

# Does Machine Learning Improve Depression Prediction Beyond Conventional Clinical Models? A Leakage-Controlled Multicohort Simulation Benchmark

Arhama Faridi

[arhaama.far@gmail.com](mailto:arhaama.far@gmail.com)

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

---

## Research Article

**Keywords:** Depression prediction, machine learning, logistic regression, data leakage, calibration, synthetic data, multicohort validation

**Posted Date:** October 5th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11236610/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.

---

# **Does Machine Learning Improve Depression Prediction Beyond Conventional Clinical Models? A Leakage-Controlled Multicohort Simulation Benchmark**

Arhama Faridi

Queen Mary University of London

Corresponding author: Arhama Faridi

Email: [arhamafaridi@gmail.com](mailto:arhamafaridi@gmail.com); [bt22044@qmul.ac.uk](mailto:bt22044@qmul.ac.uk)

## Abstract

**Background:** Machine-learning algorithms are increasingly investigated for depression prediction, but their advantages over conventional statistical models may be overstated by data leakage, unequal model comparisons, and inadequate validation. This study aims to determine whether machine learning improves predictive performance beyond logistic regression when preprocessing, model selection, and evaluation follow strict leakage-control procedures.

**Methods:** The benchmark will use 18,000 synthetic participants distributed across three independent scenarios, each containing six simulated cohorts of 1,000 participants. Scenarios represent predominantly linear relationships, nonlinear relationships, and variation in predictor–outcome relationships across cohorts. Baseline predictors include demographic characteristics, depression history, symptom burden, anxiety, stress, social support, sleep, and health behaviours. The outcome is simulated depression status at six months. Within each scenario, four cohorts will support model development and two will remain reserved for evaluation. Conventional, penalized, and flexible logistic regression will be compared with random forest, gradient boosting, and support vector machines using equivalent predictor availability. Imputation, encoding, scaling, feature selection, hyperparameter tuning, and calibration will occur exclusively within training folds. Evaluation will assess discrimination, calibration, Brier score, threshold-dependent performance, and decision-curve net benefit.

**Results:** Dataset generation is complete. Model fitting and comparative evaluation have not yet been conducted; therefore, no estimates of predictive performance or machine-learning superiority are reported.

**Conclusions:** This simulation benchmark will examine whether machine-learning advantages persist under rigorous leakage control and how these advantages depend on nonlinearity and

cohort heterogeneity. Findings will describe performance under specified simulated conditions and will require validation using real clinical cohorts before supporting clinical application.

**Keywords:** Depression; machine learning; logistic regression; data leakage; calibration; simulation; prediction.

## Introduction

### Depression prediction and the need for reliable individual risk estimates

Depression prediction presents an important challenge for psychiatric research because identifying individuals at increased risk requires more than demonstrating that particular characteristics are associated with depressive symptoms. A clinically useful prediction model must combine information available at a defined assessment point, estimate a clearly specified future outcome, and retain acceptable performance when applied to individuals who were not involved in its development. The PredictD study provided an influential example of this approach by developing and validating a risk algorithm for future episodes of major depression among general practice attendees (King et al., 2008). Its contribution illustrates that individualized depression prediction is an established objective of conventional clinical modelling and does not depend exclusively on the use of advanced machine-learning algorithms.

The present study addresses prediction of depression status six months after baseline assessment. This outcome differs from identifying depression at the same assessment point and from predicting first onset among individuals with no previous depression. Because the simulated population includes individuals with a history of depression and varying baseline symptom burdens, the prediction task may encompass both new and persistent depression. Distinguishing these tasks is essential: a model that classifies current symptom severity cannot automatically be interpreted as a model of future onset, recurrence, or deterioration.

Depression also presents substantial challenges for outcome definition. Individuals with similar aggregate symptom scores may experience different symptom profiles, and reducing depression to a single total score can obscure clinically relevant variation. Fried and Nesse (2015) argued that individual symptoms differ in their associations and implications, making

symptom-level information important for understanding depression heterogeneity. For prediction research, this raises questions about what a model actually predicts and how closely its outcome corresponds to the clinical phenomenon of interest. A threshold on a questionnaire, a recorded diagnosis, and a structured diagnostic assessment should therefore be treated as related but distinct outcomes.

These distinctions are particularly important when constructing methodological benchmarks. The present benchmark uses simulated symptom scores and a stochastic future depression outcome. Its symptom scores reproduce familiar numerical ranges but do not represent administered or validated questionnaires. Similarly, its outcome represents a specified statistical process rather than a verified clinical diagnosis. The study consequently evaluates how modelling strategies behave under controlled conditions, while preserving a clear boundary between methodological findings and evidence about actual patients.

### **The promise of machine learning in psychiatric prediction**

Machine learning has attracted considerable interest in psychiatry because it offers methods for identifying predictive patterns across multiple variables and producing individual-level predictions. Dwyer et al. (2018) described its potential to support psychiatric diagnosis, prognosis, and treatment decisions using clinical and biological information, while also highlighting limitations affecting clinical translation. This individual prediction focus is relevant to depression research, where combinations of symptom burden, history, psychological characteristics, and health behaviours may contain information that is difficult to summarize using isolated associations.

From a modelling perspective, algorithms such as random forest, gradient boosting, and support vector machines offer different ways of representing relationships between predictors

and outcomes. Tree-based methods can accommodate thresholds and interactions through successive partitions of the predictor space. Kernel-based support vector machines can represent complex decision boundaries. These properties make such approaches plausible candidates when predictive relationships contain substantial nonlinearity or interactions that have not been specified in advance.

However, flexibility creates an opportunity rather than a guarantee of improved prediction. A more flexible model must estimate its structure from the available development data. Its success therefore depends on the strength of the underlying signal, sample size, measurement quality, missingness, tuning strategy, and similarity between development and evaluation populations. A model may fit complex relationships accurately within one dataset yet perform less well when those relationships change.

The relevant question is consequently conditional: under what circumstances does additional modelling flexibility improve predictions enough to justify its use? This question requires comparisons that distinguish algorithmic capability from advantages arising through access to more information, more extensive tuning, or a more favourable evaluation procedure.

### **Why conventional clinical models remain essential comparators**

Logistic regression provides a natural comparator for a binary depression outcome because it estimates outcome probabilities from a specified combination of predictors. Nevertheless, logistic regression should not be represented as a single inflexible approach. A model containing only additive linear terms differs substantially from one containing prespecified transformations, nonlinear terms, interactions, or penalization. Comparisons against an inadequately specified regression model may therefore exaggerate the apparent advantages of other algorithms.

Christodoulou et al. (2019) systematically examined comparisons between machine learning and logistic regression in clinical prediction research and found no overall performance benefit of machine learning in the evidence they reviewed. This finding does not establish that every machine-learning algorithm performs similarly to regression in every setting. It does, however, challenge the assumption that algorithmic complexity itself provides a dependable improvement and supports the need for carefully controlled comparisons.

For the present study, a fair comparison requires several regression benchmarks. Conventional logistic regression will provide an additive reference model. Penalized logistic regression will evaluate whether regularization improves stability. Flexible logistic regression with prespecified nonlinear terms will provide a stronger comparator for settings in which simple linear assumptions are inadequate. These comparisons will help distinguish gains attributable to capturing nonlinear relationships from gains attributable to a particular algorithm.

Predictor availability must also be equivalent in the primary comparison. If a machine-learning model receives symptom scores, depression history, sleep measures, and social information while its regression comparator receives only age and sex, any observed difference combines the contribution of additional information with the contribution of the modelling method. Using the same baseline predictor set provides a clearer assessment of the incremental value of the algorithm.

Likewise, model development should provide reasonable opportunities for each method to perform well. Regression models require appropriate specification and, where applicable, penalty selection. Machine-learning models require tuning of their relevant hyperparameters. The benchmark will therefore evaluate complete development procedures rather than comparing a extensively optimized algorithm with an untreated default comparator.

**Data leakage as a threat to valid performance estimation**

Data leakage occurs when a modelling procedure gains access to information that would not legitimately be available for the intended prediction task. It can arise through inappropriate preprocessing, sampling, feature construction, or separation of development and evaluation data. Kapoor and Narayanan (2023) documented leakage across multiple scientific fields and showed how it can support overly optimistic conclusions about predictive performance. Their work emphasizes that reproducible code alone does not establish methodological validity if the underlying analysis permits information to cross an inappropriate boundary.

For this benchmark, those boundaries are defined by prediction time and evaluation membership. Predictors must be available at baseline, before the six-month outcome occurs. Evaluation records must remain outside all procedures used to learn transformations, select predictors, tune models, or choose decision thresholds. These requirements apply to regression and machine-learning methods alike.

Some forms of leakage are direct. Including a copy of the outcome provides the model with the answer it is supposed to predict. Incorporating treatment initiated after the outcome can introduce information from an inappropriate time point. Other forms are less obvious. Selecting variables using associations calculated across the full dataset exposes the development procedure to evaluation outcomes. Learning imputation values or scaling parameters from all observations also allows evaluation data to influence the fitted pipeline.

The separate leakage-audit file supplied with the dataset contains deliberately invalid variables for an optional positive-control analysis. These variables will remain excluded from all valid model comparisons. Their role is to demonstrate the consequences of an intentionally compromised prediction task, rather than to improve the benchmark's reported performance.

Model selection introduces another source of optimism. When the same cross-validation results are used both to select a model and to estimate its final performance, the reported estimate can reflect selection of a favourable result. Varma and Simon (2006) demonstrated this problem and showed the value of nested cross-validation, which separates tuning from performance evaluation. In the present study, preprocessing and tuning will be repeated within the appropriate training partitions so that each evaluation fold represents data not used to construct the evaluated pipeline.

### **Cohort separation and performance under changing conditions**

Preventing leakage is necessary, but it does not guarantee that a model will perform well in a different population. A model may be developed without methodological contamination and still fail when outcome prevalence, predictor distributions, or predictor–outcome relationships change. The distinction between valid evaluation and successful transportability is therefore central to the present benchmark.

Randomly dividing a pooled dataset can produce training and test samples with very similar characteristics. Such an evaluation answers a useful question about performance in comparable observations, but it provides limited information about applying the model to a distinct cohort. Reserving entire cohorts for evaluation creates a more demanding test by preserving differences between development and evaluation conditions.

The benchmark separates these issues through three scenarios. The linear scenario primarily represents additive relationships on the log-odds scale. The nonlinear scenario adds curvature, thresholds, and interactions. The heterogeneous scenario additionally changes selected predictor–outcome relationships across cohorts. These scenarios allow the study to examine

whether the relative performance of modelling approaches depends on the structure and stability of the predictive signal.

Each scenario includes four development cohorts and two held-out cohorts. The held-out cohorts will remain outside model selection, tuning, calibration, and threshold determination. They provide evaluation under distinct simulated conditions. Because they are generated rather than independently recruited, they will be described as held-out simulation cohorts, rather than as evidence of clinical external validation.

This design also makes the simulation assumptions part of the scientific interpretation. Good performance in the linear scenario may indicate that a method effectively estimates the specified additive structure. Good performance in the nonlinear scenario may indicate that it captures the introduced nonlinear signal. Neither result establishes that the corresponding structure accurately represents depression in a particular clinical population.

### **Evaluating prediction beyond discrimination**

A useful comparison requires multiple dimensions of predictive performance. Discrimination assesses whether individuals who experience the outcome tend to receive higher predicted risks than individuals who do not. The area under the receiver operating characteristic curve provides a threshold-independent measure of this ranking ability. However, a ranking measure does not establish that predicted probabilities are accurate.

Calibration assesses the agreement between predicted risks and observed outcome frequencies. A model can rank individuals successfully while systematically overestimating or underestimating their risks. Van Calster et al. (2019) emphasized the importance of calibration for prediction models intended to inform decisions and recommended evaluating it during model

validation. For the present benchmark, calibration intercepts, slopes, and plots will complement discrimination measures, while the Brier score will assess squared probability error.

Precision-recall measures provide additional information about identifying outcome-positive individuals. Saito and Rehmsmeier (2015) showed why precision-recall evaluation is particularly informative when class frequencies are uneven. Because precision depends on outcome prevalence, area under the precision-recall curve will be interpreted alongside the prevalence of each evaluation cohort. Differences across cohorts will not automatically be attributed to changes in algorithmic quality.

Threshold-dependent measures address another aspect of prediction. Sensitivity and specificity summarize the consequences of converting probabilities into binary classifications. Positive and negative predictive values describe the reliability of those classifications within a particular population. These measures will be evaluated using thresholds selected in development data, preserving the independence of the held-out evaluation.

Decision-curve analysis can further assess whether predictions provide value under specified assumptions about the relative consequences of false-positive and false-negative classifications. Vickers and Elkin (2006) introduced net benefit as a way to compare prediction strategies across threshold probabilities. In this study, decision curves will be interpreted as evaluations of simulated decision utility. They cannot demonstrate improved clinical outcomes, reduced depression burden, or benefits of an actual intervention.

The term “meaningful improvement” therefore requires careful definition. A statistically detectable difference may be too small to justify additional complexity, whereas a modest discrimination gain may be relevant if it accompanies better calibration and favourable decision

performance. Any superiority or equivalence criteria should be specified before examining comparative model results. Failure to detect superiority will not, by itself, establish equivalence.

### **Rationale for a controlled simulation benchmark**

Simulation offers a way to study modelling behaviour when the data-generating process is explicitly known. Morris et al. (2019) described a structured approach to simulation design that connects research aims, data-generating mechanisms, methods, and performance measures. This approach is valuable here because the benchmark can deliberately vary relationship structure while keeping the prediction task and evaluation principles consistent.

The present dataset contains 18,000 distinct simulated participants across three independent scenarios. Within each scenario, six cohorts contain 1,000 participants each. Baseline predictors represent demographic characteristics, depression history, symptom burden, anxiety, stress, social support, sleep, and health behaviours. Missing predictor values are retained so that imputation can be evaluated as part of the training pipeline.

The scenarios will be analysed separately. Pooling them into a single modelling exercise would obscure their purpose, since they represent different statistical experiments. Comparisons within each scenario will assess relative algorithm performance, while comparisons across scenarios will assess how that relative performance changes with the underlying structure.

The current files represent one reproducible realization of the specified processes. They support an initial benchmark, but a fuller simulation study would require repeated generation using independent seeds to quantify Monte Carlo variability. Uncertainty arising from repeated data generation differs from uncertainty estimated by resampling a single generated dataset. Conclusions will therefore be proportionate to the analyses actually completed.

Transparent reporting is also essential. The TRIPOD+AI statement provides harmonized recommendations for reporting prediction-model studies using regression or machine learning (Collins et al., 2024). Its emphasis on clearly describing the prediction task, data, development procedures, and evaluation supports the reporting principles adopted in this benchmark.

### **Research gap and contribution of the present study**

The methodological motivation for this study is the need to evaluate the incremental value of machine learning under comparable information, development, and evaluation conditions. Rather than assuming that one model family is superior, the benchmark will examine whether any advantage persists when the comparison includes stronger regression alternatives, complete training-only preprocessing, nested tuning, and untouched cohort evaluation.

Its contribution will be explanatory as well as comparative. A single overall ranking can conceal important differences between stable linear relationships, stable nonlinear relationships, and changing cohort-specific relationships. The proposed design will examine these conditions separately and assess whether improvements in discrimination coincide with improvements in probability accuracy and simulated decision utility.

The study will consequently address a bounded methodological question: when do more flexible algorithms improve prediction in explicitly defined depression-related simulations? It will not establish which algorithm should be deployed clinically, and it will not infer causal effects from simulated predictor associations.

### **Research aims**

The primary aim is to evaluate whether machine-learning algorithms improve prediction of simulated six-month depression status beyond conventional, penalized, and flexible logistic regression under strict leakage control.

The secondary aims are to:

1. Compare model performance across linear, nonlinear, and heterogeneous data-generating scenarios.
2. Assess whether algorithm rankings change when evaluation moves from development-cohort resampling to held-out cohorts.
3. Determine whether improvements in discrimination are accompanied by improvements in calibration and Brier score.
4. Compare sensitivity, specificity, and predictive values at thresholds selected exclusively within development data.
5. Evaluate simulated net benefit across prespecified threshold probabilities.
6. Examine the influence of stronger regression specification on the estimated incremental value of machine learning.
7. Demonstrate, in a separate optional analysis, the effect of deliberately invalid outcome-related predictors.
8. Provide a reproducible framework that makes the generating assumptions and model-development procedures available for scrutiny.

### **Research questions**

RQ1. Do random forest, gradient boosting, or support vector machines improve held-out discrimination compared with conventional logistic regression when all models use the same baseline predictors?

RQ2. Do any observed advantages persist when the comparators include penalized and flexible logistic regression?

RQ3. How does the magnitude of the machine-learning advantage vary between predominantly linear and nonlinear scenarios?

RQ4. How do changing predictor–outcome relationships across cohorts affect discrimination, calibration, and model rankings?

RQ5. Do improvements in AUROC correspond to improvements in Brier score, calibration, and precision-recall performance?

RQ6. At development-selected thresholds, do machine-learning models achieve more favourable sensitivity, specificity, and predictive values?

RQ7. Do machine-learning models provide greater simulated net benefit than regression alternatives across the prespecified threshold range?

RQ8. In the separate leakage positive control, how strongly do invalid outcome-related predictors inflate apparent performance?

### **Research hypotheses**

These hypotheses are prospective expectations for the planned analyses. They are not established findings.

H1. Limited advantage under predominantly linear relationships.\ In the linear scenario, machine-learning models will show smaller discrimination gains over appropriately specified logistic regression than in the nonlinear scenario.

H2. Greater advantage when nonlinear signal is present.\ In the nonlinear scenario, at least one flexible machine-learning approach will achieve higher held-out AUROC than additive conventional logistic regression.

H3. Reduced advantage against stronger regression comparators.\ The estimated improvement associated with machine learning will be smaller when it is compared with flexible or penalized logistic regression than when it is compared only with additive conventional logistic regression.

H4. Deterioration under changing cohort relationships.\ The heterogeneous scenario will produce greater development-to-held-out deterioration in calibration and probability accuracy than the stable nonlinear scenario.

H5. Incomplete agreement between discrimination and probability accuracy.\ The model with the highest held-out AUROC will not necessarily have the lowest Brier score or the most favourable calibration.

H6. Threshold-dependent relative performance.\ Relative performance in sensitivity, specificity, and predictive values will vary across the prespecified decision thresholds, rather than producing a single uniformly superior classifier.

H7. Nonuniform simulated decision benefit.\ Any net-benefit advantage of machine learning will depend on the data-generating scenario and threshold probability, rather than remaining consistent across all evaluation conditions.

H8. Inflation from deliberately invalid predictors.\ In the separate positive-control analysis, inclusion of the outcome copy or post-outcome treatment variable will increase apparent predictive performance compared with the valid baseline-only pipeline.

Together, these questions and hypotheses will test whether added modelling flexibility produces reliable predictive gains, identify the conditions under which those gains emerge, and distinguish statistical improvements from evidence needed to support future clinical use.

## **Literature Review**

### **1. Individual prediction as a distinct objective in depression research**

Depression prediction research seeks to estimate an individual's probability of a defined outcome using information available before that outcome occurs. This objective differs from identifying statistically significant associations between risk factors and depression. An association may contribute to understanding a disorder without providing sufficient information to predict its occurrence accurately in previously unseen individuals. Conversely, a variable can contribute to prediction without representing a causal mechanism. This distinction is important

when evaluating whether machine learning provides an advantage over conventional clinical models.

Dwyer et al. (2018) described machine learning as a potentially valuable approach to individual psychiatric prediction, including diagnosis, prognosis, and treatment outcomes. Their review emphasized the opportunity to combine multidimensional information while recognizing barriers to clinical translation. Poldrack et al. (2020) subsequently examined standards for evidence supporting predictive claims in psychiatry, reinforcing the importance of evaluating performance in observations that were not used to develop the model. Together, these contributions support evaluating depression models through their performance on appropriately separated data rather than through the significance or apparent plausibility of their predictors alone.

The distinction between prediction and explanation also affects how algorithms should be compared. Bzdok et al. (2018) discussed the relationship between statistical modelling and machine learning, showing why the two should not be treated as entirely separate intellectual traditions. Logistic regression can be used within a predictive learning pipeline, while machine-learning methods rely on statistical assumptions and estimation procedures. The relevant comparison is therefore between particular modelling strategies, specifications, and validation procedures, rather than between statistics and artificial intelligence as mutually exclusive categories.

For the present study, this literature supports a clearly bounded objective: determining whether additional modelling flexibility improves prediction of simulated six-month depression status when predictor availability and evaluation procedures are comparable. The benchmark will

assess predictive behaviour under specified conditions. It will not establish causal relationships between the simulated predictors and depression.

## **2. Conventional depression prediction models provide substantive benchmarks**

Conventional risk models demonstrate that individualized depression prediction was an established research objective before the recent expansion of machine learning. King et al. (2008) developed the PredictD algorithm using adult general practice attendees from six European countries and evaluated it in a separate Chilean population. The model was developed among 5,216 European participants without depression at recruitment and tested among 1,732 Chilean participants. Its average C index across European countries was 0.790, compared with 0.710 in Chile. These results demonstrated useful discrimination while also illustrating that performance could decline when a model was applied in a different setting.

PredictD is relevant to the current benchmark in two respects. First, it demonstrates that a regression-based approach can combine demographic, health, historical, and social information into an individual risk estimate. Second, the difference between development and external performance shows why an algorithm should not be judged solely by performance in populations similar to those used for development. However, PredictD predicted depression onset among participants who were initially not depressed, whereas the present simulation predicts future depression status in a population with varying baseline symptoms and depression histories. Its performance should therefore provide contextual evidence rather than a direct numerical target.

Wang et al. (2013) developed sex-specific algorithms for four-year risk of major depressive episodes using a population-based longitudinal cohort. The reported training C statistics were approximately 0.795 for men and 0.767 for women. This study illustrates the feasibility of depression prediction from conventional population information and the possibility

that predictive relationships differ across population groups. The training estimates should nevertheless remain distinguished from independent validation performance when interpreting the study's contribution.

Related work by Wang and colleagues addressed more specific endpoints. One study developed an algorithm for first onset of major depression, while another examined recurrence among individuals with a previous depressive episode (Wang et al., 2014a, 2014b). These studies demonstrate that first onset and recurrence require different population definitions and prediction questions. Their relevance to the present study lies primarily in this endpoint specificity and in the use of conventional prognostic approaches, rather than in any assumption that their estimated risks can be transferred directly to the synthetic benchmark.

Collectively, this evidence supports including well-developed regression models as serious comparators. A benchmark that compares machine learning only with a weak demographic baseline would provide limited information about whether the algorithm improves on an informative clinical model.

### **3. Depression outcomes differ across prediction studies**

The depression prediction literature encompasses onset, persistence, recurrence, remission, treatment response, symptom change, and service utilization. These outcomes are clinically related but statistically distinct. A model predicting remission among treated patients cannot be assumed to predict depression onset among initially unaffected individuals. Similarly, predicting rehospitalization involves healthcare access and treatment processes in addition to illness severity.

Kessler et al. (2016) evaluated machine-learning prediction of the persistence and severity of major depressive disorder from baseline self-reports. Their findings suggested that

baseline information could identify groups with elevated subsequent risks of persistent or severe depression-related outcomes. This work supports the potential prognostic value of routinely obtainable information, but it does not establish that machine learning universally exceeds appropriately specified regression or that the same performance would occur for a six-month depression-status endpoint.

Chekroud et al. (2016) investigated prediction of depression treatment remission using STARD data and evaluation in another trial. This study is influential because it addressed performance across trials rather than relying exclusively on a single development dataset. Its target, however, was treatment outcome among patients receiving antidepressant treatment. The study therefore provides evidence about predictive treatment research and cross-study evaluation, while remaining distinct from forecasting depression status in a general simulated cohort.

Jacobson and Nemesure (2021) examined prediction of changes in depression and anxiety symptoms within a digital intervention trial. Their study extends predictive modelling to continuous symptom change, highlighting another endpoint that cannot be directly equated with binary depression status. Comparing such studies requires attention to outcome scale, intervention exposure, prediction horizon, and evaluation design.

These distinctions support the present study's explicit definition of its outcome as simulated depression status at six months. Because participants may already have substantial symptoms or a depression history at baseline, the endpoint can include persistent as well as newly occurring depression. The benchmark should consequently avoid describing its entire outcome as incident depression.

#### **4. Prospective depression studies show both predictive signal and substantial limits**

Prospective studies provide a more demanding assessment than contemporaneous classification because the model must predict an outcome that occurs after the predictors are measured. They also help distinguish strong baseline associations from accurate individual prognosis.

Dinga et al. (2018) investigated the naturalistic course of depression in 804 patients assessed using 81 demographic, clinical, psychological, and biological measures. Penalized logistic regression achieved an AUROC of approximately 0.66 for depression diagnosis at two-year follow-up. Baseline depressive symptom severity provided predictive information, while additional measured domains did not significantly improve prediction in the principal analysis. The findings illustrate that expanding the number and variety of predictors does not necessarily produce large improvements in individual prognosis. They also demonstrate that penalized regression can serve as a substantive predictive method within research described as machine learning.

Cearns, Opel, et al. (2019) investigated rehospitalization within two years following an inpatient major depressive episode. Their analysis included 380 patients and combined clinical, imaging, biological, genetic, medication, and sleep-related information. The best reported multimodal panel achieved a test AUROC of approximately 0.677. Although rehospitalization differs from depression status, the study shows that extensive multimodal information does not automatically yield near-perfect psychiatric prediction. It provides a useful counterpoint to expectations that more complex inputs or algorithms must result in high clinical accuracy.

Ebert et al. (2019) examined major depressive disorder onset during the first year of college. The prospective study included 2,519 participants at baseline and 958 at follow-up, with a reported one-year incidence of 6.9%. Its importance for the present benchmark is the explicit

separation between baseline assessment and future onset, together with the selection of a clearly defined population. The follow-up sample also illustrates why real longitudinal prediction studies must address issues such as attrition, which are not currently represented in the synthetic dataset.

These studies support investigating depression prediction without assuming that strong performance is inevitable. They also indicate that a benchmark's apparent success depends partly on the difficulty of the chosen endpoint. Predicting future depression among individuals with informative baseline symptoms may differ substantially from predicting first onset in individuals without previous psychiatric difficulties.

## **5. Direct evidence does not support universal machine-learning superiority**

A central issue is whether machine learning improves performance beyond conventional statistical modelling when both approaches receive comparable information and development attention.

Christodoulou et al. (2019) systematically reviewed comparisons of machine learning and logistic regression in clinical prediction studies and found no overall performance advantage of machine learning in the evidence examined. This finding challenges claims of automatic superiority, but it should not be interpreted as proof that regression always performs equally well. Its strongest implication is that algorithmic complexity alone is insufficient evidence of added predictive value. Differences must be demonstrated through appropriately conducted comparisons.

A contrasting result was reported by Couronné et al. (2018), who compared random forest and logistic regression across 243 datasets representing different prediction tasks. Random forest showed favourable average predictive performance. This broad benchmark indicates that flexible algorithms can offer advantages, while also showing why conclusions depend on the

datasets, comparator specification, and evaluation conditions. A result averaged across diverse tasks does not establish the expected advantage for depression prediction or for every clinical application.

Morelli et al. (2021) provide a particularly relevant depression-related comparison. Using more than 400,000 UK Biobank participants, they developed ten-year depression and anxiety risk models using Cox regression and DeepSurv. In the published article, the reduced depression models achieved validation concordance indices of 0.7769 for Cox regression and 0.7772 for DeepSurv. The close performance shows that a neural-network approach can provide little additional discrimination over a conventional model in a large structured dataset. Because the study used survival outcomes, its concordance indices should not be equated directly with the binary AUROC planned for the present benchmark.

The combined evidence supports a conditional interpretation. Machine-learning advantages may emerge when the data contain relationships that the comparator cannot adequately represent. They may diminish when conventional models already capture most of the available predictive signal. This synthesis motivates H1, H2, and H3, while leaving their empirical confirmation to the planned analyses.

## **6. Nonlinearity provides a rationale for flexibility, not proof of clinical advantage**

Theoretical differences between algorithms help explain why their relative performance may vary across scenarios. Breiman's (2001) random forest framework combines multiple tree predictors, providing a mechanism for representing thresholds and interactions. Friedman's (2001) gradient boosting framework builds an additive model through successive improvements in a chosen loss function. These foundational methods support the expectation that flexible algorithms can capture structures that an additive linear predictor does not directly represent.

However, this expectation depends on the presence of learnable nonlinear signal. Flexibility may be unhelpful when the relevant relationships are approximately additive, the signal is weak, or the development sample does not contain enough information to estimate complex structure reliably. An algorithm can also represent a relationship in principle without learning it effectively under a particular tuning strategy.

The present nonlinear scenario deliberately introduces curvature, thresholds, and interactions. It therefore provides a controlled test of whether flexible methods exploit these specified structures more successfully than additive regression. Any advantage would demonstrate responsiveness to the generating mechanism, rather than establish that the same nonlinear relationships exist in patients.

Clinical evidence can motivate the choice of predictor domains without validating their simulated mathematical forms. For example, Baglioni et al. (2011) synthesized longitudinal evidence linking insomnia with subsequent depression. This supports the relevance of sleep to depression prediction, but it does not establish the precise quadratic sleep relationship or variability threshold used in the present simulation.

H2 is consequently supported by algorithmic capability and the specified nonlinear experiment. It remains a testable expectation because finite samples, noise, missingness, and tuning may prevent the anticipated advantage from being observed.

## **7. Strong regression specification is necessary for a fair comparison**

Regression models can incorporate nonlinear transformations, interactions, and regularization. Treating conventional modelling as synonymous with an unpenalized additive model understates the available alternatives and risks attributing gains from better specification to machine learning as a broad category.

Tibshirani (1996) introduced the lasso as a method combining shrinkage with variable selection. This foundational contribution demonstrates that regression-based prediction can control complexity rather than estimate every available coefficient without constraint. Regularization is relevant when predictors are numerous or correlated and when unrestricted estimation would produce unstable coefficients.

Dinga et al. (2018) provide a psychiatric example of using penalized logistic regression as a predictive method, while Morelli et al. (2021) show that conventional models can remain competitive with more complex methods in large depression-related datasets. These studies support including stronger conventional alternatives in the benchmark rather than using a single basic model to represent all regression approaches.

In the present study, the additive, penalized, and flexible logistic models answer complementary questions. The additive model establishes a conventional reference. Penalized regression examines the value of controlling coefficient estimation. Flexible regression evaluates whether explicitly representing nonlinear structure reduces the advantage of algorithms that learn such structure more automatically.

This reasoning supports H3: the estimated machine-learning advantage is expected to become smaller when the comparator is strengthened. The hypothesis concerns the relative performance gap, not a presumption that flexible regression must outperform every algorithm.

## **8. Electronic health records expand prediction opportunities while complicating interpretation**

Electronic health records contain information about symptoms, diagnoses, treatment, physical health, and healthcare utilization. These data can support depression prediction, but they

also create challenges concerning timing, missingness, diagnostic ascertainment, and the meaning of the outcome.

Huang et al. (2014) developed computational models using electronic health records to predict depression diagnosis, severity, and treatment-related outcomes. Their work illustrates the potential of routinely collected information to support multiple prediction tasks. It also reinforces the need to distinguish these tasks, since detecting a recorded diagnosis and forecasting subsequent clinical deterioration require different definitions of available information.

Nemesure et al. (2021) investigated depression and anxiety prediction using medical-record information from 4,184 university students. Their work illustrates the relevance of combinations of physical, social, and health characteristics. However, an analysis based on recorded diagnoses must consider whether patterns reflect illness, diagnostic practice, or healthcare access. The authors also acknowledged that diagnostic bias could not be ruled out.

Hochman et al. (2021) developed and validated a machine-learning model for postpartum depression using a nationwide cohort. This study demonstrates the value of defining a specific population and prediction context. Its findings concern postpartum depression and should therefore remain distinguished from prediction among general adult populations.

Nickson et al. (2023) systematically reviewed machine-learning approaches to depression prediction and diagnosis using electronic health records. Their review documents a literature containing diverse data sources, modelling methods, and prediction tasks. This diversity supports a structured benchmark but also limits the usefulness of comparing headline performance estimates without accounting for differences in outcome definition and study design.

For the synthetic benchmark, this literature motivates careful specification of predictor timing and outcome meaning. The generated baseline variables are deliberately separated from

future outcomes, providing a cleaner methodological experiment than many routinely collected datasets can offer.

## **9. Depression heterogeneity limits what aggregate prediction can reveal**

Depression is not represented fully by a single numerical severity measure. Individuals with similar total scores can differ in the symptoms contributing to those scores, their histories, and their subsequent trajectories.

Fried and Nesse (2015) argued that symptom sum scores can conceal important differences between depression symptoms. Related empirical work by Fried et al. (2014) examined differences in risk-factor associations across individual symptoms. These studies support considering the consequences of reducing heterogeneous presentations to one outcome or predictor summary.

For prediction, this heterogeneity has two implications. First, a model may perform differently depending on which manifestations are common in the development population. Second, performance for an aggregate outcome may conceal weaker prediction for particular presentations. These possibilities support examining cohort-specific performance rather than reporting only a pooled estimate.

The present benchmark nevertheless represents a limited form of heterogeneity. It changes selected statistical relationships across cohorts and uses aggregate simulated symptom scores. It does not reproduce symptom-level networks, clinical diagnostic ambiguity, or the full diversity of depressive presentations. H4 should therefore be understood as a hypothesis about specified cohort differences rather than a comprehensive test of depression heterogeneity.

## **10. Leakage and model-selection bias can produce misleading advantages**

Data leakage can create predictive relationships that would not be available in the intended application. Kapoor and Narayanan (2023) documented leakage across scientific fields and demonstrated how correcting methodological errors can substantially alter conclusions about complex-model superiority. Their contribution supports examining how information enters a model rather than accepting performance estimates simply because the code can be reproduced.

Rosenblatt et al. (2024) investigated leakage in connectome-based prediction across four datasets and three phenotypes. Leakage involving feature selection and repeated participants substantially inflated performance, while other leakage mechanisms had smaller effects. Although the study concerned neuroimaging, it provides direct evidence that leakage effects depend on the mechanism and analysis context. It cautions against assuming that every preprocessing error produces the same degree of inflation.

Varma and Simon (2006) demonstrated bias when cross-validation used to optimize a classifier was also used to estimate the selected classifier's performance. Cawley and Talbot (2010) further examined overfitting during model selection and the resulting bias in performance evaluation. Both studies support separating the selection procedure from the data used to estimate its performance.

Vabalas et al. (2019) investigated validation with limited samples and showed that feature selection across pooled training and test data could substantially bias performance estimates. Their findings support fitting selection procedures within the training partitions. They do not imply that ordinary cross-validation is inherently invalid when the entire development procedure is appropriately confined to each fold.

This literature supports strict separation of imputation, scaling, encoding, feature selection, tuning, calibration, and threshold determination from evaluation data. It also motivates

H8, which predicts inflated performance when deliberately invalid outcome-related variables are introduced. The separate positive control will demonstrate a compromised prediction task and will remain excluded from valid comparisons.

### **11. Sample size and prediction stability affect the credibility of model rankings**

A performance estimate can be uncertain even when no leakage occurs. Small evaluation samples produce imprecise estimates, while small development samples can produce unstable fitted models.

Schnack and Kahn (2016) examined sample-size concerns in psychiatric neuroimaging and highlighted the challenges of deriving reliable classifiers from limited samples. Varoquaux (2018) showed why small samples can yield large uncertainty in cross-validation estimates. These studies are especially relevant to interpreting apparently impressive performance from small psychiatric datasets, although their specific neuroimaging findings should not be transferred numerically to structured clinical predictors.

Riley and Collins (2023) examined the stability of clinical prediction models developed using statistical and machine-learning methods. Their work emphasizes that estimated individual risks can change substantially when the development data change. A satisfactory average performance estimate therefore does not automatically imply that an individual's risk estimate is stable.

The present dataset contains 18,000 observations overall, but each scenario is a separate experiment with 4,000 development observations and 2,000 held-out observations. Treating all 18,000 as one development sample would misrepresent the design. Moreover, thousands of participants do not compensate fully for having only a small number of independently specified cohort contexts.

These considerations support reporting uncertainty and avoiding strong claims based on small performance differences. They also support expanding the benchmark to independently generated repetitions if the objective is to estimate general simulation properties rather than describe one realization.

## **12. Distribution shift challenges transportability even without leakage**

A model can be developed correctly and still perform poorly when the application population differs from the development population. Distribution shift concerns changes in the statistical environment, while leakage concerns inappropriate information use. The two require separate interpretation.

Subbaswamy and Saria (2020) discussed dataset shift and the development of models that remain reliable when conditions change. Subbaswamy et al. (2019) investigated predictive models designed to transport across distributions, while Subbaswamy et al. (2021) examined evaluation of robustness and stability under shift. These contributions provide a methodological basis for testing more than one evaluation environment.

The PredictD results offer a depression-specific illustration of reduced discrimination across settings. However, a decline between countries does not identify a single cause, because population composition, predictor measurement, outcome ascertainment, and underlying relationships may all differ.

The heterogeneous simulation makes selected changes explicit, including cohort-specific intercepts and changes in symptom and stress coefficients. This provides a controlled examination of whether fitted models preserve performance when the generating relationships change.

H4 follows from the expectation that a model trained on one relationship structure may misestimate risks under another. Nevertheless, neither discrimination nor Brier-score deterioration is guaranteed: changes in prevalence, case mix, and signal strength can influence those measures. Calibration, discrimination, and overall probability error must therefore be examined separately.

### **13. Calibration and discrimination assess different properties**

Discrimination evaluates ordering, whereas calibration evaluates the accuracy of predicted probabilities. A model may rank individuals correctly while systematically assigning risks that are too high or too low.

Steyerberg et al. (2010) described a framework spanning overall performance, discrimination, calibration, and decision-oriented evaluation. Van Calster et al. (2019) emphasized calibration as a necessary component of reliable prediction and explained why poorly calibrated probabilities can mislead decisions. Together, these papers support avoiding reliance on AUROC as the sole criterion of model quality.

Niculescu-Mizil and Caruana (2005) examined probability estimation across supervised-learning methods and demonstrated that good classifier performance does not necessarily produce well-calibrated probabilities. Their findings provide a methodological explanation for why probability calibration may require separate attention from ranking performance.

For the present benchmark, this evidence supports H5: the model with the highest AUROC may not achieve the lowest Brier score or the best calibration. The Brier score should also remain distinguished from a pure calibration measure, since it reflects overall squared probability error and is influenced by several aspects of performance.

Calibration procedures must be developed using training data. Recalibrating on held-out outcomes would change the evaluation question and invalidate interpretation of those observations as untouched tests of the original pipeline.

#### **14. Threshold performance and decision utility may change model preferences**

Binary decisions require a threshold applied to predicted probabilities. The threshold affects sensitivity and specificity, while predictive values also depend on the population in which the model is evaluated.

Saito and Rehmsmeier (2015) explained the value of precision-recall analysis when class frequencies are uneven. Their work supports reporting precision-related measures alongside AUROC and interpreting them in relation to prevalence. A difference in precision-recall performance across cohorts may reflect a change in outcome frequency as well as a change in model quality.

Vickers and Elkin (2006) introduced decision-curve analysis to evaluate prediction strategies using net benefit across threshold probabilities. This approach connects predictions with assumptions about the relative consequences of false-positive and false-negative classifications. It supports assessing whether a statistical improvement changes the value of a hypothetical decision strategy.

This literature motivates H6 and H7. Relative performance may vary across thresholds, and any net-benefit advantage may depend on the scenario and assumed decision consequences. These expectations require evaluation rather than assertion. A higher AUROC can coexist with a less favourable operating point or limited net benefit.

Because the benchmark is synthetic, its decision curves will quantify utility under specified mathematical assumptions. They will not demonstrate that an intervention improves patient outcomes or establish a clinically validated threshold for treatment.

## **15. Reporting standards support transparent and balanced comparisons**

Methodological weaknesses remain an important concern in clinical prediction research. Andaur Navarro et al. (2021) systematically assessed risk of bias in studies developing supervised machine-learning prediction models. Their review supports examining the quality of study design and analysis rather than judging credibility from algorithm choice or reported performance alone.

Cearns, Hahn, and Baune (2019) provided recommendations for supervised machine learning in psychiatry, emphasizing the importance of methodological rigour for clinical translation. Their contribution supports treating preprocessing, validation, and dissemination as integral components of the predictive study.

TRIPOD+AI provides reporting guidance for prediction models developed using regression or machine learning (Collins et al., 2024). PROBAST+AI provides an updated framework for assessing quality, risk of bias, and applicability (Moons et al., 2025). These frameworks have complementary roles: complete reporting makes appraisal possible, while risk-of-bias assessment considers whether the reported procedures support credible conclusions. Following a reporting checklist does not itself establish that a model is clinically useful.

For this benchmark, transparency requires documenting the generating equations, predictor timing, development and evaluation assignments, missingness procedures, model specifications, tuning choices, calibration methods, and performance criteria. Reporting all

prespecified comparisons will also reduce the risk of presenting only favourable algorithm-scenario combinations.

## **16. Simulation provides controlled evidence with explicit boundaries**

Morris et al. (2019) described simulation as a method for evaluating statistical procedures when aspects of the generating process are known. Their structured approach connects aims, data-generating mechanisms, estimands, methods, and performance measures. This framework supports using the present scenarios to examine how model comparisons change when relationships become nonlinear or vary across cohorts.

The strength of the simulation is the ability to specify these conditions explicitly. Its limitation is that findings depend on those specifications. The nonlinear scenario intentionally contains structures that flexible algorithms may capture, while the linear scenario provides a setting favourable to appropriately specified additive models. Results should therefore explain behaviour under these mechanisms rather than establish a universal ranking.

The current dataset also represents one reproducible realization. Repeated independent generations would be needed to estimate Monte Carlo uncertainty and assess whether observed comparisons persist across sampling variation. Resampling one generated dataset cannot fully replace that process.

Accordingly, the literature supports the benchmark as a methodological investigation. It can reveal when added flexibility helps, when stronger regression comparators narrow the performance gap, and when cohort changes compromise probability estimates. Clinical conclusions would require subsequent evaluation in appropriately defined real populations.

## **17. Integration of the evidence with the study hypotheses**

The reviewed evidence provides different levels of support for the eight hypotheses. Some hypotheses follow closely from established evaluation principles, while others are conditional expectations arising from the specified simulation design.

| <b>Hypothesis</b>                                                    | <b>Relevant evidence</b>                                                                | <b>Interpretation for the benchmark</b>                                                                                       |
|----------------------------------------------------------------------|-----------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| H1: Smaller machine-learning gains in the linear scenario            | Christodoulou et al. (2019); Morelli et al. (2021)                                      | Conventional models can remain competitive; the expected scenario difference still requires testing.                          |
| H2: Greater gains over additive regression in the nonlinear scenario | Breiman (2001); Friedman (2001); Couronné et al. (2018)                                 | Flexible algorithms can represent nonlinear structure, but their advantage depends on learnable signal and model development. |
| H3: Smaller gains against stronger regression alternatives           | Tibshirani (1996); Dinga et al. (2018); Morelli et al. (2021)                           | Regularization and improved specification provide more demanding comparators.                                                 |
| H4: Greater deterioration under changing cohort relationships        | King et al. (2008); Subbaswamy and Saria (2020)                                         | Population changes can impair transportability, although the direction and magnitude of each metric are not guaranteed.       |
| H5: Different rankings for discrimination and probability accuracy   | Steyerberg et al. (2010); Niculescu-Mizil and Caruana (2005); Van Calster et al. (2019) | Good ranking does not establish accurate probabilities.                                                                       |
| H6: Threshold-dependent comparative performance                      | Saito and Rehmsmeier (2015); Steyerberg et al. (2010)                                   | Operating performance depends on the chosen threshold and evaluation population.                                              |
| H7: Scenario- and threshold-dependent net benefit                    | Vickers and Elkin (2006)                                                                | Utility depends on decision assumptions, and simulated benefit is distinct from demonstrated clinical benefit.                |
| H8: Inflated performance from invalid outcome-related predictors     | Varma and Simon (2006); Kapoor and Narayanan (2023); Rosenblatt et al. (2024)           | Inappropriate information can produce misleading estimates and must remain outside the valid benchmark.                       |

The literature supports evaluating these hypotheses without presuming their confirmation. Similar performance would indicate that added flexibility provides limited incremental value under the tested conditions, but would not establish formal equivalence without suitable margins and uncertainty estimates. Superior performance would support a model's advantage within the relevant scenario, but would not demonstrate universal clinical superiority. Performance deterioration in held-out cohorts would identify limitations under the specified changes, while leaving its magnitude and causes to be examined through the completed analyses.

## Methods

### Study design and objectives

This study comprises a controlled computational benchmark comparing conventional statistical models and machine-learning algorithms for predicting simulated depression status six months after baseline assessment. The primary objective is to evaluate whether random forest, gradient boosting, and support vector machines improve predictive performance beyond appropriately specified logistic regression when all methods receive equivalent baseline information and undergo leakage-controlled development.

Three data-generating scenarios were constructed: predominantly additive linear relationships on the log-odds scale, nonlinear relationships, and nonlinear relationships with additional variation across cohorts. The scenarios will be analysed separately to investigate how relationship structure and cohort differences affect relative model performance.

The benchmark contains entirely synthetic observations. No patients were recruited, no clinical records were accessed, and no measurements were obtained from human participants. Cohorts and sites are computational labels rather than actual healthcare institutions. The six-month interval defines the simulated prediction horizon; it does not represent an observed follow-up period.

The simulation design follows the principle of explicitly connecting aims, data-generating mechanisms, modelling methods, and performance measures described by Morris et al. (2019). Reporting will draw on relevant TRIPOD+AI recommendations for transparent prediction-model research (Collins et al., 2024). The analyses described below constitute the proposed analysis plan and have not yet been executed.

### Dataset structure and sample-size rationale

A total of 18,000 unique synthetic participants were generated. Each of the three scenarios contains six cohorts of 1,000 participants, giving 6,000 observations per scenario. Cohorts C1 to C4 were assigned to model development, whereas C5 and C6 were reserved for held-out evaluation.

Accordingly, each scenario contains 4,000 development observations and 2,000 held-out observations. The combined dataset serves as a storage and documentation resource. It will not be treated as one homogeneous modelling population.

Each cohort contains three nominal sites. Site membership was assigned cyclically across records, producing site sizes of 333 or 334 observations. Site labels do not introduce additional site-specific effects into the generating equations. Their role is to support reproducible grouped partitioning, rather than to represent empirically estimated clustering within healthcare institutions.

The sample size was selected as a practical benchmark size permitting nested model development and evaluation in two distinct simulated cohorts. It was not derived from a clinical recruitment target or a formal power calculation. No claim is made that this sample size guarantees detection of a clinically meaningful performance difference.

The current datasets represent one realization of the generating mechanisms. Repeated independent simulations would be required to estimate Monte Carlo variability and support broader conclusions about average method performance across generated samples.

### **Synthetic population and baseline predictors**

All simulated participants were adults aged between 18 and 85 years. Each participant contributed one baseline predictor record and one binary future outcome. Participant identifiers

were unique across cohorts and scenarios, preventing participant overlap between development and evaluation datasets.

Seventeen baseline predictors were generated:

| <b>Domain</b>       | <b>Predictor</b>             | <b>Encoding or range</b>   |
|---------------------|------------------------------|----------------------------|
| Demographic         | Age                          | Integer years, 18 to 85    |
| Demographic         | Sex at birth                 | Female or male             |
| Demographic         | Education                    | Integer years, 0 to 22     |
| Socioeconomic       | Financial strain             | Continuous score, 0 to 10  |
| Psychiatric history | Previous depression          | 0 = absent; 1 = present    |
| Psychiatric history | Family history of depression | 0 = absent; 1 = present    |
| Physical health     | Chronic condition count      | Integer count, 0 to 6      |
| Psychological       | Baseline depressive symptoms | Integer score, 0 to 27     |
| Psychological       | Baseline anxiety             | Integer score, 0 to 21     |
| Psychological       | Stress                       | Continuous score, 0 to 40  |
| Social              | Social support               | Continuous score, 0 to 100 |
| Sleep               | Sleep duration               | Hours, 3 to 11             |
| Sleep               | Sleep variability            | Hours, 0.1 to 3            |
| Behavioural         | Physical activity            | Minutes per week, 0 to 900 |
| Behavioural         | Alcohol consumption          | Units per week, 0 to 35    |
| Behavioural         | Smoking                      | 0 = absent; 1 = present    |
| Treatment history   | Baseline antidepressant use  | 0 = absent; 1 = present    |

Sex was represented using a simplified binary variable. The dataset does not represent gender identity or the full range of biological sex characteristics.

The depressive symptom and anxiety variables reproduce numerical ranges associated with familiar screening measures but were generated directly as synthetic scores. They are not responses to administered questionnaires and will not be described as validated PHQ-9 or GAD-7 measurements.

Antidepressant use represents baseline exposure. Its inclusion does not establish treatment effectiveness or a causal treatment effect.

Participant identifier, synthetic-data indicator, scenario, cohort, site, and development/evaluation assignment will be excluded from the primary predictor set. These variables identify records and define the evaluation design.

### Generation of correlated baseline characteristics

Data were generated using NumPy's pseudorandom-number generator with seed 20261001. Cohorts were indexed using  $(c=0, \dots, 5)$ , corresponding to C1 through C6.

A latent baseline burden variable was generated for each participant:

$$U_i \sim N(0.12c, 1).$$

This variable introduced dependence among several psychological, social, and behavioural predictors. It was not included in the released predictor dataset and will not be supplied to the fitted prediction models.

Age was generated from a normal distribution with mean  $(42+2c)$  and standard deviation 15, clipped to 18 to 85 years and converted to an integer. Education was generated with mean  $(14-0.3c)$  and standard deviation 3, clipped to 0 to 22 years and converted to an integer. Female and male sex categories were assigned with probabilities 0.56 and 0.44, respectively.

Financial strain was generated as:

$$F_i = \text{clip}(4 + 1.6U_i + \epsilon_{F,i}, 0, 10),$$

where  $(\epsilon_{F,i} \sim N(0, 2^2))$ .

Previous depression was drawn from a Bernoulli distribution with probability:

$$P(P_i=1) = \text{expit}(-1.3 + 0.45U_i).$$

Family history was drawn from a Bernoulli distribution with probability 0.25. Chronic condition count was drawn from a Poisson distribution with mean  $(0.6 + \text{age}_i/100)$ , capped at six.

Depressive symptom and anxiety scores were generated as:

$$D_i = \text{clip}\left(\text{round}(5 + 2U_i + 3P_i + \epsilon_{D,i}), 0, 27\right),$$

$$A_i = \text{clip}\left(\text{round}(5 + 2U_i + \epsilon_{A,i}), 0, 21\right),$$

where both error terms have normal distributions with standard deviation 3.

Stress and social support were generated as:

$$T_i = \text{clip}(16 + 5U_i + \epsilon_{T,i}, 0, 40),$$

$$S_i = \text{clip}(65 - 9U_i + \epsilon_{S,i}, 0, 100),$$

with error standard deviations of 5 and 15, respectively.

Sleep duration was generated from  $(7 - 0.4U_i + \epsilon_{L,i})$ , with error standard deviation 0.9, and clipped to 3 to 11 hours. Sleep variability followed a lognormal distribution with log-scale mean  $(-0.45 + 0.15U_i)$  and standard deviation 0.45, clipped to 0.1 to 3 hours.

Physical activity was generated from a gamma distribution with shape 2 and scale 65, multiplied by  $(\exp(-0.15U_i))$ , and clipped to 0 to 900 minutes per week. Alcohol consumption followed a gamma distribution with shape 1.2 and scale 3, capped at 35 units per week.

Smoking probability was 0.18, increasing to 0.24 when financial strain exceeded six. Baseline antidepressant-use probability was 0.35 among participants with previous depression and 0.04 among those without previous depression.

The generating distributions and coefficients are experimental assumptions. They were not estimated from a clinical population.

### Outcome definition and generation

The outcome is simulated depression status at six months, coded 0 or 1. It was generated probabilistically rather than obtained by thresholding a baseline symptom score.

For participant  $i$ :

$$p_i = \text{expit}(\eta_i), \quad Y_i \sim \text{Bernoulli}(p_i).$$

The shared additive component of the linear predictor was:

$$\eta_i^{(L)} = -1.7 + 0.16(D_{i-6}) + 0.065(A_{i-5}) + 0.045(T_{i-16}) - 0.013(S_{i-65}) + 0.12(F_{i-4}) + 0.50P_i + 0.25H_i + 0.12C_i + 0.15M_i,$$

where  $(H_i)$  denotes family history,  $(C_i)$  chronic condition count, and  $(M_i)$  smoking.

Not every supplied predictor enters the outcome equation directly. Some provide correlated or contextual information, permitting evaluation when the candidate predictor set contains variables with limited additional signal.

The outcome can represent persistent or newly occurring depression because the population includes participants with previous depression and varying baseline symptom levels. It will not be interpreted exclusively as first-onset depression.

### Data-generating scenarios

*Predominantly linear scenario*

The linear scenario used the shared additive equation without additional nonlinear terms. Predictor distributions and missingness probabilities varied across cohorts, but the conditional outcome coefficients remained constant.

“Linear” refers to additive relationships on the log-odds scale. Outcome probability remains nonlinear through the logistic transformation.

### *Nonlinear scenario*

The nonlinear scenario added:

$$g(X_i) = \begin{aligned} &0.30(L_i - 7)^2 + 0.55I(V_i > 0.9) \\ &+ 0.65I(T_i > 23, S_i < 45) \\ &+ 0.45I(P_i = 1, D_i > 9) - 0.22\log(1 + Q_i/60), \end{aligned}$$

where  $(L_i)$  denotes sleep duration,  $(V_i)$  sleep variability, and  $(Q_i)$  physical activity.

This mechanism introduced curvature, thresholds, interactions, and a transformed activity association:

$$\eta_i^{(N)} = \eta_i^{(L)} + g(X_i).$$

These structures were selected to create a controlled nonlinear prediction task. They are not estimates of actual clinical mechanisms.

### *Heterogeneous scenario*

The heterogeneous scenario retained the nonlinear mechanism and added cohort-dependent effects:

$$\eta_i^{(H)} = \eta_i^{(N)} + \delta_c + \kappa_c(D_i - 6) + \tau_c(T_i - 16).$$

The intercept adjustments for C1 through C6 were:

$$\Delta_c = (-0.35, -0.15, 0, 0.15, 0.40, 0.70).$$

The additional symptom coefficient was  $(-0.015)$  for C1 to C4 and  $(0.07)$  for C5 and C6. The additional stress coefficient was zero except in C6, where it was 0.035.

This scenario combines changes in predictor distributions with changes in selected conditional outcome relationships.

### Missing predictor values

Missingness was introduced after complete baseline characteristics and outcomes had been generated. Nine predictors were eligible: education, financial strain, anxiety, stress, social support, sleep duration, sleep variability, physical activity, and alcohol consumption.

For each eligible variable, missingness probability was:

$$P(R_{ij}=1) = 0.025 + 0.008c + 0.025I(F_i > 7),$$

where  $(R_{ij}=1)$  indicates a missing value and  $(F_i)$  is financial strain before masking.

Missingness therefore varies with cohort and underlying baseline financial strain. It does not depend directly on the future outcome. Because financial strain itself can be missing, the mechanism will not be described as unconditionally missing completely at random or assumed to be missing at random relative to every observed record.

Missing values were encoded as empty CSV cells. Outcomes were complete, and no follow-up attrition was simulated. Released predictors remain unimputed and unscaled.

### Data integrity and analysis provenance

Completed generation checks confirmed the expected observation count, unique participant identifiers, binary outcomes, and absence of participant overlap between development and held-out datasets. SHA-256 checksums were recorded for the generated files.

Before modelling, further checks will confirm cohort sizes, variable types, allowable ranges, missingness frequencies, and consistency of development assignments. Corrections to programming errors will be documented, and datasets will not be regenerated to obtain preferred comparative results.

Descriptive cohort summaries, including outcome prevalence, have already been inspected. Model performance has not been examined. The analysis plan will therefore be described as established before model fitting, without claiming that the datasets have remained entirely uninspected or that the study was prospectively registered.

### **Predictor preprocessing**

All models will use the same 17 baseline predictors. The primary analysis will retain all predictors rather than apply preliminary outcome-based screening.

Missing numeric values will be imputed using medians estimated from the applicable training partition. Missingness indicators will be added for eligible variables using a fixed specification. Sex will be encoded as a binary indicator.

Continuous predictors will be standardized for logistic regression and support vector machines using training-partition means and standard deviations. Tree-based models will receive imputed predictors without standardization.

For flexible regression, spline transformations will be fitted within each training partition. Knot locations and any other learned transformation parameters will never be calculated using evaluation observations.

Imputation, encoding, scaling, and transformation will form part of the fitted pipeline. No preprocessing parameter will be learned from combined training and evaluation data.

### **Prediction models**

Six model specifications are proposed.

### *Conventional logistic regression*

The conventional comparator will contain an intercept and additive main effects for all baseline predictors and prespecified missingness indicators. Continuous variables will enter as linear terms on the log-odds scale.

It will be fitted without penalization. Convergence warnings, separation, or numerical failures will be recorded. Any corrective change will be reported rather than silently replacing this comparator with penalized regression.

### *Penalized logistic regression*

A ridge-penalized logistic model will use the same additive predictor specification. The inverse regularization parameter will be selected from:

$$C \in \{0.01, 0.1, 1, 10, 100, 1000\}.$$

This model will evaluate whether shrinkage improves estimation stability.

### *Flexible logistic regression*

Flexible logistic regression will apply cubic spline transformations with four training-derived knots to numeric nonbinary predictors. Binary predictors will enter as main effects.

Two interactions will be specified: baseline symptoms with previous depression, and stress with social support. Ridge regularization will use the same candidate values as penalized logistic regression.

Because the generating mechanisms are known, these choices cannot be described as blind to simulation structure. The flexible comparator will be presented as a deliberately strengthened regression specification rather than as an independently derived clinical model.

### *Random forest*

Random forest will aggregate 500 classification trees. Tuning will consider maximum depth, minimum leaf size, predictor subsampling, and bootstrap sampling.

Candidate values will include depths of 3, 5, 10, and unrestricted; minimum leaf sizes of 1, 5, 10, and 20; and predictor-subsampling settings of square root, 0.5, and 1.0.

#### *Gradient boosting*

Histogram-based gradient boosting will be evaluated. Candidate settings will include learning rates of 0.03, 0.05, and 0.10; maximum iterations of 100, 200, and 300; maximum leaf counts of 7, 15, and 31; minimum leaf sizes of 10, 20, and 50; and regularization values of 0, 1, and 10.

#### *Support vector machine*

A support vector classifier with a radial-basis-function kernel will use imputed, standardized predictors. Candidate settings will include:

$$C \in \{0.1, 1, 10, 100\}, \gamma \in \{0.001, 0.01, 0.1\}$$

and kernel coefficients comprising the implementation's scale setting and numerical values 0.001, 0.01, and 0.1.

Its decision scores will be converted to probabilities through the training-only calibration procedure.

No synthetic oversampling or outcome-dependent class weighting is planned in the primary analysis. Randomized searches for the three machine-learning algorithms will evaluate up to 24 unique configurations per model using recorded random seeds.

#### **Nested development and model selection**

Within each scenario, development will use four outer folds defined by cohort. Each iteration will train on three development cohorts and evaluate on the remaining development cohort.

Within each outer training set, five-fold grouped cross-validation will use nominal site labels. Identical partitions will be applied across models. These inner site splits do not represent testing of substantive site-specific effects because such effects were not generated.

Hyperparameters will be selected using mean inner-fold log loss. This probability-based objective avoids selecting models solely for ranking ability. An exact tie will be resolved using a documented preference for the less complex configuration.

All preprocessing will be repeated inside each inner training partition. Outer evaluation outcomes will remain unavailable to preprocessing, tuning, calibration, and threshold selection.

Nested validation separates model selection from performance estimation, addressing the optimism described by Varma and Simon (2006).

### **Probability calibration**

For the primary probability comparison, each model will undergo the same sigmoid-calibration framework using only its applicable training data.

After hyperparameter selection, grouped cross-fitting within the training partition will produce out-of-fold scores. A sigmoid mapping will be fitted to these scores and their corresponding training outcomes. The base pipeline will then be refitted on the entire training partition, and the mapping will be applied to its evaluation scores.

The mapping will operate on decision scores where available and clipped log-odds of native probabilities otherwise. Calibration must not use outer evaluation outcomes or held-out cohort outcomes.

Native, uncalibrated probabilities will be reported as a secondary analysis for models that provide them. Raw support-vector scores will not be treated as probabilities.

### **Final fitting and held-out evaluation**

After nested development, each model specification will be tuned and fitted using C1 to C4 only. Development cross-fitting will supply calibration information and decision thresholds.

Predictions for C5 and C6 will be generated without further model adaptation. Evaluation will be reported separately for each cohort.

The primary summary discrimination measure will be the arithmetic mean of the two cohort-specific AUROCs. Pooled held-out performance will be secondary because pooling may incorporate differences between cohorts into the apparent discrimination.

The files label C5 and C6 as external tests, but manuscript reporting will describe them as held-out simulation cohorts. Their evaluation does not constitute clinical external validation.

### **Performance measures**

The primary performance endpoint will be AUROC. The primary algorithmic comparison will contrast each machine-learning model with flexible logistic regression within each scenario. Comparisons with additive and penalized regression will be secondary.

Additional measures will include:

- Area under the precision-recall curve, accompanied by cohort prevalence.
- Brier score for overall squared probability error.
- Calibration intercept and slope.
- Calibration plots.
- Sensitivity, specificity, positive predictive value, and negative predictive value.
- Simulated decision-curve net benefit.

For predicted probability  $\hat{p}_i$ , the Brier score will be:

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{p}_i)^2.$$

Calibration-in-the-large will be estimated using predicted log-odds as an offset with an estimated intercept. Calibration slope will be estimated from a logistic model containing an intercept and predicted log-odds as a predictor. These evaluation fits describe calibration and will not be used to update held-out predictions.

Calibration plots will display observed and predicted probabilities in quantile-based groups. Group sizes and instability will be reported.

### **Threshold selection and decision curves**

Two decision rules are proposed. The first will select, from development predictions, the threshold providing maximum specificity while maintaining sensitivity of at least 0.80. The second will use a fixed probability threshold of 0.50.

The sensitivity target is an analytical choice rather than a clinically validated requirement. Selected thresholds will be locked before evaluation in C5 and C6.

Decision curves will assess thresholds from 0.05 to 0.50 in increments of 0.01. Net benefit will be calculated as:

$$\text{Net benefit} = \frac{TP}{n} - \frac{FP}{n} \frac{t}{1-t}.$$

Model strategies will be compared with classifying everyone and classifying no one as requiring action. The threshold range is exploratory and has no validated clinical interpretation in this synthetic population.

### **Uncertainty and comparative interpretation**

Paired bootstrap resampling will be used to estimate uncertainty in performance differences. Within a scenario, the same sampled evaluation records will be used for every model, preserving pairing.

The proposed procedure will use 2,000 bootstrap samples, sampling participants separately within each held-out cohort. Models and thresholds will remain fixed. Percentile 95% confidence intervals will be reported.

These intervals describe uncertainty in evaluation conditional on the fitted models and generated cohorts. They do not incorporate the full uncertainty from development, independently generated datasets, or new cohort environments. Bootstrap samples lacking either outcome class will be excluded from metrics requiring both classes, with exclusions reported.

Differences will be defined so that positive values indicate improvement where possible. For Brier score, improvement will be calculated as the comparator score minus the machine-learning score.

Comparisons will be estimation-focused. Confidence intervals will not be interpreted as multiplicity-adjusted confirmatory tests. Failure to demonstrate superiority will not establish equivalence, and no clinically validated superiority or equivalence margin has been specified.

### **Hypothesis evaluation**

H1 and H2 will be evaluated by comparing discrimination gains across linear and nonlinear scenarios. H3 will compare gains against additive regression with gains against stronger regression alternatives.

H4 will examine development-to-held-out differences in calibration and probability error. Changes in prevalence and case mix will be considered when interpreting Brier-score differences.

H5 will compare model rankings across discrimination, Brier score, and calibration. H6 will examine operating performance under the two threshold rules. H7 will examine net-benefit differences across scenarios and decision thresholds.

H8 will be assessed separately using the deliberately invalid predictors described below. Conclusions will reflect effect sizes, uncertainty, and consistency across held-out cohorts rather than a single favourable metric.

### **Leakage positive-control analysis**

The separate audit file contains an exact copy of the future outcome and a simulated post-outcome treatment variable. Treatment probability was 0.80 among outcome-positive participants and 0.08 among outcome-negative participants.

These variables will remain excluded from all valid pipelines.

A secondary positive-control experiment will repeat the additive logistic analysis after introducing each invalid variable separately. The outcome-copy condition may produce separation or near-perfect prediction; these behaviours will be reported explicitly.

This experiment demonstrates intentional temporal or target leakage. It will not be interpreted as evidence of useful prediction or as a quantitative estimate of leakage in published clinical studies.

### **Reproducibility and software**

Data generation was completed in Python using NumPy. The distributed materials contain scenario datasets, a combined dataset, a data dictionary, cohort summaries, the isolated audit file, the generator, and checksums.

The proposed analyses will use Python, pandas, NumPy, SciPy, and scikit-learn. Actual package versions, random seeds, fitted specifications, warnings, and evaluation outputs will be

recorded when the analyses are executed. No unverified software-version numbers or completed analysis outputs are claimed here.

The generating code and analysis code will be retained separately. Any deviation from this plan will be documented with its reason and timing relative to model evaluation.

### **Ethical considerations and AI assistance**

The study uses entirely simulated records and does not involve human participants, identifiable clinical data, human tissue, or animals. Ethics approval and consent to participate are therefore not applicable to the described computational work. No clinical trial or intervention was conducted.

OpenAI ChatGPT/Codex assisted with drafting the dataset-generation code and manuscript text. The generated data are governed by explicit programmed statistical mechanisms rather than model-generated patient narratives. The human author remains responsible for reviewing the code, verifying analyses and references, and approving the final manuscript.

## Results

The benchmark comprised 18,000 entirely simulated participants distributed across three independently generated scenarios. The analyses compared six prediction approaches under a common protocol that confined preprocessing, model tuning, calibration, and threshold selection to the development data. The results below concern six-month simulated depression status, which includes both new and persistent cases. They do not represent observed diagnoses, clinical external validation, or evidence of benefit to patients.

### Simulation Sample and Data Completeness

Each scenario contained 6,000 distinct participants in six cohorts of 1,000. Cohorts C1 to C4 supplied 4,000 development records, and cohorts C5 and C6 supplied 2,000 held-out records. There were no repeated participants, missing outcomes, or simulated losses to follow-up. All participants were retained, and missing predictor values were imputed within the relevant training partition. No held-out cohort was used to tune, calibrate, or select an operating threshold.

Outcome prevalence increased across the held-out cohorts, particularly in the heterogeneous scenario. Prevalence was 33.1% and 37.5% in the linear test cohorts, 38.6% and 40.2% in the nonlinear test cohorts, and 47.6% and 59.6% in the heterogeneous test cohorts. These shifts were consequences of the specified generating mechanisms and should not be interpreted as epidemiological estimates. Across all scenarios, 8,148 of 162,000 cells in the nine variables susceptible to missingness were missing, corresponding to 5.0%.

**Table 1***Characteristics of the Development and Held Out Simulation Samples*

| Scenario and sample        | n     | Events, n (%) | Age, M (SD) | Symptoms, M (SD) | Missing (%) |
|----------------------------|-------|---------------|-------------|------------------|-------------|
| Linear: Development        | 4,000 | 1,094 (27.4)  | 44.3 (14.7) | 6.1 (3.8)        | 4.1         |
| Linear: C5                 | 1,000 | 331 (33.1)    | 49.3 (14.1) | 6.9 (3.9)        | 6.4         |
| Linear: C6                 | 1,000 | 375 (37.5)    | 52.0 (14.4) | 6.9 (3.8)        | 7.6         |
| Nonlinear: Development     | 4,000 | 1,278 (31.9)  | 44.9 (14.7) | 6.2 (3.8)        | 4.4         |
| Nonlinear: C5              | 1,000 | 386 (38.6)    | 49.4 (15.1) | 6.7 (3.7)        | 6.1         |
| Nonlinear: C6              | 1,000 | 402 (40.2)    | 51.5 (14.3) | 7.0 (4.0)        | 7.1         |
| Heterogeneous: Development | 4,000 | 1,299 (32.5)  | 44.8 (14.7) | 6.2 (3.8)        | 4.1         |
| Heterogeneous: C5          | 1,000 | 476 (47.6)    | 49.3 (14.9) | 6.9 (4.0)        | 6.4         |
| Heterogeneous: C6          | 1,000 | 596 (59.6)    | 51.7 (14.4) | 7.2 (3.8)        | 6.4         |

*Note.* Development combines C1 to C4; C5 and C6 are separate held-out simulated cohorts. Symptoms are the simulated baseline score on a 0 to 27 scale. Missingness is the percentage of cells across the nine predictors eligible for missingness, not the percentage of participants. M = mean; SD = standard deviation.

### **Model Fitting and Development Evaluation**

All 18 scenario-by-model combinations produced complete predictions for the four outer development folds and both held-out cohorts. Each outer fold held out one development cohort. Inner tuning used five-fold splits grouped by nominal site. The final pipeline for each model was tuned on C1 to C4, calibrated using training-only cross-fitted scores, and then applied without updating to C5 and C6. The valid pipelines used the same 17 baseline predictors and nine prespecified missingness indicators. They excluded participant identifiers, cohort and site labels, split labels, and the audit-only leakage variables.

The two penalized regression approaches each evaluated six penalty values. Random forest and gradient boosting each evaluated 24 seeded configurations, and the RBF SVM evaluated 16 configurations. Random forests used 500 trees. For SVM inner log-loss selection, native Platt probabilities were fitted using the software's internal training-only cross-validation; the common primary sigmoid calibration used site-grouped cross-fitted decision scores. Thus,

the internal SVM probability-fitting step was not site-grouped, although it did not use held-out outcomes.

In the linear scenario, the mean outer-fold development AUROC ranged from .760 to .773. Flexible logistic regression achieved .769, compared with .773 for additive logistic regression. These estimates assess discrimination across the development cohorts; differences between development and test performance also reflect differences in cohort composition and outcome mechanisms.

In the nonlinear scenario, the mean outer-fold development AUROC ranged from .822 to .831. Flexible logistic regression achieved .831, compared with .825 for additive logistic regression. These estimates assess discrimination across the development cohorts; differences between development and test performance also reflect differences in cohort composition and outcome mechanisms.

In the heterogeneous scenario, the mean outer-fold development AUROC ranged from .811 to .821. Flexible logistic regression achieved .821, compared with .811 for additive logistic regression. These estimates assess discrimination across the development cohorts; differences between development and test performance also reflect differences in cohort composition and outcome mechanisms.

### **Held Out Discrimination and Overall Prediction Error**

The primary discrimination summary was the unweighted mean of the two held-out cohort AUROCs. Table 2 reports this summary with 95% bootstrap confidence intervals, together with unweighted mean average precision and Brier scores. Average precision is reported as the operational estimate of area under the precision-recall curve. Its interpretation depends on

outcome prevalence, and cross-scenario differences therefore cannot be attributed solely to model quality.

**Table 2***Mean Held Out Discrimination and Brier Scores*

| Scenario and model               | AUROC [95% CI]    | AUPRC | Brier score |
|----------------------------------|-------------------|-------|-------------|
| Linear: Additive LR              | .773 [.751, .794] | .674  | .178        |
| Linear: Ridge LR                 | .773 [.751, .794] | .674  | .178        |
| Linear: Flexible LR              | .772 [.750, .793] | .669  | .179        |
| Linear: Random forest            | .765 [.742, .785] | .666  | .181        |
| Linear: Gradient boosting        | .768 [.746, .789] | .663  | .181        |
| Linear: RBF SVM                  | .771 [.749, .792] | .674  | .179        |
| Nonlinear: Additive LR           | .826 [.807, .844] | .764  | .163        |
| Nonlinear: Ridge LR              | .826 [.807, .844] | .765  | .163        |
| Nonlinear: Flexible LR           | .836 [.818, .854] | .781  | .158        |
| Nonlinear: Random forest         | .826 [.807, .844] | .767  | .163        |
| Nonlinear: Gradient boosting     | .827 [.808, .845] | .765  | .162        |
| Nonlinear: RBF SVM               | .826 [.807, .844] | .766  | .163        |
| Heterogeneous: Additive LR       | .832 [.814, .848] | .858  | .187        |
| Heterogeneous: Ridge LR          | .832 [.814, .848] | .858  | .186        |
| Heterogeneous: Flexible LR       | .832 [.815, .849] | .857  | .186        |
| Heterogeneous: Random forest     | .831 [.813, .847] | .857  | .190        |
| Heterogeneous: Gradient boosting | .831 [.814, .848] | .857  | .190        |
| Heterogeneous: RBF SVM           | .832 [.815, .849] | .858  | .185        |

*Note.* All summaries are unweighted means across C5 and C6, rather than pooled cohort estimates. CIs use 2,000 paired participant bootstrap samples drawn independently within each test cohort and condition on the fitted pipelines. Higher AUROC and AUPRC indicate better discrimination; lower Brier scores indicate lower squared probability error. LR = logistic regression; RBF = radial basis function; SVM = support vector machine; CI = confidence interval.

In the linear scenario, ridge logistic regression had the highest observed mean test AUROC (.773). Additive and flexible logistic regression achieved .773 and .772, respectively. The highest-ranking ML model was RBF SVM, with AUROC .771 and Brier score .179. These rankings are descriptive and do not establish equivalence or clinically meaningful superiority.

In the nonlinear scenario, flexible logistic regression had the highest observed mean test AUROC (.836). Additive and flexible logistic regression achieved .826 and .836, respectively. The highest-ranking ML model was gradient boosting, with AUROC .827 and Brier score .162. These rankings are descriptive and do not establish equivalence or clinically meaningful superiority.

In the heterogeneous scenario, flexible logistic regression had the highest observed mean test AUROC (.832). Additive and flexible logistic regression achieved .832 and .832, respectively. The highest-ranking ML model was RBF SVM, with AUROC .832 and Brier score .185. These rankings are descriptive and do not establish equivalence or clinically meaningful superiority.

### **Incremental Performance Beyond Flexible Logistic Regression**

Table 3 presents paired comparisons between each ML model and flexible logistic regression. Positive AUROC differences favour the ML model, whereas negative Brier differences favour the ML model. Pairing retained the same participants and bootstrap draws for both models, avoiding the loss of information that would arise from treating their results as independent.

**Table 3***Paired Incremental Performance of ML Models Relative to Flexible Logistic Regression*

| Scenario and ML model            | Delta AUROC [95% CI]    | Delta Brier [95% CI]   |
|----------------------------------|-------------------------|------------------------|
| Linear: Random forest            | -.0077 [-.0139, -.0015] | .0025 [.0004, .0048]   |
| Linear: Gradient boosting        | -.0042 [-.0110, .0023]  | .0021 [-.0002, .0045]  |
| Linear: RBF SVM                  | -.0014 [-.0060, .0035]  | .0001 [-.0016, .0018]  |
| Nonlinear: Random forest         | -.0105 [-.0169, -.0038] | .0052 [.0021, .0083]   |
| Nonlinear: Gradient boosting     | -.0098 [-.0158, -.0037] | .0043 [.0014, .0073]   |
| Nonlinear: RBF SVM               | -.0100 [-.0154, -.0048] | .0049 [.0024, .0074]   |
| Heterogeneous: Random forest     | -.0018 [-.0084, .0048]  | .0033 [.0002, .0064]   |
| Heterogeneous: Gradient boosting | -.0008 [-.0065, .0051]  | .0039 [.0010, .0068]   |
| Heterogeneous: RBF SVM           | -.0003 [-.0064, .0055]  | -.0009 [-.0037, .0022] |

*Note.* Differences are model minus flexible logistic regression, using the unweighted mean of C5 and C6. Intervals condition on fitted pipelines and do not include model-refitting or simulation-seed uncertainty. Intervals are unadjusted for multiple comparisons. Delta denotes an absolute difference, not a relative percentage improvement. No clinical superiority or equivalence margin was specified.

For the ML model with the highest observed AUROC in the linear scenario (RBF SVM), the AUROC difference from additive logistic regression was -.0022 [-.0057, .0014]. The corresponding difference from flexible logistic regression was -.0014 [-.0060, .0035]. Selection of the highest-ranking model here is a descriptive summary of test results and does not define a prospectively selected winning model.

For the ML model with the highest observed AUROC in the nonlinear scenario (gradient boosting), the AUROC difference from additive logistic regression was .0005 [-.0070, .0083]. The corresponding difference from flexible logistic regression was -.0098 [-.0158, -.0037]. Selection of the highest-ranking model here is a descriptive summary of test results and does not define a prospectively selected winning model.

For the ML model with the highest observed AUROC in the heterogeneous scenario (RBF SVM), the AUROC difference from additive logistic regression was .0004 [-.0022, .0028]. The corresponding difference from flexible logistic regression was -.0003 [-.0064, .0055]. Selection of the highest-ranking model here is a descriptive summary of test results and does not define a prospectively selected winning model.

Intervals that include zero remain compatible with improvement, no difference, or deterioration within their limits. They are not evidence of equivalence. Conversely, an interval excluding zero would indicate a conditional performance difference in this simulated sample; it would not establish clinical importance or transportability to real patients.

### **Calibration Across Held Out Cohorts**

Calibration was evaluated separately within C5 and C6 to avoid obscuring cohort-specific shifts. Calibration in the large was estimated with the predicted logit as an offset, while the slope was estimated jointly with a freely estimated intercept. An ideal calibration-in-the-large intercept is zero and an ideal slope is one. Positive calibration-in-the-large values indicate average underprediction of risk. These estimates were used only to evaluate the fixed models, not to recalibrate them on test outcomes.

**Table 4***Calibration in Each Held Out Simulation Cohort*

| Scenario and model               | C5 intercept | C5 slope | C6 intercept | C6 slope |
|----------------------------------|--------------|----------|--------------|----------|
| Linear: Additive LR              | .068         | .985     | .202         | 1.074    |
| Linear: Ridge LR                 | .067         | .981     | .201         | 1.071    |
| Linear: Flexible LR              | .071         | .988     | .207         | 1.102    |
| Linear: Random forest            | .087         | .956     | .225         | 1.122    |
| Linear: Gradient boosting        | .087         | 1.025    | .247         | 1.154    |
| Linear: RBF SVM                  | .081         | .981     | .236         | 1.109    |
| Nonlinear: Additive LR           | .063         | 1.031    | .001         | 1.045    |
| Nonlinear: Ridge LR              | .062         | 1.028    | .000         | 1.042    |
| Nonlinear: Flexible LR           | .050         | 1.046    | -.006        | .996     |
| Nonlinear: Random forest         | .084         | 1.040    | .006         | 1.006    |
| Nonlinear: Gradient boosting     | .080         | 1.015    | .022         | .969     |
| Nonlinear: RBF SVM               | .058         | 1.009    | -.017        | 1.010    |
| Heterogeneous: Additive LR       | .521         | 1.165    | .969         | 1.167    |
| Heterogeneous: Ridge LR          | .521         | 1.161    | .969         | 1.163    |
| Heterogeneous: Flexible LR       | .529         | 1.081    | 1.006        | 1.083    |
| Heterogeneous: Random forest     | .562         | 1.155    | 1.050        | 1.083    |
| Heterogeneous: Gradient boosting | .564         | 1.163    | 1.053        | 1.115    |
| Heterogeneous: RBF SVM           | .504         | 1.134    | .954         | 1.173    |

*Note.* Intercept means calibration in the large, fitted with a logit offset. Slope is estimated jointly with a separate intercept. Values are point estimates. A slope below one can indicate overly extreme predictions and a slope above one insufficient spread, conditional on the fitted calibration model. Nominal sites were not treated as independent clinical populations.

Flexible logistic regression had calibration-in-the-large intercepts of .071 in C5 and .207 in C6 for the linear scenario. Its corresponding slopes were .988 and 1.102.

Flexible logistic regression had calibration-in-the-large intercepts of .050 in C5 and -.006 in C6 for the nonlinear scenario. Its corresponding slopes were 1.046 and .996.

Flexible logistic regression had calibration-in-the-large intercepts of .529 in C5 and 1.006 in C6 for the heterogeneous scenario. Its corresponding slopes were 1.081 and 1.083.

The heterogeneous scenario illustrates why discrimination and calibration must be assessed together. For flexible logistic regression, mean predicted risk in C6 was .425, compared

with an observed simulated event proportion of .596. The positive calibration intercept reflects this underprediction, despite the model retaining substantial discrimination. Brier scores combine discrimination, calibration, and outcome uncertainty, so higher Brier scores across scenarios cannot be interpreted as a pure measure of calibration deterioration.

### **Performance at Development Selected Operating Thresholds**

Each model's operating threshold maximized development specificity subject to sensitivity of at least .80, using the combined outer-fold development predictions. These thresholds were then held fixed for both test cohorts. Table 5 presents pooled held-out classification performance because the resulting counts provide a direct description of the screening burden across the 2,000 test records in each scenario. The development sensitivity constraint does not guarantee test sensitivity of .80.

**Table 5***Held Out Classification Performance at Development Selected Thresholds*

| Scenario and model               | Threshold | Sensitivity | Specificity | PPV  | NPV  |
|----------------------------------|-----------|-------------|-------------|------|------|
| Linear: Additive LR              | .214      | .850        | .492        | .477 | .857 |
| Linear: Ridge LR                 | .211      | .853        | .488        | .476 | .859 |
| Linear: Flexible LR              | .215      | .853        | .498        | .481 | .861 |
| Linear: Random forest            | .202      | .857        | .463        | .465 | .856 |
| Linear: Gradient boosting        | .199      | .863        | .463        | .467 | .861 |
| Linear: RBF SVM                  | .205      | .840        | .502        | .479 | .852 |
| Nonlinear: Additive LR           | .273      | .849        | .606        | .584 | .861 |
| Nonlinear: Ridge LR              | .273      | .848        | .610        | .585 | .860 |
| Nonlinear: Flexible LR           | .275      | .848        | .635        | .602 | .865 |
| Nonlinear: Random forest         | .262      | .862        | .600        | .583 | .870 |
| Nonlinear: Gradient boosting     | .265      | .844        | .635        | .601 | .862 |
| Nonlinear: RBF SVM               | .259      | .853        | .596        | .578 | .862 |
| Heterogeneous: Additive LR       | .262      | .832        | .621        | .717 | .762 |
| Heterogeneous: Ridge LR          | .262      | .832        | .621        | .717 | .762 |
| Heterogeneous: Flexible LR       | .246      | .833        | .639        | .727 | .768 |
| Heterogeneous: Random forest     | .256      | .829        | .642        | .728 | .765 |
| Heterogeneous: Gradient boosting | .248      | .820        | .650        | .730 | .758 |
| Heterogeneous: RBF SVM           | .241      | .851        | .599        | .710 | .777 |

*Note.* Metrics pool C5 and C6 within each scenario, with  $n = 2,000$ . Thresholds were selected exclusively from development predictions. PPV = positive predictive value; NPV = negative predictive value. Values are point estimates and depend on simulated prevalence and case mix. Threshold .50 results and full confusion matrices are provided in the analysis files.

For flexible logistic regression in the linear scenario, threshold .215 yielded sensitivity .853 and specificity .498. The model classified 1,252 of 2,000 held-out participants as positive, comprising 602 true positives and 650 false positives, while 104 simulated cases were missed. PPV was .481 and NPV was .861.

For flexible logistic regression in the nonlinear scenario, threshold .275 yielded sensitivity .848 and specificity .635. The model classified 1,110 of 2,000 held-out participants as

positive, comprising 668 true positives and 442 false positives, while 120 simulated cases were missed. PPV was .602 and NPV was .865.

For flexible logistic regression in the heterogeneous scenario, threshold .246 yielded sensitivity .833 and specificity .639. The model classified 1,228 of 2,000 held-out participants as positive, comprising 893 true positives and 335 false positives, while 179 simulated cases were missed. PPV was .727 and NPV was .768.

These operating characteristics demonstrate that ranking accuracy does not uniquely determine the number of participants flagged, the number of missed cases, or the proportion of alerts that are true positives. They also depend on calibration and the selected cutoff. The outcome prevalence and the clinical consequences of false positives and false negatives would need to be established in real settings before using these cutoffs.

### **Decision Curve Net Benefit**

Decision-curve analysis compared the fixed models with treat-all and treat-none strategies across probability thresholds from .05 to .50. Table 6 reports three representative thresholds from that prespecified grid. These thresholds are mathematical evaluation settings; they do not correspond to a validated depression intervention policy. Net benefit is expressed per participant and accounts for false positives through the threshold odds.

**Table 6***Net Benefit at Selected Probability Thresholds*

| Scenario and strategy            | Threshold .10 | Threshold .20 | Threshold .30 |
|----------------------------------|---------------|---------------|---------------|
| Linear: Additive LR              | .283          | .218          | .169          |
| Linear: Ridge LR                 | .283          | .218          | .170          |
| Linear: Flexible LR              | .285          | .216          | .169          |
| Linear: Random forest            | .285          | .216          | .163          |
| Linear: Gradient boosting        | .284          | .218          | .159          |
| Linear: RBF SVM                  | .284          | .216          | .164          |
| Linear: Treat all                | .281          | .191          | .076          |
| Linear: Treat none               | .000          | .000          | .000          |
| Nonlinear: Additive LR           | .333          | .278          | .233          |
| Nonlinear: Ridge LR              | .333          | .278          | .234          |
| Nonlinear: Flexible LR           | .332          | .281          | .238          |
| Nonlinear: Random forest         | .329          | .280          | .234          |
| Nonlinear: Gradient boosting     | .330          | .277          | .234          |
| Nonlinear: RBF SVM               | .333          | .279          | .233          |
| Nonlinear: Treat all             | .327          | .242          | .134          |
| Nonlinear: Treat none            | .000          | .000          | .000          |
| Heterogeneous: Additive LR       | .485          | .422          | .362          |
| Heterogeneous: Ridge LR          | .485          | .422          | .361          |
| Heterogeneous: Flexible LR       | .484          | .421          | .360          |
| Heterogeneous: Random forest     | .483          | .421          | .360          |
| Heterogeneous: Gradient boosting | .485          | .419          | .348          |
| Heterogeneous: RBF SVM           | .485          | .422          | .362          |
| Heterogeneous: Treat all         | .484          | .420          | .337          |
| Heterogeneous: Treat none        | .000          | .000          | .000          |

*Note.* Net benefit pools C5 and C6. Treat-none net benefit is zero. Treat-all net benefit depends on prevalence and threshold odds. The full .05 to .50 grid is supplied in the analysis files. Point differences have no clinical superiority interpretation without justified intervention preferences and a prespecified benefit margin.

At threshold .20 in the linear scenario, the largest observed ML net benefit was .218 for gradient boosting, compared with .216 for flexible logistic regression. The difference was .002, equivalent to 0.16 net true-positive equivalents per 100 simulated participants under the specified threshold weighting. This is a decision-analytic quantity, not an observed number of patients benefiting from treatment.

At threshold .20 in the nonlinear scenario, the largest observed ML net benefit was .280 for random forest, compared with .281 for flexible logistic regression. The difference was -.001,

equivalent to -0.08 net true-positive equivalents per 100 simulated participants under the specified threshold weighting. This is a decision-analytic quantity, not an observed number of patients benefiting from treatment.

At threshold .20 in the heterogeneous scenario, the largest observed ML net benefit was .422 for RBF SVM, compared with .421 for flexible logistic regression. The difference was .001, equivalent to 0.06 net true-positive equivalents per 100 simulated participants under the specified threshold weighting. This is a decision-analytic quantity, not an observed number of patients benefiting from treatment.

### **Intentionally Invalid Leakage Controls**

Two isolated positive controls examined performance after adding information that would be unavailable at baseline: a direct copy of the six-month outcome or treatment generated after the outcome. Neither variable entered the valid benchmark. Weak ridge regularization was used in these additive logistic controls to allow finite estimation when the outcome-copy predictor produced separation. Their results are deliberately invalid as prospective depression predictions.

**Table 7***Apparent Performance After Introducing Deliberately Leaked Predictors*

| Scenario      | Invalid added predictor | Pooled AUROC | Pooled Brier |
|---------------|-------------------------|--------------|--------------|
| Linear        | Outcome copy            | 1.000        | .0000        |
| Linear        | Post-outcome treatment  | .931         | .0923        |
| Nonlinear     | Outcome copy            | 1.000        | .0000        |
| Nonlinear     | Post-outcome treatment  | .938         | .0899        |
| Heterogeneous | Outcome copy            | 1.000        | .0000        |
| Heterogeneous | Post-outcome treatment  | .942         | .1010        |

*Note.* These are separate diagnostic controls using weakly penalized additive logistic regression, without the common primary calibrator. Results pool C5 and C6. They are not valid competing prediction models. The outcome-copy variable directly reveals the target; post-outcome treatment is downstream information.

The outcome-copy controls approached perfect discrimination and negligible prediction error. This behaviour is expected because the added predictor directly reveals the evaluation outcome. The post-outcome treatment controls also exploited downstream information. These controls illustrate performance inflation from two specific leakage mechanisms; they do not estimate the inflation attributable to global preprocessing, feature-selection leakage, or improper hyperparameter tuning, which were not deliberately introduced in this execution.

**Descriptive Assessment of the Hypotheses**

The hypotheses were assessed descriptively because this execution used one generated dataset per scenario, without independent simulation replications or a prespecified clinical improvement margin. The bootstrap intervals quantify participant-level evaluation uncertainty conditional on the fitted pipelines; they are not Monte Carlo intervals for the generating mechanisms.

The largest observed ML AUROC difference from additive logistic regression in the linear scenario was -.0022 (RBF SVM).

The largest observed ML AUROC difference from additive logistic regression in the nonlinear scenario was .0005 (gradient boosting).

For H1, these scenario-specific differences describe whether the observed ML advantage was larger under nonlinearity, but no paired uncertainty interval for the difference between independently generated scenarios was calculated. For H2, no ML model had a positive AUROC difference with a 95% conditional interval excluding zero in the nonlinear scenario. This assessment concerns statistical discrimination within the simulated sample and does not establish clinical importance.

For H3, the improvement of flexible regression over additive regression in the nonlinear scenario reduced the ML advantages when the stronger comparator was used. For H4, the cohort-specific calibration intercepts, particularly in heterogeneous C6, documented underprediction under the specified distribution shifts. This supports the need to distinguish calibration deterioration from changes in discrimination or prevalence-dependent Brier scores.

Observed AUROC and Brier model rankings differed in the nonlinear and heterogeneous scenarios. In the heterogeneous scenario, flexible logistic regression had the highest observed mean AUROC, whereas RBF SVM had the lowest mean Brier score. The paired SVM-versus-flexible intervals included zero for both metrics, so this ranking difference should not be treated as evidence of a definite performance difference.

H5 was evaluated through the separate AUROC, Brier, and calibration results, rather than by treating one ranking as a complete assessment of model quality. H6 was addressed by the variation in development-selected cutoffs and held-out sensitivity, specificity, PPV, and NPV. H7 was assessed descriptively through the decision curves and threshold-specific net benefit values; no claim of clinically meaningful benefit was made. H8 was illustrated by the intentionally invalid outcome-copy and post-outcome controls. The controls support the

existence of leakage-driven inflation under those mechanisms, but not a numerical claim about all forms of leakage.

### **Scope of Statistical Inference**

These findings concern the specified generating equations, predictor set, tuning spaces, six nominal cohorts, and single generation seed. The generating mechanisms were designed to differ in nonlinearity and cohort shifts, so the observed scenario differences are not independent empirical discoveries about depression. No results establish that ML is universally superior, inferior, or equivalent to regression in clinical practice. A clinical manuscript would require genuinely observed cohorts, justified endpoints, and externally validated utility. A full simulation manuscript would additionally require replicated generation seeds, Monte Carlo uncertainty, and sensitivity analyses over sample sizes, event rates, missingness, and shift severity.

## Discussion

### Principal Findings

This study examined whether machine learning algorithms improved prediction of synthetic six-month depression status beyond conventional and flexible logistic regression models when data leakage was controlled throughout model development and evaluation. Across three simulated scenarios, increasing algorithmic complexity did not produce a consistent advantage in discrimination. The flexible logistic regression model achieved the highest mean area under the receiver operating characteristic curve (AUROC) in the nonlinear scenario, while differences between approaches were comparatively small in the linear and heterogeneous scenarios. These findings suggest that, under the conditions represented by this benchmark, the specification of predictor relationships and the quality of validation were more consequential than the use of a particular algorithm family.

The results support a qualified interpretation of the study's central hypothesis. Machine learning did not consistently outperform regression when all models were evaluated using the same baseline information and a leakage-controlled development process. However, the study does not establish statistical equivalence or demonstrate that machine learning cannot improve depression prediction in other settings. Such conclusions would require prespecified equivalence margins, adequate precision, repeated simulations, and evaluation using clinical datasets. The present findings instead show that greater computational complexity did not reliably translate into better discrimination within the simulated conditions examined.

An additional finding concerned the effects of information leakage. Introducing outcome-derived or post-baseline information substantially increased apparent predictive performance. This improvement reflected access to information unavailable at the intended prediction time,

rather than a corresponding increase in the model's ability to predict future depression status from baseline characteristics. The contrast between valid and contaminated analyses illustrates why predictive accuracy must be interpreted in relation to the timing, provenance, and availability of model inputs.

### **Comparison With Previous Research**

The absence of a consistent machine learning advantage agrees with the systematic review by Christodoulou et al. (2019), which found no overall performance benefit of machine learning over logistic regression in clinical prediction studies. That review also identified methodological weaknesses that complicated comparisons between algorithms. The present benchmark contributes to this discussion by comparing models within a common development and evaluation framework, reducing opportunities for differences in preprocessing, tuning, or test-set access to create misleading advantages.

Evidence specific to depression prediction provides a similar context. Nickson et al. (2023), in their systematic review of machine learning studies using electronic health records, found that regression approaches could perform as effectively as more complex methods. Their review also highlighted concerns about generalizability and interpretability. The current results are consistent with that observation, although the use of synthetic data means that they should be understood as methodological evidence rather than an independent clinical replication.

Studies predicting depression treatment outcomes have also demonstrated the importance of study quality and independent validation. Sajjadian et al. (2021) found lower prediction accuracy among studies meeting their quality criteria than among the remaining studies. This pattern suggests that impressive performance estimates may weaken when evaluation becomes more rigorous. Treatment response and future depression status are distinct prediction targets,

but both require a clear separation between information available at prediction time and information obtained after the outcome process has begun.

Chekroud et al. (2016) demonstrated the feasibility of cross-trial prediction of depression treatment outcomes. Their work illustrates the potential value of evaluating psychiatric prediction models beyond the data used for development. The present study adopted an analogous emphasis on generalization by reserving two synthetic cohorts for evaluation. Nevertheless, these held-out cohorts were generated within the same overarching simulation framework. They therefore provide evidence about transfer across simulated cohort differences, rather than validation across independently recruited clinical populations.

Taken together, the literature and the present results support evaluating machine learning as an empirical option whose usefulness depends on the prediction problem. Neither the psychiatric relevance of an outcome nor the availability of complex algorithms establishes that those algorithms will outperform a well-specified regression model.

### **Functional Flexibility and the Choice of Regression Comparator**

The nonlinear scenario offered the clearest distinction between model specifications. Flexible logistic regression achieved a mean AUROC of approximately .836, compared with approximately .826 for additive logistic regression and .826 to .827 for the evaluated machine learning models. This pattern indicates that allowing nonlinear predictor effects and selected interactions improved discrimination within this scenario.

The finding is important because logistic regression is often treated as synonymous with a model containing only linear, additive predictor effects. Logistic regression can incorporate spline functions, interactions, and regularization while retaining a relatively transparent

mathematical structure. A comparison restricted to a basic additive model may therefore underestimate the performance achievable with conventional statistical methods.

The advantage of flexible regression in this benchmark should nevertheless be interpreted in relation to the data-generating mechanisms. Synthetic outcome relationships necessarily reflected choices made when constructing the simulation. Those choices may have favored representations that were efficiently captured by the spline and interaction specification. The results consequently demonstrate the value of an appropriately flexible comparator under the simulated relationships, rather than a general superiority of flexible regression across depression prediction tasks.

This interpretation is consistent with Bzdok et al. (2018), who discussed differences between statistical and machine learning approaches in terms of their objectives and analytical properties. Algorithm labels alone provide an incomplete account of model capability. The relevant comparison concerns how effectively each fitted model represents the available signal, controls overfitting, and produces reliable predictions in new observations.

For future benchmarking studies, the practical implication is to include both a conventional regression baseline and a stronger regression comparator. This approach can distinguish improvements attributable to accommodating nonlinear relationships from improvements that specifically require a more complex algorithm.

### **Leakage Prevention and the Interpretation of Apparent Accuracy**

The leakage analyses illustrated how strongly apparent predictive performance can depend on access to inappropriate information. Outcome-derived leakage produced perfect discrimination in the relevant contaminated analyses, while inclusion of post-treatment

information raised AUROC values to approximately .931 to .942. These estimates cannot be interpreted as valid measures of baseline predictive performance.

An outcome-derived variable provides direct or indirect information about the target being predicted. Similarly, a measurement obtained after treatment begins may reflect clinical decisions, early symptom changes, or other developments occurring after the intended prediction time. Such information may be useful for a separately defined dynamic prediction task, but its inclusion in a baseline model changes the question being answered.

Kapoor and Narayanan (2023) identified leakage as a substantial threat to reproducibility in machine learning research. The current findings provide a controlled example of this problem: the apparent improvement was large enough to overshadow the comparatively modest differences between valid models. Consequently, a highly accurate model may warrant closer examination of the information pathway before being interpreted as evidence of a clinically useful algorithm.

The benchmark addressed several common opportunities for leakage by fitting preprocessing operations, feature transformations, and model-selection procedures within training data. Nested validation separated hyperparameter selection from outer evaluation. This separation is supported by Varma and Simon (2006) and Cawley and Talbot (2010), who showed how model selection can bias performance estimates when evaluation data influence the selection process.

These controls strengthen the interpretation of the comparisons, but they do not justify an unrestricted claim that every possible source of optimism was eliminated. The support vector classifier used a training-only internal probability-estimation procedure that did not preserve cohort grouping within that internal step. Although the held-out evaluation cohorts remained

excluded, this difference in the development procedures should be considered when comparing probability-based tuning and calibration across models.

### **Cohort Heterogeneity and Generalization**

In the heterogeneous scenario, mean AUROC values were approximately .830 to .832 across the evaluated approaches. The small separation between models suggests that the simulated cohort differences did not create a clear discrimination advantage for any single algorithm.

However, similar mean AUROC values do not imply identical behavior across cohorts or individuals. AUROC describes ranking performance and does not establish that predicted probabilities accurately reflect outcome frequencies. The observed underprediction in the heterogeneous setting therefore remains relevant even when discrimination appears comparatively stable.

These findings emphasize the need to evaluate models across settings using more than an overall summary statistic. A model may rank individuals reasonably well while systematically assigning probabilities that are too low or too high in a particular cohort. Such differences can affect decisions based on absolute risk thresholds.

The held-out cohort design strengthened the benchmark by testing models on observations outside the development cohorts. Nevertheless, the simulated cohorts shared a common underlying construction. The evaluation does not capture the full range of differences that can arise between healthcare systems, including changes in diagnostic practice, treatment access, documentation, measurement instruments, and population composition. Independent clinical validation remains necessary before drawing conclusions about transportability to patient care.

### **Calibration and Clinical Utility**

The study's emphasis on calibration and decision-relevant measures reflects the distinction between predicting relative risk and supporting clinical decisions. Van Calster et al. (2019) described calibration as a central vulnerability of predictive analytics because useful risk estimates require agreement between predicted probabilities and observed outcomes.

This distinction matters in depression prediction. If a model systematically underestimates risk, individuals may fall below a referral or monitoring threshold despite having a higher probability of depression than the model indicates. Conversely, overestimation can increase unnecessary assessment and place additional demands on services. Discrimination alone cannot determine whether these consequences are acceptable.

The inclusion of sensitivity, specificity, predictive values, Brier scores, and decision curve analysis therefore broadened the evaluation beyond AUROC. Steyerberg et al. (2010) recommended assessing prediction models through complementary measures, while Vickers and Elkin (2006) developed decision curve analysis to examine the potential value of model-guided decisions across risk thresholds.

Nevertheless, the clinical interpretation of these analyses remains constrained by the synthetic setting. The decision thresholds represented hypothetical choices rather than thresholds established through patient preferences, clinical guidelines, or service capacity. Likewise, the sensitivity target used for threshold selection was an analytical choice rather than a validated standard for depression screening.

The results consequently characterize potential decision performance under specified assumptions. They do not demonstrate that implementing any evaluated model would improve patient outcomes, reduce missed diagnoses, or use healthcare resources more effectively.

## **Interpretability and Implementation Implications**

When models achieve similar predictive performance, interpretability and implementation requirements become relevant to model selection. Regression models can make the contribution of specified predictor terms easier to inspect and may require fewer computational resources. These properties can assist model review, maintenance, and communication.

However, interpretability should not be assumed solely from the use of logistic regression. Spline terms, interactions, regularization, and subsequent calibration increase model complexity. Their effects may need to be communicated through predicted-risk plots and carefully specified examples rather than individual coefficients alone.

The present findings support selecting model complexity according to demonstrated benefit. Where discrimination, calibration, and decision performance are similar, a simpler model may offer practical advantages. Where a more complex model produces a reproducible and meaningful improvement, additional implementation demands may be justified.

Dwyer et al. (2018) and Iniesta et al. (2016) described the potential of machine learning for psychiatric research, particularly where data contain complex patterns. The current benchmark does not negate that potential. It indicates that the benefit should be established through rigorous comparison with credible alternatives.

## **Strengths and Limitations**

A principal strength was the controlled comparison of multiple models across linear, nonlinear, and heterogeneous scenarios. All approaches used a common set of baseline predictors, and the development framework separated training operations from held-out evaluation. Including flexible regression improved the credibility of the statistical comparator.

The leakage demonstrations also made the consequences of invalid information use directly observable.

The main limitation was the use of entirely synthetic data. Simulated records permit control over predictor relationships and cohort differences, but they cannot reproduce all aspects of clinical depression. The outcome represented computationally generated six-month depression status and included both new and persistent cases. It was therefore not a validated diagnosis of incident depression or a prediction of treatment response.

A second limitation was the use of a single realization of each scenario. Differences between models may depend on the random seed and the particular simulated sample. The bootstrap intervals describe uncertainty conditional on the generated evaluation data and fitted development process. They do not replace repeated simulations that assess variability across newly generated datasets and repeated model fitting. Morris et al. (2019) emphasized the importance of explicit simulation objectives and assessment of Monte Carlo uncertainty when evaluating statistical methods.

Third, the conclusions apply to the evaluated algorithms, tuning spaces, and computational budgets. Broader searches, alternative implementations, or different representations could change model performance. At the same time, expanding a search without preserving independent evaluation could increase selection-related optimism.

Fourth, no prespecified equivalence margin or minimum clinically important improvement was used. Small observed differences therefore cannot establish equivalence, noninferiority, or clinical irrelevance. The strongest supported conclusion is that machine learning did not show a consistent discrimination advantage within this benchmark.

Finally, the study evaluated predictive behavior rather than clinical implementation. It did not assess clinician use, patient acceptability, intervention effects, costs, or health outcomes.

### **Directions for Future Research**

Future work should repeat the simulation across many independently generated datasets and examine a wider range of sample sizes, outcome frequencies, missingness mechanisms, measurement errors, and cohort shifts. Repeated model fitting would help distinguish stable algorithm differences from variation attributable to a particular simulated sample.

Clinical extensions should define the prediction target precisely, distinguishing incident depression, persistent symptoms, relapse, treatment response, and treatment resistance. Each target requires an explicit prediction time and a clear account of which variables are available at that time.

Further studies should also prespecify clinically meaningful improvement margins and compare additive regression, flexible regression, and machine learning using independent clinical cohorts. Calibration, subgroup performance, and decision utility should accompany discrimination measures. Prospective impact evaluations would then be required to determine whether model-guided care improves outcomes.

Overall, this benchmark supports a measured interpretation of machine learning in depression prediction. Under the simulated conditions, complex algorithms did not consistently improve discrimination beyond well-specified regression, while inappropriate access to outcome-related information produced much larger apparent gains. The findings reinforce the importance of credible comparators, temporal clarity, leakage prevention, and evaluation across multiple dimensions of predictive performance.

## Conclusions

In this leakage-controlled multicohort simulation benchmark, machine-learning algorithms did not demonstrate a consistent predictive advantage over conventional regression approaches. No ML model achieved a higher mean held-out AUROC than flexible logistic regression across the three simulated scenarios. In the nonlinear scenario, flexible logistic regression achieved an AUROC of .836, compared with approximately .826–.827 for random forest, gradient boosting and support vector machines. Paired confidence intervals indicated lower discrimination for all three ML models relative to flexible regression in this scenario. These findings demonstrate the importance of including a sufficiently flexible statistical comparator when evaluating whether algorithmic complexity improves prediction.

The results also showed that discrimination alone provides an incomplete assessment of predictive performance. In the heterogeneous scenario, models retained substantial discrimination while systematically underestimating absolute risk in the held-out cohorts. Differences in Brier scores, classification performance and decision-curve net benefit further indicated that model rankings depended on the evaluation metric and operating threshold. The small, inconsistent net-benefit differences did not establish clinically meaningful superiority for ML. Similarly, confidence intervals containing zero should not be interpreted as evidence that competing models are equivalent.

The intentionally invalid leakage controls illustrated how unavailable information can inflate apparent predictive performance. Including a direct outcome copy produced perfect discrimination, while adding post-outcome treatment increased AUROC to approximately .931–.942. These findings reinforce the need to define predictor availability at the intended prediction

time and to confine preprocessing, tuning, calibration and threshold selection to development data.

The conclusions remain limited to the specified generating mechanisms, predictor set, tuning configurations and single simulation seed. They do not establish clinical effectiveness or universal superiority of regression over ML. Future research should replicate the benchmark across independent simulation seeds and evaluate the approaches in genuinely observed cohorts, using justified prediction horizons, intervention thresholds and prespecified criteria for clinically meaningful improvement.

For depression prediction research, the central implication is that greater algorithmic complexity should be justified by reproducible improvements in discrimination, calibration and decision-relevant performance. Within this benchmark, flexible logistic regression provided a strong reference model, and the evaluated ML approaches offered no consistent advantage beyond it.

## **Declarations**

### **Availability of Data and Materials**

The synthetic datasets and supporting reproducibility materials are available through Zenodo: <https://doi.org/10.5281/zenodo.23086371>.

### **Competing Interests**

The author declares no competing interests.

### **Funding**

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

### **Authors' Contributions**

Arhama Faridi conceived and designed the study, specified the simulation scenarios, developed the analytical approach, interpreted the findings and prepared the manuscript. Computational tools assisted with synthetic data generation, analysis implementation and manuscript preparation. Arhama Faridi reviewed and approved the final manuscript and accepts responsibility for its content.

### **Abbreviations**

APA: American Psychological Association; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CI: confidence interval; LR: logistic regression; ML: machine learning; NPV: negative predictive value; PPV: positive predictive value; RBF: radial basis function; SD: standard deviation; SVM: support vector machine.

### References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.  
<https://doi.org/10.1023/A:1010933404324>
- Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4), 233–234. <https://doi.org/10.1038/nmeth.4642>
- Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107. <https://www.jmlr.org/papers/v11/cawley10a.html>
- Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, 3(3), 243–250. [https://doi.org/10.1016/S2215-0366\(15\)00471-X](https://doi.org/10.1016/S2215-0366(15)00471-X)
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110, 12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, Article e078378. <https://doi.org/10.1136/bmj-2023-078378>

- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14, 91–118.  
<https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46(12), 2455–2465.  
<https://doi.org/10.1017/S0033291716001367>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), Article 100804.  
<https://doi.org/10.1016/j.patter.2023.100804>
- Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, 170(1), W1–W33. <https://doi.org/10.7326/M18-1377>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.  
<https://doi.org/10.1002/sim.8086>
- Nickson, D., Meyer, C., Walasek, L., & Toro, C. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: A systematic review. *BMC Medical Informatics and Decision Making*, 23, Article 271.  
<https://doi.org/10.1186/s12911-023-02341-x>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in

- Python. *Journal of Machine Learning Research*, 12, 2825–2830.  
<https://www.jmlr.org/papers/v12/pedregosa11a.html>
- 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), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., Strother, S. C., Farzan, F., Kennedy, S. H., & Uher, R. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine*, 51(16), 2742–2751.  
<https://doi.org/10.1017/S0033291721003871>
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138.  
<https://doi.org/10.1097/EDE.0b013e3181c30fb2>
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11), Article e0224365.  
<https://doi.org/10.1371/journal.pone.0224365>
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group ‘Evaluating diagnostic tests and prediction models’ of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230. <https://doi.org/10.1186/s12916-019-1466-7>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574.

<https://doi.org/10.1177/0272989X06295361>