*. 2024;41:157-158. [CrossRef]
9. Sarıkaya Solak S, Göktay F. The promising role of artificial intelligence in nail diseases. *Balkan Med J*. 2024;41:234-235. [CrossRef]
10. Erdoğan S. Integration of Artificial intelligence and genome editing system for determining the treatment of genetic disorders. *Balkan Med J*. 2024;41:419-420. [CrossRef]
11. Koçak D. Examination of ChatGPT's performance as a data analysis tool. *Educ Psychol Meas*. 2025:00131644241302721. [CrossRef]
12. Korkmaz S. Artificial intelligence in healthcare: a revolutionary ally or an ethical dilemma? *Balkan Med J*. 2024;41:87-88. [CrossRef]
13. Mondillo G, Frattolillo V, Colosimo S, Perrotta A. Artificial intelligence in pediatric nail diseases: limitations and prospects. *Balkan Med J*. 2025;42:86. [CrossRef]
14. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. Erratum in: *BMJ*. 2024;385:q902. [CrossRef]
15. Koçak B, Cuocolo R, dos Santos DP, Stanzione A, Ugga L. Must-have Qualities of clinical research on artificial intelligence and machine learning. *Balkan Med J*. 2023;40:3-12. [CrossRef]
16. Hayıroğlu Mİ, Altay S. The role of artificial intelligence in coronary artery disease and atrial fibrillation. *Balkan Med J*. 2023;40:151-152. [CrossRef]