DOI https://doi.org/10.70280/njph(2025)v2i1.31

Why Traditional Statistical Methods Need to Evolve in the Age of Artificial Intelligence: A Biostatistical Perspective

Deepak Raj Joshi[1]

ABSTRACT

Traditional statistical methods, basically the frequentist approach, must evolve to remain relevant in the age of Artificial Intelligence (AI). While Conventional statistical methods work under theoretical assumptions, they struggle to handle the complexities of modern biomedical data, including high dimensionality, non-linearity, and violations of key assumptions. However, this is not a problem for the newer machine learning models like support vector machines. There are new techniques like regularization (ridge, lasso) to handle many of the assumptions in traditional statistical methods, which can be implemented and automated using software like R and Python. Machine learning as a part of AI offers solutions by handling large-scale complex datasets, uncovering hidden patterns, and improving prediction power. They are based on the foundation models where statistics and mathematics meet. So, just talking about the limitations of the statistical methods is half true. The viewpoint tries to explain why to integrate AI with traditional biostatistics, creating hybrid models that combine statistical rigor with AI flexibility. Integration can enhance data analysis, causal inference, and decision-making, ultimately advancing personalized medicine and public health, ethically and transparently.

Key words: Traditional statistics, evolve, artificial intelligence, machine learning, biostatistics

Received on: 9 January 2025
Accepted on: 15 February 2025

Check for updates

[1]Department of Community Medicine and Public Health, Maharajgunj Medical Campus, Institute of Medicine, Tribhuvan University, Kathmandu, Nepal

Correspondence to:
Deepak Raj Joshi
Department of Community Medicine and Public Health, Maharajgunj Medical Campus, Institute of Medicine, Tribhuvan University, Kathmandu, Nepal
Email: raazdpk@gmail.com

INTRODUCTION

The swift advancement of Artificial Intelligence (AI) with Machine Learning (ML) has been transforming numerous fields, including biostatistics, by initiating various advanced tools like Galaxy-ML.(1) However, traditional statistical methods have long been the foundation of data analysis in Biostatistics, focusing on hypothesis testing, model-based inference, and parameter estimation. The AI can analyze big data and discover hidden patterns, changing genomics, drug discovery, and clinical trials.(2)

ML, a branch of AI, empowers systems to learn independently from data and past experiences by detecting patterns to make predictions with low or no human effort. This includes supervised learning (using labeled training data), unsupervised learning (extracting insights from unlabeled datasets), and reinforcement learning (improving decisions through feedback).(3)

AI offers a paradigm shift in data analysis by integrating automation (a concept from computer science) within the broader framework of data science, where statistics, mathematics, and computer science converge to facilitate efficient analytical solutions. The traditional statistical methods provide interpretability and strong theoretical foundations, while AI enhances these capabilities through advanced computational techniques. (4) However, this shift also highlights the need for biostatisticians to adapt their methodologies to ensure the interpretability, robustness, and, most importantly, ethical use of AI models. (5)

Integrating AI and Biostatistics can create new opportunities in medical research and public health. It raises questions regarding the limitations of traditional methods in an era of rapidly advancing technologies.(5–8) Integrating AI and biostatistics unlocks new avenues in medical and public health research by analyzing large and complex datasets and uncovering patterns. While AI serves as ML, deep learning approaches such as multilayer perceptrons (MLPs) and deep neural networks (DNNs) explore deeper by modeling complex, non-linear relationships in biomedical data. These architectures of advanced neural networks (NNs) can enhance the detection of disease biomarkers, improve patient outcome predictions, and accelerate drug discovery by learning from various sources of clinical and molecular data. The combination of ML models like MLPs and DNNs with biostatistical design and validation ensures robustness and interpretability of these innovations, addressing the limitations of traditional statistical approaches in the era of rapidly advancing technologies.(5)

This viewpoint article discusses that traditional statistical methods in biostatistics need to evolve to address the challenges posed by AI. Indeed, traditional statistical methods have a strong theoretical basis, but they struggle with complex, high-dimensional biomedical data, assumption dependencies, and limited predictive power. AI offers solutions by handling large datasets, identifying hidden patterns, and improving prediction. It efficiently manages, stores, and analyzes big data with the support of technologies like distributed computing and cloud platforms, which divide large datasets into smaller segments and process them in parallel to enhance efficiency.(9)

Machine learning algorithms like deep learning and neural networks can effectively find hidden patterns in large and complex datasets, often missed by researchers, enabling better decisions, optimization, and risk reduction.(10) AI models improve their predictive accuracy over time by continuously learning from new data. Using both past and present information, they can forecast trends, identify anomalies, and generate insights across fields like healthcare, finance, etc.(11)

This article advocates for integrating traditional biostatistical tools and AI to create hybrid models, enhance causal inference, and ensure ethical use of AI. This integration will give more precise, interpretable, and actionable insights, and will develop personalized medicine and public health.

THE ROLE OF ARTIFICIAL INTELLIGENCE IN BIOSTATISTICS

AI and ML algorithms, such as neural networks (NNs), support vector machines (SVMs), and ensemble learning methods like bagging (Bootstrap Aggregating – combines predictions from multiple models trained on random subsets of the data to reduce variance and improve accuracy), boosting(Sequentially builds models that correct errors of previous ones to improve overall performance and reduce bias), and stacking(Combines multiple different models and uses a meta-model to learn how best to blend their predictions) give scalable, flexible, and high-performance solutions with greater flexibility and predictive power in data analysis and research. The Key roles include:

Handling high-cimensional and complex data: AI can process large-scale biological datasets without relying on restrictive assumptions. The basic assumptions of classical statistics, like normality, linearity, and homoscedasticity, may not always be held with high-dimensional and heterogeneous real-world biological data like genomics and proteomics. AI models based on machine learning and deep learning are data-driven, and they do not require strict distributional and structural assumptions. They handle non-linear relationships, adapt to missing, noisy, and unstructured data, and perform an in-depth mining of potential information in data.(12)

Data analysis and pattern recognition: AI algorithms can analyze extensive molecular datasets to accelerate drug discovery, identify promising drug candidates, and simulate interactions between drugs.(5) ML algorithms are used to identify hidden patterns in medical images, genetic markers, and patient records, thereby improving diagnostic accuracy.(13) Deep Neural Networks (DNNs), which consist of multiple hidden layers between input and output, are effective at capturing complex and classified patterns in large datasets. These models learn through forward (process of data flow from input to prediction) and backward (process of learning from the error by adjusting the internal weights to improve future predictions) propagation, during which model weights the core parameters of learning are continuously updated to minimize prediction error. This adaptability makes DNNs highly suitable for diverse biomedical data types.(14)

Specific architectures like Convolutional Neural Networks (CNNs) are used in medical image analysis (e.g., detecting tumors or classifying X-rays). Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) models are well-suited for sequential data such as time-series patient records or genomic sequences. Additionally, Transformers, originally developed for natural language processing, are gaining momentum in biomedical research for their ability to capture long-range dependencies and contextual relationships in

data. These advanced neural architectures enhance diagnostic accuracy, support precision medicine, and uncover novel insights in health sciences.(13,15)

Causal Inference and prediction: AI can predict causal relationships between variables, such as exposure and outcome, risk factors and diseases, or interventions and effects. ML models like supervised learning can detect data patterns to identify the relation between exposure and outcomes, such as predicting whether long-term exposure to air pollution raises the risk of respiratory diseases, and also can suggest personalized treatment options and detect potential drug interactions. For example, an AI system might recommend the most effective medication for a diabetic patient by learning from combinations of their medical history, lab data, and real-time monitoring, while also identifying possible adverse reactions with other prescribed medications.(5)

Automation and efficiency: AI systematizes the analysis of large datasets, helping biostatisticians and epidemiologists discover new patterns, generate hypotheses, test causal relationships, and improve decision-making.(5) AI-driven methods automate data processing, reducing human intervention and potential biases.(16) However, the automation relies heavily on coding skills in computers. So, statisticians need to be proficient in programming or, at least, understand how to apply automated workflows using code-based tools and platforms.

Image analysis and diagnostics: AI-driven diagnostic systems can analyze medical imaging like X-rays, MRIs, CT scans, and pathology slides to identify cancer, fractures, and some neurological disorders.(5) For this, CNN models are used to extract important features from images like tumors, fractures, and other abnormalities with high accuracy. Advanced CNN architectures are commonly used in disease classification, segmentation of affected areas, and anomaly detection.(17)

Drug discovery and development: AI is used in drug discovery to analyze datasets related to chemical compounds, biological activity, and pharmacological profiles to predict potential drug candidates, and assess their safety and efficacy.(5) Supervised ML models, such as Decision Trees(a flowchart-like structure for making decisions by splitting data based on features), Random Forests(an advanced method that builds many decision trees and combines their answers to make more accurate and reliable predictions), SVMs, and Gradient Boosting, predict drug-target interactions and classify compound activity. Also, deep learning models like Graph Neural Networks (GNNs) and RNNs are used in molecular structures modeling and chemical sequences. These models help detect potential drug candidates, optimize molecular properties, and predict safety and efficacy profiles to accelerate early-stage drug development.(18)

Genomics and proteomics research: AI is widely used in genomics and proteomics research for tasks like DNA sequence analysis, protein structure prediction, and identification of genetic markers associated with diseases. Biostatistics is essential for designing experiments, conducting genetic association studies, and assessing the statistical significance of genetic findings.(5) Deep learning models like CNNs, RNNs, GNNs, and Transformers are applied to genomic sequences and proteomic data to uncover complex, non-linear patterns.(19)

INTEGRATING ARTIFICIAL INTELLIGENCE WITH TRADITIONAL BIOSTATISTICS

Despite AI's advantages, it should not entirely replace traditional statistical approaches because of its solid theoretical foundation. Instead, an integration of both can enhance interpretability and predictive power. Some strategies for the integration include:

Hybrid models: Combining traditional statistical methods with ML techniques, such as regularized regression and Bayesian statistics with machine learning and deep learning, allows for better models, robustness, and interpretability.(20)

Causal inference with AI: AI can assist in identifying causal relationships (like structural equation modeling with machine learning) rather than mere associations, aligning with the goals of traditional biostatistics.(21) It can support the confirmation of causality by systematically applying Bradford Hill's criteria, like; strength, consistency, specificity, temporality, and biological gradient through advanced analytics and machine learning techniques. For instance, a recent study utilized AI-powered multiple regression models on global health data to evaluate risk factors influencing BMI, and the analysis fulfilled all nine Hill's criteria, providing strong evidence for a causal link between those risk factors and BMI outcomes. This demonstrates how AI can go beyond detecting associations to offering robust support for causal inference in public health research.(22)

Ethical and transparent AI: Ensuring AI models are explainable and validated using classical statistical principles enhances their acceptance in medical research.(23) Web scraping is an example of an ethical data collection using AI. It collects openly available

datasets for text mining by adhering to the terms of service, privacy policies, and avoiding sensitive or private information from the website. For instance, to build datasets for natural language processing models, researchers may scrape open-access medical data literature. But, the highly used large AI models like OpenAI's ChatGPT have trained large amount of text available on the internet and scraped without explicit permission, raising concerns on transparency, right of intellectual property, and ethical data collection.(24)

AI is transforming biostatistics by automating the analysis of large and complex biological and medical datasets, identifying patterns and insights that may not be apparent through traditional statistical methods. AI techniques, underpinned by statistical principles, enhance data analysis, causal inference, and decision-making in biostatistics and epidemiology.

Moreover, AI supports the integration of data from different sources, such as genomic, clinical, environmental, and social data, enhancing the completeness, consistency, and comparability of data for more comprehensive analyses. Explainable AI and Ethical AI aim to make AI systems transparent, interpretable, accountable, and aligned with ethical principles and human values. AI's integration with biostatistics not only enhances the efficiency and accuracy of data analysis but also opens new avenues to solve complex public health issues and advance personalized medicine.

CONCLUSION

The integration of AI can significantly enhance traditional statistical methodologies. Rather than replacing established techniques, AI can serve as a complementary tool, addressing the complexities of modern biomedical data analysis. By merging AI-driven approaches with classical statistical reasoning, researchers can attain more effective and actionable insights in healthcare.

Biostatisticians must integrate AI thoughtfully, maintaining statistical rigor, interpretability, and ethics. This careful approach will ensure that advancements in personalized medicine and public health are achieved ethically and transparently. Ultimately, the combination of conventional statistical expertise and innovative AI techniques will empower biostatisticians to drive significant improvements in health outcomes with a deeper understanding of biological processes.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

1. Gu Q, Kumar A, Bray S, Creason A, Khanteymoori A, Jalili V, et al. Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine. PLoS Comput Biol. 2021 Jun 1;17(6):e1009014.
2. 2Serrano DR, Luciano FC, Anaya BJ, Ongoren B, Kara A, Molina G, et al. Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine. Pharmaceutics. 2024 Oct 14;16(10):1328.
3. 3What is machine learning? Understanding types & applications - Spiceworks [Internet]. [cited 2025 Apr 19]. Available from: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/
4. 4Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, et al. Artificial intelligence: A powerful paradigm for scientific research. The Innovation. 2021 Nov 28;2(4):100179.
5. Zhao Y. Artificial Intelligence and Biostatistics: Revolutionizing Medical Research.
6. Bhandari DR, Baron M, Shah K, Kandel S. Role of Statistics in Artificial Intelligence Technology. NCC J. 2024 Dec 6;9(1):133–9.
7. Min J, Song X, Zheng S, King CB, Deng X, Hong Y. Applied Statistics in the Era of Artificial Intelligence: A Review and Vision [Internet]. arXiv; 2024 [cited 2025 Feb 26]. Available from: http://arxiv.org/abs/2412.10331
8. Faes L, Sim DA, van Smeden M, Held U, Bossuyt PM, Bachmann LM. Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins? Front Digit Health. 2022 Jan 26;4:833912.
9. Hadoop vs Spark: Big Data Showdown [Internet]. [cited 2025 Apr 19]. Available from: https://www.wallarm.com/cloud-native-products-101/hadoop-vs-spark-big-data-processing
10. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021 Mar 22;2(3):160.
11. Sarker IH. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. Sn Comput Sci. 2022;3(2):158.
12. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019 Apr 16;20(1):76.
13. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2017 May 6;19(6):1236–46.
14. Mall PK, Singh PK, Srivastav S, Narayan V, Paprzycki M, Jaworska T, et al. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. Healthc Anal. 2023 Dec 1;4:100216.

15. Choi SR, Lee M. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. Biology. 2023 Jul;12(7):1033.

16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56.

17. Mienye ID, Swart TG, Obaido G, Jordan M, Ilono P. Deep Convolutional Neural Networks in Medical Image Analysis: A Review. Information. 2025 Mar;16(3):195.

18. Obaido G, Mienye ID, Egbelowo OF, Emmanuel ID, Ogunleye A, Ogbuokiri B, et al. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. Mach Learn Appl. 2024 Sep 1;17:100576.

19. Li Z, Gao E, Zhou J, Han W, Xu X, Gao X. Applications of deep learning in understanding gene regulation. Cell Rep Methods. 2023 Jan 23;3(1):100384.

20. Bzdok D, Krzywinski M, Altman N. Machine learning: A primer. Nat Methods. 2017 Nov 30;14(12):1119–20.

21. Pearl J. Causality. Cambridge University Press; 2009. 487 p.

22. Cundiff DK, Wu C. Artificial intelligence analytics applied to body mass index global burden of disease worldwide cohort data derives a multiple regression formula with population attributable fraction risk factor coefficients testable by all nine Bradford Hill causality criteria [Internet]. medRxiv; 2021 [cited 2025 Apr 19]. p. 2020.07.27.20162487. Available from: https://www.medrxiv.org/content/10.1101/2020.07.27.20162487v4

23. Lipton ZC. The Mythos of Model Interpretability [Internet]. arXiv; 2017 [cited 2025 Feb 26]. Available from: http://arxiv.org/abs/1606.03490

24. Jayachandran J, Arni V. Traversing the Ethical Landscape of Data Scraping for AI [Internet]. Rochester, NY: Social Science Research Network; 2023 [cited 2025 Apr 19]. Available from: https://papers.ssrn.com/abstract=4666354