https://www.mediresonline.org/

MEDIRES Symbol of Quality Research

ISSN: 2833-0951

Open Access

Biomedical and Biotechnological Sciences

Short Communication

Data Mining and Statistical Modelling In Biomedical Research: Concepts, Applications, and Future Perspectives

[1]Obasi Miriam Oluchi, [2]Emereole Chinwe Nelly, [2]Iwuji Joy Chidinma and [1]Okafor Onyeka Edmond

[1.]Department of Medical Laboratory Science, Imo State University, Imo State-460222,Owerri.
[2.]Department of Science Laboratory Technology, Federal University of Technology, Imo State-460222,Owerri.

***Corresponding Author:** Nnodim Johnkennedy, Department of Medical Laboratory science, Imo state university, Owerri- 460222, Nigeria

Received Date: 18 March 2026; **Accepted Date:** 20 April 2026; **Published Date:** 23 May 2026

Abstract

Nowadays, biomedical studies swim in data, needing tools like computing methods, stats, and algorithms just to make sense of messy health and biology information. At the core sit two main approaches - data mining digs up unseen trends inside massive piles of numbers; meanwhile, statistics build frameworks to check assumptions, spot relationships, then forecast outcomes or summarize findings. Here's a look at those concepts: their meaning, inner workings, real-world role, plus where they show up across medical science efforts. Later on, it touches on today's struggles, ethical puzzles, followed by shifts expected beyond 2025. Through combining data-guided findings with pattern analysis, progress in personalized treatment and applied studies becomes clearer.

Keywords: Biomedical research, Data mining, Statistical modelling, Machine learning, Predictive analytics, Bioinformatics, precision medicine, Casual inference.

Introduction

Fast changes in biotech, genomics, and digital health have flooded science with piles of biological details. Thanks to tools like high-speed gene reading, protein tracking, medical scans, along with computerized patient files, huge messy collections of facts now pile up faster than old math tricks can handle. Spotting trends, pulling out key traits, letting systems learn on their own - that is where digging into data helps. Yet firm number work still matters, offering ways to measure doubt, draw lines between cause and effect, reach decisions backed by proof. Together, these paths help find signs of illness, name diseases, judge patient dangers, choose which medicines deserve attention sooner. Lately, smart calculations plus careful stats show up more often in clinics, fueled by sharper software and an ever-growing flood of shared measurements [1].

Peeking into piles of information, data mining spots fresh trends where none were seen before. Instead of guesses, it relies on what shows up across heaps of records. On the flip side, statistical mo9++delling leans on numbers and chance to make sense of events or peer ahead. It frames ideas through equations tied to real-world behavior. Together they've reshaped medical research - uncovering signals in biology, shaping outlooks for sick individuals, fine-tuning care built around personal traits.

Mining Data

Data mining is the process of finding new, useful, and hidden patterns or relationships in massive databases using computers and statistics. It is a necessary part of the knowledge discovery in databases (KDD) process, which includes

collecting, cleaning, transforming, finding patterns, and making sense of data. In the biomedical field, data mining finds different types of diseases, finds links between genes and diseases, and predicts how well patients will do. Data mining employs machine learning and statistical algorithms to reveal covert or concealed patterns within extensive datasets [2].

Statistical Modelling

Numbers help spot patterns between changing parts when scientists build a kind of blueprint using real-world observations. This framework rests on ideas about where data comes from, opening paths to forecast outcomes or weigh possible truths [3]. Biomedical work leans on styles like regression or survival methods, each suited to judge treatment effects or gauge uncertainty with precision. Some approaches mix group trends with individual details, others adjust beliefs as new evidence appears. These tools shape understanding without claiming absolute certainty.

Though they share ground, data mining leans toward spotting trends in massive datasets while statistical modelling focuses more on structured reasoning and measuring doubt. Methods like clustering show up often in both fields, yet their goals shift depending on context. Instead of strict rules, one builds broad predictions, the other tests narrow ideas with care. Feature selection appears across practices, but purpose shapes how it fits within each approach. Regression trees serve discovery just as dimensionality reduction aids precision. Clarity of assumption matters little when scaling detection, yet becomes central during validation. Patterns found through sweeping searches gain strength only after rigorous checking. Biomedical advances grow best not from either alone, but by starting wide then narrowing down. Findings emerge first through exploration, later confirmed through disciplined analysis. One reveals what might matter, the next judges whether it truly does.

Types of data in biomedicine and how they affect approaches

Genomic and Multi-omics: sequencing DNA and RNA, studying epigenomics and proteomics. High dimensional (millions of features possible), frequently sparse, with intricate measurement error frameworks. A lot of people use public resources like TCGA and ENCODE [4].

Medical Imaging: Radiology (CT, MRI), digital pathology (whole slide pictures). These images are very high-dimensional, spatially structured, and typically need sophisticated pre-processing and deep learning [5].

Electronic Health Records (EHRs): Longitudinal and multimodal (structured labs, vitals, and unstructured clinical notes). Some problems are missingness, uneven sampling, and a coding system that changes [6].

Wearable and longitudinal public health data: continuous time series from sensors; helpful for monitoring and intervention research.

These data sets vary in size, noise levels, and privacy and ethical concerns, all of which affect the choice of algorithms and validation methods [7].

The main ways to mine data and how they are used in biomedical research

Some common ways to mine data in biomedical research are:

Clustering (K-means hierarchical spectral density based) is used to locate illness classes from expression data or to find a patient's phenotype from multimodal EHR information.

Dimensionality reduction (PCA, t-SNE, UMAP, autoencoders) is necessary for visualising and cleaning up noise before modelling high-dimensional omics or picture features [8].

Models for supervised prediction

Deep learning - like CNNs or transformers - shines when handling images or sequences. Though kernel tricks help where clear dividing lines hide, support vector machines step in then. Clinical tabular data? Gradient boosting usually keeps up just fine, while random forests hold their ground too. Instead of going all-in on one method, many setups mix deep features with older statistical tools, simply because clarity matters more than complexity [9].

Some tools - like autoencoders or prior-trained networks - help shrink data size while keeping useful patterns intact. Transfer learning steps in when data gets sparse, especially with things like gene readings or medical scans. Instead of packing every detail, these systems highlight what matters most. Feature picking becomes key under such conditions. Methods such as LASSO filter out noise by favoring stable signals across samples. Stability selection does this more carefully, checking consistency over many tries. This cuts down clutter without losing meaning. Less guesswork happens later because inputs are cleaner upfront.

Most times, statistical models help draw conclusions from data. When questions about cause come up, these methods dig into what drives change. Checking if a model works well often means testing it against real results.

Important examples of statistical models include these

Estimation and Hypothesis Testing

Even today, tools like GLMs, Cox models, or mixed-effects setups remain key whenever research aims to measure effects, check hypotheses, or handle influencing factors. Getting the structure right, validating choices, then sharing how sure you are matters deeply in biology [3].

Loose Takeaways from Real World Health Watch

Most health data sits out there, yet making sense of it isn't straightforward. Figuring out if a treatment works often leans on tools like propensity scores - though these rely

heavily on unproven guesses layered with modeling tricks. Some approaches juggle inverse probability weights or lean on instrumental variables, each demanding close coordination with specialists. Thoughtful design matters, so does checking results under shifting conditions [10].

Uncertainty and Calibration in Clinical Practice

Most tools used in medicine need clear ranges showing how sure they can be about results. Just looking at how well a model tells groups apart isn't enough on its own. Without checks like accuracy plots or range estimates, trust drops fast. Outside testing matters just as much - research shows that skipping it risks real-world failure [11].

Medical Research Applications

Patterns, along with gene groups sharing activity levels, emerge through data mining in individual cell measurements. Following that, models based on statistics judge how strong certain impacts might be, weighing ideas about living systems while measuring links - like changes in gene output, corrected for repeated checks. When strict number tests join forces with smart algorithms spotting traits, results grow more trustworthy. These combined methods lead to firmer conclusions in biology [9].

From messy clinical notes, hidden patterns emerge when natural language tools sift through over time. Following this step, models calculate risk levels while measuring how much confidence sits behind each result. Because of these layers, drawing reliable conclusions becomes possible without controlled trials. Tools built this way support real-world research by turning raw text into structured insights. Uncertainty stays visible instead of ignored along the journey. Each piece feeds into forecasts that hold up under scrutiny. Picture-based data sets help forecast results, divide body parts, or spot injuries through deep learning. When models mix scan findings with patient details, they adjust predictions - weighing each factor carefully - to match real medical choices. For these systems to pass review boards, number checks and clear reasoning must back every step. Proof methods show how trustworthy the outputs really are.

Piles of possible compounds come from chemical collections, combined biological data, plus fast testing methods - these get sorted using pattern-finding tools. Once flagged, those picks face number-based checks, dose tests, and repeated trials to pin down real impact. When tight filters meet smart grouping tricks - like sorting top hits - the outcomes often surprise, showing just how much quicker answers can arrive [9].

Though numbers help spot trends, predict outcomes, or measure what might have happened without treatment - like how many people survived thanks to medicine - digging through raw data reveals unusual groups of cases. Put together, these methods let health teams get support fast when outbreaks strike.

Data Mining Challenges

One reason studies fail to match results? Poor tracking of how data was handled. When code updates get logged properly, it helps others follow along. Running analyses inside isolated digital spaces keeps conditions stable across tries. If datasets can be shared openly, that openness supports verification by peers. Explaining each cleaning step and model choice clearly adds clarity. Lately experts proposed fresh blueprints aimed at building trustworthy health software tools [12].

When data does not reflect real-world diversity, health gaps can grow worse. Before rolling out systems, checks for fairness plus testing across different groups become necessary [13].

Some doctors feel uncertain about black-box models. Tools like SHAP, LIME, or Saliency maps might help - yet they need careful handling. Even so, understanding how a model works does not take the place of testing it in real clinical settings [14].

Stopping some data from moving around helps keep patient details private under HIPAA and GDPR. Even though methods like shared model training and fake datasets show promise, they come with hurdles in testing and reliability. Government groups now issue more rules about using artificial intelligence and machine learning in health tools. People building these systems have to meet standards for real-world tests, handling risks, and watching performance after release [14]. When setting up a study, figure out the main medical question, who it affects, what results matter, then pick how to judge success - like accuracy scores, probability matching, or impact on choices.

From the start, keeping track of where data comes from matters just as much as checking how well it works. Quality stays high because strict checks happen at every step. Missing pieces get addressed fairly - often filled in thoughtfully when methods like multiple imputation fit best. Consistency shows up clearly in how codes are applied across the board [7].

Pre-trained models come first - include them alongside your code, random seeds, and exact setup. Reporting? Stick to standards like TRIPOD or CONSORT-AI where they fit; toss in notebooks and containers too [12]. Performance splits by age, gender, ethnicity pop up next - look closely there. Bias lurking around? Apply fixes, track fairness metrics if needed [13].

Checking how well a product works during real medical use begins with designing around patient needs. Yet feedback keeps shaping changes even after launch. Though testing starts early, learning never stops once live. Because watching actual performance reveals what trials miss. So ongoing

observation supports better outcomes over time.
Right now, the path forward for biomedical data analysis is taking shape. Instead of just predicting outcomes, models mix machine learning with cause-and-effect reasoning to better understand treatments.
These systems learn patterns while also tracing biological mechanisms behind them. Alongside this, managing how models work in hospitals means using MLOps practices - like tracking changes, spotting data shifts, and keeping versions updated. Tools that monitor performance over time help keep algorithms useful in actual medical settings. This blend supports longer-term reliability without constant manual oversight. Progress continues through tighter feedback loops between clinics and computational pipelines.

Conclusion

Nowhere else has change been sharper than in biomedical research, thanks to data mining paired with statistical models sparking fresh insights grounded in evidence. Progress in genetics ties closely to gains in clinical informatics, linked through advances in imaging alongside strides in drug design. A steady path forward means leaning on data mining to uncover patterns while using statistics to confirm them. Without both working together, trustworthiness, ethics, and consistency could slip away.

Acknowledgement

The authors express their sincere gratitude to the Department of Medical Laboratory Science, for providing institutional support internet facilities throughout the study.

Author contributions:

Concept and study design: Obasi Miriam Oluchi,

Manuscript drafting: Emereole Chinwe Nelly,

Critical revision of manuscript: Iwuji Joy Chidinma and Okafor Onyeka Edmond

Final approval of manuscript: All authors

Funding: No funding.

Conflict of interest: The author declared no conflict of interest.

Ethics approval: The protocol adhered to the tenets of the Declaration of Helsinki and was approved by the Department of Medical Laboratory Ethical Committee.

AI tool usage declaration: No AI tool was used in manuscript preparation.

References

1. Ajibade, Samuel-Soma M., Gloria Nnadwa Alhassan, Abdelhamid Zaidi, Olukayode Ayodele Oki, Joseph Bamidele Awotunde, Emeka Ogbuju, and Kayode A. Akintoye. “Evolution of machine learning applications in medical and healthcare analytics research: A bibliometric analysis.” Intelligent Systems with Applications 24 (2024): 200441.
2. Streiber, Anna Maria, Sanne JW Hoepel, Elisabet Blok, Frank JA Van Rooij, Julia Neitzel, Jeremy Labrecque, M. Kamram Ikram, and Daniel Bos. “Improving reproducibility of data analysis and code in medical research: 5 recommendations to get started.” BMJ open 15, no. 10 (2025): e104691.
3. Olson, David L. “Data mining in business services.” Service Business 1, no. 3 (2007): 181-193.
4. Matsui, Shigeyuki, Jennifer Le-Rademacher, and Sumithra J. Mandrekar. “Statistical models in clinical studies.” Journal of Thoracic Oncology 16, no. 5 (2021): 734-739.
5. Pickard, Joshua, Victoria E. Sturgess, Katherine O. McDonald, Nicholas Rossiter, Kelly B. Arnold, Yatrik M. Shah, Indika Rajapakse, and Daniel A. Beard. “A Hands-On Introduction to Data Analytics for Biomedical Research.” Function 6, no. 2 (2025): zqaf015.
6. Admassu, Bisrat Yeshewas, Malwina Hołownia, Katarzyna Kędzior, Jean-Etienne Poirrier, and Stefano Perni. “Systematic Reviews of Machine Learning in Healthcare: A Literature Review.” (2023).
7. Pinto da Costa, Joaquim Fernando, and Manuel Cabral. “Statistical methods with applications in data mining: A review of the most recent works.” Mathematics 10, no. 6 (2022): 993.
8. Yang, Xue, Kexin Huang, Dewei Yang, Weiling Zhao, and Xiaobo Zhou. “Biomedical big data technologies, applications, and challenges for precision medicine: a review.” Global Challenges 8, no. 1 (2024): 2300163.
9. Daoud, Adel, and Devdatt Dubhashi. “Statistical modeling: the three cultures.” Harvard Data Science Review 5, no. 1 (2023).
10. Preti, Luigi M., Vittoria Ardito, Amelia Compagni, Francesco Petracca, and Giulia Cappellaro. “Implementation of machine learning applications in health care organizations: systematic review of empirical studies.” Journal of medical Internet research 26 (2024): e55897.
11. Sukri, N.F.A.A., Fazamin, A., Yusof, K., Ismail, I., Yusoff, H.M. and Yacob, A. (2025) ‘A systematic literature review on machine learning in healthcare prediction’, International Journal of Online and Biomedical Engineering (iJOE), 21(6), pp. 155–177.
12. Patharkar, Abhidnya, Fulin Cai, Firas Al-Hindawi, and Teresa Wu. “Predictive modeling of biomedical temporal data in healthcare applications: review and future directions.” Frontiers in physiology 15 (2024): 1386760.
13. Završnik, Jernej, Peter Kokol, Bojan Žlahtič, and Helena Blažun Vošner. “Machine Learning in Primary Health Care: The Research Landscape.” In Healthcare, vol. 13, no. 13, p. 1629. MDPI, 2025.

14. Ghofrani, Alireza, and Hamed Taherdoost. “Biomedical data analytics for better patient outcomes.” Drug Discovery Today 30, no. 2 (2025): 104280.
15. Polce, Evan M., and Kyle N. Kunze. “A guide for the application of statistics in biomedical studies concerning machine learning and artificial intelligence.” Arthroscopy: The Journal of Arthroscopic & Related Surgery 39, no. 2 (2023): 151-158.

Citation: Nnodim Johnkennedy, Data Mining and Statistical Modelling In Biomedical Research: Concepts, Applications, and Future Perspectives, *Biomed. Biotechnol.Sci.* Vol. 2 Iss. 1. (2026) DOI: 10.58489/2833-0951.011