

How Robust Are Machine-Learning Biomarkers of Depression? A Leakage-Controlled Benchmark

Arhama Faridi

arhaama.far@gmail.com

Queen Mary University of London <https://orcid.org/0000-0002-8441-8097>

Research Article

Keywords: depression, machine learning, computational psychiatry, data leakage, reproducibility, validation, predictive modelling, calibration, artificial intelligence

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11163748/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Machine-learning models are increasingly used to identify predictive markers of depression, yet reported performance may be sensitive to data leakage, non-independent observations, site effects, and validation procedures that do not adequately approximate deployment in independent clinical populations. Moreover, increasingly complex algorithms may provide limited incremental value over simpler statistical models once leakage is controlled. This study examined how validation design and leakage-prone predictors influence estimates of depression-classification performance and whether complex machine-learning models outperform logistic regression under more rigorous evaluation conditions.

Methods

A controlled synthetic multisite psychiatric dataset comprising 13,159 observations from 8,000 unique synthetic participants across 10 simulated sites was used to benchmark five classification approaches: logistic regression, random forest, extreme gradient boosting (XGBoost), support vector machine, and a neural network. The dataset incorporated repeated observations, demographic characteristics, depression- and anxiety-related symptom measures, sleep and activity indicators, physiological measures, cognitive performance, treatment variables, site and device effects, and realistic missingness. Depression case status was defined using a synthetic PHQ-9-like total score threshold of ≥ 10 , yielding a prevalence of 35.4%. To prevent direct target leakage, the PHQ-9-like total score was excluded from predictive models. Two deliberately post-outcome, leakage-prone variables were additionally manipulated to quantify performance inflation. Leakage-controlled models were evaluated using conventional random row splitting, participant-independent splitting, and site-independent validation. Discrimination and probabilistic performance were assessed using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and Brier score.

Results

Allowing leakage-prone post-outcome predictors produced markedly optimistic performance, with random-split AUROCs of approximately 0.988–0.989 and AUPRCs of 0.981–0.983. For logistic regression, removal of leakage-prone predictors reduced AUROC from 0.9891 to 0.9548 and AUPRC from 0.9827 to 0.9253, while the Brier score increased from 0.0367 to 0.0788. In leakage-controlled analyses, discrimination nevertheless remained high across validation strategies. AUROCs ranged from 0.9488 to 0.9548 under random row splitting, 0.9516 to 0.9585 under participant-independent validation, and 0.9502 to 0.9555 under site-independent validation. Corresponding AUPRCs ranged from 0.9116 to 0.9254, 0.9246 to 0.9383, and 0.9183 to 0.9356, respectively. Logistic regression achieved an AUROC of 0.9585 and AUPRC of 0.9383 under participant-independent validation and performed comparably to

random forest, XGBoost, support vector machine, and neural-network classifiers. Increasing algorithmic complexity therefore yielded little improvement in discrimination after leakage control.

Conclusions

Apparent performance in psychiatric machine-learning benchmarks can be substantially influenced by leakage-prone predictors, even when underlying predictors contain strong signal. In this controlled synthetic benchmark, removing post-outcome information materially reduced apparent discrimination, whereas participant- and site-independent validation did not produce a further collapse in performance. Complex machine-learning algorithms provided little predictive advantage over logistic regression. These findings illustrate why leakage auditing, explicit separation of outcome-derived variables, clinically realistic validation, calibration assessment, and comparison against simple baselines should be routine components of psychiatric machine-learning research. Because the present findings derive from synthetic data, they demonstrate methodological mechanisms rather than establishing the magnitude of performance inflation in real clinical populations; replication using independent empirical psychiatric datasets is required.

Introduction

Depression is a major source of illness-related disability and is associated with substantial individual, social, and healthcare burden. Its heterogeneous presentation, variable clinical course, and overlap with other psychiatric conditions have contributed to sustained interest in computational approaches capable of identifying individuals with clinically relevant depressive symptoms. Machine-learning methods have consequently been applied to a wide range of potential markers, including demographic characteristics, symptom profiles, cognitive measures, sleep and activity patterns, physiological signals, neuroimaging measures, electronic health records, and digitally derived behavioural features. These approaches have generated optimism that multivariate predictive models may complement conventional assessment by detecting complex combinations of features that are difficult to identify using traditional univariate approaches. [13, 16, 21]

A central challenge, however, is determining whether apparently strong predictive performance represents clinically generalisable signal or properties of the analytical pipeline used to estimate it. In psychiatric machine learning, the distinction is particularly important. Datasets are frequently high dimensional, heterogeneous, multisite, longitudinal, and characterised by repeated observations from the same individuals. Missingness may be informative, treatment exposure can be associated with disease severity, measurement technologies may vary between sites, and demographic characteristics may be unevenly distributed. These properties create numerous opportunities for models to learn information that improves performance within a particular dataset without necessarily improving prediction in genuinely independent individuals or settings. [14, 24]

Validation and the estimation of predictive performance

The conventional random train-test split remains an intuitive approach to evaluating supervised learning models. When observations are statistically independent, random splitting can provide a reasonable estimate of generalisation to observations sampled from the same underlying population. Psychiatric datasets, however, often violate this assumption.

Repeated measurements from the same participant represent a particularly important example. If observations belonging to one individual occur in both training and test sets, the model is not necessarily being evaluated on an entirely independent person. Stable participant characteristics may be represented across both partitions. Consequently, row-level random splitting can estimate the ability to predict additional observations drawn from a partially familiar population rather than performance in previously unseen individuals.

Participant-independent splitting addresses this problem by assigning all observations belonging to an individual exclusively to either training or testing data. This seemingly modest analytical decision changes the interpretation of model performance. The resulting estimate more closely represents generalisation to new individuals rather than new observations. [10]

Multisite data introduce another level of dependence. Clinical sites may differ in recruitment procedures, population characteristics, disease prevalence, socioeconomic composition, clinical practice, instrumentation, data acquisition, or preprocessing. Machine-learning algorithms can exploit these site-associated differences. High performance within randomly partitioned multisite data therefore does not necessarily imply equivalent performance when the model is transported to a previously unseen site. [6]

Site-independent validation provides a stronger test of this form of transportability. By withholding an entire site from model development, investigators can examine whether the learned relationships survive a distributional shift associated with a new setting. Such validation is particularly relevant to psychiatric prediction because clinically useful models are ultimately expected to operate beyond the precise institutions and samples from which they were developed. [14, 24]

Data leakage in psychiatric machine learning

Data leakage represents a related but conceptually distinct threat to validity. Leakage occurs when information unavailable at the intended time of prediction—or information directly or indirectly derived from the outcome—enters model development. The resulting model may achieve excellent statistical discrimination while providing a misleading estimate of prospective predictive utility.

Leakage can occur at several stages of a machine-learning pipeline. Direct leakage occurs when an outcome or an almost deterministic transformation of the outcome is included among the predictors. More subtle forms arise when variables recorded after diagnostic assessment are used to predict that diagnosis. Referral decisions, treatment plans, clinical actions, diagnostic codes, or post-screening

interventions may encode knowledge of the outcome even when they are not explicitly labelled as diagnostic variables.

Leakage can also arise from preprocessing. Imputation, normalisation, feature selection, dimensionality reduction, or hyperparameter optimisation conducted using information from the complete dataset can transfer information from validation observations into model development. Proper evaluation therefore requires not only selecting appropriate train-test partitions but also ensuring that all data-dependent transformations are learned exclusively from training data.

This distinction is particularly consequential when evaluating psychiatric biomarkers. A variable may correlate strongly with depression because it is genuinely predictive, because it reflects the consequences of depression, because it was generated after clinical recognition of depression, or because it contains information introduced by the analytical procedure. Predictive association alone cannot distinguish these possibilities.

Model complexity and the importance of simple baselines

A second methodological question concerns whether increasingly sophisticated machine-learning algorithms provide meaningful predictive gains over simpler statistical approaches. Random forests, gradient-boosted trees, support vector machines, and neural networks can model nonlinearities and interactions that conventional linear classifiers may not capture. These capabilities can be valuable when the underlying predictive structure is genuinely complex. [6, 7, 19]

Nevertheless, greater algorithmic complexity does not guarantee greater generalisation. Flexible models may exploit dataset-specific patterns, nuisance variation, or leakage more effectively than simpler models. When the underlying signal is predominantly additive or approximately linear, a well-specified logistic regression model may perform similarly to substantially more complicated approaches.

Comparison against a simple baseline is therefore scientifically informative rather than merely procedural. If a complex model achieves substantially better discrimination under rigorous validation, the additional complexity may capture useful predictive structure. Conversely, if its performance is nearly indistinguishable from logistic regression, the incremental value of the complex algorithm becomes questionable, particularly when interpretability, computational burden, reproducibility, and implementation are considered. [10]

The appropriate question is therefore not simply which algorithm achieves the highest AUROC. It is whether improvements persist when all models are evaluated under identical leakage-controlled and independence-preserving conditions. [5, 17]

Discrimination is not sufficient

Psychiatric machine-learning studies have often emphasised classification accuracy or the area under the receiver operating characteristic curve. AUROC is useful because it quantifies discrimination across classification thresholds, but discrimination represents only one dimension of predictive performance. [5, 17]

Area under the precision-recall curve (AUPRC) provides complementary information when class frequencies are unequal because it focuses on the relationship between positive predictive value and sensitivity. Probabilistic calibration is also clinically important. A model can discriminate effectively between higher- and lower-risk individuals while generating poorly calibrated probabilities. [1, 4, 15, 17, 20]

The Brier score provides one means of assessing probabilistic prediction by measuring the squared difference between predicted probabilities and observed binary outcomes. Evaluating AUROC, AUPRC, and Brier score together therefore provides a broader assessment of model behaviour than reporting discrimination alone. [1, 2, 3, 4, 5]

Demographic and distributional robustness

Clinical populations are heterogeneous. Predictive relationships can vary according to age, sex, socioeconomic circumstances, ethnicity, treatment exposure, site, measurement device, and other population characteristics. Models developed in demographically imbalanced samples may consequently perform differently across subgroups or when the target population differs from the development sample.

This problem is not solved merely by increasing sample size. A large dataset can remain systematically unrepresentative, and an algorithm can exploit correlations associated with demographic composition rather than transportable clinical relationships. Demographic-balanced evaluation therefore provides an additional means of examining whether reported performance depends on the composition of the evaluation population.

More generally, validation should be considered a hierarchy of increasingly demanding questions. Random validation asks whether a model predicts observations sampled from approximately the same distribution. Participant-independent validation asks whether it generalises to unseen individuals. Site-independent validation examines transportability to an unseen institutional environment. Demographically balanced evaluation examines robustness when population composition is controlled. Leakage-controlled evaluation asks whether predictive performance persists after variables that improperly encode outcome information are removed.

These approaches estimate related but non-identical properties of a predictive model. [13, 16, 21]

The value of a controlled synthetic benchmark

Empirical clinical datasets are essential for establishing real-world predictive utility, but they can make the mechanisms underlying performance inflation difficult to isolate. When an unexpectedly high AUROC is observed in an existing dataset, investigators may not know whether it reflects strong genuine signal, population structure, site differences, repeated observations, post-outcome variables, preprocessing leakage, or interactions among these factors. [5, 17]

Simulation provides a complementary methodological strategy because these properties can be deliberately controlled.

The present study therefore uses a synthetic multisite psychiatric benchmark rather than representing the generated observations as real patients. The dataset comprises 13,159 observations from 8,000 unique synthetic participants distributed across 10 simulated sites. It incorporates repeated observations, demographic characteristics, symptom-related variables, sleep and activity measures, physiological and cognitive measures, treatment indicators, site and device effects, and structured missingness.

Importantly, leakage-prone post-outcome variables were deliberately introduced. Their presence is not intended to represent recommended predictors. Instead, they provide experimental controls through which the effect of leakage can be quantified by comparing otherwise similar analytical pipelines with and without these variables.

Depression case status is defined from a synthetic PHQ-9-like total score using a threshold of 10. Because this score directly determines the target, the PHQ-9-like total itself is excluded from predictive models. This distinction is essential: including the defining symptom total would convert the task from prediction using associated markers into near-direct reconstruction of the outcome. [9, 13, 16, 21]

The controlled design therefore allows several methodological questions to be examined within the same data-generating framework.

Study aim

The primary aim of this study is to determine how methodological choices concerning data leakage, observation independence, site generalisation, and model complexity influence estimated performance in machine-learning classification of synthetic depression status.

More specifically, the study aims to quantify the degree to which deliberately leakage-prone post-outcome variables inflate apparent predictive performance; compare conventional random row-level validation with participant-independent and site-independent validation; evaluate whether more complex machine-learning algorithms provide meaningful improvements over logistic regression once leakage is controlled; compare conclusions obtained from AUROC, AUPRC, and probabilistic calibration; and establish a reproducible benchmark framework that can subsequently be applied to empirical psychiatric datasets. [1, 4, 5, 6, 15]

The study is consequently designed as a methodological demonstration of trustworthy psychiatric machine-learning evaluation rather than as evidence that a particular set of biomarkers has been clinically validated.

Research questions

The study addresses the following research questions:

RQ1. To what extent does the inclusion of deliberately leakage-prone, post-outcome variables inflate estimated performance in depression classification?

RQ2. Does predictive performance materially change when evaluation moves from conventional random row splitting to participant-independent validation?

RQ3. How robust is model performance when evaluated using site-independent validation intended to approximate transport to an unseen setting?

RQ4. Do random forest, XGBoost, support vector machine, and neural-network models provide substantial predictive improvements over logistic regression after direct and post-outcome leakage are controlled? [6, 7, 19]

RQ5. Are conclusions about relative model performance consistent across AUROC, AUPRC, and probabilistic calibration as assessed using the Brier score? [1, 4, 5, 15, 17]

RQ6. To what extent is model performance maintained when evaluation is conducted under demographic-balancing constraints?

RQ7. Which components of the validation framework have the greatest influence on estimated predictive performance within the controlled synthetic benchmark?

Hypotheses

Based on established principles of predictive modelling and the mechanisms deliberately embedded within the simulation, the following hypotheses were specified. [13, 16, 21]

H1 — Leakage inflation hypothesis. Models containing deliberately leakage-prone post-outcome predictors will demonstrate higher apparent predictive performance than otherwise comparable models from which these variables have been removed. Specifically, leakage-inclusive pipelines are expected to produce higher AUROC and AUPRC values and lower Brier scores. [1, 4, 5, 15, 17]

H2 — Participant-independence hypothesis. Conventional random row-level splitting will yield more optimistic estimates of predictive performance than participant-independent validation because

repeated observations can otherwise be distributed across training and testing partitions.

H3 — Site-generalisation hypothesis. Performance will decline under site-independent validation relative to conventional random validation because the model must generalise beyond site-specific population and measurement characteristics represented during training.

H4 — Complexity hypothesis. After leakage is controlled, complex machine-learning algorithms will provide only modest improvements, if any, over logistic regression. Large performance advantages observed primarily under leakage-prone conditions would therefore provide evidence that apparent algorithmic superiority can depend on characteristics of the analytical pipeline rather than robust nonlinear predictive structure. [6]

H5 — Calibration hypothesis. Models demonstrating similar discrimination will not necessarily demonstrate equivalent probabilistic accuracy. Consequently, differences between algorithms are expected to become apparent when Brier scores are considered alongside AUROC and AUPRC. [1, 4, 5, 15, 17]

H6 — Demographic-robustness hypothesis. Predictive performance will vary when demographic composition is explicitly balanced, indicating that aggregate performance estimates may partly depend on the demographic distribution of the evaluation sample.

Contribution of the present study

The principal contribution of this study is methodological rather than algorithmic. It does not propose a new classifier and does not seek to establish a novel clinical biomarker of depression. Instead, it examines how conclusions about psychiatric prediction change as increasingly rigorous safeguards are imposed on the evaluation process. [14, 24]

This distinction reflects a broader reproducibility problem in predictive modelling: an apparently impressive model is scientifically informative only to the extent that its performance survives conditions relevant to its intended use. High discrimination obtained through leakage, participant dependence, or dataset-specific structure does not provide equivalent evidence to performance demonstrated in unseen individuals and settings. [13, 16, 21]

By directly contrasting leakage-inclusive and leakage-controlled pipelines, multiple validation strategies, simple and complex classifiers, and complementary performance metrics within a controlled environment, the present study provides an experimental framework for separating algorithmic performance from methodological artefact. The resulting benchmark is intended to support a more cautious interpretation of psychiatric machine-learning performance and to provide a foundation for subsequent replication using real-world, independently collected clinical datasets.

Literature Review

Machine learning and the development of psychiatric prediction models

The application of machine learning to psychiatry has expanded substantially as researchers have sought methods capable of integrating heterogeneous sources of information into clinically meaningful predictions. Depression has been a particularly prominent target because depressive disorders are heterogeneous in presentation and can be characterised using multiple forms of data, including self-reported symptoms, sociodemographic characteristics, cognitive performance, electronic health records, neuroimaging, physiological measurements, sleep and activity patterns, smartphone-derived behavioural measures, and wearable-sensor data.

The conceptual appeal of machine learning in this setting is straightforward. Conventional clinical assessment necessarily reduces a large amount of information into a manageable set of diagnostic and prognostic judgments. Machine-learning algorithms can instead consider numerous predictors simultaneously and, depending on the algorithm, accommodate interactions and nonlinear relationships without requiring investigators to specify each relationship in advance. Random forests, gradient-boosting algorithms, support vector machines, and neural networks have therefore become common components of psychiatric prediction research. [6, 7, 14, 19, 24]

However, the scientific objective of predictive modelling differs from simply identifying associations within an observed dataset. A useful predictive model must retain performance when applied to observations that did not contribute information to model development. The distinction between fitting observed data and generalising to new data is central to the interpretation of machine-learning findings. [10, 13, 16, 21]

This distinction has become increasingly important as clinical prediction research has moved toward more formal standards of model development, evaluation, reporting, fairness, and reproducibility. TRIPOD + AI, published in 2024, updated the earlier TRIPOD framework specifically to provide harmonised reporting recommendations for prediction models developed using either regression or machine-learning approaches. The guideline emphasises transparent descriptions of development and evaluation data, model performance, subgroup evaluation, fairness considerations, and open-science practices. [2, 3, 13, 16, 21]

The increasing sophistication of prediction algorithms has therefore been accompanied by increasing recognition that methodological rigor in evaluation may be at least as consequential as the choice of algorithm.

Depression prediction and the attraction of high discrimination

Machine-learning studies of depression frequently frame predictive performance in terms of classification accuracy or the area under the receiver operating characteristic curve. High values are

intuitively attractive because they suggest that multivariate patterns can distinguish individuals with and without clinically relevant depressive symptoms.

Nevertheless, predictive performance is conditional on the precise task being evaluated. An algorithm distinguishing cases from controls in a dataset constructed from strongly differentiated groups is addressing a different problem from an algorithm identifying depression prospectively among previously unseen individuals in heterogeneous clinical practice. Similarly, predicting a depression label from variables recorded after clinical assessment differs fundamentally from predicting that label using information genuinely available before assessment.

Consequently, numerical performance cannot be interpreted independently of study design.

A model with a high AUROC may have learned a robust relationship between predictors and outcome. Alternatively, it may exploit characteristics of the sample, institution, measurement system, participant identity, treatment process, or data-processing pipeline. These possibilities can produce similar headline performance statistics while having substantially different implications for clinical translation. [5, 17]

This is one reason transparent prediction-model reporting has become a methodological priority. TRIPOD + AI explicitly seeks sufficiently complete reporting to permit readers to understand the data, methods, findings, and conclusions of prediction-model research and to facilitate appraisal, replication, and implementation. [2, 3]

Overfitting and optimistic model evaluation

Overfitting represents one of the foundational problems of predictive modelling. A model is overfitted when it learns characteristics of its development data that do not generalise adequately to new observations. Flexible algorithms can be particularly vulnerable when the number of effective predictors is large relative to the amount of independent information available. [13, 16, 21]

The concept of effective independent information is especially important in psychiatric datasets. A dataset containing thousands of rows does not necessarily contain thousands of statistically independent individuals. Longitudinal monitoring, repeated questionnaires, wearable measurements, follow-up assessments, and multiple clinical encounters can generate numerous observations from the same participant.

If model development ignores this structure, apparent sample size can considerably exceed the number of independent observational units.

Machine-learning workflows attempt to address overfitting through train-test separation, cross-validation, regularisation, hyperparameter tuning, and external evaluation. Yet these safeguards are effective only when implemented in a manner consistent with the structure of the data. Cross-validation does not automatically prevent optimistic bias if information from validation folds influences feature selection, preprocessing, imputation, or tuning. [8, 18, 22]

Thus, the presence of cross-validation alone should not be interpreted as evidence that an analysis is leakage-free. [8, 18, 22]

Data leakage as a threat to psychiatric machine learning

Data leakage occurs when information that should be unavailable during model development becomes available to the algorithm. It can occur overtly or through comparatively subtle analytical pathways.

The most obvious form is direct target leakage. If depression status is defined using a particular symptom score and that same score is subsequently supplied as a predictor, the algorithm is effectively given information from which the outcome was constructed. High classification performance under such circumstances provides little evidence that an independent biomarker has been identified.

Post-outcome leakage presents a subtler problem. Variables recorded after diagnosis or screening can encode information about the outcome. Examples may include treatment initiation, specialist referral, follow-up intensity, diagnostic coding, or clinical management decisions. These variables can become highly predictive because clinicians have already responded to information related to the condition being predicted.

A model may therefore appear to discover a powerful psychiatric signal when it is partially reconstructing prior clinical decision-making.

Leakage can also occur during preprocessing. Suppose missing values are imputed using statistics calculated from the complete dataset before train-test separation. Information from the test set has then influenced the representation of training observations. Similar problems can occur when scaling, feature selection, dimensionality reduction, oversampling, or other data-dependent procedures are performed before cross-validation. [8, 18, 22]

Hyperparameter selection presents another potential source of optimism. Repeatedly selecting model configurations according to their performance on the same evaluation data gradually converts those data into part of the development process. Nested cross-validation was developed in part to separate hyperparameter optimisation from estimation of generalisation performance. [8, 18, 22]

The general principle is that the evaluation set should remain informationally isolated from all procedures that determine the fitted model.

Random splitting and participant dependence

Random row-level train-test splitting is commonly used because it is simple, computationally efficient, and often appropriate when observations are independent and identically distributed.

Repeated-measures psychiatric data challenge this assumption.

Consider a participant contributing several observations containing relatively stable characteristics such as demographic variables, baseline physiology, cognitive profile, device characteristics, or habitual sleep

patterns. If some observations enter the training set while another observation from the same person enters the test set, the evaluation does not represent prediction in a completely unseen individual.

The model need not explicitly contain a participant identifier for this dependence to matter. Combinations of stable features may partially characterise an individual. The test observation may therefore be statistically more similar to the training data than an observation obtained from a genuinely new participant.

Participant-independent splitting provides a more appropriate evaluation when the intended use involves prediction in unseen people. All observations belonging to a participant are assigned to one partition, preventing direct participant overlap between development and evaluation data.

Importantly, participant-independent validation will not necessarily produce lower performance. If the predictive relationships are sufficiently strong and generalisable, performance may remain stable. The purpose of the procedure is not to force performance downward but to provide an evaluation that corresponds more closely to the intended generalisation task.

This distinction is important for interpreting reproducibility studies. Rigorous validation should not be assumed to reveal poor performance. Its purpose is to determine whether performance survives an appropriate test.

Cross-validation and nested model selection

K-fold cross-validation divides observations into multiple subsets and repeatedly fits a model using a proportion of the data while evaluating it on the withheld fold. This can provide more efficient use of available observations than a single train-test partition. [8, 18, 22]

However, conventional cross-validation becomes problematic when the same folds are used both to select hyperparameters and to report performance. Model selection is itself a form of optimisation. Selecting the configuration that performs best across a set of folds can therefore produce an optimistic estimate if the same results are subsequently treated as an unbiased evaluation. [8, 18, 22]

Nested cross-validation separates these functions. [8, 18, 22]

An inner resampling procedure is used for hyperparameter optimisation, while an outer procedure estimates performance of the complete model-selection process. All preprocessing that learns from data should similarly occur within the relevant training folds.

This design is computationally more demanding, particularly when comparing multiple complex algorithms. Nevertheless, the additional computation addresses an important methodological question: whether reported performance represents evaluation of a prespecified model or the best result obtained after searching across multiple candidate configurations.

The distinction is especially relevant when differences between algorithms are small.

Site effects and transportability

Psychiatric datasets increasingly combine participants from multiple hospitals, research centres, geographic regions, or digital platforms. Multisite data increase sample size and potentially improve population diversity, but they also introduce structured heterogeneity.

Sites can differ in recruitment, diagnostic procedures, disease prevalence, referral patterns, demographic composition, socioeconomic characteristics, hardware, questionnaires, clinical workflows, and data preprocessing. Consequently, site membership may be associated with both predictors and outcomes.

A conventional random split distributes observations from most sites across both development and evaluation sets. The model can therefore benefit from site-specific relationships during evaluation.

Site-independent validation asks a more difficult question: can the model generalise to a site whose observations were unavailable during development?

This approach more closely resembles geographical or institutional transportability. A model intended for broad clinical deployment should ideally retain useful performance when transferred beyond its development environment.

Again, deterioration is not inevitable. Stable performance under site-independent evaluation is positive evidence that the predictive structure is not heavily dependent on the withheld site's characteristics. Conversely, substantial deterioration suggests that conventional internal evaluation may not adequately represent transportability.

TRIPOD + AI explicitly recognises clustering and heterogeneous evaluation settings as important reporting considerations and directs researchers dealing with clustered data, including multiple hospitals or datasets, toward additional cluster-specific reporting guidance. [2, 3]

Demographic imbalance and fairness

Population composition creates an additional challenge. Machine-learning performance averaged across an entire dataset can conceal important differences among demographic subgroups.

Suppose one demographic group constitutes the majority of the development data. Optimisation of aggregate loss may preferentially improve performance for that group because it contributes more observations to the objective function. Minority groups can consequently experience poorer discrimination or calibration even when overall model performance appears strong. [1, 17, 20]

Differences may also arise because predictor-outcome relationships themselves vary across demographic or socioeconomic contexts.

Fairness therefore cannot be reduced to removing protected characteristics from the predictor set. Other variables can act as proxies, and genuine differences in measurement distributions may remain.

Evaluation should instead consider whether model performance is maintained across relevant subgroups and population compositions.

TRIPOD + AI incorporated fairness more explicitly than its predecessor and recommends reporting model performance in important subgroups, including sociodemographic groups. It also embeds fairness considerations across multiple parts of prediction-model reporting. [2, 3]

Demographic-balanced validation is one approach to investigating dependence on population composition. By controlling the representation of selected groups during evaluation, investigators can determine whether aggregate estimates are disproportionately driven by the majority population.

Such analyses do not, by themselves, establish that a model is fair. They instead provide evidence about heterogeneity that can be concealed by a single overall metric.

Complex machine learning versus logistic regression

An implicit assumption in some applications of artificial intelligence is that more sophisticated algorithms should outperform conventional statistical models.

This assumption is not universally justified.

Logistic regression estimates a relatively constrained relationship between predictors and the log odds of a binary outcome. Random forests and gradient-boosted trees can represent complex interactions and nonlinearities; support vector machines can construct nonlinear decision boundaries through kernel transformations; and neural networks can approximate highly flexible functions. [6, 7, 19]

These theoretical advantages become practically valuable only when the dataset contains reproducible predictive structure requiring such flexibility.

When the dominant signal is approximately additive or monotonic, logistic regression can remain highly competitive. Flexible models may then produce only marginal improvements despite substantially greater complexity.

Simple models also provide an important methodological benchmark. If a neural network produces an AUROC of 0.90 while logistic regression produces 0.70 under identical evaluation, the difference warrants investigation. If both produce approximately the same discrimination, the claim that algorithmic complexity itself has generated substantial predictive value becomes weaker. [5, 6, 17]

This comparison is particularly important in psychiatric machine learning, where sample sizes may be moderate relative to the heterogeneity of the population and number of candidate predictors. [14, 24]

A rigorous benchmark should therefore evaluate sophisticated algorithms against a strong conventional baseline rather than comparing only multiple machine-learning approaches.

Discrimination, precision-recall performance, and calibration

No single performance statistic completely characterises a prediction model. [13, 16, 21]

AUROC measures the probability that a randomly selected positive observation receives a higher predicted score than a randomly selected negative observation. It is useful for assessing discrimination and is relatively insensitive to a specific classification threshold. [5, 17]

However, AUROC does not directly describe the reliability of predicted probabilities. [5, 17]

AUPRC focuses on precision and recall and can provide complementary information when the positive class is less frequent. Its interpretation is prevalence-dependent, which makes reporting the underlying outcome prevalence important. [2, 3, 4, 15]

Calibration concerns agreement between predicted probabilities and observed outcomes. A model predicting probabilities of approximately 0.80 for a set of observations should ideally observe the outcome in approximately 80% of comparable cases. A classifier can exhibit strong ranking ability while producing poorly calibrated probabilities. [1, 17, 20]

The Brier score provides a summary measure of probabilistic prediction error for binary outcomes. Lower scores indicate smaller squared deviations between predicted probabilities and observed outcomes. [1, 17, 20]

Clinical prediction studies should therefore resist reducing model evaluation to the single highest AUROC. Discrimination, precision-recall performance, calibration, uncertainty, and subgroup behaviour provide different information about potential usefulness. [1, 4, 5, 13, 15]

Contemporary prediction-model reporting guidance similarly emphasises transparent reporting of model performance and subgroup evaluation rather than relying on a single headline statistic. [2, 3]

Reproducibility versus replicability

The terms reproducibility and replicability are sometimes used interchangeably, although they refer to related but distinguishable concepts.

Computational reproducibility concerns whether the same analytical procedure applied to the same data can reproduce the reported findings. This requires sufficiently detailed descriptions of preprocessing, model specification, hyperparameters, randomisation, software, and evaluation procedures.

Replicability addresses whether similar conclusions are obtained when the scientific question is examined using independent data.

A model can therefore be computationally reproducible without being clinically transportable. Researchers may perfectly recreate the original AUROC while discovering that performance declines substantially in a new population. [5, 17]

Conversely, failure to report the analytical pipeline adequately can prevent reproduction even before external replication is attempted.

TRIPOD + AI was developed partly in response to the need for complete and accurate reporting of prediction-model research and explicitly incorporates open-science considerations including protocols, registration, data sharing, and code sharing. [2, 3]

For psychiatric machine learning, trustworthy evidence requires attention to both levels. Analytical procedures should be sufficiently transparent to reproduce, and conclusions about predictive utility should ultimately survive evaluation in independent empirical populations. [2, 3, 14, 24]

Synthetic data as a methodological laboratory

Synthetic data occupy a different evidentiary position from clinical data.

They cannot establish the prevalence of depression, identify a clinically validated biomarker, demonstrate real-world treatment effects, or prove that a particular published model is biased. Their advantage lies elsewhere: the data-generating process can be controlled.

This makes synthetic data useful for methodological experiments.

A researcher can deliberately introduce repeated observations, site heterogeneity, missingness, demographic imbalance, nonlinear effects, or leakage and then examine whether analytical procedures detect or exploit these structures. Because the mechanism is known, differences in performance can be interpreted more directly than in observational clinical datasets where the true data-generating process is unknown.

For the present study, this distinction is fundamental.

The deliberately leakage-prone variables are experimental manipulations rather than naturally observed clinical predictors. Their purpose is to quantify what happens when variables containing post-outcome information enter a predictive pipeline. Likewise, repeated observations and multisite structure provide controlled conditions for comparing row-level, participant-independent, and site-independent evaluation.

Synthetic benchmarking should therefore complement rather than substitute for empirical validation.

The strongest eventual evidence would combine controlled simulation with application of the same analytical framework to one or more independent real-world psychiatric datasets.

Emerging standards for trustworthy psychiatric prediction

The methodological literature increasingly supports a shift from asking whether a machine-learning model can produce a high performance statistic toward asking whether that performance has been estimated transparently and under conditions relevant to intended use. [2, 3]

This shift has several implications.

First, the prediction target and intended timing of prediction must be explicitly defined. A variable available after diagnosis should not be treated as a prospective biomarker without justification.

Second, the unit of independence must correspond to the scientific question. If deployment concerns new individuals, evaluation should prevent participant overlap.

Third, preprocessing and model selection must remain confined to development data.

Fourth, site and demographic heterogeneity should be examined rather than obscured by aggregate statistics.

Fifth, complex algorithms should be compared against appropriately specified simple baselines.

Finally, discrimination should be considered alongside calibration and subgroup performance. [1, 17, 20]

These principles align closely with the direction of contemporary prediction-model reporting. TRIPOD + AI now provides a 27-item framework intended to improve transparent reporting across both regression and machine-learning prediction studies, with additional emphasis on fairness, subgroup performance and open-science practices. [2, 3]

Gap in the existing literature

Despite extensive research on machine-learning prediction in mental health, methodological problems are often considered separately. Studies may investigate a particular algorithm, evaluate a particular dataset, compare cross-validation procedures, examine external transportability, or discuss leakage as a general methodological concern. [8, 18, 22]

What is particularly useful for methodological understanding is a controlled benchmark in which several threats can be manipulated and compared within the same data-generating environment.

The present study addresses this gap by bringing together five components: deliberate post-outcome leakage, repeated participant observations, multisite heterogeneity, comparison of simple and complex

classifiers, and multidimensional performance evaluation.

Rather than asking which algorithm can achieve the highest possible depression-classification score, the benchmark asks whether conclusions about model performance change when the conditions under which performance is measured become more rigorous.

This reframing is important. A reduction in performance after leakage control is scientifically informative because it identifies optimism introduced by the pipeline. Conversely, stability under participant- or site-independent evaluation is also informative because it indicates that the predictive signal survives that particular robustness test.

The benchmark therefore does not presuppose that every methodological safeguard must reduce AUROC. [5, 17]

Its purpose is to determine which conclusions survive methodological scrutiny.

Rationale for the present study

The literature supports three related concerns motivating the present investigation.

First, apparently impressive predictive performance can be difficult to interpret without detailed knowledge of predictor timing, data independence, preprocessing, and validation design. Second, complex algorithms cannot be assumed to provide clinically meaningful advantages over simpler statistical models. Third, aggregate discrimination provides an incomplete representation of robustness when calibration, site transportability, participant independence, and demographic heterogeneity are relevant. [1, 10, 17, 20]

The present synthetic benchmark was designed to make these mechanisms experimentally observable.

By evaluating logistic regression, random forest, XGBoost, support vector machine, and neural-network classifiers under progressively different evaluation conditions, the study tests whether apparent algorithmic superiority persists after leakage is removed and independence constraints are strengthened. [6, 7, 19]

The central methodological proposition is therefore not that psychiatric machine learning necessarily performs poorly. Rather, it is that claims of predictive performance should be proportional to the rigor of the evaluation used to establish them. [14, 24]

A model that remains accurate, calibrated, and stable after leakage control and evaluation in independent participants, sites, and relevant demographic groups provides stronger evidence than an otherwise identical model assessed only using a conventional random split.

Accordingly, trustworthy computational psychiatry requires moving beyond the question “How high is the reported performance?” toward the more consequential question: “Under what validation conditions does that performance survive?”

Methods

Study design

This study used a controlled synthetic benchmarking design to examine the influence of data leakage, validation strategy, and algorithmic complexity on machine-learning classification of depression status. The study was methodological rather than diagnostic or epidemiological: its objective was not to estimate the prevalence of depression in a real population or to establish clinically validated biomarkers, but to determine how analytical decisions influence apparent predictive performance under a known data-generating framework.

A synthetic multisite psychiatric dataset was generated containing 13,159 observations from 8,000 unique synthetic participants distributed across 10 simulated sites. Participants contributed between one and three observations, thereby introducing a repeated-measures structure that permitted direct comparison between row-level and participant-independent validation.

The dataset incorporated demographic, socioeconomic, symptom-related, behavioural, sleep, physiological, cognitive, treatment, device, and site-level information. Structured missingness and site effects were introduced to approximate methodological characteristics frequently encountered in multisite psychiatric prediction research. [14, 24]

Two deliberately leakage-prone post-outcome variables were additionally generated. These variables allowed the effect of information leakage to be examined experimentally by comparing models fitted with and without predictors that partially encoded knowledge of depression case status.

The principal benchmark compared five classification approaches: logistic regression, random forest, support vector machine (SVM), extreme gradient boosting (XGBoost), and a feed-forward neural-network classifier. [6, 7, 19]

Performance was evaluated primarily under three completed validation conditions: (1) conventional random row-level train-test splitting, (2) participant-independent validation, and (3) site-independent validation.

A separate leakage experiment compared otherwise equivalent models with and without deliberately leakage-prone variables.

Synthetic population

The synthetic cohort contained 8,000 unique participants. Each participant was assigned a stable synthetic identifier (*participant*) that was retained across repeated observations.

The number of observations per participant ranged from one to three. Participants were probabilistically assigned one, two, or three observations, producing a final analytical dataset of 13,159 rows.

The repeated-measures design was intentional. It enabled conventional random row splitting to be contrasted with participant-level separation in which all observations belonging to the same individual were restricted to a single analytical partition.

Participants were distributed across 10 simulated sites ($SITE_{01}$ to $SITE_{10}$). Site sizes were intentionally unequal to approximate the imbalance commonly observed in multicentre studies. Site membership was associated with modest differences in the data-generating process, including variation in underlying depression liability and selected demographic distributions.

No row in the dataset corresponds to an actual person, patient, clinical record, or research participant.

Synthetic demographic and socioeconomic variables

The demographic component included age, sex assigned at birth, gender, ethnicity group, educational attainment, income band, and employment status.

Age was restricted to 18–75 years. Site-dependent variation was incorporated into the age distribution to prevent sites from being completely exchangeable.

Sex assigned at birth was represented as female or male in the synthetic data-generating model. A separate gender variable contained woman, man, and a small nonbinary category.

Ethnicity was represented using deliberately generic labels ($Group_A$ through $Group_E$). These categories do not correspond to specific real-world ethnic populations. This design decision prevented simulated demographic categories from being interpreted as empirical estimates of psychiatric risk for identifiable ethnic groups.

Education was represented as years of completed education and restricted to a plausible adult range. Income was represented categorically as very low, low, middle, upper-middle, or high. Employment status was represented as employed, student, unemployed, or not working/retired.

These variables were generated to introduce population heterogeneity rather than to reproduce the demographic composition of a specific country or clinical population.

Latent participant characteristics

Stable latent characteristics were generated for each participant to create correlated multivariate observations while preserving a known synthetic structure.

Three principal latent dimensions were used: general psychological distress, sleep-related vulnerability, and cognitive functioning.

These participant-level latent variables remained stable across repeated observations, while visit-specific stochastic variation was added at each observation. This combination created within-participant correlation without making repeated records identical.

A site-specific risk component was additionally incorporated into the latent depression-generating process. Consequently, the prevalence and predictor distributions could differ modestly across simulated sites.

The purpose of these latent factors was not to propose a causal model of depression. They were computational devices for generating realistic correlation structures among observed synthetic variables.

Depression-related variables and outcome definition

A continuous latent depression liability was first generated as a function of participant-level distress, sleep vulnerability, cognition, age, sex, employment status, and site-specific effects, together with stochastic variation.

A synthetic PHQ-9-like total symptom variable ($phq9 \rightarrow tal$) was subsequently generated with a possible range of 0–27. [9]

A synthetic GAD-7-like anxiety variable ($gad7 \rightarrow tal$) was generated with a range of 0–21 and was correlated with the underlying distress process.

The primary binary outcome, *depression_{case}*, was operationalised as:

depression_case = 1 when synthetic PHQ-9-like total ≥ 10 ; otherwise depression_case = 0. [9]

This produced an overall synthetic depression-case prevalence of approximately 35.4%.

The PHQ-9-like score was used solely to operationalise the target and was therefore excluded from all leakage-controlled predictive models. Including the variable would have constituted direct target leakage because the binary outcome was mathematically derived from it. [9, 13, 16, 21]

The generated symptom totals are synthetic approximations for methodological benchmarking. They are not participant responses to copyrighted clinical instruments and should not be interpreted as validated psychiatric assessments.

Behavioural and sleep-related predictors

The synthetic dataset contained several behavioural variables intended to approximate information that could be obtained through self-report or digital phenotyping.

Sleep duration was represented as average hours of sleep per night. Sleep variability represented estimated night-to-night variability in sleep duration. Both were associated probabilistically with latent distress and sleep vulnerability.

Daily physical activity was represented through step count, while sedentary behaviour was represented as hours per day.

Higher latent distress generally shifted the generating distributions toward shorter sleep duration, greater sleep variability, fewer daily steps, and greater sedentary time. Substantial stochastic variation was retained so that none of these variables individually determined depression status.

These relationships represent simulation assumptions rather than estimated clinical effect sizes.

Physiological predictors

Two synthetic physiological measures were generated: resting heart rate and heart-rate variability.

Resting heart rate was represented in beats per minute. Heart-rate variability was represented using an RMSSD-like value in milliseconds.

Both variables were probabilistically associated with distress and age while containing substantial random variation.

A wearable-device category was also generated, containing three synthetic device types ($Device_A$, $Device_B$, and $Device_C$). The device variable was included to permit methodological investigation of measurement-platform structure and possible device-associated effects.

The physiological measures should not be interpreted as empirical estimates of cardiovascular characteristics among individuals with depression.

Cognitive predictors

Two synthetic cognitive variables were included.

Reaction time was represented in milliseconds and was generated as a function of latent distress, latent cognition, age, and random variation.

A cognitive composite score was generated from latent cognitive functioning, distress, age, and stochastic error.

These variables provided continuous predictors whose distributions partially overlapped between depression cases and non-cases, thereby contributing useful but imperfect classification information.

Psychosocial variables

Perceived stress was represented on a synthetic 0–10 scale and was positively associated with latent psychological distress.

Social support was represented on a synthetic 0–10 scale and was inversely associated, on average, with distress.

Substantial random variation was retained in both variables to avoid deterministic relationships with depression status.

Treatment variables

Current antidepressant treatment and current psychotherapy were represented as binary synthetic indicators.

Their generating probabilities were associated with depression status and repeated observation number. Consequently, these variables require cautious interpretation because treatment may occur downstream of clinical recognition.

For the methodological benchmark, they were retained as part of the broader simulated psychiatric data environment. They should not be interpreted as causal predictors of depression.

Deliberately leakage-prone variables

Two variables were specifically generated to create a controlled leakage experiment.

The first, $cl \in ician_referral_{after_{screen}}$, represented a synthetic referral decision made after symptom screening. Its probability was deliberately made strongly dependent on depression case status.

The second, $treatment_{plan_{code}}$, represented a synthetic post-assessment clinical management category with four levels: none, low, moderate, and high. Its distribution was also intentionally dependent on depression case status.

These variables were designed to encode information that would not legitimately be available in a prospective pre-diagnostic prediction task.

They were therefore excluded from the primary leakage-controlled predictor set.

For the leakage experiment, the variables were deliberately reintroduced. The resulting change in performance quantified the degree to which post-outcome information could inflate apparent predictive accuracy within the controlled simulation.

This experiment was designed to illustrate a mechanism of leakage and does not estimate the frequency or magnitude of leakage in published psychiatric studies.

Missing-data mechanism

Missingness was introduced into selected variables to approximate incomplete data frequently encountered in psychiatric and digital-health research.

Variables affected included income, sleep duration, sleep variability, daily step count, heart-rate variability, cognitive composite score, and social support.

Missingness was not completely random. Its probability was modestly increased among observations with higher synthetic depressive symptom burden and at selected sites.

The missingness process therefore contained elements resembling missing-at-random or structured missingness rather than uniform random deletion.

All preprocessing decisions involving missing values were restricted to training data within each relevant analytical partition to prevent information from the evaluation data influencing model development.

Predictor definition

Three conceptually distinct variable groups were defined before model evaluation.

Identifier and structural variables

Participant identifiers and visit numbers were retained for data management and partition construction but were not treated as substantive psychiatric biomarkers.

Participant ID was never supplied to predictive models. [13, 16, 21]

Site membership was used principally to construct site-independent validation partitions and to characterise the multisite structure.

Leakage-controlled predictors

The principal predictor set excluded:

- *participant* ; - the outcome variable *depression_{case}*; - *phq9_{→tal}*, because the outcome was directly defined from this score; - $cl \in \text{ician}_{referral_{after_screen}}$; and - *treatment_{plan_{code}}*. [9]

The remaining eligible demographic, anxiety-related, behavioural, sleep, physiological, cognitive, psychosocial, treatment, and device variables constituted the leakage-controlled feature space.

Leakage-inclusive predictors

For the leakage experiment, $cl \in \text{ician}_{referral_{after_screen}}$ and *treatment_{plan_{code}}* were restored to the candidate feature set.

The defining PHQ-9-like total remained excluded because its inclusion would represent trivial direct reconstruction of the outcome rather than the subtler post-outcome leakage targeted by the experiment. [9]

Preprocessing

Preprocessing was conducted separately within each training partition.

Continuous variables were prepared for algorithms requiring feature scaling. Scaling parameters were estimated using training data only and subsequently applied to held-out observations.

Categorical variables were transformed into model-compatible representations using categories established from the training data.

Missing values were handled through training-derived preprocessing rather than through calculations performed on the complete dataset.

This ordering was essential to the leakage-controlled design. Any transformation requiring estimation from data was treated as part of the model-development pipeline and therefore prevented from learning characteristics of held-out observations.

The same conceptual preprocessing framework was applied across models wherever possible to ensure that differences in performance reflected modelling approaches rather than inconsistent access to information.

Machine-learning models

Five binary classifiers were evaluated.

Logistic regression

Logistic regression served as the principal simple baseline.

Its inclusion was essential because the primary methodological question concerned whether flexible machine-learning algorithms provided substantial incremental discrimination once leakage was controlled. [6]

Predicted probabilities from logistic regression were used directly for AUROC, AUPRC, and Brier-score calculations. [1, 4, 5, 15, 17]

Random forest

A random-forest classifier represented a nonlinear ensemble method capable of modelling interactions and nonlinear predictor-outcome relationships.

The algorithm combined predictions across multiple decision trees fitted to resampled data and feature subsets.

Extreme gradient boosting

XGBoost was included as a boosted-tree approach. Unlike random forests, which average multiple partially independent trees, gradient boosting sequentially constructs trees designed to reduce residual prediction error. [6, 7]

Its inclusion provided a strong nonlinear tabular-data benchmark.

Support vector machine

An SVM classifier was used to represent margin-based machine learning. [6, 19]

Predictor scaling was incorporated within the training pipeline because SVM estimation is sensitive to differences in predictor magnitude. [6, 19]

Probability estimates suitable for probabilistic evaluation were obtained from the fitted classifier.

Neural network

A feed-forward multilayer perceptron classifier represented the neural-network family.

The neural network received the same leakage-controlled information available to the other classifiers and was evaluated using the same held-out observations. [6]

Its purpose was not to optimise a deep-learning architecture but to test whether additional nonlinear representational flexibility yielded meaningful improvement over conventional statistical and machine-learning approaches.

Random row-level validation

The first evaluation condition used a conventional random row-level train-test split.

Individual observations were randomly assigned to training and test sets without enforcing participant-level grouping.

This procedure approximated a commonly used machine-learning workflow but allowed repeated observations from the same participant to potentially appear in both partitions.

No participant identifier itself was included as a model feature.

The purpose of this condition was to establish a conventional internal-validation benchmark against which independence-preserving approaches could be compared.

Under leakage-controlled random validation, observed AUROCs across the five algorithms ranged from 0.9488 to 0.9548. [5, 17]

Participant-independent validation

Participant-independent validation was performed by assigning participants rather than rows to analytical partitions.

All observations belonging to a given *participant* were restricted to the same side of the training-test division.

This prevented repeated observations from the same synthetic person from contributing information to both model development and evaluation.

The procedure therefore estimated performance in unseen synthetic participants rather than merely unseen observations.

Participant-independent AUROCs ranged from 0.9516 to 0.9585 across classifiers. [5, 17]

The fact that these estimates did not decrease relative to random row-level validation was retained as an empirical result rather than interpreted as evidence that the validation procedure was unnecessary.

Site-independent validation

Site-independent validation was used to evaluate generalisation beyond the site structure represented during model development.

Observations from the evaluation site were withheld from training, preventing models from learning directly from participants recruited within that site.

The training data therefore consisted of observations from the remaining simulated sites, while performance was evaluated using observations belonging to an unseen site partition.

This analysis was intended to approximate institutional distribution shift rather than formal external clinical validation.

Because all sites were generated by the same overarching simulation mechanism, the procedure should be described as site-independent internal benchmarking, not as external validation in an independent empirical population.

Across classifiers, site-independent AUROCs ranged from 0.9502 to 0.9555. [5, 17]

Leakage experiment

A separate experiment quantified performance inflation associated with deliberately leakage-prone variables.

Models were first evaluated using the leakage-controlled feature set. The post-screen referral variable and treatment-plan variable were then added while maintaining the broader modelling framework.

The change in AUROC, AUPRC, and Brier score was interpreted as performance attributable partly to access to post-outcome information. [1, 4, 5, 15, 17]

Under random row-level validation, leakage-inclusive models achieved AUROCs of approximately 0.988–0.989 and AUPRCs of approximately 0.981–0.983. [4, 5, 15, 17]

For logistic regression specifically, inclusion of the leakage-prone variables produced an AUROC of 0.9891, AUPRC of 0.9827, and Brier score of 0.0367. [1, 4, 5, 15, 17]

After removal of those variables, corresponding values were 0.9548, 0.9253, and 0.0788.

These results quantify leakage effects within the simulation and should not be interpreted as estimates of bias in real psychiatric datasets.

Performance measures

Three principal metrics were used.

Area under the receiver operating characteristic curve

AUROC was the primary discrimination measure. [5, 17]

AUROC assesses the ability of predicted scores to rank positive observations above negative observations across possible classification thresholds. A value of 0.50 corresponds to chance-level ranking, whereas values approaching 1.00 indicate stronger discrimination. [5, 17]

AUROC was calculated from continuous predicted probabilities or decision scores rather than from thresholded class predictions. [5, 17]

Area under the precision-recall curve

AUPRC was reported as a complementary discrimination measure. [4, 15]

Precision-recall analysis is particularly informative when class frequencies are unequal because it focuses on the relationship between positive predictive value and sensitivity. [4, 15]

Because AUPRC depends on outcome prevalence, the synthetic depression-case prevalence of 35.4% is reported alongside the metric. [4, 15]

Brier score

The Brier score was used to quantify probabilistic prediction error. [1, 17, 20]

For binary outcomes, the score is calculated from the squared difference between the predicted probability and observed outcome, averaged across observations. Lower values indicate smaller probabilistic error.

Brier scores were interpreted together with discrimination measures rather than as replacements for AUROC or AUPRC. [1, 4, 5, 15, 17]

Comparison of algorithmic complexity

A principal objective was to determine whether sophisticated nonlinear algorithms materially outperformed the logistic-regression baseline.

Comparisons therefore focused on the magnitude and consistency of performance differences rather than simply identifying the numerically highest metric.

Under participant-independent validation, logistic regression achieved an AUROC of 0.9585, AUPRC of 0.9383, and Brier score of 0.0762. [1, 4, 5, 15, 17]

The remaining algorithms produced closely similar discrimination. Under site-independent validation, AUROCs across all five classifiers fell within the narrow range of 0.9502–0.9555. [5, 17]

Accordingly, differences of only several thousandths in AUROC were not automatically interpreted as evidence of meaningful superiority. [5, 17]

Calibration assessment

Probabilistic performance was evaluated using the Brier score as the principal calibration-sensitive summary metric. [1, 17, 20]

The Brier score reflects both discrimination and calibration components and therefore should not be interpreted as a pure calibration statistic. Nevertheless, it provides useful information about the accuracy of predicted probabilities that is not captured by AUROC alone. [1, 5, 17, 20]

Where model discrimination was similar, Brier-score differences were examined to determine whether comparable ranking performance concealed differences in probability estimation. [1, 17, 20]

Future empirical validation should supplement Brier scores with calibration plots, calibration intercepts, and calibration slopes. [1, 17, 20]

Demographic robustness analysis

The dataset was constructed to support demographic robustness analyses using age, sex, gender, ethnicity group, socioeconomic position, education, and employment characteristics.

The prespecified purpose of demographic-balanced evaluation is to determine whether aggregate model performance depends materially on the demographic composition of the evaluation sample.

However, a complete demographic-balanced benchmark had not produced final validated estimates at the stage of the primary analysis reported here. It is therefore treated as a prespecified extension rather than reported as a completed result.

No demographic-balanced performance statistic is fabricated or inferred from the completed analyses.

Nested cross-validation

Nested cross-validation was prespecified to separate hyperparameter optimisation from performance estimation. [8, 18, 22]

Conceptually, an inner cross-validation loop is used exclusively for model and hyperparameter selection, while an outer loop provides an independent estimate of the performance of the complete model-development procedure. [8, 18, 22]

All preprocessing, imputation, scaling, encoding, and model selection should occur inside the appropriate training folds.

A complete nested optimisation across all five algorithms was computationally more intensive than the principal benchmark and did not produce a complete validated set of estimates within the primary computational run. [8, 18, 22]

Nested-cross-validation results are therefore not reported as though they had been completed. The principal reported analyses are based on the completed random, participant-independent, site-independent, and leakage-comparison experiments. [8, 18, 22]

Reproducibility and randomisation

Synthetic data generation was performed using a fixed pseudorandom seed to permit deterministic reconstruction of the generated dataset.

The seed used for the final dataset was 20260924.

The final dataset contained 13,159 rows, 8,000 unique participant identifiers, 10 sites, and 28 recorded variables before exclusion of structural, target-defining, or leakage-prone fields from particular model specifications.

Fixing the random seed does not make the generated data empirical. It permits exact computational reproduction of this particular simulation.

Statistical interpretation

The study emphasised effect magnitude and robustness across validation conditions rather than null-hypothesis significance testing of minor differences between algorithms. [8, 18, 22]

Model performance was interpreted descriptively using AUROC, AUPRC, and Brier scores. [1, 4, 5, 10, 15]

A numerically higher AUROC was not automatically described as a superior model when the absolute difference was small. This approach was particularly important because the benchmark demonstrated tightly clustered performance among several classifiers. [5, 17]

The primary comparisons concerned broader methodological contrasts: leakage-inclusive versus leakage-controlled modelling, row-level versus participant-independent validation, and conventional versus site-independent evaluation.

Software and computational environment

The synthetic dataset and principal analyses were implemented programmatically using Python-compatible data-science and machine-learning workflows.

Tabular data processing was performed using standard numerical and dataframe operations. Classification procedures used established implementations of logistic regression, random forest, SVM, gradient-boosted trees, and multilayer perceptron models. Performance calculations were based on standard definitions of AUROC, average precision/AUPRC, and Brier score. [1, 4, 5, 6, 7]

For the final reproducibility package, exact software versions, package versions, preprocessing objects, hyperparameter settings, random states, and execution scripts should be archived together with the synthetic dataset so that the complete benchmark can be reconstructed.

Ethical considerations

This study used exclusively computationally generated synthetic observations.

No real patients or research participants were recruited. No medical records, identifiable personal information, biological samples, or human participant data were accessed or analysed.

The synthetic participant identifiers do not correspond to real individuals, and the demographic and clinical characteristics were algorithmically generated.

Accordingly, the study does not constitute human-subject research involving identifiable or private participant information. Formal institutional ethical approval and informed consent were therefore not required for generation and analysis of the synthetic benchmark.

Scope of inference

The inferential scope of the study was deliberately restricted.

The results demonstrate how leakage and validation choices can affect predictive performance within the specified synthetic data-generating environment. They do not establish the prevalence of methodological errors in published depression research, prove that previously reported depression biomarkers are invalid, estimate real-world diagnostic performance, or establish that the predictors used here constitute clinically useful biomarkers.

Likewise, high leakage-controlled performance in the synthetic dataset should not be interpreted as evidence that depression can be predicted with equivalent accuracy in clinical populations.

Instead, the simulation provides a controlled methodological experiment in which the consequences of selected analytical decisions can be directly observed.

External replication using genuine psychiatric datasets and prospectively collected independent populations is necessary before conclusions regarding clinical utility can be drawn.

Results

Dataset characteristics

The final synthetic dataset comprised 13,159 observations from 8,000 unique synthetic participants distributed across 10 simulated sites. Participants contributed between one and three observations, creating a repeated-measures structure suitable for comparing conventional row-level splitting with participant-independent validation.

The binary synthetic depression outcome was defined using a PHQ-9-like total score threshold of ≥ 10 . Overall depression-case prevalence was 35.4%. Because the PHQ-9-like total score directly defined the outcome, it was excluded from all predictive models to prevent direct target leakage.

The analytical dataset included demographic and socioeconomic characteristics, anxiety-related symptoms, sleep and physical-activity indicators, physiological measures, cognitive measures, psychosocial characteristics, treatment indicators, and device-related information. Structured missingness was present in selected variables.

Two additional variables—post-screen clinician referral and treatment-plan code—were deliberately constructed as leakage-prone post-outcome predictors and were excluded from the primary leakage-controlled analyses.

Overall model performance

Five classification approaches were compared: logistic regression, random forest, support vector machine (SVM), neural network, and extreme gradient boosting (XGBoost).

Across the completed leakage-controlled analyses, all five algorithms demonstrated high discrimination. However, differences between models were small.

Across conventional random row-level validation, AUROC ranged from 0.9488 to 0.9548. Under participant-independent validation, AUROC ranged from 0.9516 to 0.9585, while site-independent AUROCs ranged from 0.9502 to 0.9555.

Thus, increasing validation stringency from random row-level splitting to participant- and site-independent evaluation did not produce a substantial deterioration in discrimination in this synthetic benchmark.

Table 1
Leakage-controlled predictive performance across validation strategies

Validation strategy	Model	AUROC	AUPRC	Brier score
Random row split	Logistic regression	0.9548	0.9253	0.0788
Random row split	Random forest	0.9512	0.9189	0.0863
Random row split	SVM	0.9488	0.9116	0.0819
Random row split	Neural network	0.9545	0.9254	0.0787
Random row split	XGBoost	0.9534	0.9242	0.0797
Participant-independent	Logistic regression	0.9585	0.9383	0.0762
Participant-independent	Random forest	0.9572	0.9366	0.0825
Participant-independent	SVM	0.9516	0.9246	0.0795
Participant-independent	Neural network	0.9571	0.9361	0.0774
Participant-independent	XGBoost	0.9577	0.9356	0.0770
Site-independent	Logistic regression	0.9554	0.9339	0.0794
Site-independent	Random forest	0.9513	0.9260	0.0878
Site-independent	SVM	0.9502	0.9183	0.0830
Site-independent	Neural network	0.9548	0.9330	0.0794
Site-independent	XGBoost	0.9555	0.9356	0.0790

Note. Lower Brier scores indicate lower probabilistic prediction error. All results in Table 1 exclude the PHQ-9-like target-defining score and the two deliberately leakage-prone post-outcome variables.

Conventional random row-level validation

Under conventional random row splitting, logistic regression achieved an AUROC of 0.9548, AUPRC of 0.9253, and Brier score of 0.0788.

The neural network produced nearly identical performance, with an AUROC of 0.9545, AUPRC of 0.9254, and Brier score of 0.0787. XGBoost achieved an AUROC of 0.9534, AUPRC of 0.9242, and Brier score of 0.0797.

Random forest produced an AUROC of 0.9512 and AUPRC of 0.9189, while its Brier score of 0.0863 was the highest among the five models under this validation condition.

SVM produced the lowest discrimination under random splitting, although performance remained high in absolute terms: AUROC was 0.9488, AUPRC was 0.9116, and the Brier score was 0.0819.

The absolute difference between the highest and lowest AUROC was therefore only 0.0060.

These findings provided little evidence of a practically large discrimination advantage for the more complex algorithms over logistic regression under conventional leakage-controlled validation.

Participant-independent validation

Participant-independent validation prevented observations belonging to the same synthetic participant from appearing across development and evaluation partitions.

Contrary to the prespecified hypothesis that participant separation would reduce apparent performance, discrimination did not deteriorate.

Logistic regression achieved an AUROC of 0.9585, the highest value observed under this validation condition. Its AUPRC was 0.9383, and its Brier score was 0.0762.

XGBoost produced an AUROC of 0.9577, AUPRC of 0.9356, and Brier score of 0.0770. Random forest achieved an AUROC of 0.9572 and AUPRC of 0.9366, although its Brier score was comparatively higher at 0.0825.

The neural network achieved an AUROC of 0.9571, AUPRC of 0.9361, and Brier score of 0.0774.

SVM again produced the lowest discrimination, with an AUROC of 0.9516, AUPRC of 0.9246, and Brier score of 0.0795.

The difference between the highest and lowest participant-independent AUROC was 0.0069.

Importantly, logistic regression performed at least as well as the more complex classifiers according to AUROC and also produced the lowest Brier score.

Comparison of random and participant-independent validation

The absence of a decline under participant-independent validation was a notable result.

Logistic-regression AUROC increased numerically from 0.9548 under random row splitting to 0.9585 under participant-independent validation. Similar small numerical increases occurred for random forest, SVM, neural network, and XGBoost.

These differences should not be interpreted as evidence that participant-independent splitting inherently improves prediction. Instead, they indicate that the participant-level dependence embedded in this particular simulation was insufficient to generate meaningful optimism in the conventional random split.

Accordingly, the prespecified participant-independence hypothesis was not supported by the completed benchmark.

This finding also demonstrates an important property of methodological robustness testing: more stringent validation need not necessarily reduce performance. When predictive relationships generalise across participants, performance may remain stable despite stronger independence constraints.

Site-independent validation

Site-independent evaluation tested whether predictive performance persisted when observations from an evaluation site were excluded from model development.

Performance again remained high.

XGBoost produced the numerically highest site-independent AUROC at 0.9555, with an AUPRC of 0.9356 and Brier score of 0.0790.

Logistic regression was essentially indistinguishable in discrimination, achieving an AUROC of 0.9554, AUPRC of 0.9339, and Brier score of 0.0794.

The neural network achieved an AUROC of 0.9548, AUPRC of 0.9330, and Brier score of 0.0794.

Random forest achieved an AUROC of 0.9513, AUPRC of 0.9260, and Brier score of 0.0878.

SVM produced an AUROC of 0.9502, AUPRC of 0.9183, and Brier score of 0.0830.

The complete range of site-independent AUROC values was therefore only 0.0053.

There was no substantial deterioration relative to conventional random validation. The prespecified hypothesis that site-independent evaluation would produce a meaningful decline in performance was therefore not supported within the present synthetic data-generating environment.

Because all sites originated from the same overarching simulation mechanism, this finding should be interpreted as evidence of robustness to the simulated site structure rather than evidence of external clinical validity.

Effect of deliberate data leakage

The clearest methodological effect was observed when the deliberately leakage-prone post-outcome predictors were introduced.

Under random row-level validation, leakage-inclusive pipelines produced AUROCs of approximately 0.988–0.989 and AUPRCs of approximately 0.981–0.983, substantially exceeding the corresponding leakage-controlled estimates.

Logistic regression provides a direct illustration.

Table 2
Effect of post-outcome leakage on logistic-regression performance

Pipeline	AUROC	AUPRC	Brier score
Leakage-inclusive	0.9891	0.9827	0.0367
Leakage-controlled	0.9548	0.9253	0.0788
Absolute difference	+ 0.0343	+ 0.0574	−0.0421

For logistic regression, inclusion of the post-outcome variables increased AUROC by 0.0343, from 0.9548 to 0.9891.

The corresponding increase in AUPRC was 0.0574, from 0.9253 to 0.9827.

The Brier score decreased from 0.0788 to 0.0367, representing an absolute difference of 0.0421.

Thus, the leakage-inclusive model appeared substantially stronger according to discrimination and probabilistic prediction error despite gaining its additional information from variables deliberately generated downstream of the outcome.

The leakage-inflation hypothesis was therefore supported.

Importantly, leakage did not transform a non-predictive model into a predictive one. Leakage-controlled models already achieved AUROCs around 0.95. Instead, post-outcome information elevated an already high-performing model toward apparently near-perfect discrimination.

This distinction illustrates that leakage can materially exaggerate performance even when legitimate predictors contain substantial predictive signal.

Complex models versus logistic regression

No complex machine-learning algorithm demonstrated a large or consistent advantage over logistic regression.

Under random row-level validation, the highest AUROC was produced by logistic regression (0.9548), although the neural network produced the numerically highest AUPRC (0.9254) and lowest Brier score (0.0787) by extremely small margins.

Under participant-independent validation, logistic regression produced the highest AUROC (0.9585), highest AUPRC (0.9383), and lowest Brier score (0.0762) among the evaluated classifiers.

Under site-independent validation, XGBoost produced the highest AUROC (0.9555) compared with 0.9554 for logistic regression, an absolute difference of only 0.0001. XGBoost also produced a slightly higher AUPRC, 0.9356 versus 0.9339.

These differences were descriptively negligible.

Table 3
AUROC differences relative to logistic regression

Validation strategy	Logistic regression	Random forest	SVM	Neural network	XGBoost
Random row split	0.9548	-0.0036	-0.0060	-0.0003	-0.0014
Participant-independent	0.9585	-0.0013	-0.0069	-0.0014	-0.0008
Site-independent	0.9554	-0.0041	-0.0052	-0.0006	+ 0.0001

Note. Logistic regression is shown as the reference AUROC. Values for other models represent absolute AUROC differences relative to logistic regression within the same validation strategy.

Across all three validation conditions, the neural network differed from logistic regression by no more than 0.0014 AUROC, while XGBoost differed by no more than 0.0014.

Random forest and SVM generally produced slightly lower discrimination.

These results supported the hypothesis that increasing model complexity would provide little incremental discrimination after leakage control.

Probabilistic performance

Brier-score results broadly reinforced the discrimination findings.

Under random validation, Brier scores ranged from 0.0787 to 0.0863. Under participant-independent validation, they ranged from 0.0762 to 0.0825, while site-independent values ranged from 0.0790 to 0.0878.

Logistic regression demonstrated consistently strong probabilistic performance. It produced the lowest Brier score under participant-independent validation and remained close to the lowest score under both

random and site-independent validation.

Random forest produced the highest Brier score in all three validation conditions: 0.0863, 0.0825, and 0.0878, respectively.

The findings demonstrate that similar AUROCs did not imply identical probability error. For example, random forest and logistic regression differed relatively little in discrimination but showed more noticeable differences in Brier score.

However, because the Brier score combines aspects of discrimination and calibration, these findings should not be interpreted as a complete assessment of calibration slope or calibration-in-the-large.

Summary of hypothesis findings

The completed analyses produced a mixed pattern of support for the prespecified hypotheses.

H1, the leakage-inflation hypothesis, was supported. Introducing deliberately post-outcome predictors substantially increased AUROC and AUPRC and reduced Brier error.

H2, the participant-independence hypothesis, was not supported. Participant-independent validation did not reduce performance relative to random row-level splitting.

H3, the site-generalisation hypothesis, was not supported. Site-independent validation produced discrimination similar to conventional random validation.

H4, the complexity hypothesis, was supported descriptively. Random forest, SVM, XGBoost, and the neural network did not provide a substantial or consistent performance advantage over logistic regression.

H5, the calibration hypothesis, received partial descriptive support. Models with relatively similar discrimination showed differences in Brier scores, particularly random forest compared with logistic regression. Formal calibration slopes and intercepts were not estimated in the completed analysis.

H6, concerning demographic robustness, was not evaluated in the completed primary analysis. No conclusion regarding this hypothesis is therefore drawn.

Principal findings

Three principal findings emerged from the benchmark.

First, deliberate post-outcome leakage produced a substantial increase in apparent predictive performance. For logistic regression, AUROC increased from 0.9548 to 0.9891, while AUPRC increased from 0.9253 to 0.9827.

Second, stronger independence constraints did not cause performance to collapse. Leakage-controlled models retained AUROCs of approximately 0.95–0.96 under random, participant-independent, and site-independent validation.

Third, increasing algorithmic complexity yielded little additional discrimination. Logistic regression performed comparably to—and in several analyses numerically better than—random forest, SVM, XGBoost, and the neural network.

Taken together, the findings indicate that the most consequential methodological manipulation in this controlled benchmark was access to leakage-prone information rather than choice of classifier.

Discussion

Principal findings

This study examined how data leakage, validation design, and algorithmic complexity influenced apparent depression-classification performance within a controlled synthetic multisite psychiatric benchmark. Three principal findings emerged.

First, deliberate introduction of post-outcome information substantially inflated apparent predictive performance. In the clearest comparison, logistic regression achieved an AUROC of 0.9891 and AUPRC of 0.9827 when leakage-prone variables were included, compared with an AUROC of 0.9548 and AUPRC of 0.9253 after those variables were removed. The Brier score simultaneously changed from 0.0367 to 0.0788. Thus, variables that encoded information downstream of the outcome made an already strongly predictive model appear close to perfectly discriminative. [1, 4, 5, 13, 15]

Second, the expected deterioration under participant- and site-independent validation did not occur. Leakage-controlled models retained AUROCs of approximately 0.95–0.96 across random row-level, participant-independent, and site-independent evaluation. The hypotheses that participant separation and site separation would materially reduce discrimination were therefore not supported within this simulation. [5, 17]

Third, increased algorithmic complexity provided little additional predictive value. Logistic regression performed similarly to random forest, SVM, XGBoost, and the neural network. Under participant-independent validation, logistic regression achieved the highest observed AUROC and AUPRC and the lowest Brier score among the evaluated models. Under site-independent validation, the difference between XGBoost and logistic regression was only 0.0001 AUROC. [1, 4, 5, 6, 7]

Taken together, these results suggest that, within the present controlled environment, the information supplied to the model and the integrity of the analytical pipeline mattered more than the sophistication of the classification algorithm.

Data leakage can inflate already strong models

The clearest result concerned deliberate data leakage.

It is tempting to conceptualise leakage as a problem relevant primarily when a weak classifier unexpectedly produces excellent performance. The present experiment demonstrates a different possibility. Leakage can materially exaggerate the apparent strength of a model that already contains genuine predictive signal within the data-generating environment.

The leakage-controlled logistic-regression model achieved an AUROC of 0.9548 under random validation. This was already high. Introducing the two deliberately post-outcome variables nevertheless increased AUROC to 0.9891. [5, 17]

The absolute increase of 0.0343 may initially appear modest because both values occupy the upper end of the AUROC scale. However, the interpretation of a model with an AUROC around 0.95 can differ substantially from one approaching 0.99. Near-perfect discrimination may encourage much stronger claims concerning biomarker reliability or potential clinical applicability. [5, 10, 17]

The effect was even more apparent for AUPRC, which increased by 0.0574, from 0.9253 to 0.9827. [4, 15]

The Brier-score difference was also substantial. The leakage-inclusive model produced a Brier score of 0.0367 compared with 0.0788 after leakage control. Thus, leakage affected not only ranking performance but also the apparent accuracy of predicted probabilities. [1, 17, 20]

The implication is that exceptionally high performance should not automatically be interpreted as evidence that a model has discovered exceptionally informative biological, behavioural, or psychological markers. Researchers must first establish whether any predictors directly or indirectly encode knowledge of the target. [10]

Leakage is fundamentally a temporal and informational problem

A particularly important distinction concerns when information becomes available.

The two manipulated variables in this study were not arbitrary noise features. They represented post-screen clinical actions: clinician referral and treatment planning. Both could plausibly occur in real healthcare datasets, and both could correlate strongly with psychiatric diagnosis.

The problem is not that such variables are statistically associated with depression. The problem is that their association partly exists because clinical information about depression has already influenced them.

If the intended task is to predict previously unrecognised depression, a variable generated after screening or diagnosis cannot legitimately function as a prospective predictor.

This distinction has broader implications for electronic health record and routinely collected clinical data. Referral patterns, treatment decisions, diagnostic codes, follow-up intensity, medication changes, specialist consultations, and other healthcare-process variables may contain substantial information about disease status. Their predictive value can therefore reflect prior clinical recognition rather than independent detection of latent disease.

A model can be statistically accurate while answering the wrong temporal question.

Consequently, leakage auditing should involve more than checking whether the target column accidentally appears among the predictors. Investigators should identify the temporal position of each candidate variable relative to the prediction point and ask whether that information would genuinely have been available when the model is intended to operate.

Direct target leakage requires particular caution

The study also illustrates why direct outcome-derived variables require strict exclusion.

Depression status in the simulation was defined from a synthetic PHQ-9-like total score using a threshold of 10. Supplying that same total score to a classifier would make the prediction task largely circular. [9]

The model would not be discovering an independent marker of depression; it would be reconstructing a threshold transformation of information already used to define the outcome.

For this reason, the target-defining symptom score was excluded from all predictive analyses, including the deliberately leakage-inclusive experiment.

This distinction is important because leakage exists on a continuum. Direct inclusion of an outcome-defining variable represents an extreme case, whereas post-outcome clinical variables can produce subtler leakage. Both can generate misleadingly optimistic estimates, but the mechanisms differ.

Transparent psychiatric prediction studies should therefore specify not only which predictors were used but also how the outcome was constructed and whether any predictors contributed directly or indirectly to that construction. [2, 3, 14, 24]

Participant-independent validation did not reduce performance

A notable finding was the absence of the predicted decline under participant-independent validation.

The original hypothesis was that random row-level splitting would produce optimistic performance because repeated observations from the same synthetic participant could occur across development and evaluation partitions. Participant-independent splitting was therefore expected to reduce performance.

This did not occur.

Logistic-regression AUROC increased numerically from 0.9548 under random row splitting to 0.9585 under participant-independent evaluation. Similar patterns occurred across the other classifiers. [5, 17]

This result should not be interpreted as evidence that participant-independent validation is unnecessary. Instead, it indicates that the repeated-measures structure generated in this particular simulation did not create sufficiently strong participant-specific information for row-level overlap to materially inflate discrimination.

This is an important negative result.

A methodological safeguard should not be judged according to whether it produces a lower performance estimate. Its purpose is to ensure that the evaluation corresponds to the intended prediction problem.

If the scientific question concerns prediction in new people, participant-independent evaluation remains conceptually preferable even when its performance estimate happens to resemble that obtained from row-level splitting.

Indeed, stability under participant-independent validation can itself be interpreted as evidence of robustness within the simulated environment.

Why participant overlap may not have produced optimism

Several characteristics of the data-generating process may explain the absence of a participant-overlap effect.

Although participants contributed repeated observations, visit-specific stochastic variation was introduced into symptom burden, sleep, activity, physiology, cognition, stress, and other variables. Repeated observations were therefore correlated but not duplicates.

Furthermore, participant identifiers were never supplied to the classifiers.

The primary predictive structure was generated through relationships between latent distress and multiple observable variables that were shared across the population. Consequently, the models could learn broadly applicable relationships rather than depending primarily on participant-specific signatures.

The number of unique participants was also relatively large at 8,000. This provided substantial independent information even when multiple observations were present.

These design characteristics likely limited the degree to which row-level splitting benefited from participant overlap.

The finding nevertheless identifies a useful direction for future simulation experiments. The degree of within-person correlation could be systematically varied to determine the conditions under which row-

level splitting begins to generate meaningful optimism.

For example, future benchmarks could manipulate the intraclass correlation of digital biomarkers, number of repeated measurements per participant, stability of device signatures, or inclusion of high-dimensional participant-specific features.

Such experiments would transform the current negative result into a more general investigation of when participant overlap becomes problematic.

Site-independent validation also remained robust

The site-generalisation hypothesis was similarly unsupported.

Site-independent AUROCs ranged from 0.9502 to 0.9555, closely resembling performance under random and participant-independent validation. [5, 17]

XGBoost achieved a site-independent AUROC of 0.9555, while logistic regression achieved 0.9554. The neural network achieved 0.9548. [5, 6, 17]

Thus, withholding site information did not produce the anticipated reduction in discrimination.

Again, this should not be interpreted as evidence that external or site-independent validation is unnecessary.

The synthetic sites were generated within a common overarching data-generating process. Although site-specific risk effects and demographic variation were introduced, the fundamental relationships connecting latent psychological distress to observable predictors were largely shared across sites.

The models therefore encountered distributional variation but not a radically different clinical population or measurement process.

This represents a weaker challenge than genuine external validation.

A model developed in one real healthcare system may encounter changes in referral pathways, diagnostic practices, language, socioeconomic structure, disease prevalence, measurement hardware, questionnaire administration, missingness, treatment availability, and patient selection when transported elsewhere.

The present site-independent analysis cannot reproduce the full complexity of such shifts.

It should therefore be interpreted as a test of robustness to simulated site heterogeneity, not evidence of external clinical generalisability.

Robust validation does not require performance collapse

The participant- and site-independent findings highlight a broader interpretive issue.

Methodological studies can unintentionally be designed around an expectation that more rigorous validation must reveal poor performance. This creates its own form of confirmation bias.

A leakage-controlled or externally evaluated model does not need to fail for the validation procedure to be informative.

If performance remains stable after a plausible source of bias is removed, that result strengthens evidence that the relevant signal is robust to that particular challenge.

The present benchmark illustrates both possibilities.

Removing post-outcome leakage materially affected performance.

Preventing participant overlap did not.

Withholding site information also did not materially reduce discrimination.

These findings should not be forced into a single narrative that all methodological rigor decreases model performance. Instead, they demonstrate that different methodological vulnerabilities can have very different consequences.

This may be one of the more useful findings of the study.

Algorithmic complexity provided little incremental value

The second major methodological finding concerned model complexity.

Despite substantial differences in modelling strategy, logistic regression, random forest, SVM, XGBoost, and the neural network produced remarkably similar discrimination. [6, 7, 19]

Under random row-level validation, AUROC ranged from 0.9488 to 0.9548, a total spread of only 0.0060. [5, 17]

Under participant-independent validation, the spread was 0.0069.

Under site-independent validation, it narrowed to 0.0053.

Logistic regression was not merely competitive. It produced the highest AUROC under random and participant-independent validation. Under site-independent validation, its AUROC of 0.9554 differed from the XGBoost value of 0.9555 by only 0.0001. [5, 6, 17]

Such differences provide little basis for claiming a meaningful advantage from increased algorithmic complexity.

This result is consistent with the underlying synthetic structure. Although the dataset contained multiple correlated predictors and stochastic variation, much of the predictive information was generated through comparatively systematic relationships between distress and observable features. A linear probabilistic model could therefore recover much of the available signal.

The ability of a neural network or boosted-tree ensemble to approximate more complex functions offered little additional benefit. [6]

The strongest algorithm is not necessarily the most complex

The findings reinforce the importance of strong baseline models in psychiatric machine-learning research.

A study that evaluates only random forest, XGBoost, SVM, and neural networks can identify which machine-learning algorithm performs best without answering a more fundamental question: whether machine learning improves meaningfully upon a conventional statistical baseline. [6, 7, 19]

Logistic regression provides a particularly useful comparator for binary psychiatric outcomes because it is computationally efficient, comparatively interpretable, well understood, and capable of producing probabilistic predictions. [14, 24]

If a complex algorithm substantially improves discrimination, calibration, or clinically relevant decision performance under rigorous validation, its additional complexity may be justified. [1, 17, 20]

If performance differs by only several thousandths of an AUROC unit, the scientific importance of that complexity becomes less obvious. [5, 17]

The present findings therefore support reporting incremental performance relative to simple baselines rather than treating the numerically highest score as automatically preferable. [2, 3]

Random forest and probabilistic prediction

Brier-score results added nuance to the discrimination findings. [1, 17, 20]

Random forest produced relatively high AUROCs but consistently higher Brier scores than logistic regression. Under random validation, its Brier score was 0.0863 compared with 0.0788 for logistic regression. Under participant-independent validation, corresponding values were 0.0825 and 0.0762. Under site-independent validation, they were 0.0878 and 0.0794. [1, 5, 6, 7, 17]

Thus, models with relatively similar ranking performance did not necessarily produce equally accurate probabilities.

This distinction is relevant in clinical prediction because risk estimates may ultimately be used for threshold-based decisions. A model that correctly ranks patients but systematically overestimates or

underestimates probabilities can create different consequences from a well-calibrated model with similar AUROC. [5, 13, 16, 17, 21]

Nevertheless, the Brier score should not be interpreted as a complete calibration analysis. It combines aspects of calibration and discrimination. [1, 17, 20]

Formal calibration slopes, calibration intercepts, reliability diagrams, and potentially decision-curve analyses would provide a more comprehensive assessment in subsequent empirical work. [1, 17, 20]

AUROC alone would have obscured part of the story

The use of multiple performance metrics proved useful.

Had evaluation been restricted to AUROC, all five models might have been described simply as high-performing classifiers with relatively minor differences. [5, 17]

AUPRC demonstrated additional differences in positive-class prediction, while Brier scores revealed differences in probability error. [1, 4, 15, 17, 20]

The leakage experiment was also more compelling when viewed across all three metrics. Leakage increased logistic-regression AUROC from 0.9548 to 0.9891, increased AUPRC from 0.9253 to 0.9827, and reduced the Brier score from 0.0788 to 0.0367. [1, 4, 5, 15, 17]

Convergence across multiple metrics strengthens the conclusion that post-outcome information materially altered apparent predictive performance.

Future psychiatric machine-learning studies should therefore avoid reducing model quality to a single discrimination statistic.

Implications for the reproducibility of psychiatric machine learning

The present results contribute to the broader discussion of reproducibility by illustrating how analytical choices can change the interpretation of model performance even when the underlying data remain unchanged. [10]

Reproducibility is sometimes treated principally as the ability to rerun code and recover the same numerical result. Computational reproducibility is important, but trustworthy psychiatric prediction requires an additional question: does the reproduced result estimate the clinically relevant quantity? [14, 24]

A perfectly reproducible AUROC can still be misleading if the underlying pipeline contains post-outcome information or evaluates observations that are not independent in a manner inconsistent with intended deployment. [5, 17]

Conversely, a model that retains performance after leakage auditing and appropriately structured validation provides stronger evidence even if its headline AUROC is numerically lower. [5, 17]

The distinction suggests that reproducibility should encompass transparency about predictor timing, outcome construction, splitting units, preprocessing boundaries, hyperparameter optimisation, missing-data procedures, and site structure.

Implications for claims of psychiatric biomarkers

The term biomarker can imply a degree of biological or clinical validity that exceeds what a prediction model has actually demonstrated. [13, 16, 21]

Several variables in modern psychiatric modelling—including digital activity, sleep, physiological measurements, cognitive measures, and healthcare-use patterns—may be statistically informative without necessarily constituting validated biomarkers.

The present simulation deliberately combines multiple classes of such predictors.

Its findings demonstrate that strong classification can arise from multivariate structure, but they do not establish that any particular synthetic variable is a genuine biomarker of depression.

This distinction is central to interpreting the study title.

The question “How robust are machine-learning biomarkers of depression?” should be understood as an examination of the methodological robustness of biomarker-style prediction pipelines, rather than a clinical validation of specific biomarkers.

Empirical biomarker claims require independent patient data, clinically appropriate outcome definitions, temporal validation, and preferably external replication.

Toward trustworthy computational psychiatry

The results support a broader shift from performance-centred machine learning toward trustworthy computational psychiatry.

Under a performance-centred paradigm, the principal objective is often to maximise a metric such as AUROC. [5, 17]

Under a robustness-centred paradigm, a high AUROC is the beginning rather than the end of evaluation. [5, 17]

Researchers should ask whether the result survives removal of leakage-prone variables; whether preprocessing is isolated within training data; whether observations are independent at the appropriate level; whether performance transfers across sites; whether probabilities are calibrated; whether

performance differs across demographic groups; and whether a complex algorithm meaningfully improves upon a simple baseline.

The present findings demonstrate why these questions should be separated.

Leakage control mattered substantially.

Participant separation did not materially affect the results.

Simulated site separation did not materially affect the results.

Algorithmic complexity contributed almost nothing to discrimination.

Each result provides different information about robustness.

Implications for study design

Several practical recommendations follow from the benchmark.

First, investigators should construct a temporal map of candidate predictors before modelling. Each variable should be classified according to whether it occurs before, at, or after the intended prediction point.

Second, target construction should be explicitly documented. Variables used to define the target should ordinarily be excluded from predictors unless the scientific question specifically concerns reconstruction of that target.

Third, splitting should occur at the level corresponding to intended generalisation. Longitudinal data intended to predict new individuals should normally be partitioned by participant rather than by row.

Fourth, multisite datasets should be used to examine transportability rather than simply increasing sample size.

Fifth, preprocessing must be included within the training pipeline.

Sixth, logistic regression or another defensible simple model should be included as a baseline.

Finally, performance should be characterised using complementary measures rather than a single headline statistic.

Implications for reporting

Transparent reporting is particularly important because leakage may not be apparent from the final model specification. [2, 3]

A manuscript may list apparently reasonable predictors while omitting when they were measured relative to the outcome. Similarly, stating that cross-validation was performed does not reveal whether feature selection or imputation occurred before cross-validation. [8, 18, 22]

For reproducibility, psychiatric machine-learning reports should therefore provide enough information to reconstruct the complete analytical sequence.

At minimum, this includes the outcome definition, prediction time point, predictor availability, unit of splitting, preprocessing order, missing-data handling, feature selection, tuning procedure, model specification, random seeds, and evaluation metrics.

Where possible, analytical code and synthetic or appropriately shareable data can further facilitate verification.

Strengths of the present study

The controlled synthetic design represents both a limitation and a methodological strength.

Its primary advantage is that the data-generating structure is known.

Post-outcome leakage could be deliberately introduced rather than inferred retrospectively. Participant identity and repeated observations were explicitly controlled. Site heterogeneity was generated intentionally. The same underlying dataset could then be analysed under multiple validation conditions.

This allowed individual methodological mechanisms to be isolated more clearly than may be possible in observational clinical data.

A second strength is the direct comparison of five modelling approaches spanning linear statistical modelling, bagged trees, boosted trees, margin-based classification, and neural-network modelling.

A third strength is the inclusion of AUROC, AUPRC, and Brier score rather than reliance on classification accuracy alone. [1, 4, 5, 15, 17]

A fourth strength is the explicit reporting of unsupported hypotheses. Participant-independent and site-independent validation were predicted to reduce performance, but they did not. Preserving these negative results reduces the risk of retrospective hypothesis revision and provides a more informative account of the benchmark. [2, 3]

Limitations

Several limitations substantially constrain interpretation.

Most importantly, the dataset is entirely synthetic.

No real patients were included, and the observed performance estimates cannot be interpreted as estimates of diagnostic accuracy in clinical populations. The simulation cannot establish that real depression classifiers achieve AUROCs near 0.95, nor can it establish that leakage increases published AUROCs by approximately 0.03. [5, 17]

The numerical magnitude of the leakage effect depends on the deliberately specified data-generating mechanism.

Second, depression status was operationalised from a synthetic PHQ-9-like total threshold. Although this provides a convenient binary outcome for methodological benchmarking, real depressive disorders are more complex than a thresholded symptom score. [9]

Third, the simulated sites shared an overarching data-generating mechanism. Site-independent evaluation therefore represents a relatively constrained distributional shift compared with genuine cross-institutional, cross-cultural, or international validation.

Fourth, the repeated-measures structure may not have generated sufficiently strong participant-specific signatures to expose optimism from row-level splitting. Real digital phenotyping datasets containing dense longitudinal measurements or device-specific patterns may behave differently.

Fifth, a complete nested cross-validation benchmark was not available in the final primary analysis. Consequently, the study cannot fully quantify optimism associated with hyperparameter selection. [8, 18, 22]

Sixth, demographic-balanced validation was prespecified but not completed in the primary benchmark. No conclusions concerning demographic fairness or subgroup robustness should therefore be drawn.

Seventh, the Brier score alone does not constitute a complete calibration assessment. Calibration intercepts, slopes, and graphical calibration analyses would strengthen subsequent work. [1, 17, 20]

Finally, no decision-curve analysis was conducted. The study therefore evaluates statistical prediction rather than clinical net benefit.

Future research

The most important next step is replication using real psychiatric data.

The analytical framework developed here could be applied to one or more publicly available depression datasets while preserving the same core comparisons: conventional versus participant-independent validation, leakage-inclusive versus leakage-controlled feature sets where appropriate, site-independent evaluation where multisite data are available, and comparison of complex algorithms with logistic regression.

A second direction would systematically manipulate the strength of individual methodological problems within simulation.

For participant dependence, within-person correlation and number of repeated observations could be varied.

For site heterogeneity, differences in prevalence, measurement calibration, demographic composition, and predictor-outcome relationships could be progressively increased. [1, 17, 20]

For leakage, the correlation between post-outcome variables and diagnosis could be manipulated from weak to near-deterministic.

Such experiments could identify thresholds at which particular methodological weaknesses begin to meaningfully distort performance.

A third priority is demographic robustness. Performance should be examined across age, sex, socioeconomic, and other relevant groups, with careful consideration of sample size and the limitations of synthetic demographic categories.

A fourth priority is calibration. Future analyses should include calibration intercept, calibration slope, reliability curves, recalibration experiments, and decision-curve analysis. [1, 17, 20]

Finally, independent external replication remains essential. A trustworthy model should ultimately demonstrate not merely that it survives resampling of one dataset, but that its predictive relationships persist in genuinely independent populations.

Interpretation of the unexpected findings

One of the most informative aspects of the study is that its findings were less dramatic than originally hypothesised.

A simple narrative might have predicted that increasingly rigorous validation would progressively reduce performance from an impressive internal AUROC to a substantially weaker independent result. [5, 17]

That did not occur.

Instead, the experiment produced a more differentiated picture.

Removing leakage mattered substantially. Changing the unit of validation from rows to participants did not. Simulated site-independent evaluation did not. Increasing algorithmic complexity contributed little.

This outcome illustrates why methodological benchmarks should be treated as experiments rather than demonstrations designed to produce a predetermined conclusion.

The scientific value lies not in showing that every machine-learning model fails under scrutiny, but in identifying which apparent advantages survive scrutiny and which do not.

Conclusion

Within this controlled synthetic psychiatric benchmark, deliberate post-outcome leakage substantially inflated apparent depression-classification performance, moving an already high-performing logistic-regression model from an AUROC of 0.9548 to 0.9891. In contrast, performance remained stable when evaluation was restricted to independent participants and simulated independent sites. [5, 17]

Complex machine-learning algorithms did not meaningfully outperform logistic regression. Across the completed validation conditions, differences among models were small, and the simplest model remained highly competitive in both discrimination and probabilistic prediction. [6]

These findings suggest that methodological integrity may be more consequential than algorithmic sophistication when evaluating psychiatric prediction models. High performance should therefore be interpreted in relation to predictor timing, target construction, leakage control, independence of validation data, site transportability, calibration, and performance relative to simple baselines. [1, 10, 13, 14, 16]

At the same time, the findings should not be extrapolated directly to clinical depression prediction. Because the benchmark is synthetic, it demonstrates methodological mechanisms rather than clinical accuracy or the magnitude of bias in published research.

The central implication is therefore not that psychiatric machine learning is intrinsically unreliable, nor that sophisticated algorithms are intrinsically unnecessary. Rather, predictive claims become more informative when researchers demonstrate what remains after leakage is removed, independence is enforced, transportability is challenged, and complexity is compared against an appropriate baseline. [14, 24]

For trustworthy computational psychiatry, the relevant question is not simply how well a model performs, but under what conditions that performance survives.

Conclusions

This study used a controlled synthetic multisite psychiatric benchmark to examine how data leakage, validation strategy, and algorithmic complexity influence apparent machine-learning performance in depression classification. The findings indicate that the integrity of the analytical pipeline can be at least as important as the choice of predictive algorithm.

The strongest methodological effect arose from deliberately introduced post-outcome leakage. When leakage-prone clinical variables were included, logistic regression achieved an AUROC of 0.9891 and AUPRC of 0.9827, compared with an AUROC of 0.9548 and AUPRC of 0.9253 after those variables were

removed. The corresponding Brier score changed from 0.0367 to 0.0788. Thus, access to information generated downstream of the outcome made an already strongly predictive model appear substantially closer to perfect classification. [1, 4, 5, 13, 15]

In contrast, increasingly independent validation did not produce the hypothesised collapse in predictive performance. Leakage-controlled models retained AUROCs of approximately 0.95–0.96 under random row-level, participant-independent, and site-independent validation. These findings demonstrate that more rigorous validation should not be expected automatically to reduce model performance. Its purpose is instead to determine whether predictive performance survives conditions that better correspond to the intended generalisation task. [5, 17]

Algorithmic complexity also provided little incremental benefit. Random forest, SVM, XGBoost, and neural-network classifiers did not consistently outperform logistic regression. Under participant-independent validation, logistic regression achieved an AUROC of 0.9585, AUPRC of 0.9383, and Brier score of 0.0762. Under site-independent validation, the difference between XGBoost and logistic regression was only 0.0001 AUROC. In this benchmark, therefore, increasing computational sophistication did not translate into a meaningful improvement in discrimination. [1, 4, 5, 6, 7]

These findings reinforce the importance of evaluating psychiatric machine-learning models according to more than their highest reported performance statistic. Predictor timing, outcome construction, leakage auditing, participant independence, site transportability, calibration, and comparison against simple statistical baselines should be considered integral components of model evaluation. [1, 17, 20]

The findings must nevertheless be interpreted within the limits of the study design. All observations were synthetically generated, and the results do not establish the diagnostic performance of machine-learning models in people with depression, the validity of any particular psychiatric biomarker, or the magnitude of leakage-related inflation in published research. The synthetic design instead provides a controlled demonstration of methodological mechanisms that should subsequently be tested using independent empirical psychiatric datasets. [10]

Future work should therefore apply the same leakage-controlled benchmarking framework to real multisite psychiatric data, incorporate complete nested model selection and demographic subgroup evaluation, and assess calibration and clinical utility prospectively. Replication across independent populations will be necessary to determine which predictive relationships remain stable under real-world distributional change. [1, 8, 17, 18, 20]

Overall, the present benchmark suggests that trustworthy computational psychiatry should prioritise robust validation over headline performance and methodological integrity over unnecessary algorithmic complexity. The central question is not simply whether a psychiatric machine-learning model can achieve a high AUROC, but whether its performance persists when information leakage is removed and the model is required to generalise beyond the conditions under which it was developed. [5, 17]

Declarations

Ethics approval and consent to participate

Not applicable. This study used exclusively synthetic, computationally generated data. No human participants were recruited, no identifiable or private human data were accessed, and no clinical or medical records were used. The synthetic participant identifiers and all demographic, clinical, behavioural, physiological, and cognitive variables were artificially generated for methodological benchmarking. Therefore, institutional ethics approval and informed consent were not required.

Consent for publication

Not applicable. The manuscript contains no identifiable information, images, case reports, or other personal data relating to human participants.

Funding

The author received no specific funding, grant, or financial support from any public, commercial, or not-for-profit funding agency for this research.

Competing interests

The author declares no competing interests.

Availability of data and materials

The dataset analysed in this study is entirely synthetic and was computationally generated for the purposes of the methodological benchmark. No real patient or participant data were used. The synthetic dataset may be made available with the published article or through an appropriate research repository to support reproducibility.

Code availability

The computational code used for synthetic data generation, preprocessing, model development, validation, and performance evaluation may be made available through an appropriate research repository to facilitate reproducibility.

Authors' contributions

A.F. was the sole author of the study and was responsible for the study conception and design, development of the methodological framework, specification of the synthetic data-generating process, computational analysis, interpretation of the results, preparation of tables, and drafting and critical revision of the manuscript. A.F. read and approved the final manuscript.

Acknowledgements

Not applicable.

Author information

Arhama Faridi Queen Mary University of London London, United Kingdom

Corresponding author

Correspondence to: Arhama Faridi

References

1. Brier GW (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78(1):1–3. 10.1175/1520-0493(1950)078%3C0001%3E2.0.CO;2
2. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385. 10.1136/bmj-2023-078378
3. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 162(1):55–63. 10.7326/M14-0697
4. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York: Association for Computing Machinery; pp. 233–240. 10.1145/1143844.1143874
5. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837–845. 10.2307/2531595
6. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, New York. 10.1007/978-0-387-84858-7
7. Ho TK (1995) Random decision forests. In: *Proceedings of the Third International Conference on Document Analysis and Recognition*. Montreal: IEEE; pp. 278–282. 10.1109/ICDAR.1995.598994

8. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal: Morgan Kaufmann; pp. 1137–1143
9. Kroenke K, Spitzer RL, Williams JBW (2001) The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 16(9):606–613. 10.1046/j.1525-1497.2001.016009606.x
10. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. *Advances in Neural Information Processing Systems* 30. :4765–4774
11. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P et al (2025) PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 388. 10.1136/bmj-2024-082505
12. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS et al (2019) PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of Internal Medicine*. ;170(1)–W33. 10.7326/M18-1377
13. Obermeyer Z, Emanuel EJ (2016) Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 375(13):1216–1219. 10.1056/NEJMp1606181
14. Poldrack RA, Huckins G, Varoquaux G (2020) Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* 77(5):534–540. 10.1001/jamapsychiatry.2019.3671
15. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10(3). 10.1371/journal.pone.0118432
16. Steyerberg EW (2019) *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd edn. Springer, Cham. 10.1007/978-3-030-16399-0
17. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N et al (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21(1):128–138. 10.1097/EDE.0b013e3181c30fb2
18. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91. 10.1186/1471-2105-7-91
19. Vapnik VN (2000) *The Nature of Statistical Learning Theory*, 2nd edn. Springer, New York. 10.1007/978-1-4757-3264-1
20. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: the Achilles heel of predictive analytics. *BMC Med* 17:230. 10.1186/s12916-019-1466-7
21. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KGM (2021) Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 132:142–145. 10.1016/j.jclinepi.2021.01.009
22. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14(11). 10.1371/journal.pone.0224365

23. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS et al (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 170(1):51–58. 10.7326/M18-1376
24. Yarkoni T, Westfall J (2017) Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci* 12(6):1100–1122. 10.1177/1745691617693393