STATISTICS MADE EASY

Traditional statistics versus machine learning in clinical registries: A pragmatic workflow for matching methods to data and clinical questions

A Wentzel,[1,2] E Schaafsma[3] and M Blignaut[4]

[1] Hypertension in Africa Research Team (HART), 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] SHARE Registry Projects, South African Heart Association, Johannesburg, South Africa
[4] Centre for Cardio-Metabolic Research in Africa (CARMA), Division of Medical Physiology, Department of Biomedical Sciences, Stellenbosch University, Tygerberg, South Africa

Email:
annemarie.wentzel@nwu.ac.za

A Wentzel https://orcid.org/0000-0002-3765-2040
E Schaafsma https://orcid.org/0009-0003-4263-6883
M Blignaut https://orcid.org/0000-0001-7645-9780
DOI: https://doi.org/10.24170/23-2-8319

ABSTRACT

This piece discusses the importance of data type, identification, and organisation for machine learning (ML) and neural network (NN) development, and the applicability of ML for statistical analysis in large clinical and physiological datasets, such as the South African Heart Association Registry (SHARE).

Core outcomes/key lessons

To enable clinicians and researchers to:

- **Systematically assess their clinical dataset (registry data, e.g. SHARE) for variable types, dimensionality, sample size, missingness, and event rates.**
- **Understand when traditional statistical methods are sufficient, when regularised regression is preferable, and when more complex ML approaches are justified.**
- **Recognise common pitfalls (overfitting, multicollinearity, data leakage, mis-specified outcomes), and how to avoid them in both "classic" and ML settings.**
- **Apply a staged workflow to their own data, using the SHARE-transcatheter aortic valve implantation (TAVI) registry as an illustrative case.**

SA Heart® 2026;23:97-102

INTRODUCTION

Clinical and physiological datasets, such as national and institutional cardiovascular registries, are increasingly high-dimensional and heterogeneous, combining demographic, biometric, diagnostic, and laboratory data with longitudinal outcomes. In this context, there is growing enthusiasm for ML and NNs as powerful tools to complement traditional statistical approaches, particularly because they can, in principle, capture complex, nonlinear relationships and interactions when many covariates are present. We recently proposed using ML and NNs to augment the conventional analysis of clinical and physiological data, highlighting their potential to improve prediction and risk stratification in such multidimensional settings.[1] These developments have led to a perception that any large, clinically rich dataset is "ideal" for advanced ML or NN models.

However, in practice, the most appropriate analytic strategy depends critically on the data's structure and quality, as well as the clinical question being asked. The type of outcome (continuous, binary, time-to-event), the balance between sample size and the number of predictors, the degree of multicollinearity, the mix of categorical and continuous variables, missing data patterns, and the intended use of the model (prediction versus explanation) should be considered. All of the above influence whether traditional regression, regularised regression, or more complex ML methods are likely to be optimal. Simply applying an ML technique because a dataset is "large" or "multidimensional" risks overfitting, loss of interpretability, and, ultimately, limited clinical impact.

In this article, we investigate a general, practice-oriented perspective on an existing clinical database. Using a contemporary TAVI (Transcatheter Aortic Valve Implantation) registry as an exemplar, we suggest a structured way for clinicians and researchers to evaluate their data and align analytic choices with data characteristics and clinical goals.[2] Instead of focusing on the performance of one particular algorithm, we propose a pragmatic workflow that spans traditional statistical models, regularised regression, and more flexible ML approaches. We also discuss when each step along this continuum is likely to be appropriate. Overall, we aimed to provide a practical guide for clinicians, physiologists, and applied researchers working with registries and routinely collected clinical data. By navigating the considerations that arose in our own work, we illustrate how to move from an initial enthusiasm for "doing ML" to a more disciplined, question-driven, and data-aware selection of methods.

A MOTIVATING CASE: THE SHARE-TAVI REGISTRY

SHARE is a multi-center, prospective, web-based registry of TAVI procedures in South Africa. It captures detailed clinical

histories, demographics, comorbidities, procedural characteristics, TAVI outcomes, complications (classified according to Valve Academic Research Consortium [VARC] definitions), and follow up information, including 30-day and longer term outcomes. This rich, routinely collected dataset was established to inform health policy, promote equitable access to TAVI, and benchmark local practice against international standards.

TAVI is increasingly offered to older patients with multiple comorbidities, in whom post-procedural risk stratification remains challenging. Although procedural success rates in South African TAVI programmes are high, existing risk scores demonstrate only modest discrimination for post-TAVI mortality and other clinically relevant outcomes. Traditional analyses of registry data, typically based on linear or logistic regression, make only partial use of the available high-dimensional information and often assume linear relationships between predictors and outcomes. This has led to interest in applying more flexible approaches, including regularised regression and ML methods, to improve the prediction of outcomes, such as pre-TAVI aortic valve area and 1 year mortality.

Our original analytic plan for the SHARE reflected this enthusiasm: we aimed to develop and internally validate regularised regression models (ridge, lasso, elastic net) to predict pre-TAVI valve area and 1-year mortality, and to explore nonlinear relationships between demographic and clinical variables (including sex and comorbidities) and these outcomes. As we engaged more deeply with the data and the modelling challenges, it became clear that the key questions were not only "which algorithm performs best?", but more fundamentally "what does the structure of this dataset permit, and what is the most appropriate level of model complexity given our goals?"

Specifically, before choosing a method, one should systematically evaluate:

- The number of events available per candidate predictor for the outcome of interest.
- The extent of correlation and multicollinearity among predictors (e.g. between related echocardiographic or laboratory variables).
- How variables are measured and coded (binary, ordinal, continuous, raw versus transformed, derived scores).
- Whether the dataset is truly "big" in a modelling sense, with many more predictors than observations, or primarily "clinically rich" but of moderate size.

These considerations can lead to analytic choices that differ, and sometimes are simpler, than initially envisaged. In the case of SHARE, we propose a general, stepwise workflow for moving from traditional statistics to more advanced methods, while respecting both the data and the clinical questions at hand. This is complemented with practical recommendations for clinicians and researchers using registries (Table I).

Step 1: Know your data – the taxonomy of clinical datasets

Any analytic journey should begin with a clear understanding of the data at hand. Clinical registries like the SHARE-TAVI registry

TABLE I: Practical steps and recommendations.

Step	Recommendation	Key points
1	Create a "data and question sheet".	Document outcomes (type, prevalence, followup), sample size and events, predictor types and counts, missingness/data quality, and the intended clinical use of any model.
2	Match method complexity to data and goals.	Start with descriptive statistics and simple regression. Use regularised regression when predictors are many/correlated. Reserve complex ML (ensembles, NNs) for problems where data volume and use case truly justify it.
3	Respect sample size and EPV constraints.	Calculate EPV for binary/time-to-event outcomes. Limit candidate predictors, interactions, and nonlinear terms to what the data can support and use penalisation and prespecification of key variables.
4	Prevent data leakage.	Define prediction time point clinically (e.g. preTAVI decision). Exclude variables only known after the outcome. Ensure feature construction does not inadvertently use future information.
5	Plan validation and transportability early.	Use internal validation (cross-validation, bootstrapping). Choose predictors that are routinely available across centres. Anticipate how casemix, coding, and practice differences may affect performance elsewhere.
6	Prioritise interpretability and clinical usefulness.	In moderatesized registries, treat regularised regression with thoughtful nonlinear terms as the "advanced default". Prefer models that clinicians can understand and implement, and link outputs to concrete decisions (risk thresholds, example scenarios).
7	Co-design with a multidisciplinary team.	Involve clinicians, statisticians, and data scientists from planning onwards. Use their combined expertise to define predictors, interactions, and use cases, and to iteratively refine models as data and practice evolve.
8	Report transparently and support reproducibility.	Clearly describe the cohort, predictors, outcomes, handling of missing data, model specification, validation, and metrics. Provide enough detail (e.g. code lists, transformations) for others to replicate and externally validate the work.

EPV: events-per-variable, ML: machine learning, NN: neural network, TAVI: transcatheter aortic valve implantation.

can feel intuitively "large" and "complex". However, their statistical properties often place them in a very specific niche between small, singlecentre cohorts and true "big data". A simple taxonomy of data structure helps to avoid mismatches between methods and data.

The outcome type strongly shapes the analytic options. In SHARE, pre-TAVI aortic valve area is a continuous outcome, while 1-year mortality is a binary event, and other followup measures could be considered time-to-event or longitudinal outcomes. Each of these outcomes raises different questions about scale, distribution, and appropriate performance metrics, naturally aligning them with different modelling families.[3]

Clinical registries often contain a mix of predictor types. SHARE includes continuous predictors (age and biochemical markers), ordinal variables (the New York Heart Association [NYHA] Functional Classification class or frailty scales), and nominal variables (valve type, sex, and centre), alongside counts (number of binomially labelled comorbidities) and composite scores. Each measurement scale carries implicit assumptions about how differences between categories should be treated, and these assumptions must be honoured or explicitly transformed when moving from traditional regression to ML methods.[4,5]

Dimensionality matters more than sheer row count.[6] A registry may enrol a few thousand patients, but if it captures dozens of demographic, clinical, imaging, and laboratory variables, the ratio of sample size to predictors, and, crucially, events-per-predictor for binary outcomes, may still be modest. This has direct implications for whether complex models can be stably fitted, and for the risk of overfitting when many predictors compete for a relatively small number of events.

Patterns of missing data and measurement error must be examined before selecting an analytic method. Routine registries often show non-random missingness (e.g. laboratory tests ordered only in more severe patients) and variable data quality across centres. While regularised regression and ML methods are sometimes perceived as robust "out of the box", they are no substitute for understanding and, where possible, addressing missingness and misclassification by design or through appropriate imputation strategies.

Combined, these dimensions (i.e. outcome type, predictor mix, dimensionality, and missingness) define a "data profile" that can guide method selection.[7] Many cardiovascular and procedural registries occupy a space that might be called "moderately high-dimensionality but not truly big data". They are rich in variables and clinically valuable, yet remain constrained by sample size, event counts, and data quality. Recognising this helps temper expectations that very deep or highly flexible ML models will necessarily be appropriate or outperform well-tuned regression-based approaches.

Step 2: Map clinical questions to analytic goals

The second step is to articulate what the analysis is meant to achieve. Different clinical questions map to different analytic goals, and not all goals require – or are even compatible with – highly complex ML models.

A useful starting distinction is between prediction and explanation. In the SHARE context, predicting 1-year mortality after TAVI for individual patients is a prototypical predictive goal, with a focus on accurate risk estimates, and the tolerance for some loss of transparency may be higher if prediction improves meaningfully. Conversely, understanding how sex, renal function, or specific comorbidities relate to mortality, or how valve area changes with age and anatomical features, reflects an explanatory or mechanistic interest, in which the interpretability of effect estimates and their uncertainty are central. Methods that excel at prediction are not always the most informative for explanation, and vice versa.[3]

The outcome focus also matters. Continuous functional outcomes (e.g. pre-TAVI valve area) invite questions about calibration and mean error, whereas binary or time-to-event outcomes highlight discrimination, event prediction, and competing risks. Longitudinal trajectories (e.g. serial echocardiographic measurements or repeated quality-of-life scores) add further complexity and may require hierarchical or time-series approaches.

Finally, the intended use of the results influences the level of complexity deemed desirable. A bedside risk score integrated into clinical workflow demands parsimony and transparency, a model designed to identify quality improvement targets at the programme level can tolerate more complexity, and a model intended to generate mechanistic hypotheses may prioritise clear and robust associations over marginal gains in predictive accuracy – it all depends on the intended use and outcome. In SHARE, one might reasonably wish to develop a simple, implementable risk score for 1-year mortality, an institutional benchmarking tool, or exploratory analyses to determine which patient features drive adverse outcomes.

These distinctions (prediction versus explanation, outcome structure, and intended use) help determine where along the spectrum – from simple regression to complex ML – a given analysis should sit. In many registry settings, the combination of moderate sample size, mixed predictor types, and a strong need for clinical interpretability will favour methods that extend traditional models (e.g. regularised regression with carefully chosen non-linear terms) rather than immediate recourse to opaque algorithms.[6]

Step 3: A continuum of methods – from classic to machine learning

Once the data profile and analytic goals are clear, methods can be viewed along a continuum rather than as competing camps. A simple conceptual "ladder" can help situate a registry with SHARE-like data.

The base of this ladder consists of descriptive and univariate analyses. These include distributions of key variables, cross-

tabulations, unadjusted associations between predictors and outcomes, and basic checks for missingness and outliers. In SHARE, such analyses might show, for example, the distribution of age and valve area, crude mortality rates by NYHA class, or simple differences between centres.[3] Although sometimes dismissed as preliminary, this level is essential to understand the data and prevent later models from encoding artefacts.

The next rung comprises traditional multivariable models with linear regression for continuous outcomes, logistic regression for binary outcomes, and Cox models for time-to-event data. These are familiar to clinicians, conceptually straightforward, and can often be implemented with standard statistical software. Their strengths include transparency, direct estimates of effect sizes and confidence intervals, and well-understood diagnostics. Their limitations emerge in high-dimensional or highly correlated settings, where linearity assumptions may be implausible, multi-collinearity can destabilise estimates, and the number of candidate predictors threatens to overwhelm the number of events.

Regularised regression methods (ridge regression, lasso, and elastic net) occupy a middle rung and can be viewed as a bridge between classical regression and more flexible ML.[8] They retain the basic regression framework but add penalties on model complexity, shrinking coefficient estimates, and, in some cases, selecting variables. For SHARE, with a large number of patients and many potentially correlated clinical and echocardiographic variables, these methods are particularly attractive, as they address multi-collinearity, reduce overfitting risk, and can yield parsimonious models that remain interpretable for individual predictors.[9]

At the top of the ladder are more flexible ML approaches, including tree-based ensembles (random forests, gradient boosting), support vector machines, and various NN architectures. These methods can, in principle, capture intricate nonlinear relationships and high order interactions without explicit specification. Their advantages become most evident in truly large datasets, in problems involving unstructured data (images, waveforms), or where interactions are numerous and difficult to predefine. However, they often come at the cost of reduced transparency, more complex tuning, and sensitivity to data idiosyncrasies when sample sizes and event counts are modest.

Placed on this ladder, SHARE likely sits in the zone where enhanced regression through thoughtful variable encoding, non-linear terms, and regularisation is the most natural next step beyond traditional modelling. The registry's size and richness justify methods that can handle correlated predictors and mild non-linearities but may not support the stable estimation and validation of very highcapacity ML models (e.g. deep NNs), especially when the number of events is limited and external validation cohorts are not yet available.

Step 4: A practical workflow for choosing methods

To make these ideas actionable for clinicians and applied researchers, it is helpful to distil them into a simple, reproducible workflow applicable to any clinical dataset, with SHARE as an illustrative case.

First, characterise the data using the taxonomy in step 1. For SHARE, this would mean documenting the total sample size, number of events for each outcome of interest, distributions and types of all candidate predictors, patterns of incompleteness, and any known measurement limitations.[9] This step results in a concise "data profile" that can be shared among collaborators.

Second, define the primary analytic goals and constraints as in step 2. For example, in SHARE, one might prioritise development of an internally validated prediction model for 1-year mortality, with secondary goals of understanding the role of specific comorbidities and generating a simple risk score suitable for routine use. Constraints might include limited events per predictor, absence of external validation data, and the need for a model that can be implemented without complex infrastructure.

In step 3, start with the simplest method that can reasonably address the question. For a continuous outcome (e.g. valve area), this might mean linear regression with a modest set of clinically selected predictors. For 1-year mortality, a logistic regression model with prespecified core predictors might serve as a baseline. These initial models set a reference point for performance and interpretability.

Evaluate model performance and calibration using appropriate internal validation as step 4. Cross-validation, bootstrapping, or split-sample approaches can be used to estimate outofsample discrimination, calibration, and error metrics. In SHARE, this step might show that a simple logistic model achieves only modest discrimination in mortality, suggesting room for improvement.

As a possible step 5, escalate complexity in a controlled fashion. If simple models underperform meaningfully, one can introduce regularised regression to handle larger predictor sets and multi-collinearity, or add carefully chosen nonlinear terms and interactions. For SHARE, this might involve moving from a small logistic regression model to an elastic net model, including a broader set of clinical and echocardiographic variables, while maintaining internal validation at each stage. More flexible ML methods should be considered only if (1) the data volume and signal-to-noise ratio justify their capacity; (2) the incremental gains in performance are likely and clinically meaningful; and (3) the implications for interpretability and implementation are acceptable.

At each step, regardless of the method, it is important to monitor for signs of overfitting, such as overly optimistic apparent performance relative to cross-validated estimates, unstable variable selection across resamples, or implausibly large effect sizes. Multi-collinearity should be assessed, at least

informally, to avoid misinterpreting coefficients. Finally, the clinical plausibility and implementation of model outputs should be scrutinised, and questions such as "Do the identified predictors and their directions of effect make sense?" and "Can the model be used at the bedside or in a policy context without specialised infrastructure?" should be asked. If the answer to these questions becomes less clear as complexity increases, this may be a signal to pause or reconsider. Characterise, clarify goals, start simple, validate, then escalate complexity only when justified by both data and clinical need.[4]

Step 5: Handling nonlinearities and interactions without overcomplicating

Nonlinear relationships and interactions are often invoked as reasons to adopt complex ML models. Yet, many of the clinically important forms of nonlinearity and interaction can be accommodated within regression or regularised regression frameworks, preserving interpretability while capturing richer patterns.

In a SHARElike registry, one might suspect that age has a nonlinear association with 1-year mortality, that valve area exhibits threshold effects, or that renal function and frailty interact with procedural risk. These hypotheses can be addressed by incorporating transformations and spline functions into regression models. For example, age and key biomarkers can be modelled using restricted cubic splines, allowing their relationship with the outcome to bend smoothly without imposing a single global linear relationship.[10] Valve area could be transformed or segmented around clinically meaningful thresholds to reflect known physiology.

Interactions of clinical interest, such as sex by age, sex by renal function, or comorbidity burden (e.g. hypertension or diabetes) by frailty, can be prespecified based on prior knowledge and encoded as product terms in the model. When combined with regularisation, a model can include a richer set of plausible nonlinear and interaction terms while controlling for overfitting. This strategy allows the analysis to remain grounded in clinically interpretable quantities (e.g. how the risk gradient with age differs between men and women), rather than relying on opaque interaction structures learned automatically by an algorithm.

Whether the added complexity of these terms is justified should be evaluated empirically. In SHARE, one might compare a baseline logistic regression model with linear terms to an extended model including splines and key interactions, using internal validation to assess changes in discrimination, calibration, and clinical utility. If the extended model delivers only marginal gains at the cost of greater complexity, it may be preferable to retain the simpler specification. Conversely, if nonlinear terms substantially improve calibration across the age spectrum or better capture risk in specific subgroups, this justifies their inclusion without necessitating a shift to blackbox ML.

Importantly, this approach underscores that "handling nonlinearity" is not synonymous with "using complex ML".[4] Many clinically relevant nonlinearities and interactions can be captured by thoughtful modelling within regression-based frameworks, particularly when supported by regularisation and rigorous validation. For registries like SHARE, this may offer a balanced path that respects both the richness of the data and the practical needs of clinicians who must interpret and use the results.

TABLE II: Common pitfalls and how to avoid them.

	Pitfall	Alternatives
1	Overinterpreting small, noisy ML models as "AI-driven" applications.	Clinical datasets are often smaller and noisier than they appear, with modest event counts and heterogeneous measurement quality. In such settings, highly flexible ML models can fit idiosyncrasies of the sample rather than the true signal, particularly when internal validation is weak or absent. Transparent reporting of sample size, event counts, and the validation strategy, combined with comparisons to simpler baselines, is essential to avoid overstating these models.
2	Ignoring EPV constraints.	For binary outcomes, too few EPV is associated with unstable estimates, overfitting, and overly optimistic apparent performance, regardless of whether conventional regression or ML is used. Explicitly calculating EPV and tailoring model complexity (including the number of candidate variables and interaction terms) to this constraint is a critical step in responsible model development.
3	Data leakage.	Data leakage occurs when information that would not be available at the intended time of prediction is inadvertently used during model training. In clinical prediction models, a classic example is including diagnostic codes or post-procedural variables that are only finalised after discharge in a model intended to predict in-hospital or early post-procedural outcomes. Careful temporal alignment of predictors and outcomes, along with a clear definition of the prediction time point, is needed to avoid this subtle but pervasive error.
4	Neglecting transportability and generalisability.	Models that perform well in their development registry may fare poorly when applied to new settings, populations, or time periods. Differences in case-mix, practice patterns, data coding, and measurement frequency can all compromise transportability. When external validation is not yet feasible, sensitivity analyses, causal reasoning about predictors and outcomes, and cautious claims about scope are important safeguards.
5	Presenting models without clear clinical use cases.	Explicitly defining who will use the model, at what point in the care pathway, and how the predictions will change decisions helps align methodological choices with clinical value.

AI: artificial intelligence, EPV: events-per-variable, ML: machine learning.

COMMON PITFALLS AND HOW TO AVOID THEM

Even when data and questions are well characterised, several recurring pitfalls can undermine the validity and usefulness of statistical and ML models in clinical registries (Table II). These pitfalls include the over-interpretation of data, where apparent "artificial intelligence-driven" performance becomes an illusion of overfitting and noise exploitation.[11-13]

For binary outcomes, it is easy to ignore events-per-variable, as in the case of SHARE, where 1-year deaths are relatively few compared to the dozens of available predictors. Attempts to fit very detailed models, particularly without regularisation, risk producing unstable coefficients and spurious variable importance. Similarly, data leakage can lead to inflated performance estimates.[14] For example, in a SHARE-based context, incorporating variables that are recorded after the TAVI procedure or using follow-up information to engineer predictors for a baseline risk model would constitute leakage, inflating performance estimates.

It is also possible to neglect a registry's generalisability. For instance, a SHARE-derived mortality model might rely on variables whose distributions or meanings differ across other centres or countries, leading to miscalibration and degraded performance elsewhere. Emphasising internal performance alone, without explicit consideration of external validation or likely shifts in population characteristics, can give a misleading sense of robustness.[15]

Lastly, a common pitfall is to develop and report models without articulating a specific clinical use case, mechanistic pathway, or implementation strategy. A complex TAVI risk model may be technically impressive, but if it is too cumbersome for bedside use, lacks clear decision thresholds, or does not address a clinical question, it is unlikely to influence practice. In a SHARE-like scenario, a model that predicts 1 year mortality but does not inform procedural selection, follow up intensity, or patient counselling risks, remains an academic exercise.

Combined, these pitfalls underscore that the value of a model lies not in its algorithmic sophistication, but in the way it respects data limitations, avoids methodological traps, and serves clearly defined clinical purposes.

CONCLUSION

Clinical registries, such as the SHARE-TAVI registry, offer rich opportunities to improve risk stratification, understand treatment outcomes, and inform policy. Their multi-dimensional nature naturally invites interest in ML and other advanced methods. Yet, as this article has argued, the true power of ML in this context lies not in complexity for its own sake, but in disciplined, question-driven, and data-aware method selection. This stepwise framework emphasises that many registries are "moderately high-dimensional but not big data", and that for such datasets, thoughtful extensions of traditional models, including regularisation and carefully specified nonlinearities, may offer the most appropriate balance between performance and interpretability. Consequently, SHARE illustrates why model choice must be grounded in data structure, event counts, measurement quality, and intended clinical use. The framework we propose is intended to help clinicians, researchers, and registry owners navigate this landscape, enabling them to harness both traditional statistics and ML in ways that respect their data and ultimately serve patients.

Conflict of interests : none declared

REFERENCES

1. Wentzel A, Blignaut M. 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. SA Heart. 2026;23(1):35-41.
2. Schaafsma E, Weich H, Scherman J, et al. Outcomes in the South African SHARE-TAVI registry: Comparison of mortality and risk between treated and untreated cohort, and recent 1-year outcomes in the maturing local TAVR programme. Eur Heart J. 2023;44(Suppl 2):ehad655.3028. https://doi.org/10.1093/eurheartj/ehad655.3028.
3. Matsui S, Le-Rademacher J, Mandrekar SJ. Statistical models in clinical studies. J Thorac Oncol. 2021;16(5):734-9. https://doi.org/10.1016/j.jtho.2021.02.021.
4. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. Medicina (Kaunas). 2020;56(9):455. https://doi.org/10.3390/medicina56090455.
5. Jing B, Boscardin WJ, Deardorff WJ, et al. Comparing machine learning to regression methods for mortality prediction using Veterans Affairs Electronic Health Record clinical data. Med Care. 2022;60(6):470-9. https://doi.org/10.1097/MLR.0000000000001720.
6. Hu Y, Zhang X, Slavin V, et al. 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.
7. Islam R, Weir C, Del Fiol G. Clinical complexity in medicine: A measurement model of task and patient complexity. Methods Inf Med. 2016;55(1):14-22. https://doi.org/10.3414/ME15-01-0031.
8. Bjerre LM, Peixoto C, Alkurd R, Talarico R, Abielmona R. Comparing AI/ML approaches and classical regression for predictive modeling using large population health databases: Applications to COVID-19 case prediction. Glob Epidemiol. 2024;8:100168. https://doi.org/10.1016/j.gloepi.2024.100168.
9. Basu S, Andrews J. Complexity in mathematical models of public health policies: A guide for consumers of models. PLoS Med. 2013;10(10):e1001540. https://doi.org/10.1371/journal.pmed.1001540.
10. Schuster NA, Rijnhart JJM, Twisk JWR, Heymans MW. Modeling non-linear relationships in epidemiological data: The application and interpretation of spline models. Front Epidemiol. 2022;2:975380. https://doi.org/10.3389/fepid.2022.975380.
11. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: A synthetic knowledge synthesis. Sci Prog. 2022;105(1):368504211029777. https://doi.org/10.1177/00368504211029777.
12. Yang HS, Rhoads DD, Sepulveda J, et al. Challenges and considerations of developing and implementing machine learning tools for clinical laboratory medicine practice. Arch Pathol Lab Med. 2023;147(7):826-36. https://doi.org/10.5858/arpa.2021-0635-RA.
13. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med. 2021;4(153). https://doi.org/10.1038/s41746-021-00521-5.
14. van de Mortel LA, van Wingen GA. Data leakage in machine learning studies creep into meta-analytic estimates of predictive performance. Mol Psychiatry. 2025;30(12):6070-1. https://doi.org/10.1038/s41380-025-03336-y.
15. Fehr J, Piccininni M, Kurth T, Konigorski S. Assessing the transportability of clinical prediction models for cognitive impairment using causal models. BMC Med Res Methodol. 2023;23(187). https://doi.org/10.1186/s12874-023-02003-6.