The new frontier of statistics: Modern machine learning approaches as alternatives to traditional statistical tests in biological, clinical, and epidemiological research with a focus on cardiac event prediction

***A Wentzel*[1,2] *and M Blignaut*[3]**

[1] Hypertension in Africa Research Team, North-West University, Potchefstroom, South Africa
[2] South African Medical Research Council, Unit for Hypertension and Cardiovascular Disease, North-West University, Potchefstroom, South Africa
[3] Centre for Cardio-Metabolic Research in Africa, Division of Medical Physiology, Department of Biomedical Sciences, Stellenbosch University, Tygerberg, South Africa

Address for correspondence:
M Blignaut
Division of Medical Physiology
Centre for Cardio-Metabolic Research in Africa (CARMA)
Biomedical Research Institute
Faculty of Medicine and Health Sciences
Level 3, Room 3023
Tygerberg Campus
Stellenbosch University
Stellenbosch
South Africa
7599

Email:
mblignaut@sun.ac.za

A Wentzel https://orcid.org/0000-0002-3765-2040
M Blignaut https://orcid.org/0000-0001-7645-9780

DOI: https://doi.org/10.24170/23-1-7883

ABSTRACT

As the complexity and volume of biological and clinical data increase, traditional statistical methods, such as logistic regression, discriminant analysis, analysis of variance (ANOVA), and multivariate analysis, often fall short of capturing the intricate patterns needed for accurate prediction and classification. Here, we explore alternative analytical frameworks rooted in modern machine learning (ML) techniques that offer enhanced capabilities for diverse biomedical applications. For example, these frameworks demonstrate superior predictive performance for cardiac events compared with classical logistic regression. However, challenges, interpretability, and future directions are important considerations when facing this new frontier. Moreover, systematically integrating these advanced computational tools into routine clinical and epidemiological research is imperative. This co-authored column forms part of the "Statistics Series" and builds on *A simple guide to analyse data* by Prof. Libhaber.[1]

Keywords: machine learning, neural networks, deep learning, high-dimensional biological data

SA Heart® 2026;23:35-41

WHAT ARE THE CHANCES?

At the heart of statistics lies the need to accurately estimate the probability that a given event, feature, or phenotype falls within or outside the expected range. Normality in biomedicine is traditionally defined as the likelihood that an observation falls within an expected range, based on sample data and probability models.[2] For example, the effect of a new treatment or therapy applied to the general population can be inferred by measuring the outcomes in a small subset of the population, and testing the null hypothesis that there will be no effect in the population. This hypothesis is tested with mathematical models based on certain features of the dataset that quantify the probability (if the study were repeated many times) that the same treatment or therapy effect observed in the study sample will be detected to the same extent in the larger, general population.

Herein lies the key difference between traditional statistical methods and ML: both inference and probability are based on mathematical models, whilst ML uses algorithms to identify patterns within existing data to make predictions about other datasets. Although traditional statistical tools, such as t-tests, simple correlations, and standard regression, remain essential for hypothesis-driven analyses, effect estimation, and transparent inference, they often assume linear relationships (never the case in biology), limited interactions, and relatively simple datasets with few features or attributes (referred to as low-dimensional datasets in ML).[3,4]

However, contemporary biological and clinical datasets are often large and high-dimensional, with numerous variables or features, where each feature becomes a dimension. These datasets can also be heterogeneous, combining omics, imaging, continuous physiological monitoring, and longitudinal electronic health record data, where the true relationships are likely nonlinear, interactive, and highly context-dependent.[5] ML methods, including neural networks (NN), tree-based ensembles, and kernel methods (see "Definitions"), are specifically designed to model such complex structures, leading to more accurate predictions and enabling the discovery of clinically relevant patterns that may not be apparent using traditional methods alone.[5,6]

DEFINITIONS OF MACHINE LEARNING (ML)

ML is a set of computational methods that actively learn and identify patterns from data to make predictions, connections, or uncover structure, without explicitly programmed, fixed decision rules. These models are trained on clinical data, including laboratory results, imaging, waveforms, and health records, to predict outcomes like disease presence, deterioration risk, and treatment response.[6] Unlike the classical inferential statistics discussed in the previous section, ML emphasises predictive performance and flexibility, allowing correlated variables and complex, multidirectional relationships to be considered simultaneously.[7]

Neural networks and deep learning

Artificial NNs are a widely used subset of ML methods inspired by how biological neurons process signals, consisting of layers of interconnected "nodes" that transform input data using learned weights and nonlinear activation functions.[8,9] When many layers are stacked, the resulting deep NNs (deep learning) can automatically learn increasingly abstract, multidirectional representations and relationships from complex, raw data without manual feature engineering.[7,10] Consequently, deep learning has exceeded expert-level performance in tasks such as medical image classification, disease detection on radiographs, and protein structure prediction, demonstrating its capacity to handle high-dimensional, unstructured biomedical data.

WHEN AND HOW TO APPLY MACHINE LEARNING IN BIOMEDICINE

ML is particularly appropriate when the primary goal is prediction, pattern recognition, or classification (predicting disease risk, identifying biological marker patterns, stratifying patients, or detecting pathology in images), especially in the presence of many predictors, potential nonlinearities, and interactions.[7,8] It is well-suited to high-dimensional omics data, medical imaging, waveform data (electrocardiography [ECG], electroencephalogram, continuous blood pressure), and rich electronic health record datasets, where traditional models may overfit or fail to capture structure, provided that sufficient sample size, careful validation, and appropriate regularisation are used.

However, when the main objective is to estimate interpretable effect sizes, test specific mechanistic hypotheses, or communicate simple associations, conventional regression and related statistical models remain preferable. In these cases, ML can be used as a complementary tool, applied initially to optimise traditional methods, rather than as a replacement. Table I provides an extensive summary of traditional statistical methods, their uses, the ML alternative, and examples in biomedical applications. For instance, a systematic review of the application of different ML approaches to analyse ECG data found that ECG deep-learning models are increasingly clinically relevant; however, their reporting is highly variable, and few publications provide sufficient detail for methodological reproduction or model validation by external groups.[11]

TYPES OF MACHINE LEARNING TASKS AND DATA

Most biomedical ML applications fall into 3 broad paradigms: (1) supervised learning, where models are trained on labelled outcomes (e.g. disease vs. no disease) to perform prediction or classification; (2) unsupervised learning, which discovers structure, such as clusters or latent patterns in unlabelled data; and (3) semi- or self-supervised approaches that leverage both labelled and unlabelled data.[8] Supervised methods are widely used for diagnostic and prognostic models, mortality risk prediction, and treatment response modelling, while unsupervised methods underpin patient subtyping, endotype discovery, and exploratory analysis of high-dimensional biological measurements.[5,6] Deep learning extends these paradigms to unstructured data, such as images, free text, and raw signals, allowing direct modelling from pixels, narrative notes, dimensional data, or waveforms when sufficient data and computational resources are available.

INTEGRATING MACHINE LEARNING IN CLINICAL PRACTICE AND CLASSIC STATISTICS

For clinical researchers and clinicians, the key is not to abandon traditional statistics, but to integrate ML and NN methods where they add clear value, such as improving risk stratification, automating image interpretation, or identifying novel patient subgroups. Rigour remains essential; model development should include careful preprocessing, transparent variable selection, appropriate cross-validation or external validation, and attention to calibration, fairness, and interpretability, particularly when models influence patient care.[6,10] As biomedical data become more complex and abundant, ML and NNs provide a powerful, complementary, informative toolkit that can augment and provide richness, rather than replace established statistical approaches, setting the stage for more precise, data-driven, and individualised medicine. There are already many easily accessible tools available, many of which are free and open source (such as Python-based libraries). However, their use depends on experience, coding capabilities, dataset type, and available hardware (Figure 1 illustrates a broad overview of the different ML tools and their applications).

Nonetheless, it is important to note that traditional statistical adjustments are often required when using ML and NNs to ensure reliable performance, generalisation, and interpretability, particularly in biomedicine. Moreover, incorporating more traditional statistical techniques may also address ML's sensitivity to data quality by focusing on issues such as overfitting through

TABLE I: Comparing traditional statistics to machine learning alternatives.[13-20]

Traditional method and aim	Typical biomedical use	ML/NN alternative for a similar aim	When the ML alternative is useful	Example biomedical application*
Two-sample t-test (continuous outcome, 2 groups for univariate analysis), multiple linear regression for continuous outcomes (if normally distributed)	Compare the mean of a biomarker or physiological measure between 2 groups (biomarker levels in cases vs. controls).	**Regularised regression (e.g. Lasso/ridge), tree-based models (e.g. random forest, gradient boosting), including group indicator plus additional covariates.**	When there are many correlated biomarkers or covariates, and the goal is to predict group membership or an outcome (rather than only test the mean difference), and to capture nonlinear relationships.	Predicting tuberculosis treatment failure using multiple clinical and demographic variables where simple mean differences are insufficient; random forest or NNs can model complex risk patterns.
One-way ANOVA (continuous outcome, multiple groups)	Compare mean outcome across several treatment or exposure groups (e.g. comparing mean blood pressure across 3 drug regimens).	**Supervised learning models with categorical group variables plus covariates (e.g. gradient-boosted trees, random forest, NNs) predicting continuous outcomes.**	When interest is in predicting the outcome under different treatment or exposure conditions, while incorporating many patient-level features, and when interactions and nonlinear dose–response relationships may exist.	Modelling systolic blood pressure response to different antihypertensive regimens using numerous baseline characteristics with gradient boosting to identify subgroups with the largest benefit.
Pearson/Spearman correlation (pairwise association) or a partial correlation	Quantify the association between 2 continuous variables (e.g. CRP and disease severity score) but does not account for the bidirectionality of biological systems.	**Nonlinear regression, kernel methods (e.g. support vector regression), or flexible feature importance measures from tree-based models.**	When relationships are suspected to be nonlinear or involve interactions with other features, ML models can estimate variable importance and partial dependence instead of a single correlation coefficient.	Exploring complex associations between continuous glucose monitor metrics and cardiovascular risk markers, using random forest to capture nonlinear effects rather than a single correlation per pair.
Simple/multiple linear regression (continuous outcome)	Model a continuous clinical outcome (e.g. lung function, ejection fraction) as a function of several predictors with interpretable coefficients.	**Regularised regression (Lasso, elastic net), random forest regression, gradient boosting machines, or feedforward NNs.**	When there are many predictors, multicollinearity, or nonlinear effects and interactions, ML can improve prediction and automatically select or weight variables while controlling overfitting.	Predicting heart disease severity or exercise capacity from numerous clinical, lab, and imaging features with random forest regression often achieves better predictive performance than standard linear models.
Logistic regression (binary outcome)	Model probability of an event (e.g. disease presence, treatment failure) as a function of predictors, with odds ratios for interpretability.	**Tree-based classifiers (random forest, gradient boosting), support vector machines, or deep NNs.**	When accurate classification or risk stratification is prioritised, particularly with many variables, nonlinear interactions, or complex feature sets (e.g. combined clinical and laboratory data).	Predicting risk of treatment failure in tuberculosis or cardiovascular disease using multiple demographic, clinical, and lab features; studies show that random forests or gradient boosting can outperform logistic regression in some datasets for discrimination metrics.
Multinomial/ordinal logistic regression (multi-class outcomes)	Model categorical outcomes with more than 2 levels (e.g. disease stage I–IV, NYHA class).	**Multi-class random forests, gradient-boosted trees (e.g. XGBoost), multi-class support vector machines, or multi-class NNs.**	When there are high-dimensional predictors and complex decision boundaries between classes, or when using heterogeneous inputs (e.g. imaging plus tabular data).	Classifying heart failure stage or severity class from combined EHR data using gradient-boosted trees to optimise multi-class discrimination beyond a parametric ordinal model.
Cox proportional hazards regression (time-to-event outcome)	Model hazard of events (e.g. time to death, time to readmission) using covariates, providing hazard ratios and survival curves.	**Random survival forests, gradient boosting for survival (e.g. survival XGBoost), and survival NNs (e.g. DeepSurv).**	When proportional hazards or linear effects may be violated, when many predictors and nonlinearities are present, or when prediction of individualised risk trajectories is prioritised over simple hazard ratios.	Predicting breast cancer survival or lung cancer prognosis using high-dimensional clinical and molecular predictors with random survival forests or survival gradient boosting, sometimes outperforming classical Cox models in discrimination.

TABLE I: Continued

Chi-squared test/Fisher's exact test (categorical association)	Test association between 2 categorical variables (e.g. genotype vs. disease status, treatment vs. response categories) in contingency tables.	**Supervised classifiers (e.g. random forest, gradient boosting, naive Bayes) using categorical predictors to model outcome probabilities directly.**	When there are multiple categorical predictors and their interactions matter for outcome prediction, rather than simply testing the independence of a single pair.	Predicting antibiotic resistance profile or treatment response category from multiple categorical predictors (e.g. pathogen type, prior exposure, comorbidities) using gradient boosting rather than separate chi-squared tests.
Principal components analysis for dimension reduction	Reduce dimensionality of correlated continuous variables (e.g. many metabolic markers) to a smaller set of uncorrelated components for visualisation or downstream modelling.	**Nonlinear dimension reduction, such as autoencoders (NNs), t-SNE, or UMAP (although not all are predictive models).**	When the underlying structure is believed to be nonlinear or manifold-like, and the aim is to discover latent patterns or clusters in high-dimensional data (e.g. omics, imaging-derived features).	Discovering patient subtypes in multi-omics cancer data using autoencoders to learn low-dimensional representations, then clustering patients to identify molecularly distinct disease phenotypes.
Cluster analysis (k-means, hierarchical clustering)	Unsupervised grouping of patients or features based on similarity, often to identify phenotypes or subgroups without outcome labels.	**Model-based clustering (Gaussian mixture models), density-based clustering (DBSCAN), or deep clustering approaches (IDEC) that couple NNs (auto-encoders) with clustering objectives.**	When clusters may be non-spherical, overlapping, or embedded in high-dimensional nonlinear spaces, a richer structure is expected than can be captured by distance-based methods alone.	Phenomapping in heart failure or sepsis, using high-dimensional clinical and biomarker data with advanced ML clustering methods to identify clinically meaningful subgroups that differ in prognosis or treatment response.
Repeated-measures ANOVA/linear mixed models (longitudinal continuous outcomes)	Analyse trajectories over time (e.g. repeated blood pressure or biomarker measures) and test group or time effects with random effects for subjects.	**Recurrent NNs, temporal convolutional networks, or sequence models (e.g. transformers) for time series; random forest or boosting with engineered longitudinal features.**	When temporal patterns are complex, sampling is irregular, or large-scale time series from wearables or ICUs are available, and the aim is to predict future events or detect deterioration rather than only test mean trajectory differences.	Early prediction of sepsis or decompensation in ICU patients using multichannel vital sign time series with recurrent or convolutional NNs, enabling continuous risk scoring beyond traditional mixed-model analyses.

** Artificial intelligence (Perplexity) was used to identify biomedical application examples (listed in the last column) in literature (cited at the top).*
ANOVA: analysis of variance, CRP: C-reactive protein, DBSCAN: Density-Based Spatial Clustering of Applications with Noise, EHR: electronic health record, ICU: intensive care unit, IDEC: Improved Deep Embedded Clustering, ML: machine learning, NN: neural network, NYHA: New York Heart Association Functional Classification, t-SNE: t-distributed stochastic neighbour embedding, UMAP: uniform manifold approximation and projection.

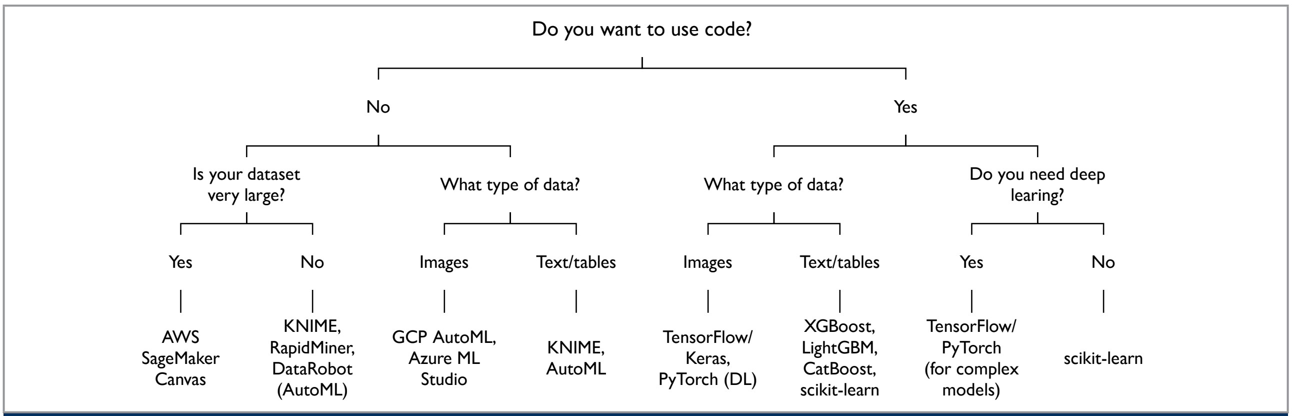
Do you want to use code?
No
Yes
Is your dataset very large?
What type of data?
What type of data?
Do you need deep learing?
Yes
No
Images
Text/tables
Images
Text/tables
Yes
No
AWS SageMaker Canvas
KNIME, RapidMiner, DataRobot (AutoML)
GCP AutoML, Azure ML Studio
KNIME, AutoML
TensorFlow/ Keras, PyTorch (DL)
XGBoost, LightGBM, CatBoost, scikit-learn
TensorFlow/ PyTorch (for complex models)
scikit-learn

FIGURE 1: This schematic (generated with ChatGPT using specific prompts to identify the most used AI tools, followed by a prompt to organise the data according to the ability to code) shows the types of ML tools that can be used based on data type and coding experience. Amazon (AWS) SageMaker Canvas is a no-code ML service for data processing and model building. Scikit-learn works very well for tabular biomedical data (e.g. clinical variables, lab tests, biomarker panels), whilst XGBoost/LightGBM is best for high-performance models on structured biomedical datasets. For DL purposes, TensorFlow with Keras is very good, with a user-friendly interface. PyTorch offers greater flexibility and is most commonly used in academic environments to build custom biomedical DL architectures. KNIME is often used as a data analytics, reporting, and integration platform, integrating various components for ML and data mining through its modular data pipeline.
AutoML: automated machine learning, AI: artificial intelligence, DL: deep learning, ML: machine learning.

TABLE II: Pros and cons of machine learning in biomedical research.[13,15,18-20]

Dimension	Pros (advantages)	Cons (limitations/risks)
Data complexity and scale	Can handle high-dimensional, heterogeneous data (imaging, omics, waveforms, EHR) without requiring simple, prespecified linear relationships, making it well-suited to modern, multimodal biomedical datasets that are complex, nonlinear, and often contain many correlated predictors.	Strongly dependent on data quality, completeness, and representativeness; noisy labels, missingness, site effects, and small or biased samples can severely degrade performance and undermine external validity, especially when training data do not reflect the target population.
Prediction and pattern recognition	Often achieves superior performance to traditional models for tasks such as disease risk prediction, image-based diagnosis, and outcome forecasting when datasets are sufficiently large and well-curated, detecting subtle patterns (e.g. in retinal or radiology images) that are difficult for humans or simple models to capture.	Prone to overfitting if model complexity is high relative to the effective sample size, if feature engineering is careless, or if validation is weak; internal metrics can appear excellent while true generalisation to new settings or hospitals is poor, leading to overly optimistic published results.
Personalisation and precision	Enables more granular risk stratification, identification of high-risk subgroups, and personalised treatment strategies by integrating many clinical, biological, and behavioural predictors, supporting precision medicine and biomarker-based trial design or response prediction.	Personalisation can exacerbate bias if subgroup models are derived from small or unbalanced strata; models may encode and amplify structural inequities (e.g. underrepresented ethnic groups or sexes) unless fairness and subgroup performance are explicitly evaluated and corrected.
Efficiency and automation	Once trained and validated, models can process large data volumes rapidly, automating repetitive tasks such as image triage, signal screening, or EHR-based early warning scores, potentially freeing clinicians and researchers to focus on interpretation, study design, and patient-facing decisions.	Automation can encourage over-reliance on algorithmic outputs, with the risk that clinicians or researchers defer to model predictions without adequate scrutiny; poorly integrated tools can increase workload or alert fatigue rather than reduce it, and errors may propagate at scale.
Discovery and hypothesis generation	Facilitates data-driven discovery of new patterns, phenotypes, and interactions (e.g. phenomapping, unsupervised clustering of omics or imaging), suggesting novel hypotheses, biomarkers, or mechanistic leads that can be tested with targeted experiments or classical statistical models.	Data-driven discoveries can be difficult to interpret mechanistically, and spurious clusters or associations are common when multiple testing and validation issues are not handled rigorously; there is a risk of "pattern hunting" without sufficient biological grounding or prespecified questions.
Interpretability and transparency	Some ML methods (e.g. regularised linear models, shallow trees, generalised additive models, or post-hoc explainability tools) can provide variable importance, partial dependence, and other insights that complement traditional effect estimates, aiding understanding of complex relationships.	Many high-performing models, especially deep NNs, behave as "black boxes" with limited transparency about how predictions are generated, which complicates mechanistic understanding, communication with clinicians and regulators, and formal adoption into guidelines or decision pathways.
Data access, privacy, and governance	Can encourage the development of high-quality, well-curated research datasets and data infrastructures, and motivate federated or privacy-preserving methods that analyse distributed data without centralising identifiable information.	Requires access to large, often linked patient-level datasets, raising substantial issues around consent, privacy, re-identification risk, data ownership, and security; regulatory and ethical constraints may limit data sharing and multicentre validation, reducing reproducibility and generalisability.
Technical and resource demands	Stimulates multidisciplinary collaboration between clinicians, statisticians, computer scientists, and engineers; open-source tools and pre-trained models (e.g. for imaging or NLP) can lower entry barriers for research groups.	Robust model development and deployment demand specialised expertise in data engineering, ML, and software practices; deep learning can be computationally expensive, requiring substantial hardware, maintenance, and monitoring infrastructure that not all groups possess.
Evaluation, validation, and reproducibility	When done well, rigorous cross-validation, temporal validation, and external validation across sites can yield models with strong, well-characterised generalisation performance and well-calibrated risk estimates that stand up to prospective testing.	Many published biomedical ML studies use small datasets, weak validation, optimistic performance metrics, and incomplete reporting, hindering reproducibility and inflating expectations; code, trained models, and data are often not fully shared, limiting independent verification and re-use.
Workflow and culture	Offers potential to streamline research pipelines (e.g. automated feature extraction from images, structured data from free text) and clinical workflows (e.g. triage, prioritisation), and can augment human expertise in a complementary way.	Integration into existing research and clinical workflows can be challenging, requiring changes in processes, training, and culture; scepticism from clinicians, concerns about medico-legal liability, and misalignment with clinical priorities can slow or block adoption, even for technically strong models.

EHR: electronic health record, ML: machine learning, NN: neural network.

TABLE III: When to use traditional methods versus machine learning.

Aspect	Traditional statistical methods (e.g. t-test, linear/logistic/Cox regression, correlations)	ML/NNs (including deep learning)
Primary goal	Estimate and test associations or effects (e.g. exposure–outcome relationships, hazard ratios), with emphasis on inference and interpretability.	Maximise predictive accuracy or pattern recognition (e.g. classify, risk-stratify, detect structure), often with less emphasis on explicit parameter interpretation.
Typical assumptions	Prespecified model form (often linear), limited interactions, relatively low number of predictors versus sample size, and structured noise assumptions (e.g. normality, proportional hazards).	Flexible, data-driven function classes that can capture nonlinearity and complex interactions, with fewer parametric assumptions but a stronger need for regularisation and validation.
Data size and dimensionality	Best when the number of observations is much larger than the number of predictors, and variables are carefully selected a priori (e.g. classical cohort studies, public health surveys).	Particularly advantageous for high-dimensional data (e.g. genomics, radiomics, multi-omics, rich EHR data) where the number of predictors can rival or exceed the sample size.
Data type	Structured, tabular data with clearly defined variables (e.g. age, blood pressure, lab values, questionnaire scores).	Can handle both structured and unstructured data, such as images, free text, waveforms, and sensor streams, often directly from raw inputs (pixels, time series).
Example: prognosis/risk prediction	Cox regression model to estimate hazard ratios for mortality in a cardiovascular cohort using a small panel of risk factors (age, blood pressure, cholesterol, smoking) and to quantify their independent effects.	Gradient boosting or deep learning model using dozens to hundreds of variables from EHRs to predict in-hospital deterioration or 30-day mortality, focusing on accurate risk stratification rather than individual effect sizes.
Example: diagnostic classification (tabular data)	Logistic regression using a limited set of clinical variables (e.g. body mass index, fasting glucose, blood pressure) to estimate odds of metabolic syndrome and test specific risk factor hypotheses.	Random forest or support vector machine using a richer set of features (e.g. labs, vitals, comorbidities, medication history) to classify patients at high risk of developing diabetes or metabolic syndrome, optimised for sensitivity/specificity.
Example: medical imaging	Linear measurements and simple thresholds (e.g. lesion size, ejection fraction) analysed with standard statistics to compare groups or assess associations with outcomes.	Convolutional NN trained on large sets of labelled computed tomography, magnetic resonance imaging, or X-ray images to detect lung nodules or classify tumours, often achieving radiologist-level accuracy in identifying malignancy.
Example: omics/ high-throughput biology	Multiple regression or univariate testing with correction for multiple comparisons to relate a small subset of preselected genes or proteins to an outcome, mainly for hypothesis-driven analysis.	Regularised models, tree-based ensembles, or deep learning applied to genome-wide or proteomic profiles to predict drug response or discover molecular subtypes of cancer or other diseases.
Example: longitudinal/ monitoring data	Mixed-effects models or repeated-measures ANOVA to test average trajectories over time (e.g. HbA1c or blood pressure trends) and assess group differences with interpretable coefficients.	Recurrent or temporal convolutional NNs trained on continuous wearables or ICU monitoring data (e.g. heart rate, rhythm, glucose sensors) to detect early signs of sepsis, arrhythmia, or decompensation in real time.
Strengths	Transparent modelling, explicit effect estimates (odds ratios, hazard ratios), strong theory for inference and uncertainty quantification, easier to audit and communicate to clinicians and regulators.	Captures complex nonlinear patterns and interactions, scales to large and heterogeneous datasets, excels at prediction and pattern discovery, and can directly ingest raw or minimally processed data.
Limitations	May underperform when relationships are nonlinear or highly interactive, or when many correlated predictors are present; performance can degrade in very high-dimensional settings.	Models can be less interpretable, prone to overfitting without rigorous validation, and resource-intensive; performance advantages over well-specified traditional models are not guaranteed in small or simple datasets.
When to prefer	Hypothesis-driven work where quantifying and testing specific associations is primary, datasets are moderate in size and dimensionality, and interpretability is paramount (e.g. guideline development, mechanistic research).	Prediction- or classification-focused problems with complex, high-dimensional, or unstructured data, where improving accuracy, risk stratification, or pattern discovery is central (e.g. imaging artificial intelligence, multi-omics risk scores, EHR-based early warning systems).

ANOVA: analysis of variance, EHR: electronic health record, ICU: intensive care unit, ML: machine learning, NN: neural network.

regularisation, where all variables are taken into account, even though they might not have the same effect (which is accounted for with Lasso and ridge regression in traditional statistics and adjusted R^2 metrics to penalise unnecessary complexity).[11] Similarly, cross-validation and bootstrapping used in traditional statistics provide confidence intervals for model predictions, mitigating variance in high-dimensional biological data, like multi-omics. At the same time, feature selection via principal component analysis or elastic nets reduces noise in heterogeneous datasets, thereby improving representativeness before NN training. For example, batch effect correction using variance analysis is essential to adjust for technical heterogeneity in large sequencing datasets, enabling accurate cell clustering with deep learning models.

PRACTICAL IMPLICATION FOR BIOMEDICAL RESEARCH

When grounded in high-quality data, rigorous validation, and well-posed clinical or biological questions, ML can markedly improve discovery, prediction, and personalisation in biomedicine. Simultaneously, it must be integrated with domain expertise and conventional statistical reasoning, rather than treated as a stand-alone solution. ML excels at modelling complex, high-dimensional heterogeneous data and can enhance disease prediction, image-based diagnosis, patient stratification, and biomarker discovery. Yet, its performance is highly dependent on data quality and representativeness, and many powerful models are opaque, limiting mechanistic insight and trust.[12] For example, the data quality of cross-sectional dataset tests used for drug efficacy prediction is of utmost importance, as batch effects and heterogeneity can decrease accuracy during training and affect real-world biomedical data.

Opaque models, like NNs, limit mechanistic insight. For instance, “black box” predictions in genomics data hinder trust and biological interpretability, prompting “visible ML” that incorporates pathways for transparency. Thus, as with traditional statistics, ensuring excellent initial data quality through rigorous preprocessing, batch effect correction, and representative sampling is paramount when applying NNs or ML techniques in biomedicine, as it directly bolsters model accuracy, generalisation to real-world scenarios, and overall trustworthiness beyond opaque predictions.

Biomedical ML also raises challenges around bias, generalisability, privacy, consent, and secure data infrastructure, and often requires specialist expertise in data engineering and model development. Without careful validation and governance, models are vulnerable to overfitting and overly optimistic performance claims, which remain major concerns in the current literature. Table II summarises the advantages and the limitations/risks of ML in biomedical research.

CONCLUSION

It is important to know when to use ML, and when traditional methods will suffice (summarised in Table III), while including the type of analysis that will be performed in the planning stages of the study, ensuring that the data complies with the analysis requirements, and improving the outcome and applications of biomedical studies, overall.

Conflict of interest: none declared.

REFERENCES

1. Libhaber E. A simple guide to analyse data: Descriptive statistics in quantitative research. SA Heart. 2025;22:188-90. https://doi.org/10.24170/22-03-7664.
2. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151-69. https://doi.org/10.1146/annurev.publhealth.23.100901.140546.
3. Jarantow SW, Pisors ED, Chiu ML. Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. Curr Protoc. 2023;e801. https://doi.org/10.1002/cpz1.801.
4. Tong, C. Statistical inference enables bad science; Statistical thinking enables good science. Am Stat. 2019;1305:246-61. https://doi.org/10.1080/00031305.2018.1518264.
5. Kufel J, Bargieł-Łączek K, Kocot S, et al. What is machine learning, artificial neural networks and deep learning? Examples of practical applications in medicine. Diagnostics. 2023;13(15):2582. https://doi.org/10.3390/diagnostics13152582.
6. Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: A review of state-of-the-art methods. Comput Biol Med. 2022;145:105458. https://doi.org/10.1016/j.compbiomed.2022.105458.
7. Yang S, Zhu F, Ling X, Liu Q, Zhao P. Intelligent health care: Applications of deep learning in computational medicine. Front Genet. 2021;12:1-21. https://doi.org/10.3389/fgene.2021.607471.
8. Weiss R, Karimijafarbigloo S, Roggenbuck D, Rödiger S. Applications of neural networks in biomedical data analysis. Biomedicines. 2022;10:1-30. https://doi.org/10.3390/biomedicines10071469.
9. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2016;16:741-58. https://doi.org/10.1002/pmic.201500396.
10. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9:1-12.
11. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser. 2019;1168:022022. https://doi.org/10.1088/1742-6596/1168/2/022022.
12. Avula V, Wu KC, Carrick RT. Clinical applications, methodology, and scientific reporting of electrocardiogram deep-learning Mmdels: A systematic review. JACC Adv. 2023.2. https://doi.org/10.1016/j.jacadv.2023.100686.
13. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics. 2018;19:270. https://doi.org/10.1186/s12859-018-2264-5.
14. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24:198-208. https://doi.org/10.1093/jamia/ocw042.
15. Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Meth. 2018;18:24. https://doi.org/10.1186/s12874-018-0482-1.
16. Hu Y, Zhang X, Slavin V, Belsti Y, Grove K. Beyond comparing machine learning and logistic regression in clinical prediction modelling: Shifting from model debate to data quality. J Med Internet Res. 2025;27:e77721. https://doi.org/10.2196/77721.
17. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Cham: Springer. 2019. https://doi.org/10.1007/978-3-030-16399-0.
18. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Dig Health. 2020;2:e179-e191. https://doi.org/10.1016/S2589-7500(20)30018-2.
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56. https://doi.org/10.1038/s41591-018-0300-7.
20. Van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Meth. 2014;14:13. https://doi.org/10.1186/1471-2288-14-137.