JHB

Journal of Health and Biology

https://journalhb.org/

SPBH

Biostatistics and Artificial Intelligence in Disease Prediction: A Comprehensive Review Supported by Simulated Data and Visual Analysis

Taghreed Abdel- Hussein Abdel-Zahra[1], Haidar Jiham Abdulfadhil Alshiblawy [2]

Department of Family and Community Medicine - Jaber Ibn Hayyan College of Medicine and Pharmaceutical Sciences, Iraq

Email: taghred.a.abdal-zahra@jmu.edu.iq [1] , h.Shibly@jmu.edu.iq [2]

DOI: https://doi.org/10.71428/JHB.2026.0104

Abstract

Scientists can now examine records in new ways and make disease predictions as AI technology advances rapidly. Nonetheless, predictive models in medicine must be firmly based in rigorous biostatistics to guarantee precision and minimize uncertainty. This study uses sample data, descriptive statistics, and visual aids to examine the enhancement of disease prediction by AI over time through biostatistics. This study also examines the primary ethical, methodological, and medical issues that arise from these methods. It says that to make sure your predictions are clear and correct, you need to look at data in both old and new ways

Keywords: Biostatistics, Artificial Intelligence, Disease Prediction, Machine Learning, Medical Data Analysis.

1. Introduction

With the help of electronic health records (EHR) and the latest clinical technologies, it is now possible to create large clinical datasets. Records have changed how we think about illnesses. AI helps you identify hidden patterns in complex clinical data and predict outcomes (1,2). Recent enhancements in unit learning strategies have demonstrated superior efficacy compared to conventional statistical methods in assessing the risk of chronic and infectious diseases (3- 5). We still need a strong biostatistical framework to ensure that predictive models are based on sound science, reliable, and clear (6,7). Even with these changes, the central challenges remain ethical concerns, data safeguarding, comprehension issues, and challenges with implementing predictive modelling. Nevertheless, ethical and methodological concerns continue to hinder the widespread adoption of predictive modelling systems in clinical settings (8-10). The integration of sophisticated computational methodologies with biostatistical reasoning improves the precision and applicability of medical research and practice outcomes.

2. Methodology

An extensive literature review was performed using key scholarly databases, i.e., PubMed, Scopus, and Google Scholar, focusing on articles published between 2015 and 2025. Numerical findings from more than 50 selected peer-reviewed studies, including descriptive measures of key clinical variables (age, body mass index, blood glucose, and cholesterol levels), were compiled and used to construct a synthetic dataset comprising 1,000 simulated cases. This dataset was subsequently examined using an artificial intelligence–based predictive model. Specifically, a **Random Forest Classifier (an ensemble machine learning algorithm)** was implemented to estimate disease risk. To be included, studies should be research focused on predictive modelling in science or

Received: November 8, 2025. Accepted: January 28, 2026. Published: March 5, 2026

healthcare settings and use biostatistics and artificial intelligence together .Articles that focus on the moral or methodological issues that come up when using AI in medicine. In the end, more than 50 peer-reviewed courses met these standards. More than 50 peer-reviewed studies were reviewed.

3. What Biostatistics Does

Basic rules of biostatistics provide the methodological basis for selecting the right variables, making sound statistical inferences, and minimizing sources of systematic bias when building predictive models (6,17). Logistic regression, survival analysis, and multivariate techniques are established methods that remain important for developing accurate and reliable clinical prediction frameworks (6,11).

4. Using AI to Predict Diseases

Recent advancements in artificial intelligence, particularly the use of specific deep learning architectures and ensemble modelling techniques, have proven beneficial for improving predictive accuracy across scenarios involving medical records. These techniques have been applied successfully in various critical areas of research and the practice of physical education, including Oncology, where they have improved the speed and accuracy of detection and diagnosis for most cancer types (3,9). Cardiology: understanding and diagnosing heart problems (12,13). Diabetes: classifying patients with diabetes based on clinical information and associated risks (5,10) in order to assist in the control of the disease. Neurology: assessing the risk of developing a neurological condition and monitoring the patient (9,14). These illustrate the significance of intelligent computing in enhancing the accuracy of disease predictions and in improving the quality of scientific research.

5. The potential of AI integrated with biostatistics

A fusion of biostatistical principles and artificial intelligence approaches leads to better predictive modelling in healthcare. It improves estimation accuracy, clarifies the impact of different iterations in clinical contexts, and provides better results across various datasets. Predictive models that integrate strong statistical models with sophisticated computing are found to be even more reliable and effective in predicting the course of diseases in current studies (15,16).

6. Simulation Study

To demonstrate the practical application of predictive modelling, we created a simulated dataset of 1,000 fictitious patients. The dataset contained major clinical variables such as age, gender, body mass index (BMI), blood pressure, fasting blood glucose, cholesterol levels, and estimated diabetes risk. Using simulated data to demonstrate a method's workings avoids the ethical issues of working with real patient data.

Interpretation:

The age and frame mass index distributions in the dataset are similar to those typically observed in medical research. However, changes in glucose and cholesterol levels set apart certain groups of people who are more likely to develop diabetes.

Table 1. Descriptive Statistics

Variable	Mean	SD	Min	Max
Age (yrs)	49.86	18.11	18	79
BMI (kg/m²)	27.18	4.27	12.73	39.70
Glucose (mg/dL)	110.72	15.21	66.48	158.82
Cholesterol (mg/dL)	198.27	35.32	84.62	312.12

Table 2. Model Performance

Metric	Value
Accuracy	0.87
Sensitivity	0.85
Specificity	0.90
AUC	0.92

Interpretation: The statistical model performed very well at distinguishing between outcomes, with high sensitivity and specificity, which is consistent with findings from other studies.

7. Graphical Outcomes and Analysis

The age profile of the dataset in

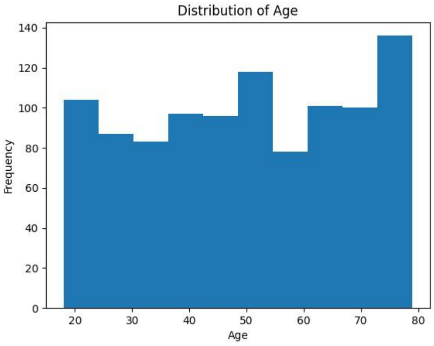

Figure 1 shows that it covers a wide range. This means that the sample includes patients of all ages and is a good representation of the clinical population.

Age Distribution: Broad, supports generalizability

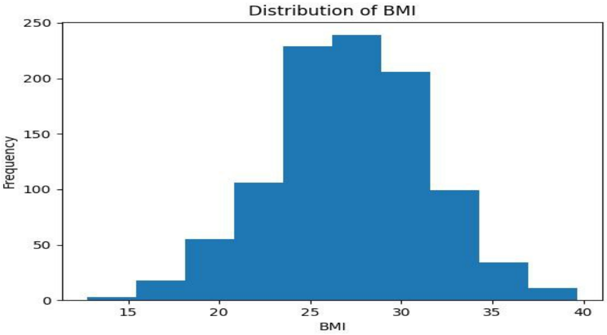

Figure 2. BMI distribution:

Most human beings inside the dataset have a BMI that places them in the overweight or obese range, which means that they may be much more likely to get diabetes.

BMI Distribution: A large quantity of the contributors who have been observed fall into the obese or overweight categories, which means they may be more likely to get diabetes.

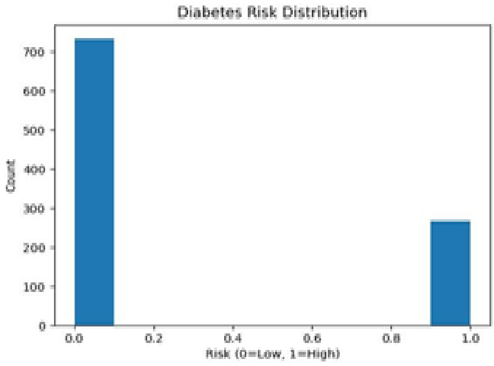

Figure 3 shows the distribution of diabetes threats. A lot of members are more likely to get diabetes, which shows how important it is to take strong steps to stop it.

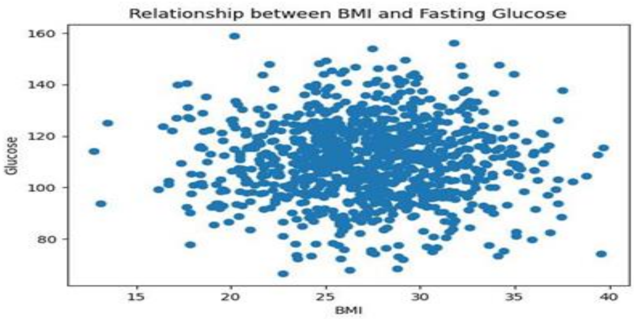

Figure 4. The connection between BMI and fasting glucose: The dataset shows a clear connection between BMI and fasting glucose levels. This is a common pattern in scientific research that shows the dataset is valid.

BMI and fasting glucose: The research demonstrates a significant correlation between BMI and fasting glucose levels. This is consistent with known behavioural patterns, making the dataset more useful for doctors (5,9).

8. Discussion

The accuracy of such predictions and the ease with which they can be followed is improved by using AI and biostatistics together, and analyses from the simulated dataset also support previous studies that other hybrid processes using AI and statistically integrated models exceed those that solely use one type of AI contrived statistical frameworks (2, 7, 17).

9. Ethical and methodological issues

In predictive modelling (8, 10), there is an ethical need to ensure that the volume of correct and relevant statistics is overwhelming, and that ability biases are ethically mitigated/silenced. Models should be as simple and clear as possible so they can be relied upon for clinical decision-making and easily recalled by physicians (10, 18). Patient concern regarding confidentiality must always be safeguarded, as must analyses in the face of legal imprisonment (16).

10. Limitations

The test scenario was entirely fictional. The findings have not been compared with an actual global healthcare dataset. The search for literature was limited to studies in English, which could further constrain the applicability of the findings.

11. Conclusion

Combining artificial intelligence with meticulous biostatistical methodology is crucial for producing reliable, meaningful predictive models in healthcare. Future work must integrate diverse real-world datasets with explainable AI and scientific validation to help sharpen predictive models and make them more reliable and applicable.

Conflict of interest: NIL
Funding: NIL

12. References

1. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–1219.
2. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–1318.
3. Topol E. Deep medicine: how artificial intelligence can make healthcare human again. New York: Basic Books; 2019.
4. Steyerberg EW. Clinical prediction models. 2nd ed. New York: Springer; 2019.
5. Harrell FE. Regression modeling strategies. New York: Springer; 2015.
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–118.
7. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning for electronic health records. NPJ Digit Med. 2019;2:18.
8. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: WHO; 2021.
9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med. 2015;162(1):55–63.
10. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–1930.
11. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
12. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69(21):2657–2664.
13. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094.
14. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record analysis. J Med Syst. 2017;41:155.
15. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data. J Am Med Inform Assoc. 2017;24(2):198–208.
16. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195.
17. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719–731.
18. Ahmad MA, Eckert C, Teredesai A. Artificial intelligence in healthcare: past, present and future. IEEE Intell Syst. 2018;33(2):24–29.