From traditional statistical modelling to machine learning algorithms:

a view from the high-dimensional data perspective

Hotly debated topics in Biostatistics: a STRATOS perspective

Federico Ambrogi[1,2], Jörg Rahnenführer[3], Riccardo De Bin[4], Willi Sauerbrei[5], Lara Lusa[6]

[1]Dept. of Clinical Sciences and Community Health, University of Milan, Italy [2]IRCCS Policlinico San Donato, Italy
[3]TU Dortmund, Germany [4]University of Oslo, Norway [5]Medical Center – University of Freiburg, Germany
[6]University of Primorska, Slovenia

Seoul, South Korea, 12-16 July 2026

The promise of machine learning

Technological drivers

"As of 2020, the volume of medical knowledge is estimated to double every 2 to 3 months, and given the manifest impossibility of one individual retaining all of that knowledge, modern medicine increasingly requires clinicians to effectively digest and make practical use of the volume of data available to them."

NAGY, et al. JCO Clinical Cancer Informatics, 2020

Precision medicine and High-dimensional data

Omics and Multi-source data integration to identify disease subtypes and enable targeted treatments

Computational power

Despite having been studied for the better part of a century, the computing power and advances in algorithms necessary for AI's broad application have only arisen recently.

- Early foundations (1970s–1980s): Recursive partitioning / decision trees (1970s); Early interpretable models (e.g. CART)
- Neural networks (1950s–1980s, revival in 1980s): Perceptron and early multilayer models; Backpropagation (1980s) led to renewed interest
- Statistical learning era (1990s): Support Vector Machines (SVMs); Boosting (e.g. AdaBoost)
- Ensemble era (2000s): Bagging (bootstrap aggregation); Random forests; Boosting refined and extended
- Deep learning era (2010s–present): Large neural networks trained on massive datasets. Enabled by GPUs

Breiman's legacy: the two cultures of data modelling

Rashomon

- many different models can fit the same dataset equally well, achieving similar predictive accuracy.
- Instability
- "Aggregating over a large set of competing models can reduce the nonuniqueness while improving accuracy"

Occam

- Accuracy vs Simplicity.
- Accuracy generally require more complex prediction methods
- "Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy, then try to understand why"

Bellman

- The curse of dimensionality ('60)
- Dimensionality could be a blessing
- "Instead of reducing dimensionality, increase it by adding many functions of the predictor variables"

Twenty-five years after The Two Cultures,

Let us revisit Breiman's key ideas in the age of machine learning and foundation models.

The Rashomon Effect

Many good models - aggregate them (Breiman) or choose the most interpretable one.

Christodoulou et al. (J Clin Epidemiol 2019)

– Systematic review shows **no performance benefit of ML** over logistic regression for clinical prediction models.
– Across 71 studies, performance measured by AUC was similar between ML and LR in low-risk-of-bias comparisons, while apparent advantages of ML were mostly observed in studies with methodological flaws, particularly poor validation

Hu et al. (J Med Internet Res. 2025)

– No universal model superiority confirmed; ML gains in structured tabular clinical data remain **inconsistent and context-dependent**.
– Performance in clinical prediction depends on dataset characteristics (e.g. sample size, nonlinearity, class imbalance), not on the algorithm alone

Kattan and Gerds: Medical Risk Prediction Models

"... we start by noting that for practical purposes it is often reasonable to expect that all sound strategies when applied to the same dataset should end-up with comparable results ..."

C. Rudin

"uncertainty arising from the data leads to a Rashomon set; a larger Rashomon set probably contains interpretable models, thus interpretable accurate models often exist".

Occam and Explainability: the black-box problem

The black-box problem

– Complex ML models may improve predictive performance but sacrifice interpretability:
– a serious barrier in clinical and regulatory contexts (despite Breiman belief).

Stop Explaining: Rudin C. Nat Mach Intell. 2019

– Especially with structured data the simplicity vs accuracy trade off is often not true
– Interpretable models can achieve comparable predictive performance while providing transparency and accountability.

Hand D. Stat Sci 2006

– Most of the predictive performance in classification problems is captured by simple models, with more complex methods yielding only marginal improvements that are often negligible in practice.
– These small gains are frequently overwhelmed by real-world uncertainties (e.g., data drift, measurement error), so the apparent superiority of complex models is often an illusion.

Combining clinical and omics covariates

– Gorfine: Flexible Deep Neural Networks for Partially Linear Survival Data: Estimation and Survival Inference;
– De Bin: Combining clinical and molecular data in regression prediction models: insights from a simulation study (2020)

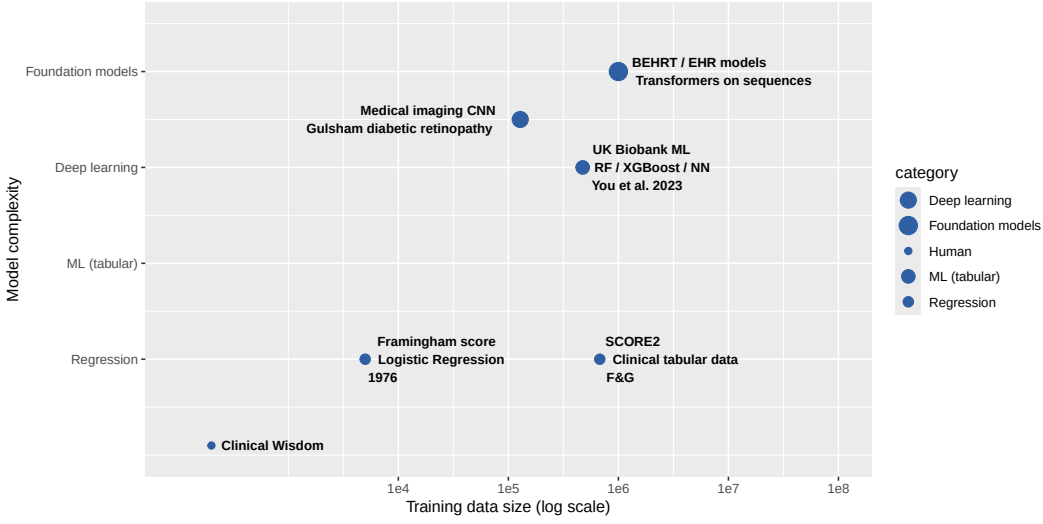
Bellman: the curse (Breiman's blessing) of dimensionality

Domain	ML advantage	Assessment
RCTs & exposure comparison **Structured data, design-driven identification**		Randomization removes confounding by design; well-developed estimators for ATE are efficient and interpretable
RCTs & treatment/exposure comparison **HD/multimodal data, EHR, imaging, omics**		Causal ML (TMLE, meta-learners) allows model-agnostic, doubly-robust estimation of heterogeneous treatment effects; HD covariates and multimodal data make flexible ML nuisance models the natural choice
Medical imaging **Unstructured data, visual patterns**		Opens previously unreachable possibilities: expert-level retinopathy screening **(Gulshan et al., JAMA 2016)**, dermatologist-level skin cancer classification **(Esteva et al., Nature 2017)**, AI surpassing radiologists in mammography **(McKinney et al., Nature 2020)**
Diagnostic / prognostic prediction **Structured, tabular clinical data, biomarkers, omics**		Classical stat excels with few known covariates: SCORE2 **(SCORE2 WG, Eur Heart J 2021)**, genomic signatures (Oncotype DX, MammaPrint). ML shows gains on HD/EHR data **(You et al., Stroke Vasc Neurol 2023)**; clinical deployment remains limited. Combining clinical and HD covariates **(De Bin et al., Stat Med 2014; Gorfine & Zucker, Ann Rev Stat 2023)**
Diagnostic / prognostic prediction **Unstructured data, clinical notes, EHR sequences**		EHR foundation models treat patient history as language. Same transformer paradigm as LLMs applied to clinical sequences. NYUTron outperforms physicians on readmission prediction **(Jiang et al., Nature 2023)**. rigorous validation needed

Key driver of ML uselfuness: data complexity (dimensionality, structure, modality), not the domain per se.

The statistics-ML spectrum: selected biomedical examples

Adapted from Beam & Kohane, JAMA 2018;319:1317 Machine learning as a continuum with traditional statistical methods, where increasing reliance on data-driven learning (rather than human knowledge) offers greater flexibility but requires more data, reduces interpretability, and does not guarantee better or more trustworthy results.



The question of the sample size

Sample size reality check

– [Dhiman, 2023] systematic review reported that 73% of binary clinical prediction models had sample sizes *below* the recommended minimum.
– Random forests may need 20× more events per predictor than logistic regression.

Warning about sample size

In a simulation study for predicting dichotomous endpoints, van der Ploeg and colleagues suggested that machine learning techniques could be data hungry
– 20 events per variable for developing well calibrated LR models
– 200 events per variable for developing well calibrated models with Random Forest
– "In the current study, we evaluated the following modelling techniques, using default settings as far as possible:. . . "

Key insights

– Tuning parameters strongly influence prediction performance, model sparsity, and stability, Overfitting and Underfitting.
– Correctly tuned simple methods often outperform sophisticated methods with poor tuning
– Default parameter values frequently perform poorly, especially in high-dimensional settings

De Bin R., Hornung R., Gandin I., Lusa L., Michiels S. ***A gentle introduction to tuning parameters: their role, importance, and how to choose them.*** **STRATOS TG9.**

Data hungriness revisited

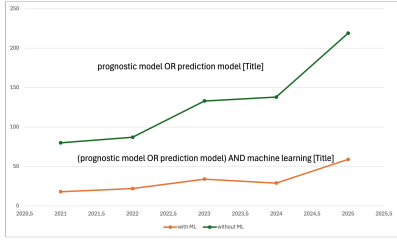
– Ceteris paribus, small or no increase in sample size is likely to be necessary, remember Rashomon.
– A separate consideration is that ML becomes advantageous when the data are complex (e.g. high-dimensional), in which case large samples are required for training and tuning.

Evidence about the use of ML from systematic reviews

STRATOS TG9

– Using systematic reviews to identify prediction models, time trends of the use of machine learning techniques, between 2020 and 2022 were explored.
– Eight reviews were identified, including 1448 prediction models published in 887 papers.
– There was no sign of an increase in using of machine learning methods.
– Poor reporting remained quite common

Lusa et al. BMC Med Res Methodol. 2025.



Limited clinical Deployment

"All the models included in our quantitative analysis were developed using traditional methods. Prognostic model development using newer techniques, such as machine learning approaches, is becoming increasingly common. We identified several such models through our search. However, none of these had yet been externally validated in the target population."

Risk prediction models for familial breast cancer, Cochrane 2026

Hu et al. (2025)

Critical barriers persist, only one study performed external validation; ... 88% of studies received a high risk of bias rating in the analysis domain under PROBAST, ... claims of ML superiority over CHA_2DS_2-VASc must be interpreted with caution.

Advancing stroke prevention in atrial fibrillation: a systematic review of machine learning-based risk prediction models. Int J Med Inform. 2026

Performance assessment

Internal Validation

– Model performance evaluated on new subjects taken from the training data using cross-validation, bootstrapping or repeated data splitting (train-test)

– Invalid method of cross-validation: It is critical that any data-driven variable selection steps are included as part of the model building process when model performance is assessed using any internal validation method. See Sachs and McShane 2016.

– Partial CV: A prediction model using proteomic data from the discovery cohort was built. First, 5,000 proteins linked to treatment response were identified. Then, 60 key proteins were selected to create a prediction model using Elastic Net, Random Forest, and XGBoost. Model's performance were tested using 10-fold cross-validation.

External Validation

– Model performance evaluated on new subjects, not included in the training data, taken from different countries, hospitals, time periods, etc.

– "Our general conclusion here is congruent with normal practice in 'hard science'; to be acceptable, a theory (model) must be confirmed by different workers in a different setting."[Altman, D.G. and Royston, P. (2000), What do we mean by validating a prognostic model?]

Performance assessment

Adapted from Van Calster et al., Lancet Digital Health 2025

Five performance domains

- **Discrimination:** AUROC
- **Calibration:** calibration plot, O:E ratio, slope/intercept
- **Overall performance:** Brier score, log-loss, R^2
- **Classification:** accuracy, sensitivity, specificity, PPV/NPV, F1, MCC
- **Clinical utility: Most Important Domain** net benefit, decision curves, expected cost

Two desirable characteristics

– **Properness**: the expected value is optimised by the correct model

– **Clear focus**: measure either *statistical performance* or *decision–analytical performance*, without conflating the two

Measures to avoid

Most classification measures (accuracy, balanced accuracy, Youden index, F1, MCC, . . .) are improper at clinically relevant decision thresholds.

Recommended core set

AUROC (discrimination); Calibration plot (preferably smoothed); Clinical utility measure: net benefit with decision-curve analysis; Risk distribution plot by outcome category

From model development to clinical readiness

Different frameworks for complementary questions related to ML or HDD

Was the prediction model, diagnostic study, clinical trial reported transparently? (Reporting frameworks)	TRIPOD+AI, STARD-AI, CONSORT-AI, SPIRIT-AI
Is the study methodologically trustworthy? (Methodological appraisal frameworks)	TREE (Transparency, Reproducibility, Ethics and Effectiveness), PROBAST+AI
Can clinicians understand and use the model? (Clinical communication framework)	Model Facts
Is the omics biomarker, the AI system, the AI biomarker ready for clinical translation? (Translational readiness frameworks)	NCI Omics, DECIDE-AI, EBAI

Common themes across frameworks

1. Clear definition of target population & intended use

All frameworks require explicit specification of the target population, users, clinical setting, intended decision, and role of the model.

2. Protection against overfitting

Performance must be assessed using data not involved in model development; internal validation alone is not evidence of clinical usefulness.

3. Generalizability through external validation

Each framework treats external validation as the critical gate before a model can be trusted for broader use. External validation is the critical step demonstrating that a model remains reliable across different populations, locations, and times

4. Clinically meaningful performance metrics

Discrimination alone is insufficient: performance should be evaluated using measures aligned with the intended clinical task, including calibration when relevant, clinical utility, and comparison with meaningful benchmarks or standards of care (Van Calster et al., Lancet Digital Health 2025).

5. Full and reproducible model specification

Inputs, preprocessing, algorithm, outputs, thresholds (when applicable), and implementation details must be described sufficiently for independent assessment and verification.

6. Transparency about limitations & inappropriate use

Frameworks consistently require disclosure of model limitations, uncertainty, and applicability. In particular TREE, Model Facts, DECIDE-AI, and EBAI go further by emphasising where the model should not be used, likely failure modes, and the need for human oversight outside validated contexts.

Conclusions

- Statistical modelling and ML form a continuum: the right tool depends on domain, data structure, and scientific question.
- Empirical evidence suggests that no algorithm is universally superior. Performance is primarily driven by the data-generating problem and the quality of model development and tuning. ML may offer advantages for high-dimensional, multimodal, or unstructured data.
- ML, Foundation models and LLMs expand the frontier but inherit the same validation and reproducibility demands: deployment principles apply equally.
- The strongest consensus across modern frameworks is that trustworthy AI/ML models requires clear intended use, independent and external validation, transparency, and clinical utility.
- The debate should move from "statistics versus machine learning" to "how to build trustworthy and clinically useful models with ML".

References

- Lusa et al. Evaluation of changes in prediction modelling in biomedicine using systematic reviews. *BMC Med Res Methodol.* 2025. doi:10.1186/s12874-025-02605-2
- Breiman L. Statistical modeling: the two cultures. *Statist Sci.* 2001;16(3):199–231. doi:10.1214/ss/1009213726
- Hu Y et al. Beyond comparing ML and logistic regression. *J Med Internet Res.* 2025;27:e77721. doi:10.2196/77721
- Christodoulou E et al. Systematic review shows no performance benefit of ML over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004
- Rudin C. Stop explaining black box ML models for high-stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206–215. doi:10.1038/s42256-019-0048-x
- Hager P et al. Evaluation and mitigation of the limitations of LLMs in clinical decision-making. *Nat Med.* 2024;30(9):2613–2622. doi:10.1038/s41591-024-03097-1
- Shool S et al. A systematic review of LLM evaluations in clinical medicine. *BMC Med Inform Decis Mak.* 2025. doi:10.1186/s12911-025-02954-4
- Zhang M et al. Integrating ML into statistical methods in disease risk prediction modelling. *Health Data Sci.* 2024;4:0165. doi:10.34133/hds.0165
- Aldea M et al. ESMO Basic Requirements for AI-based Biomarkers In Oncology (EBAI). *Ann Oncol.* 2025. doi:10.1016/j.annonc.2025.11.009
- Sachs MC, McShane LM. Issues in developing multivariable molecular signatures for guiding clinical care decisions. *J Biopharm Stat.* 2016. doi: 10.1080/10543406.2016.1226329
- Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA* 2018. doi:10.1001/jama.2017.18391


That’s all folks!
Questions welcome