Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing imbalanced hospital readmission prediction: patient separation, resampling contamination and probability calibration

Mohammad Rabiul Hossen Chowdhury
University of Alabama
Shomi Aktar
saktar@ua.edu
University of Alabama

Research Article

Keywords: hospital readmission, grouped validation, class imbalance, probability calibration, cluster bootstrap, reproducibility

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11199597/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Auditing imbalanced hospital readmission prediction: patient separation, resampling contamination and probability calibration

Mohammad Rabiul Hossen Chowdhury[1] and Shomi Aktar[2*]

[1]Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, Alabama, USA.
[2]Department of Mathematics, The University of Alabama, Tuscaloosa, Alabama, USA.

*Corresponding author(s). E-mail(s): saktar@ua.edu;
Contributing authors: mchowdhury18@crimson.ua.edu;

Abstract

Hospital readmission models require evaluation that respects repeated patients and separates training from validation. We audit a workflow predicting recorded readmission within 30 days and reanalyze 59,129 encounters from 39,726 patients using patient-disjoint fitting, calibration, and evaluation partitions. In the source split, 46.08% of test encounters belonged to patients also represented in training. A controlled 20-partition experiment additionally compares fold-local synthetic oversampling with a contaminated donor pool while holding real validation observations and synthetic training counts fixed. Under patient-level splitting, including validation donors changes mean cross-validation area under the receiver operating characteristic curve (AUC) by +0.005 for logistic regression and +0.012 for XGBoost. In the primary mixed-type analysis, logistic regression and XGBoost achieve evaluation AUCs of 0.653 and 0.657 and average precisions of 0.215 and 0.231. Their paired AUC difference is 0.0041 (95% patient-cluster interval: **−0.0042** to 0.0128). Recall is 0.255 and 0.263 at thresholds chosen on calibration patients under a 10% false-positive-rate constraint. A prior-admissions-only reference reaches AUC 0.604. Lower categorical-aware synthetic oversampling ratios improve on full balancing but do not exceed the corresponding unresampled baselines; retaining laboratory-result categories does not improve discrimination on the same evaluation patients. Exploratory subgroup estimates identify differences in recall, with limited support in small groups. Together, the controlled audits, calibration assessment, and sensitivity analyses provide a reproducible account of how evaluation choices shape performance in this historical readmission cohort.

Keywords: hospital readmission, grouped validation, class imbalance, probability calibration, cluster bootstrap, reproducibility

1 Introduction

Hospital readmission prediction is an imbalanced classification problem in which the apparent strength of a model depends on the evaluation design as well as its algorithm. A dataset can contain multiple encounters from the same patient, nominal variables represented by integer codes, and variables whose availability depends on the point in the care pathway. These features make a

simple random split of rows an incomplete specification of the intended prediction problem. A model evaluated on new encounters from previously observed patients answers a different question from a model evaluated on previously unseen patients.

The public Diabetes 130-US Hospitals dataset provides an accessible setting for examining these issues [4, 11]. Its repeated encounters and mixed predictors support useful methodological demonstrations, but its historical and observational nature constrain clinical interpretation. Published applications include machine-learning analyses of early readmission [6]. Comparisons between such studies require matching cohort construction, patient handling, outcomes, and validation procedures; a reported AUC alone does not establish comparability.

Class imbalance introduces a second evaluation challenge. Oversampling can change both the fitted decision rule and the scale of predicted probabilities. Applying oversampling before dividing cross-validation folds can also allow information from validation observations to influence training [12]. Ordinary synthetic minority oversampling interpolates numerical feature vectors [2]; applying that operation directly to arbitrary category codes does not respect nominal structure. Moreover, discrimination, calibration, and threshold-specific performance address different questions. Precision–recall analysis is particularly relevant when the positive outcome is uncommon [10], while probability calibration assesses the agreement between estimated risks and observed frequencies [7].

This study audits an initial exploratory notebook for the diabetes readmission dataset and reports an empirical reanalysis. The initial notebook and audit records, available from the corresponding author upon reasonable request, document the workflow being evaluated. The contribution is a controlled examination of evaluation choices in imbalanced prediction. We quantify patient overlap, compare resampling donor pools over repeated partitions, and evaluate logistic regression (LR) and XGBoost (XGB) with separate calibration and conditional patient-cluster uncertainty. A simple reference model, sampling-ratio checks, and alternative handling of laboratory and payer fields test how the conclusions depend on specific modeling choices. The experiments isolate donor-pool contamination under a fixed representation and complement it with a reproducible comparison of specified model configurations.

1.1 Relationship to previous work

The present audit builds on established evidence that evaluation design can alter apparent predictive performance. Vandewiele et al. [12] demonstrated optimism from oversampling before partitioning, using synthetic examples and a preterm-birth application. Here, the narrower experimental question is how admitting validation donors changes internal validation when the real validation rows, preprocessing, original training rows, and synthetic training counts remain fixed. This controlled comparison is repeated under both encounter- and patient-level partitioning. It quantifies the contribution of this specific contamination mechanism within the evaluated pipeline.

Strack et al. [11] examined early readmission in a selected diabetes cohort and excluded deaths and hospice discharges. Our historical cohort is retained to audit the initial workflow, with discharge exclusions assessed separately. Published prediction studies such as Emi-Johnson and Nkrumah [6] provide application context, but their reported scores are not direct benchmarks unless cohorts and evaluation designs match. Table 1 summarizes these distinctions.

2 Data and analytical cohort

2.1 Source and outcome

The source file contains 101,766 encounters from the UCI repository's historical diabetes dataset, covering care during 1999–2008 [4]. The binary outcome equals one for the recorded category `<30` and zero for `>30` or `NO`. Thus, the negative class combines later readmission and no recorded readmission. The endpoint is recorded readmission within 30 days, as defined by the source dataset.

The analysis evaluates retrospective prediction using encounter information available by discharge. Patient and encounter identifiers are retained for audit and partitioning but excluded from predictors. Evaluation focuses on patient-disjoint partitions of the historical cohort.

Table 1 Selected related work and the role of the present analysis. This comparison highlights the study questions and evaluation designs.

Study	Relevant focus	Relationship to the present audit
Vandewiele et al. [12]	Oversampling before partitioning and optimistic evaluation	Motivates the matched-count donor-pool contrast with real validation observations.
Strack et al. [11]	Diabetes readmission, clinical cohort selection, and HbA1c measurement	Informs the discharge-eligibility sensitivity analysis; outcome and cohort definitions remain explicit.
Emi-Johnson and Nkrumah [6]	Machine-learning prediction of 30-day readmission	Provides application context for interpreting cohort and evaluation differences.
Present analysis	Controlled donor-pool experiment and patient-disjoint workflow audit	Combines repeated contamination contrasts with separate calibration, conditional uncertainty, and cohort sensitivity checks.

2.2 Reproduction of the original cohort

To separate evaluation changes from cohort changes, we retain the initial notebook's complete-case cohort. Four fields—weight, medical specialty, maximum glucose serum result, and glycated hemoglobin (HbA1c) result—are removed. The original parsing convention treats `?` and pandas default missing-value tokens as missing; this also affects literal `None` entries in laboratory-result fields. The source documentation identifies unmeasured laboratory tests as a distinct category [11]. The dropped fields are therefore not interpreted as clinically irrelevant; a separate sensitivity analysis retains the laboratory-result categories. Complete-case filtering on the remaining variables yields 59,129 encounters from 39,726 patients, with 6,514 positive outcomes (11.02%). Table 2 records the sequential exclusions. The ordering is a reproducible accounting convention, not an attribution of independent missingness effects.

This restriction defines the historical complete-case sample, with substantial exclusions associated with missing payer information. The cohort retains the original discharge categories, including death and hospice records, to preserve comparability with the initial workflow. A separate sensitivity analysis removes death and hospice discharges while preserving the original patient partitions; the primary historical audit remains unchanged.

2.3 Descriptive encounter distributions

Figures 1–3 display the binary target, numerical predictor distributions, and the original readmission categories by age. The retained cohort contains 52,615 negative and 6,514 positive encounters. Count plots describe the selected analytical sample and do not, by themselves, establish differences in risk. The negative binary target pools later readmission and no recorded readmission; the age chart preserves the three original categories.

2.4 Predictor representation

Nine predictors are treated numerically: age-band index, length of stay, number of laboratory procedures, number of procedures, number of medications, prior outpatient visits, prior emergency visits, prior inpatient visits, and number of diagnoses. Age bands are mapped to ordered indices; this imposes an equal-spacing assumption. All remaining predictors, including integer-coded admission and discharge variables, are treated as categorical. Three diagnosis fields are mapped to the original notebook's broad diagnostic groups. The exact mapping is supplied in the analysis code; its numerical ICD-code boundaries are retained for comparability and are not presented as a clinically validated ontology.

Table 2 Sequential reproduction of the analytical cohort. Counts after each step are encounters. The four removed fields are not used for complete-case filtering.

Step	Removed at step	Remaining
Source file	—	101,766
Missing race	2,273	99,493
Missing payer code	39,712	59,781
Missing primary diagnosis	10	59,771
Missing secondary diagnosis	175	59,596
Missing third diagnosis	467	59,129

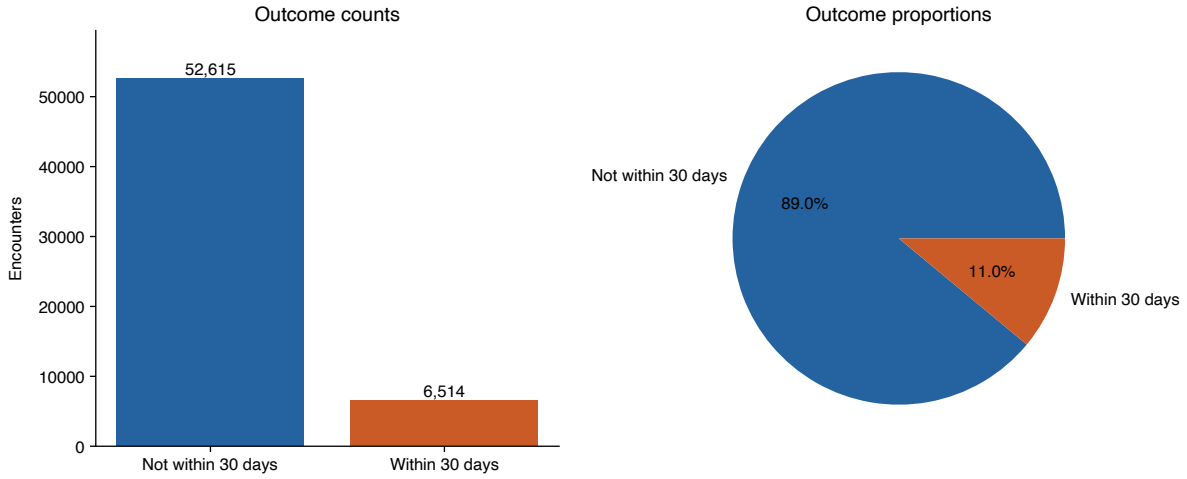


Fig. 1 Outcome distribution in the historical complete-case analytical cohort (59,129 encounters). The negative category includes later readmission and no recorded readmission

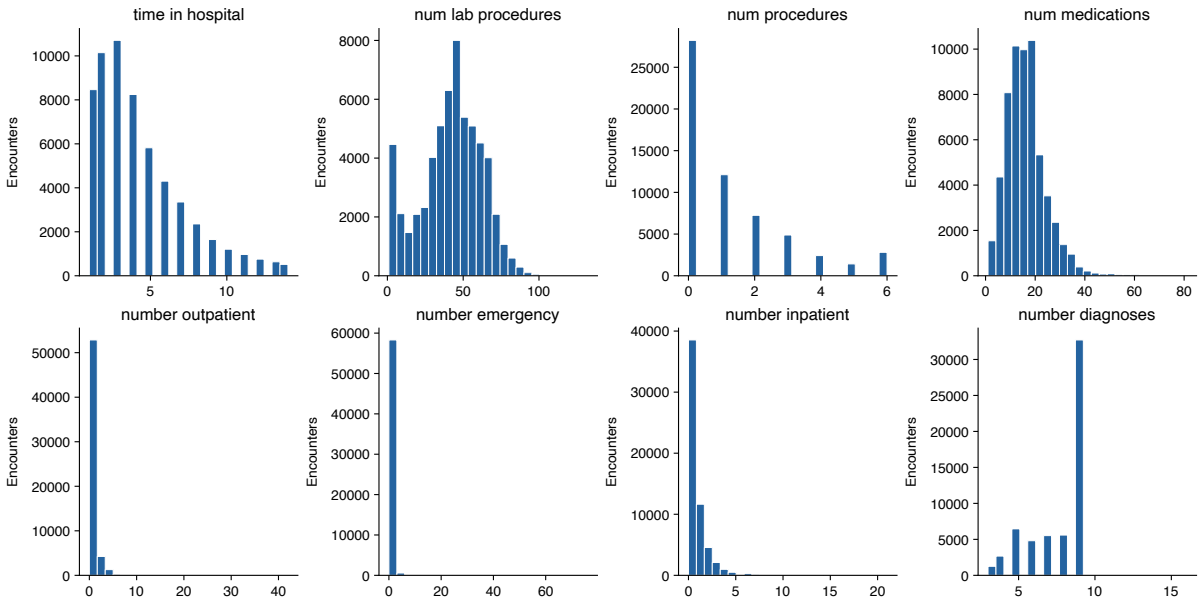


Fig. 2 Distributions of eight numerical predictors in the historical analytical cohort. Counts are encounters and may include repeat patients

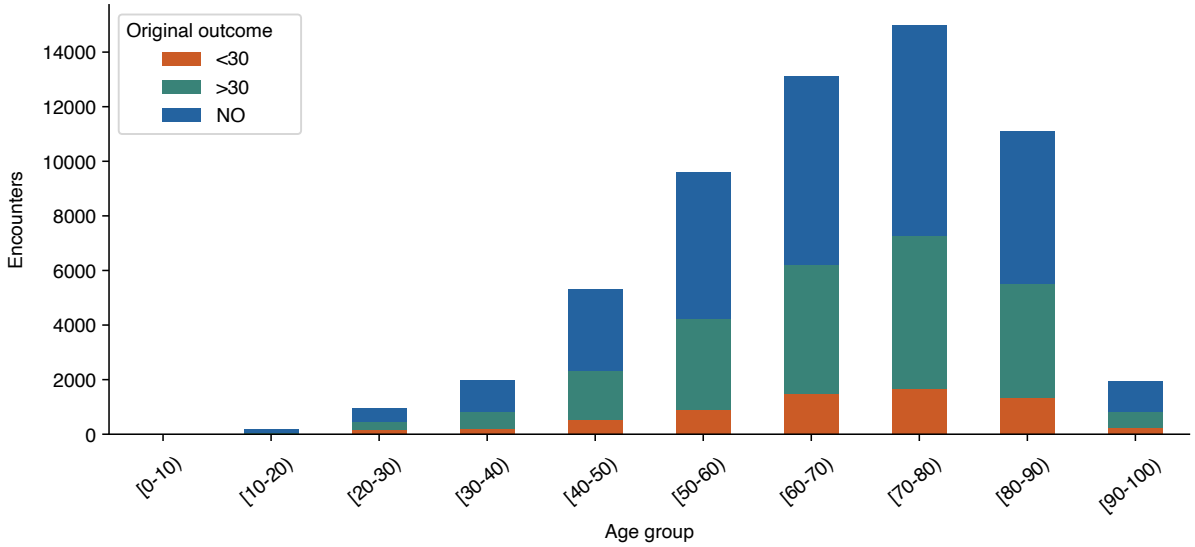


Fig. 3 Age distribution and original outcome categories in the historical cohort. This is a descriptive count plot; larger bars do not by themselves indicate higher readmission risk

3 Evaluation design

3.1 Audit of the initial workflow

The original notebook removes identifiers before an encounter-level stratified split. Reconstructing its split using retained identifiers reveals 4,374 patients represented on both sides. Of 11,826 original test encounters, 5,450 (46.08%) belong to patients also represented in training. This does not quantify the magnitude of optimism, but it prevents interpreting that split as evaluation in unseen patients.

The initial workflow fits scaling and applies ordinary SMOTE before cross-validation on the resampled training matrix. Its held-out test rows are not included in that SMOTE call, but the cross-validation folds do not isolate preprocessing and synthetic-sample generation. In addition, nominal labels are encoded as numerical values before interpolation. Finally, a proposed positive-class weighting step computes its ratio after balancing the classes, giving a weight of one. The primary corrected analysis addresses these issues jointly. The controlled experiment below isolates a narrower mechanism—validation observations entering the synthetic-sampling donor pool—without changing the original validation observations. This contrast complements the broader workflow reanalysis.

3.2 Patient-disjoint partitions

Patients, rather than encounters, define partition membership. A shuffled five-fold stratified group split with seed 2026 supplies the first held-out fold as the evaluation partition. A shuffled four-fold stratified group split with seed 2027 divides the remaining development patients into fitting and calibration partitions, again using its first held-out fold. Every encounter from a patient remains in one partition. Table 3 shows the realized sizes; the audit confirms zero patient overlap between every pair of partitions.

The two primary model families additionally undergo three-fold stratified group cross-validation within the fitting partition, using seed 2028. These folds estimate internal discrimination for the fixed model configurations. All learned preprocessing and resampling are fitted anew using only the training portion of each fold. Final base models use the full fitting partition; the calibration and evaluation partitions are never resampled.

Table 3 Realized mutually patient-disjoint partitions. Outcomes and prevalence are encounter-level quantities.

Partition	Encounters	Patients	Positive	Prevalence
Fitting	35,421	23,800	3,857	10.89%
Calibration	11,789	7,948	1,302	11.04%
Evaluation	11,919	7,978	1,355	11.37%

3.3 Models and imbalance strategies

The primary comparison uses L_2-regularized logistic regression (LR; $C = 1$, LBFGS solver, maximum 2,000 iterations) and XGBoost gradient-boosted trees (XGB; 200 trees, maximum depth 5, learning rate 0.1, row and column sampling fractions 0.8, histogram tree construction, seed 42) [3]. Four strategies are evaluated for each family: no resampling; class weighting; random oversampling (ROS); and SMOTENC, the mixed categorical/numerical variant available in imbalanced-learn [8]. ROS and SMOTENC use the library's default balancing target and seed 42; SMOTENC uses five neighbors.

For each training fit, numerical features are standardized and categorical values are encoded using training-derived category maps. The sampler, when present, operates on this mixed representation with categorical indices explicitly specified. Categorical features are then one-hot encoded before model fitting; unseen categories are handled without refitting on held-out data. SMOTENC protects nominal levels from ordinary linear interpolation, although interpolation of count-like numerical variables can still produce fractional synthetic values. These synthetic values serve as training inputs within the evaluated resampling procedure.

Logistic-regression class weights use inverse class frequencies from the actual training subset. XGBoost's positive-class weight is the ratio of negative to positive outcomes in that subset. In the final fitting partition this ratio is approximately 8.18, rather than one. Fixed settings define a reproducible comparison of the specified configurations.

Five additional unresampled models—decision tree, random forest, k-nearest neighbors, support-vector machine, and multilayer perceptron—extend the descriptive model comparison. Their precise settings and computational restrictions are given below, and their results appear in the main Results section. They share the same calibration and evaluation patients. This secondary comparison provides complementary descriptive evidence alongside the primary uncertainty analysis.

3.4 Additional exploratory model specifications

The seven-model comparison includes the preceding unresampled logistic regression (LR) and XGBoost (XGB) configurations, together with five additional models. It is a descriptive extension of the primary comparison. The five added models are summarized by evaluation-set point estimates; three-fold cross-validation and bootstrap uncertainty characterize the primary model comparison.

The decision tree (DT) uses maximum depth six and minimum leaf size 50. The random forest (RF) uses 200 trees, maximum depth ten, and minimum leaf size 20. Both use seed 42. The k-nearest-neighbor classifier (KNN) uses five neighbors, uniform weighting, and the library's default Euclidean-distance setting. All three use the same numerical scaling and one-hot categorical representation as the primary models, without resampling.

The support-vector machine (SVM) uses an RBF kernel, $C = 1$, and `gamma=scale`. To bound computation, fitting is restricted to 20,000 encounters selected by randomly ordering fitting patients with seed 42 and adding complete patient groups while respecting the encounter budget. Selection does not consult outcomes. The comparison therefore differs in fitting sample size as well as model family. SVM probability estimation is disabled; an unpenalized logistic model calibrates the decision score using the common calibration patients.

The multilayer perceptron (MLP) uses hidden layers of 128, 64, and 32 ReLU units with Adam optimization and the library's remaining default settings, including learning rate 0.001 and L_2 parameter 0.0001. Epoch selection uses a separate first fold from a five-fold stratified group

split within fitting patients (seed 2031). Preprocessing is fitted only on the inner training patients. Epoch-wise validation log loss is monitored for at most 100 epochs, with improvement tolerance 10^{-4} and patience ten. The selected epoch count is five. Preprocessing and a fresh network are then fitted on the complete fitting partition for those five epochs, with seed 42. The calibration and evaluation partitions are not used for epoch selection.

For models producing probabilities, the same logit-based logistic calibration procedure as described below is applied. Each model's operating threshold is selected on calibration patients under the empirical FPR constraint. The rebuilt LR and XGB probabilities reproduce the corresponding primary predictions within 10^{-6}. The comparison describes metric-specific performance across the seven fitted configurations.

3.5 Calibration and operating thresholds

Each primary model produces an uncalibrated probability p. We fit an unpenalized logistic calibration model on the separate calibration partition:

$$\widetilde{p} = \operatorname{logit}^{-1}\{a + b \operatorname{logit}(p)\}, \tag{1}$$

with probabilities clipped to $[10^{-7}, 1-10^{-7}]$ before transformation. Calibration estimates are held fixed during evaluation. All eight fitted slopes are positive, so calibration preserves ranking apart from numerical clipping and ties. Calibration cannot by itself rescue weak discrimination.

For an illustrative resource-constrained operating point, the decision threshold maximizes empirical calibration-set recall among ROC thresholds having false-positive rate no greater than 0.10. Formally,

$$t^* = \operatorname*{arg\,max}_{t:\,\mathrm{FPR}_{\mathrm{cal}}(t)\le 0.10} \mathrm{Recall}_{\mathrm{cal}}(t). \tag{2}$$

In a tie, the first maximizing threshold in the returned ROC sequence is retained. This rule defines the illustrative operating point in the implemented analysis. The realized evaluation false-positive rate need not be at most 0.10. We report results at this operating point and retain the conventional 0.5 threshold in Online Resource 1. Evaluation-set threshold sweeps are descriptive and do not drive selection.

3.6 Metrics and conditional uncertainty

Let patient i contribute n_i evaluation encounters with outcomes y_{ij} and calibrated predictions $\widetilde{p}_{ij}$, and let $N = \sum_i n_i$. We report ROC-AUC, average precision (AP), Brier score, recall, precision, and false-positive rate. The Brier score and AP are

$$\begin{aligned} \mathrm{BS} &= N^{-1} \sum_i \sum_{j=1}^{n_i} (\widetilde{p}_{ij} - y_{ij})^2, \\ \mathrm{AP} &= \sum_k (R_k - R_{k-1}) P_k. \end{aligned} \tag{3}$$

where P_k and R_k are precision and recall at successive operating points ordered by increasing recall. The Brier score evaluates squared probability error [1]. AP uses the stepwise recall-weighted precision definition implemented in scikit-learn, not trapezoidal integration of a precision–recall curve. Calibration is displayed in ten quantile bins. We also report diagnostic intercepts and slopes obtained by regressing evaluation outcomes on the logit of calibrated predictions; these diagnostics do not modify predictions.

Uncertainty for the eight primary configurations uses 1,000 percentile bootstrap samples of evaluation patients with seed 2029. Each sampled patient contributes all of their evaluation encounters, and repeated draws repeat the entire cluster. The estimands remain encounter-weighted: patients with more encounters contribute more observations. Let S_b be the multiset of patients drawn in replicate b. A metric M is recomputed on all encounters contributed by S_b, including repeated patient draws. The same draws are used across models, giving the paired difference $\Delta_b = M_{\mathrm{XGB},b} - M_{\mathrm{LR},b}$ between the unresampled baselines. The 2.5th and 97.5th percentiles define the intervals. These intervals quantify evaluation-patient sampling uncertainty conditional on the trained models, calibration maps, selected thresholds, and fixed partition.

Race-, recorded-gender-, and age-specific metrics are exploratory. Separate patient-cluster resampling provides subgroup intervals for the two unresampled baselines, with valid-replicate counts

recorded when resamples are single-class. The intervals are pointwise and unadjusted for multiplicity, supporting exploratory subgroup comparisons. Small strata can yield unstable or degenerate estimates, particularly for thresholded metrics. We therefore interpret subgroup estimates together with their denominators and intervals.

3.7 Controlled resampling-donor experiment

We run 20 repetitions for each of two outer split units: encounters and patients. Encounter splitting uses an 80/20 stratified random split; patient splitting uses the first test fold from shuffled five-fold stratified group splitting. The corresponding internal three-fold validation uses the same unit. Split seeds are $4100 + r$ and internal-fold seeds $5100 + r$, for $r = 0, \ldots, 19$. LR and XGB settings are held fixed. This experiment uses a fixed numerical/ordinal representation and ordinary SMOTE to examine a legacy sampling mechanism; it is distinct from the categorical-aware primary pipeline.

For each internal fold, scaling and category maps are fitted only on its original training observations and then held fixed across three arms: no resampling; SMOTE using training-fold donors only; and SMOTE whose donor pool also includes the original validation observations and their labels. The third arm deliberately introduces contamination. Its synthetic pool is generated from the outer training set after transformation with the fold-trained preprocessing, and a seeded subset supplies exactly the number of synthetic minority observations used in the fold-local arm. Original training rows, original validation rows, preprocessing, and post-sampling class counts are identical between these two arms. Original validation rows are never directly appended to training. Validation metrics are always computed on real observations at their observed prevalence, not on synthetic validation rows.

This design isolates the effect of admitting validation observations into the donor pool under the specified sampling procedure. The matched-count design evaluates donor-pool contamination under a controlled diagnostic representation. Ordinary SMOTE interpolation of ordinal category codes is retained for this diagnostic experiment; the primary models use categorical-aware preprocessing and sampling.

The final model for each outer split is fitted once without sampling and once with SMOTE on the full outer training partition. The two SMOTE arms therefore have exactly the same held-out test result by construction: their difference concerns internal validation, and neither uses the outer test set. We report means and sample standard deviations over the 20 repetitions. Repetitions reuse patients, so the standard deviations describe split sensitivity, not independent-sample confidence intervals. Encounter- versus patient-split test results concern different evaluation populations and are not interpreted as a causal leakage effect.

3.8 Reference model and sensitivity analyses

A regularized LR model using only prior inpatient admissions provides a simple reference. It uses the original patient partitions, separate logistic calibration, and the same threshold rule. Two additional SMOTENC ratios, 0.25 and 0.50 minority/majority after resampling, are compared with the primary fully balanced ratio of 1.00. The declared categorical indices are checked against the numerical-then-categorical preprocessing order.

For a same-patient sensitivity check, the raw file is parsed with pandas default missing tokens disabled and only "?" marked missing; the two laboratory-result fields are retained as categorical predictors, including their unmeasured categories. Models are refitted on the original fitting patients and evaluated on the identical original evaluation patients. A second, expanded-cohort analysis retains the laboratory fields and removes payer missingness from row eligibility. It compares a missing payer category with dropping the payer predictor entirely, while retaining the remaining complete-case restrictions. New patient-disjoint fitting, calibration, and evaluation partitions use the original splitting seeds. Because the expanded population and its evaluation patients differ, its metrics are not paired improvements over the original cohort.

For all new comparisons on the original evaluation patients, 1,000 paired patient-cluster bootstrap samples (seed 2091) estimate differences in AUC and AP from the corresponding unresampled full-feature model. Intervals remain conditional

on fitted models, calibration, and thresholds, and are exploratory without multiplicity adjustment. Expanded-cohort estimates are descriptive point estimates.

3.9 Discharge-eligibility sensitivity analysis

A post-audit sensitivity analysis excludes death and hospice discharge codes (11, 13, 14, 19, 20, and 21), following the clinical rationale of Strack et al. [11]. The remaining complete-case rules are unchanged. Exclusions are applied within the original fitting, calibration, and evaluation partitions, so retained patients keep their original roles. We refit the unresampled LR and XGB pipelines with their original settings, recalibrate using retained calibration encounters, and reselect thresholds using the same false-positive-rate rule. A second refit removes discharge disposition as a predictor.

For a comparison on identical observations, the saved historical models are also evaluated on the retained evaluation encounters, using their original calibration maps and thresholds. Paired patient-cluster intervals use 1,000 draws with seed 2092 and condition on the fitted models. Differences from the historical full evaluation population are descriptive; only comparisons within the retained evaluation set are paired. Grouped XGB SHAP summaries use 2,000 retained evaluation encounters sampled with seed 2033. These post-audit analyses assess sensitivity to discharge eligibility within the historical dataset.

3.10 Visualization and reproducibility

Risk maps summarize actual evaluation encounters grouped by prior inpatient admissions and length of stay. Cells with fewer than 100 encounters are masked. Predicted risks average each encounter's full-model probability; the mix of other predictors can differ across cells. These displays summarize observed encounter groups and their fitted predictions. A three-dimensional rendering uses the same observed bins without interpolation.

For model inspection, exact tree-path SHAP contributions are computed for a seeded random sample of 2,000 evaluation encounters [9]. They explain the raw XGBoost log-odds margin, rather than calibrated probabilities. Contributions of one-hot columns are summed within original variables before calculating grouped mean absolute attribution. Random-forest impurity importance is reported separately because it has different interpretation and potential split-selection biases. Both attribution methods describe fitted-model associations.

Machine-readable partition, prediction, metric, calibration, and interval records are available from the corresponding author upon reasonable request. The main computation uses Python 3.13.9, NumPy 2.3.5, pandas 2.3.3, scikit-learn 1.7.2, imbalanced-learn 0.14.0, and XGBoost 3.4.1. The input checksum and seeds are recorded in the analysis protocol, which is available on the same basis. Independent checks recompute metrics from saved predictions and verify patient separation. Reporting is informed by the transparency objectives of TRIPOD+AI [5].

3.11 AI assistance

AI text-editing tools were utilized solely to improve language clarity and grammatical phrasing during the drafting of this manuscript. Both authors take responsibility for the final work.

4 Results

4.1 Repeated resampling-donor audit

Table 4 and Figure 4 report the completed repeated experiment. The contaminated and fold-local arms share the same real validation observations, so their internal AUC difference cannot be explained by substituting synthetic observations into the validation metric. They also have matching synthetic sample counts.

With encounter-level splitting, the mean contaminated-minus-fold-local CV AUC differences are LR: +0.0044 (SD 0.0011), and XGB: +0.0120 (SD 0.0014). With patient-level splitting, the mean contaminated-minus-fold-local CV AUC differences are LR: +0.0052 (SD 0.0010), and XGB: +0.0122 (SD 0.0017). These contrasts quantify this particular contamination mechanism under the diagnostic representation; they

should not be substituted for the primary mixed-type model estimates. Without resampling, mean outer-test AUC is 0.6418 for LR and 0.6685 for XGB under encounter splitting, versus 0.6431 and 0.6701 under patient splitting. Patient separation therefore does not produce a systematic AUC decrease in this experiment. The overlap audit establishes that the source split is not an unseen-patient evaluation; it does not establish the size or direction of an AUC bias.

4.2 Discrimination and probability accuracy

The evaluation partition contains 1,355 positive outcomes among 11,919 encounters (11.37%). Unresampled LR achieves ROC-AUC 0.653 and AP 0.215; unresampled XGB achieves ROC-AUC 0.657 and AP 0.231 (Table 5; Figure 5). The corresponding mean fitting-partition grouped cross-validation AUCs are 0.661 and 0.662. Thus, both the internal and held-out assessments indicate modest discrimination under the implemented patient-disjoint design.

The paired XGB-minus-LR ROC-AUC difference is 0.0041, with a conditional 95% interval spanning zero (−0.0042, 0.0128). The AP difference is 0.0154 (0.0044, 0.0264), and the Brier-score difference is −0.00081 (−0.00136, −0.00029), favoring XGB for those two metrics in this fitted comparison. These metric-specific findings do not establish broad algorithmic superiority. Marginal and paired intervals are reported together in Table 6; marginal interval overlap is not used as a significance test. Figure 6 displays the corresponding configuration-specific intervals.

The calibrated Brier scores are 0.09670 for LR and 0.09589 for XGB. Evaluation calibration diagnostics are intercept 0.114 and slope 1.036 for LR, and intercept 0.031 and slope 1.002 for XGB. These descriptive point estimates and binned curves characterize probability calibration in the held-out historical evaluation sample.

4.3 Calibration diagnostics

Table 7 distinguishes calibration-map coefficients, fitted on calibration patients, from evaluation diagnostics fitted only to assess calibration. Diagnostic estimates do not change the evaluated predictions. Large changes in calibration-map intercepts after weighting or oversampling illustrate why raw scores should not automatically be interpreted as natural-prevalence probabilities. The coefficients are reported as descriptive point estimates.

4.4 Imbalance strategies

Class weighting and ROS yield LR AUCs of 0.657 and 0.657, respectively, compared with 0.653 without resampling. The corresponding XGB AUCs are 0.650 and 0.654, compared with 0.657 without resampling. These differences summarize the point-estimate performance of the specified strategies.

SMOTENC produces lower AUC and AP than the corresponding unresampled configuration in both families: LR AUC decreases to 0.578 and AP to 0.159; XGB AUC decreases to 0.619 and AP to 0.197. Its calibrated Brier scores are also larger. Figure 7 presents the full metric pattern. This result concerns the specified settings, preprocessing, and cohort. The sampling-ratio analysis further characterizes this pattern within the evaluated workflow.

Class weighting changes the fitted objective, but the two separately calibrated XGB models still have similar operating-point trade-offs. Figure 8 compares the genuine fitting-set weight calculation with the unweighted baseline. This experiment characterizes the predictive consequences of class weighting.

4.5 Operating-point trade-offs

The selected probability thresholds for unresampled LR and XGB are approximately 0.163 and 0.173, respectively. At these thresholds, evaluation recall is 0.255 and 0.263, precision is 0.243 and 0.251, and FPR is 0.102 and 0.101. Thus, about three quarters of positive encounters remain unflagged at this illustrative constraint. The bootstrap interval for the recall difference also spans zero (Table 6).

Lowering a threshold increases the fraction of encounters flagged and changes the balance between missed positives and false alerts (Figure 9). These trade-offs should be interpreted

Table 4 Controlled donor-pool audit: mean (SD) AUC over 20 outer partitions. The two SMOTE arms share their final fitted model and therefore have identical test estimates by design.

Split unit	Model	Arm	Internal CV	Outer test
Encounter	LR	Fold-local SMOTE	0.6337 (0.0023)	0.6339 (0.0090)
Encounter	LR	None	0.6418 (0.0017)	0.6418 (0.0082)
Encounter	LR	Validation-donor SMOTE	0.6382 (0.0021)	0.6339 (0.0090)
Encounter	XGB	Fold-local SMOTE	0.6595 (0.0029)	0.6652 (0.0082)
Encounter	XGB	None	0.6626 (0.0030)	0.6685 (0.0065)
Encounter	XGB	Validation-donor SMOTE	0.6715 (0.0027)	0.6652 (0.0082)
Patient	LR	Fold-local SMOTE	0.6323 (0.0023)	0.6352 (0.0073)
Patient	LR	None	0.6406 (0.0021)	0.6431 (0.0066)
Patient	LR	Validation-donor SMOTE	0.6375 (0.0020)	0.6352 (0.0073)
Patient	XGB	Fold-local SMOTE	0.6582 (0.0028)	0.6656 (0.0076)
Patient	XGB	None	0.6602 (0.0027)	0.6701 (0.0075)
Patient	XGB	Validation-donor SMOTE	0.6704 (0.0030)	0.6656 (0.0076)

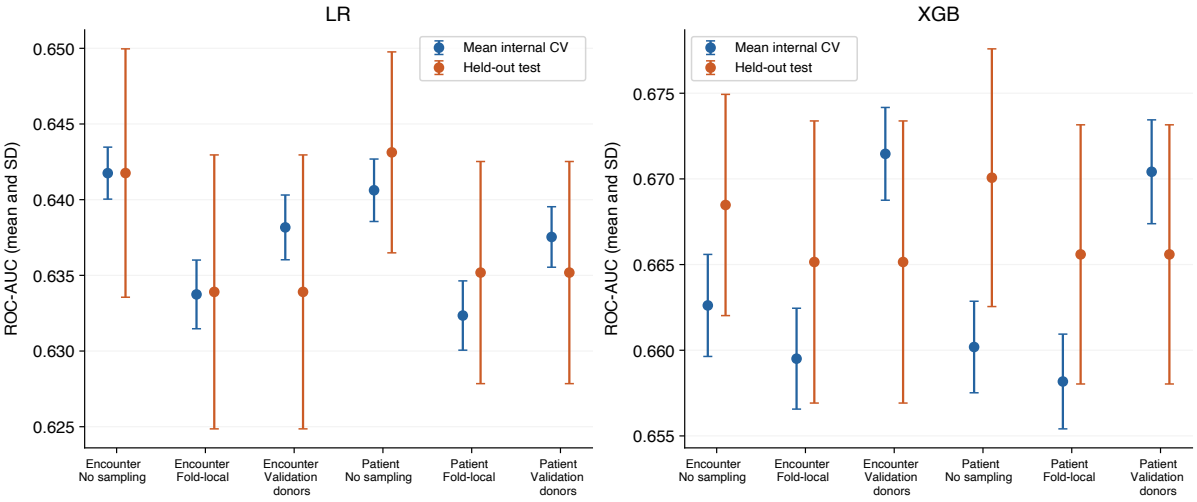


Fig. 4 Controlled donor-pool audit across 20 partitions. Error bars show between-partition standard deviations, not confidence intervals. Contaminated sampling admits validation donors but keeps validation observations real and matches the fold-local synthetic training count. The two SMOTE arms share identical outer-test results by design

Table 5 Evaluation performance after separate-patient calibration. Recall, precision and FPR use each model's calibration-selected threshold. Higher ROC-AUC, AP, recall and precision are favorable; lower Brier score and FPR are favorable, subject to threshold trade-offs.

Model / Strategy	AUC	AP	Brier	Recall	Precision	FPR
LR / None	0.653	0.215	0.0967	0.255	0.243	0.102
LR / Weights	0.657	0.217	0.0966	0.252	0.244	0.100
LR / ROS	0.657	0.216	0.0967	0.252	0.240	0.102
LR / SMOTENC	0.578	0.159	0.0999	0.181	0.186	0.101
XGB / None	0.657	0.231	0.0959	0.263	0.251	0.101
XGB / Weights	0.650	0.223	0.0963	0.276	0.256	0.103
XGB / ROS	0.654	0.222	0.0964	0.270	0.249	0.105
XGB / SMOTENC	0.619	0.197	0.0981	0.239	0.236	0.099

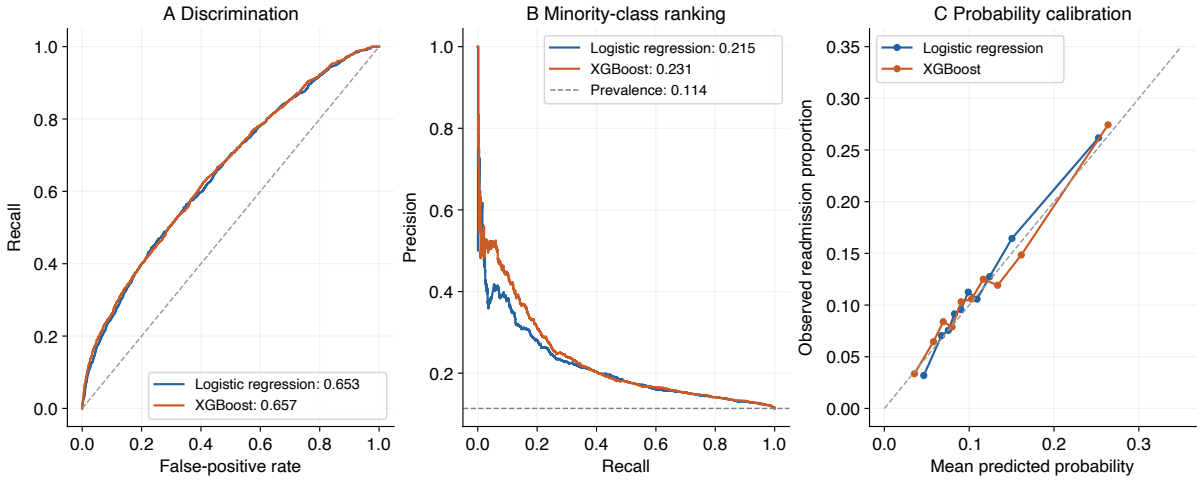


Fig. 5 Patient-disjoint evaluation of unresampled LR and XGB after separate-patient calibration. Panels show ROC, precision–recall, and ten-bin calibration curves. Dashed references indicate chance ranking, evaluation prevalence, and ideal calibration. AP is average precision; calibration connecting lines are descriptive, not uncertainty bands

Table 6 Primary baseline estimates and conditional 95% patient-cluster bootstrap intervals. The difference is XGB minus LR, calculated using paired patient draws. Values are rounded independently.

Metric	LR	XGB	Paired difference
ROC-AUC	0.653 [0.637, 0.672]	0.657 [0.640, 0.675]	+0.0041 [-0.004, 0.013]
Average precision	0.215 [0.193, 0.242]	0.231 [0.207, 0.257]	+0.0154 [0.004, 0.026]
Brier score	0.0967 [0.0920, 0.1013]	0.0959 [0.0912, 0.1005]	-0.00081 [-0.00136, -0.00029]
Recall	0.255 [0.226, 0.285]	0.263 [0.235, 0.294]	+0.0089 [-0.009, 0.025]
False-positive rate	0.102 [0.094, 0.110]	0.101 [0.094, 0.109]	-0.0010 [-0.006, 0.004]

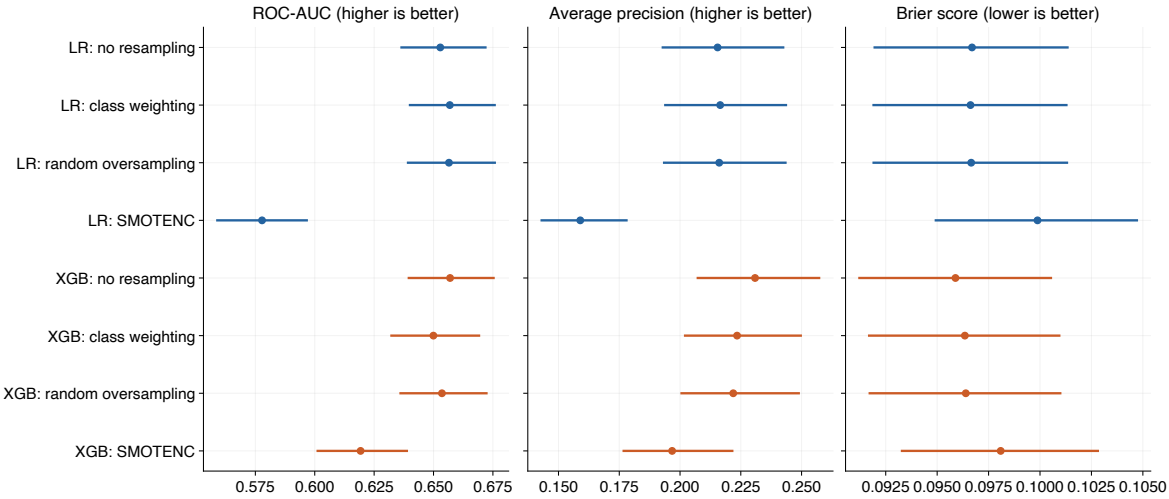


Fig. 6 Points are evaluation metrics after separate-patient calibration; lines are 95% percentile intervals from 1,000 patient-cluster bootstrap samples. Intervals quantify evaluation-patient sampling uncertainty conditional on the fitted models. Paired-difference intervals support comparison of the two baseline models

Table 7 Calibration-map intercept and slope, followed by diagnostic intercept and slope on evaluation patients. Ideal diagnostic values are zero and one.

Configuration	Calibration map Intercept	Slope	Evaluation diagnostic Intercept	Slope
LR_none	-0.138	0.920	0.114	1.036
LR_weighted	-2.064	0.833	0.142	1.051
LR_ros	-2.062	0.819	0.146	1.054
LR_smotenc	-1.929	0.191	-0.050	0.959
XGB_none	-0.263	0.844	0.031	1.002
XGB_weighted	-1.914	0.817	-0.026	0.975
XGB_ros	-1.920	0.863	-0.047	0.963
XGB_smotenc	-0.912	0.642	-0.143	0.914

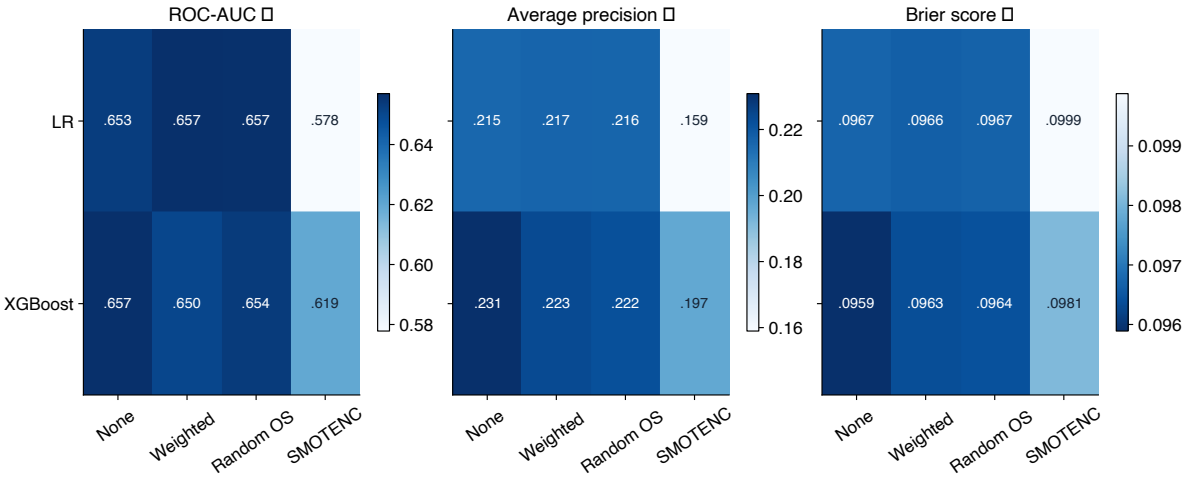


Fig. 7 Eight fixed model/imbalance configurations after calibration. Each metric has its own scale; darker shading indicates favorable direction within that metric. Color intensities cannot be compared across metrics. Numerical estimates support comparison without a composite ranking

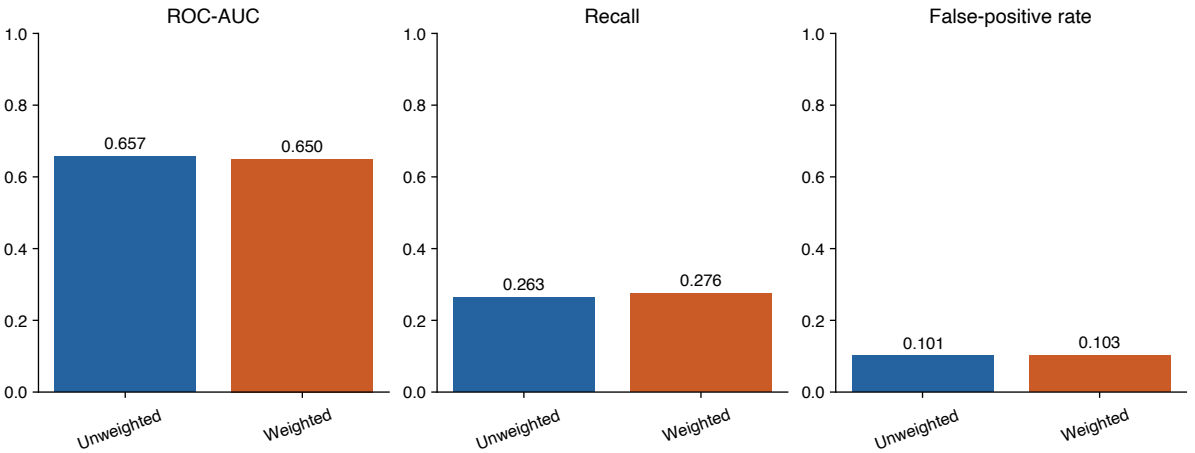


Fig. 8 Genuine class-weight comparison: XGBoost positive-class weight is calculated from the unresampled fitting data, with each model separately calibrated and assigned a threshold using calibration patients. The original notebook instead calculated weight after balancing, producing weight one. The comparison describes discrimination and operating-point performance

in relation to an intervention's costs and capacity. The analysis quantifies threshold-dependent recall, false-positive rates, and flagged encounters under an explicit calibration-based operating rule.

The confusion matrices in Figure 10 make the operating-point consequence explicit. Calibrated XGB at threshold 0.5 identifies 12 positive encounters and generates eight false positives. At the calibration-selected threshold it identifies 357 positive encounters and generates 1,066 false positives, while 998 positive encounters remain unflagged. Both matrices use the same evaluation patients.

4.6 Subgroups and supported risk displays

Subgroup sample sizes vary substantially. The evaluation set includes 9,455 encounters and 1,091 positive outcomes in the recorded Caucasian category, 2,001 and 217 in AfricanAmerican, 208 and 23 in Hispanic, 175 and 17 in Other, and 80 and 7 in Asian. We retain source category labels for reproducibility without treating them as biological classifications. The smallest groups cannot support precise comparisons, and single threshold decisions can change their estimated recall appreciably.

Figure 11 presents race-specific estimates and conditional intervals; gender and age displays are included below. At the common XGB threshold, recall is 0.171 (95% conditional interval 0.110–0.236) for the recorded AfricanAmerican group and 0.282 (0.248–0.317) for the Caucasian group, with FPRs 0.091 and 0.105. For LR, recall is 0.134 (0.089–0.186) and 0.283 (0.244–0.318), respectively. These exploratory estimates describe group-specific missed positive encounters at the common operating threshold, with pointwise intervals characterizing sampling uncertainty. A single common operating threshold, variable outcome prevalence, measurement limitations, and subgroup support all affect interpretation.

Figures 12 and 13 extend the interval displays to recorded gender and age. Figure 14 provides a compact XGB overview using specificity, so that the favorable color direction is consistent. The race-specific false-negative envelope in Figure 15 illustrates how group extremes change across thresholds; it is not a confidence band or evidence of an optimal fairness threshold.

The two-dimensional risk maps provide a supported display of how observed rates and mean predictions vary across encounter groups (Figure 16). They support inspection of local sample size and prediction patterns, rather than an adjusted estimate of the effect of prior admissions or stay duration. The three-dimensional view provides a spatial representation of the same descriptive summaries.

Figure 17 renders the same supported XGB risk bins in three dimensions. The underlying encounter counts, observed proportions, and mean predictions are also included in Online Resource 1. The spatial presentation complements the two-dimensional descriptive display.

4.7 Simple reference and sensitivity results

The prior-inpatient-admissions-only LR reference achieves AUC 0.6039 and AP 0.1760 on the same evaluation encounters (Table 8). Compared with full-feature LR, its paired AUC difference is −0.0490 with 95% conditional interval [-0.0639, -0.0337], and its AP difference is −0.0394 with interval [-0.0515, -0.0282]. The full LR model therefore adds ranking information beyond this simple utilization-history reference in the evaluated sample; this does not quantify clinical benefit.

At SMOTENC ratios 0.25 and 0.50, LR AUCs are 0.6037 and 0.5866, compared with 0.5778 at full balance and 0.6529 without resampling. XGB AUCs are 0.6447 and 0.6345, compared with 0.6193 at full balance and 0.6570 without resampling. Figure 18 shows that the poor fully balanced result is sensitive to the sampling ratio, while none of the tested ratios exceeds its unresampled baseline on the point estimate. This does not establish a universal ordering of imbalance methods.

Retaining laboratory-result categories on the same original patients gives LR AUC 0.6528 and AP 0.2153, and XGB AUC 0.6543 and AP 0.2280. Figure 19 compares these estimates with the original baselines. The XGB AUC difference from its original baseline is approximately −0.0026, with conditional interval [-0.0089, 0.0031]. Thus, preserving the categorical semantics is appropriate,

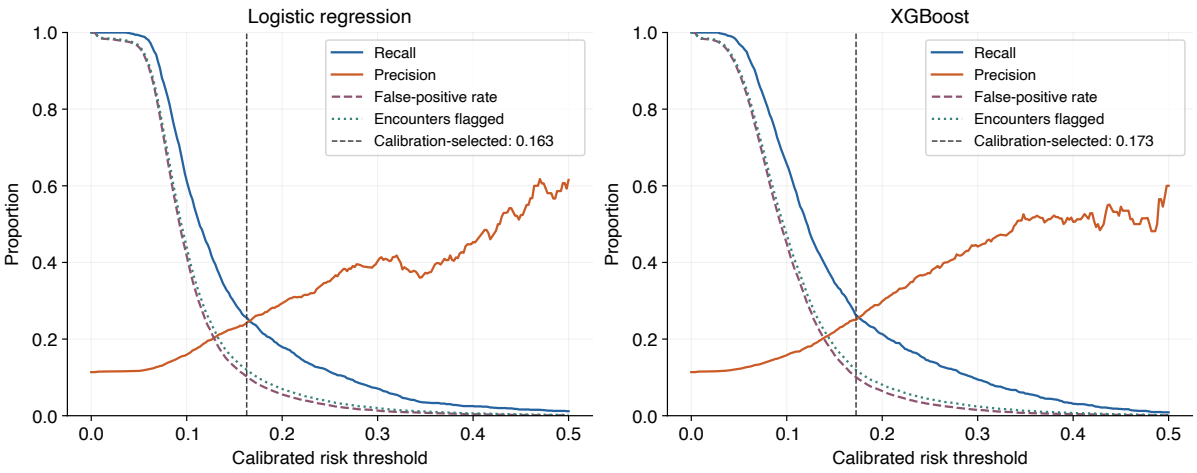


Fig. 9 Descriptive evaluation-set threshold sweeps. Vertical lines indicate thresholds selected using calibration patients to target at most 10% empirical FPR. Flagged fraction counts encounters, not unique patients. Reported thresholds follow the calibration-based selection rule

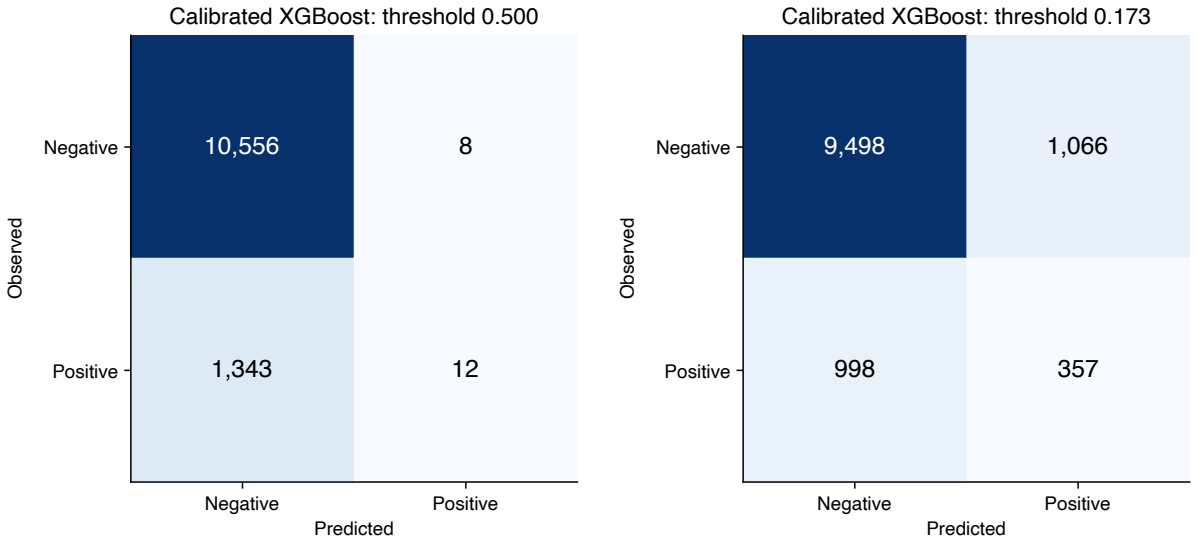


Fig. 10 Confusion matrices for calibrated XGBoost at 0.5 and at the threshold selected on calibration patients. Counts refer to the same 11,919 evaluation encounters. The operating threshold follows the illustrative calibration-based FPR constraint

but does not by itself improve discrimination in this fixed comparison.

The expanded eligibility rule yields 98,053 encounters from 68,630 patients, with 11,066 positive outcomes. Its fitting, calibration, and evaluation sizes are 58,869, 19,545, and 19,639, with no shared patients. Treating missing payer code as a category gives LR AUC/AP 0.6597/0.1934 and XGB 0.6685/0.2154; dropping payer as a predictor gives 0.6575/0.1923 and 0.6647/0.2124. These results show that many more encounters can be retained, but changing the population and partition prevents attributing the numerical differences solely to missingness handling. The expanded-cohort analysis supplies descriptive internal evaluation estimates for the broader population.

4.8 Discharge-eligibility results

Excluding 1,538 encounters leaves 57,591 encounters from 38,802 patients, including 6,492 recorded readmissions within 30 days. Exclusions comprise

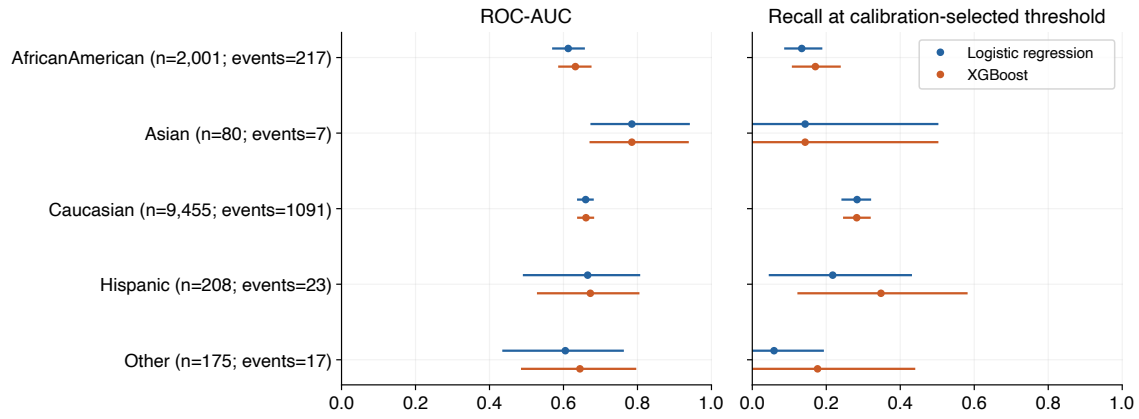


Fig. 11 Exploratory race-specific evaluation of the two unresampled baselines. Labels show encounters and positive encounters. Lines are patient-cluster percentile intervals from up to 1,000 valid replicates; excluded single-class replicates are counted in the source records. No multiple-comparison adjustment or causal fairness inference is made

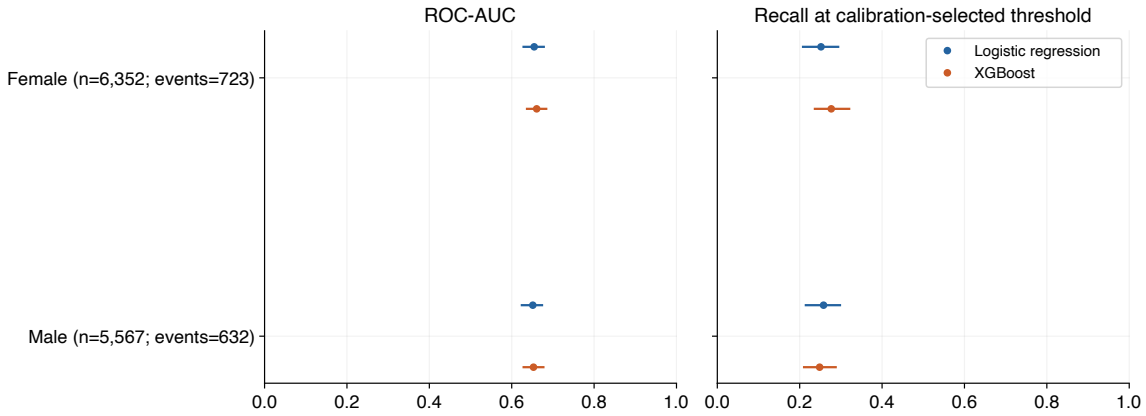


Fig. 12 Exploratory gender-specific evaluation of unresampled baseline models. Labels show encounters and positive encounters, not independent patient counts; patient counts are in the source table. Lines are patient-cluster percentile intervals from up to 1,000 valid replicates. Single-class replicates are excluded and their counts retained. Small-group and boundary estimates can be unstable or degenerate; intervals are pointwise, unadjusted, and exploratory

Table 8 Completed reference and sensitivity analyses after separate-patient calibration. Expanded-cohort rows use a different evaluation population. Recall and FPR use calibration-selected thresholds.

Experiment	Test n	AUC	AP	Brier	Recall	FPR
LR: prior inpatient only	11919	0.6039	0.1760	0.0980	0.173	0.063
LR: SMOTENC 0.25	11919	0.6037	0.1823	0.0988	0.233	0.105
LR: SMOTENC 0.50	11919	0.5866	0.1661	0.0996	0.203	0.103
XGB: SMOTENC 0.25	11919	0.6447	0.2245	0.0964	0.269	0.104
XGB: SMOTENC 0.50	11919	0.6345	0.2120	0.0972	0.250	0.106
LR: same patients + labs	11919	0.6528	0.2153	0.0967	0.255	0.102
XGB: same patients + labs	11919	0.6543	0.2280	0.0961	0.266	0.105
LR: expanded, missing payer category	19639	0.6597	0.1934	0.0946	0.243	0.098
XGB: expanded, missing payer category	19639	0.6685	0.2154	0.0933	0.267	0.101
LR: expanded, payer dropped	19639	0.6575	0.1923	0.0946	0.235	0.097
XGB: expanded, payer dropped	19639	0.6647	0.2124	0.0934	0.249	0.103

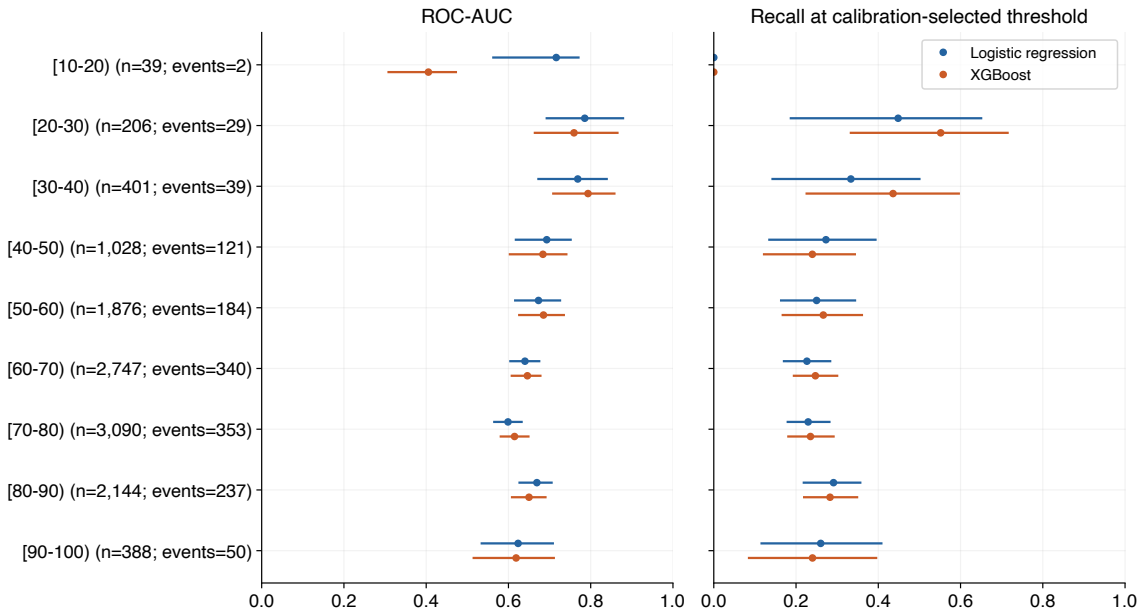


Fig. 13 Exploratory age-specific evaluation of unresampled baseline models. Labels show encounters and positive encounters, not independent patient counts; patient counts are in the source table. Lines are patient-cluster percentile intervals from up to 1,000 valid replicates. Single-class replicates are excluded and their counts retained. Small-group and boundary estimates can be unstable or degenerate; intervals are pointwise, unadjusted, and exploratory

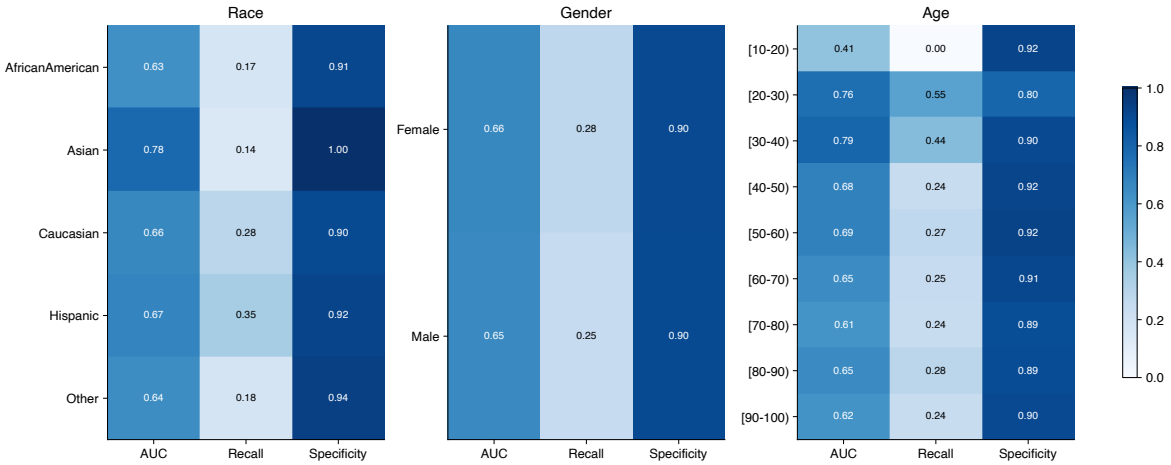


Fig. 14 Descriptive XGBoost subgroup dashboard at the calibration-selected operating point. Specificity replaces false-positive rate so the color direction is consistent. Metrics remain different quantities and must not be combined into a fairness score. Use confidence-interval plots and subgroup counts before interpreting differences

956 encounters with code 11, 300 with code 13, 273 with code 14, eight with code 19, and one with code 20; code 21 is absent from the historical cohort. The retained fitting, calibration, and evaluation sets contain 34,508, 11,473, and 11,610 encounters, respectively, with zero patient overlap. The evaluation set contains 1,348 positive outcomes.

Table 9 compares all models on those same 11,610 encounters. The refitted LR and XGB models achieve AUCs of 0.646 and 0.647, compared with 0.646 and 0.650 for the historical models

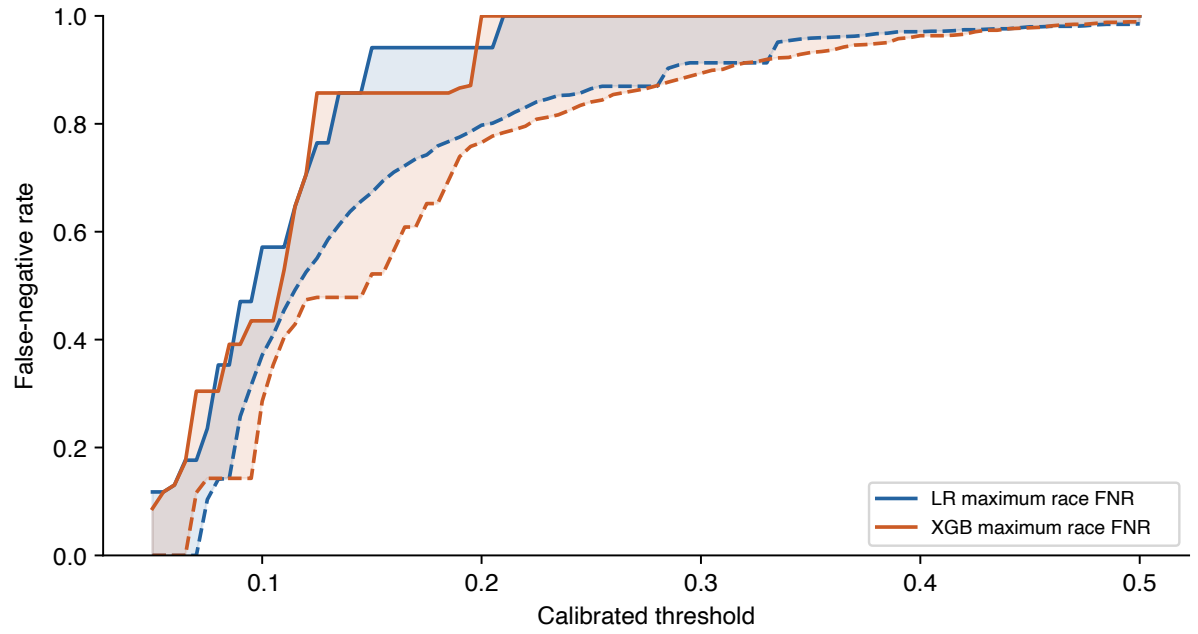


Fig. 15 Race-specific false-negative-rate envelope over evaluation thresholds. Solid curves show the maximum and dashed curves the minimum across observed race groups. Small-group sampling variation can strongly influence these extremes; the envelope is not a confidence band and does not establish a fairness-optimal threshold

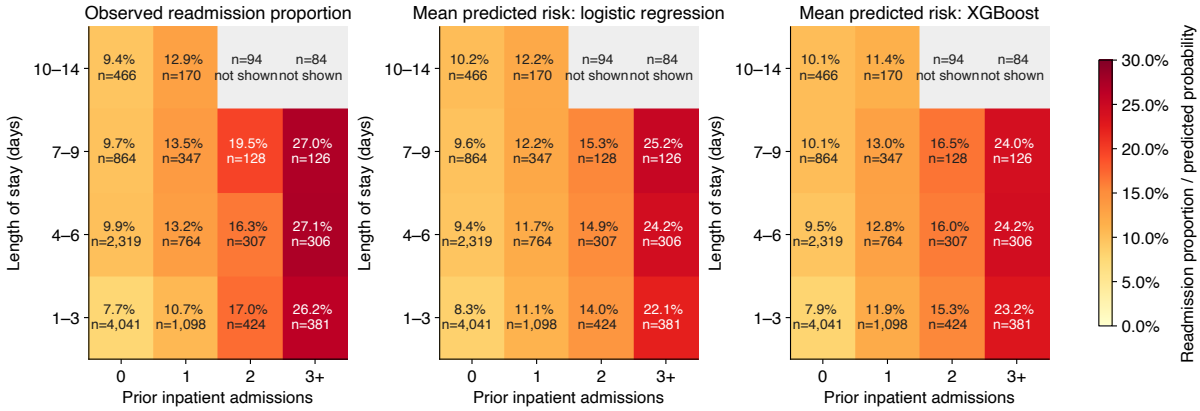


Fig. 16 Binned evaluation summaries by prior inpatient admissions and length of stay. Cells with fewer than 100 encounters are masked. Predictions average each encounter's full-model estimate; other characteristics differ between cells. Cells report descriptive point estimates for observed encounter groups

restricted to these encounters. Paired AUC differences for refitting are −0.0004 (95% conditional interval: −0.0010, 0.0004) and −0.0026 (−0.0088, 0.0038), respectively. Thus, this sensitivity check does not demonstrate a clear AUC change from refitting after the exclusions. These paired contrasts differ from comparing results across the full and restricted evaluation populations.

Removing discharge disposition lowers the refitted LR and XGB point estimates to AUCs of 0.625 and 0.626. In the eligible-cohort XGB model, prior inpatient admissions and discharge disposition remain the two largest grouped mean absolute

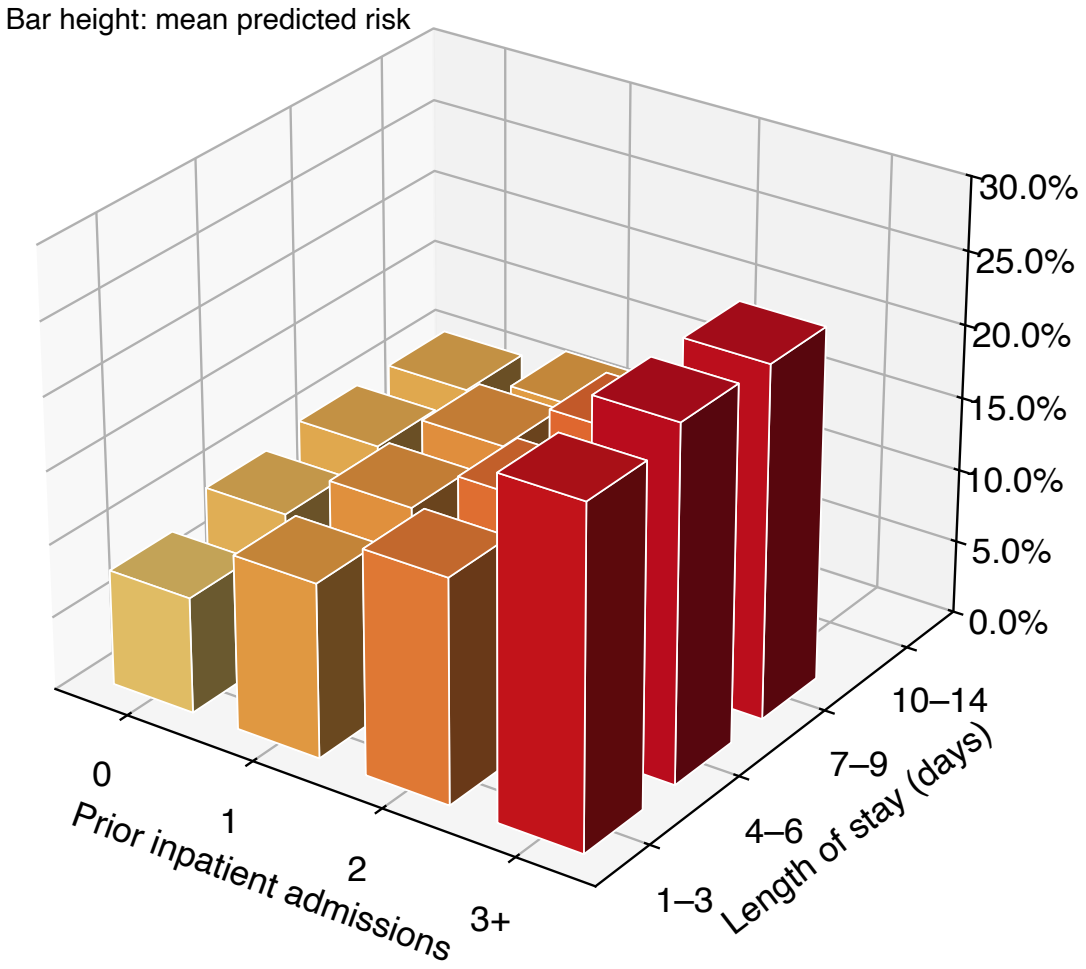


Fig. 17 Three-dimensional rendering of the XGBoost panel in the two-dimensional risk-map figure. Bar height shows mean calibrated prediction in observed groups with at least 100 encounters. Category positions are equally spaced for display; bin widths differ. Each bar represents a descriptive point estimate for its observed bin. Use the two-dimensional risk-map figure for precise comparison and sample sizes

SHAP contributions, at 0.298 and 0.180 raw log-odds units. The latter therefore reflects information beyond the excluded death and hospice categories, although attribution remains descriptive and model-dependent. Full metrics, conditional intervals, and grouped attributions are retained in Online Resource 1.

4.9 Exploratory seven-model comparison

Table 10 and Figures 20–21 summarize the seven fixed configurations on the common evaluation set. MLP has the largest point-estimate ROC-AUC (0.660), followed by XGB (0.657), RF (0.654), LR (0.653), DT (0.640), SVM (0.581), and KNN (0.555). These point estimates compare the specified configurations: SVM uses a smaller

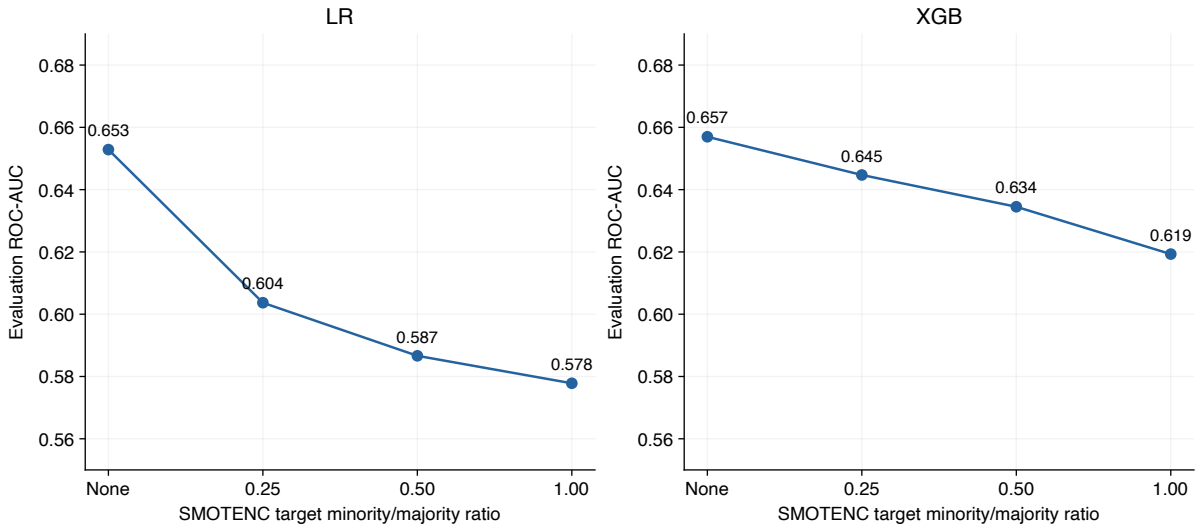


Fig. 18 Evaluation ROC-AUC under the tested SMOTENC ratios on identical original evaluation patients. None denotes the unresampled model. Points are descriptive; paired conditional intervals are provided in Online Resource 1. Ratios were not selected using evaluation performance

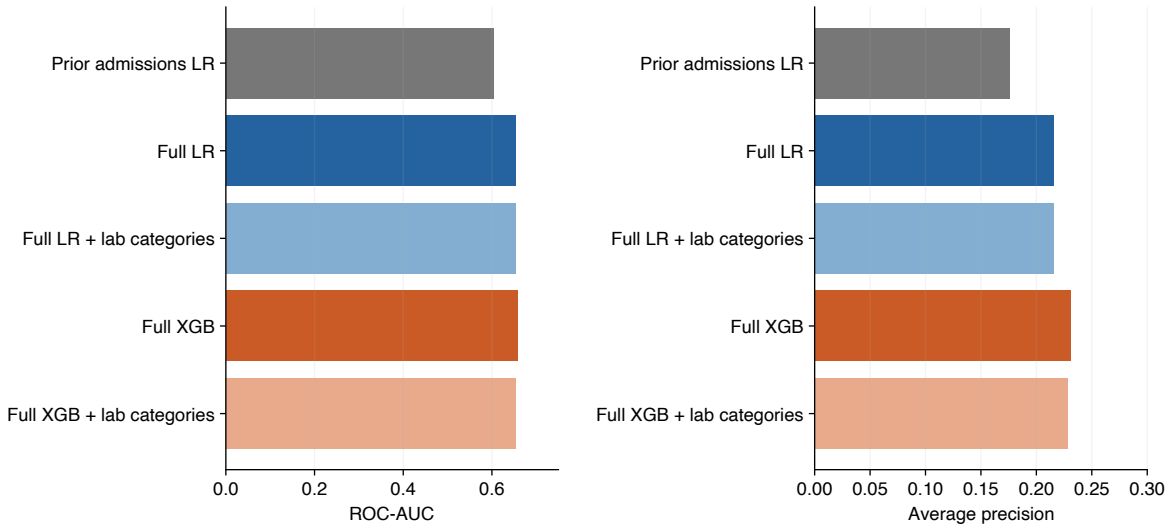


Fig. 19 Simple LR reference, full-feature baselines, and laboratory-category sensitivity on identical evaluation patients. All models are separately calibrated. Axis ranges differ between AUC and AP; point estimates alone do not replace paired uncertainty

Table 9 Discharge-eligibility sensitivity on the same 11,610 evaluation encounters. Historical models retain their original fitting and calibration; eligible models are refitted and recalibrated after exclusions. Thresholds are chosen using the corresponding calibration sets.

Model	Fitting/predictors	AUC	AP	Brier	Recall	FPR
LR	Historical, all predictors	0.646	0.217	0.0987	0.256	0.105
LR	Eligible, all predictors	0.646	0.217	0.0987	0.254	0.103
LR	Eligible, no discharge predictor	0.625	0.199	0.0997	0.222	0.101
XGB	Historical, all predictors	0.650	0.232	0.0979	0.264	0.104
XGB	Eligible, all predictors	0.647	0.235	0.0978	0.274	0.102
XGB	Eligible, no discharge predictor	0.626	0.195	0.0998	0.214	0.102

fitting sample, and the primary unresampled LR–XGB comparison includes paired uncertainty estimates.

The ordering also depends on the metric. XGB has AP 0.231, compared with 0.214 for MLP, despite MLP's slightly larger ROC-AUC. At the model-specific calibration-selected thresholds, MLP and XGB have F1 scores of 0.258 and 0.257, respectively. This illustrates why a single metric should not determine a preferred model without a defined use case. Both conventional 0.5 and calibration-selected operating points are tabulated in Online Resource 1.

4.10 Model structure and feature attribution

Figure 22 shows the first three levels of the fitted depth-six decision tree. Displayed numerical thresholds are standardized units. Category labels denote binary one-hot indicators, rather than ordered category codes. This truncated display illustrates the fitted model structure.

The RF impurity-importance plot (Figure 23) and grouped XGB SHAP plot (Figure 24) both place prior inpatient admissions and discharge disposition prominently. For XGB, the mean absolute grouped SHAP values for prior inpatient admissions, discharge disposition, laboratory-procedure count, primary diagnosis, and medication count are approximately 0.286, 0.199, 0.091, 0.081, and 0.066 raw log-odds units. The complete grouped attribution table is retained in Online Resource 1. These are descriptive properties of the fitted model and selected explanation sample.

Figure 25 displays the distribution of contributions across 2,000 evaluation encounters. Native XGBoost scoring and attribution preserve the sparse representation used during fitting; additivity is checked against the raw margin, and native predictions are checked against the saved model probabilities. The color display uses numerical values or categorical indicators. Contributions decompose the fitted raw log-odds scores into feature-level attributions.

5 Discussion

The central contribution of this study is a controlled audit of how validation and resampling choices shape reported performance in imbalanced readmission prediction. The repeated donor-pool experiment holds real validation observations and synthetic training counts fixed, allowing the effect of validation-donor contamination to be examined within the tested pipeline. Across patient-level partitions, contamination increased mean cross-validation ROC-AUC by approximately 0.005 for logistic regression and 0.012 for XGBoost. This controlled contrast complements the patient-overlap audit and separates a specific source of optimistic validation from the broader changes made during reanalysis.

Patient separation defines the population to which an evaluation applies. In the initial split, 46.08% of test encounters belonged to patients also represented in training. The corrected partitions evaluate previously unseen patients. However, the repeated audit did not demonstrate a systematic test-AUC penalty from patient separation, and the full difference from the initial workflow cannot be attributed to overlap alone. Preprocessing, resampling, and representation also changed. The contribution is therefore a documented examination of evaluation choices, rather than an assumption that every correction must reduce performance.

The primary comparison illustrates why discrimination, calibration, and operating-point performance should be assessed together. Logistic regression and XGBoost achieved ROC-AUCs of approximately 0.653 and 0.657. XGBoost improved average precision and Brier score in the conditional paired comparison, while its ROC-AUC advantage remained uncertain. At thresholds selected on calibration patients under a 10% false-positive-rate constraint, recall was approximately 0.255 and 0.263, respectively. These results distinguish the ability to rank encounters from the probability estimates and decisions obtained at a specified operating point.

Sampling-ratio and cohort analyses further show that imbalance treatment must be evaluated as part of the complete workflow. Lower SMOTENC ratios improved on full balancing but did not outperform the corresponding unresampled baselines. Retaining laboratory-result categories did not improve discrimination on the same evaluation patients. The expanded-cohort experiments assess a different evaluation population and therefore provide sensitivity evidence rather than paired improvements over the original cohort.

Table 10 Secondary seven-model point estimates on the common evaluation set. F1 and recall use the calibration-selected threshold t. The SVM has a smaller fitting sample. AP denotes average precision.

Model	Fit encounters	AUC	AP	Brier	t	Recall	F1
LR	35,421	0.653	0.215	0.0967	0.163	0.255	0.248
XGB	35,421	0.657	0.231	0.0959	0.173	0.263	0.257
DT	35,421	0.640	0.199	0.0967	0.178	0.224	0.246
RF	35,421	0.654	0.216	0.0966	0.169	0.252	0.247
KNN	35,421	0.555	0.133	0.1003	0.137	0.137	0.151
SVM	20,000	0.581	0.178	0.0999	0.115	0.208	0.209
MLP	35,421	0.660	0.214	0.0966	0.179	0.268	0.258

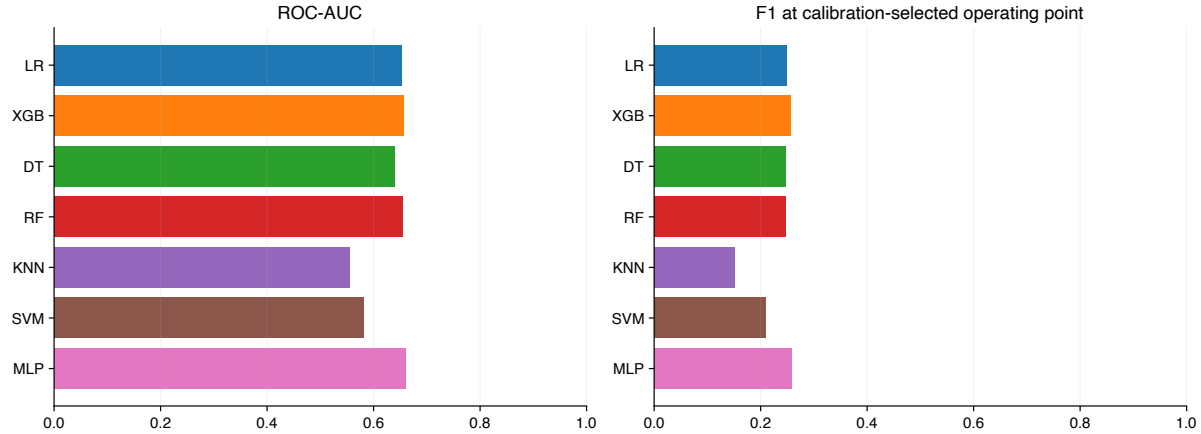


Fig. 20 Seven fixed model configurations on a common patient-disjoint evaluation set. F1 uses model-specific thresholds chosen on calibration patients. SVM uses at most 20,000 fitting encounters; MLP epochs use a grouped inner stopping partition. Results describe the seven specified configurations

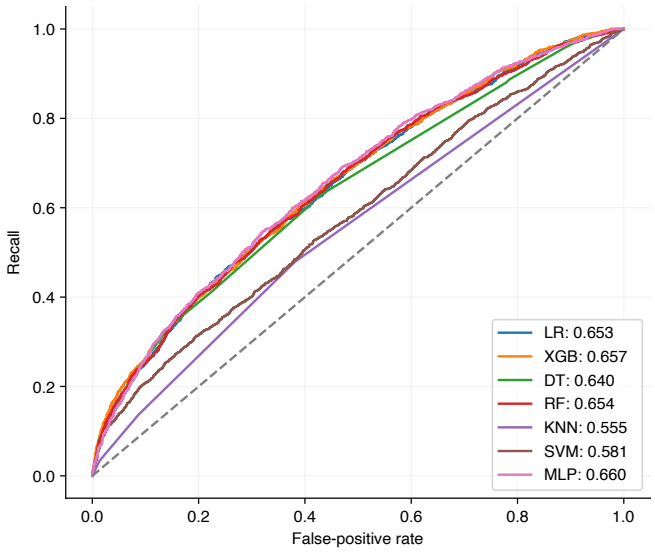


Fig. 21 ROC curves for the seven fixed configurations. Legends show evaluation ROC-AUC. Calibration is fitted on different patients from evaluation. Curves describe ranking performance in the evaluation sample

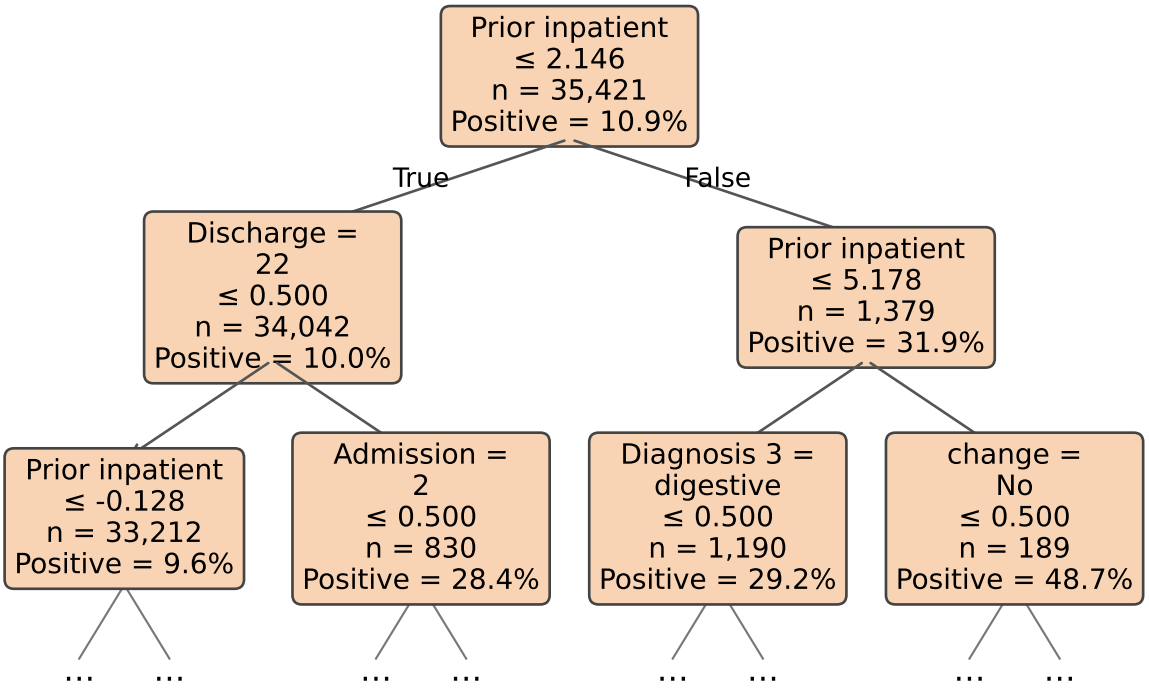


Fig. 22 Top three levels of the depth-six decision tree fitted to the training patients, with minimum leaf size 50. Split values for numerical features are standardized units. Nodes show training counts and the observed positive-outcome fraction. Display truncation does not change the fitted model. Categorical labels denote one-hot indicators; a 0.5 split tests absence versus presence of the named category

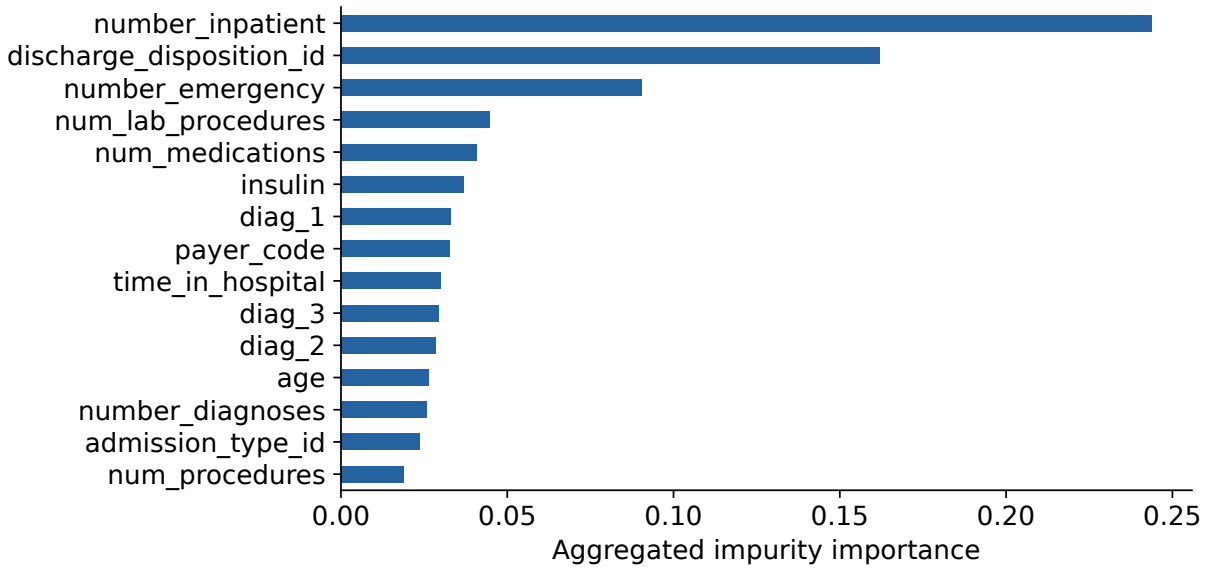


Fig. 23 Random-forest impurity-based importance, summed across one-hot columns belonging to an original feature. This training-based diagnostic describes fitted-model importance and can favor variables with many possible splits

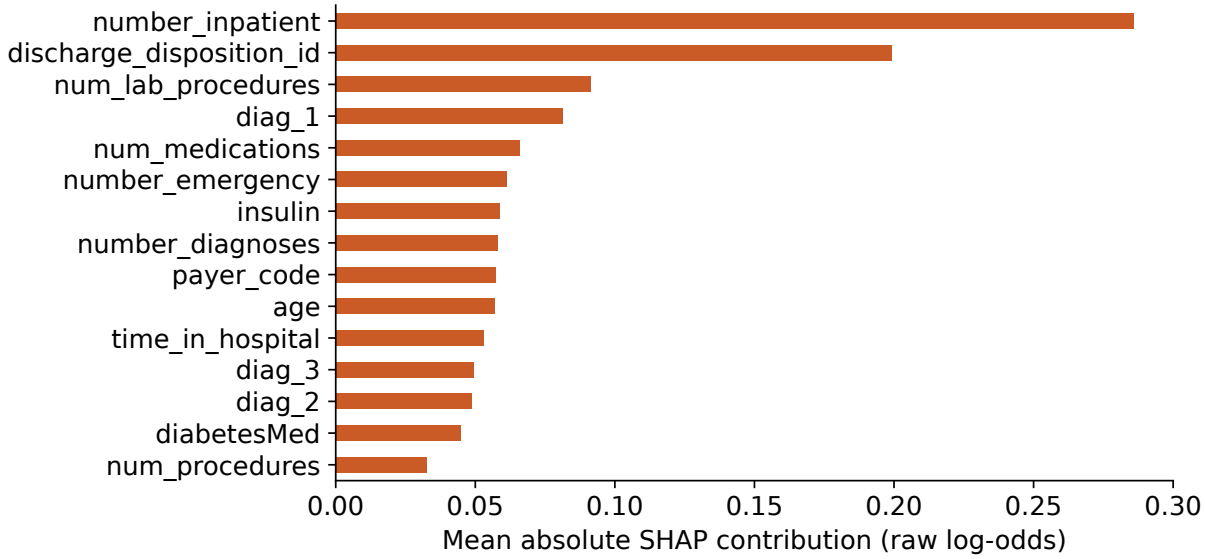


Fig. 24 Exact tree-path SHAP contributions for a seeded random sample of 2,000 evaluation encounters, explaining the uncalibrated XGBoost margin in log-odds units. One-hot contributions are summed within each original variable before taking mean absolute values. Attributions characterize feature contributions to the fitted raw margin

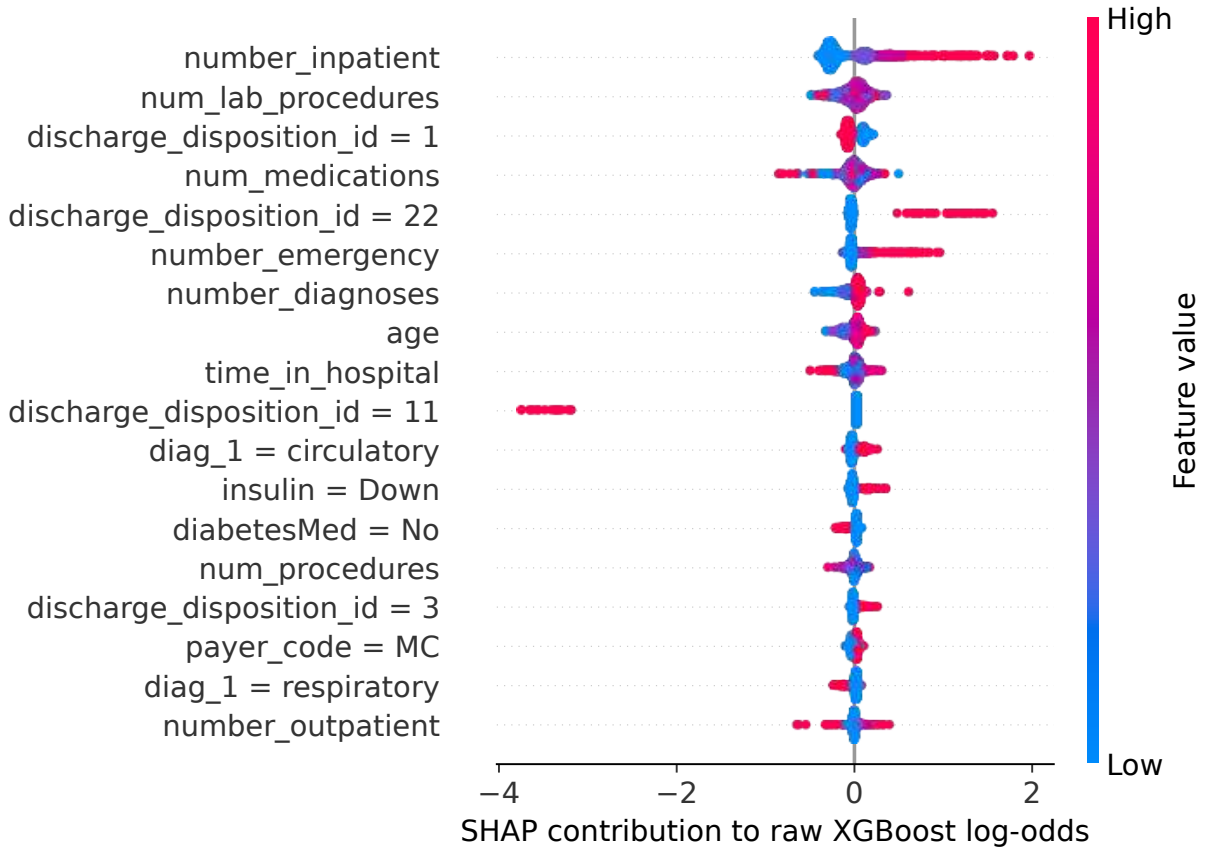


Fig. 25 SHAP distribution for the same 2,000 evaluation encounters. Features are transformed model inputs: numerical values are standardized and categorical levels are one-hot indicators. Color denotes the transformed input value; contributions explain the raw log-odds margin, not calibrated probabilities. Correlated predictors may share attribution

Figures 22–25 complement these comparisons by describing model structure and attribution. The decision-tree display illustrates how prior admissions and encounter characteristics enter the fitted decision rules. Random-forest importance and grouped XGBoost SHAP values both highlight prior inpatient admissions and discharge disposition. Their agreement indicates that these variables are prominent in the fitted models, although impurity importance and SHAP quantify different properties. The SHAP distribution additionally shows variation in feature contributions across evaluation encounters. These complementary diagnostics connect aggregate predictive performance with feature-level fitted-model behavior.

The discharge-eligibility analysis adds a distinction between changing the evaluation population and refitting the workflow. On identical eligible evaluation encounters, refitting after death and hospice exclusions yields small AUC differences with intervals spanning zero. Discharge disposition remains prominent after the exclusions, and removing it lowers discrimination. These results characterize the contribution of discharge-time information to prediction within the historical cohort.

The study integrates several complementary levels of retrospective internal evaluation. The primary comparison uses a fixed patient-disjoint partition, while the 20-partition experiment measures the sensitivity of the controlled donor-pool contrast to partition choice. Patient-cluster intervals quantify evaluation-sample uncertainty conditional on the fitted models, calibration maps, and thresholds. Pointwise subgroup estimates are interpreted alongside their sample sizes. Laboratory-category, payer-handling, and discharge-eligibility analyses characterize the dependence of results on cohort and representation choices. Together, these components support conclusions about discharge-time prediction and evaluation design within the historical populations and configurations studied.

6 Conclusion

This study contributes a reproducible audit of evaluation choices in imbalanced hospital readmission prediction. Its controlled, repeated donor-pool experiment isolates the effect of validation-donor contamination within a specified pipeline by holding validation observations and synthetic training counts fixed. Patient-disjoint evaluation, separate probability calibration, patient-cluster uncertainty estimation, and sampling and cohort sensitivity analyses provide complementary evidence on how reported performance depends on the analytical design.

The findings extend beyond classifier ranking by connecting predictive performance with validation integrity, probability estimation, operating-point behavior, and model interpretation. The discharge-eligibility sensitivity further separates cohort composition from model refitting. Together, the experiments and visual diagnostics provide a documented approach for examining imbalanced prediction workflows and interpreting their results. The resulting evidence provides a reproducible methodological foundation for evaluating imbalanced readmission workflows in historical patient data.

Statements and Declarations

Funding

The authors received no funding for this work.

Competing interests

The authors declare that they have no financial or non-financial competing interests.

Author contributions

Mohammad Rabiul Hossen Chowdhury performed the data analysis. Shomi Aktar developed the research design and modeling workflows. Both authors prepared, read, and approved the final manuscript.

Ethics approval and consent

This study uses the publicly distributed, de-identified Diabetes 130-US Hospitals dataset. No new participant recruitment or contact was undertaken. No institutional ethics approval or exemption determination is reported for this secondary analysis.

Data availability

The source dataset is available from the UCI Machine Learning Repository at https://doi.org/

10.24432/C5230J. The data-file checksum, cohort and partition records, evaluation predictions, and supporting analysis records are available from the corresponding author upon reasonable request. Complete numerical summary tables are supplied in Online Resource 1.

Code availability

The code used for data preprocessing, model development, evaluation, and figure generation is available from the corresponding author upon reasonable request.

Supplementary information

Online Resource 1 (ESM_1.zip): Complete numerical results in 16 CSV files, with a README explaining the table contents and reporting conventions.

References

[1] Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2

[2] Chawla NV, Bowyer KW, Hall LO, et al (2002) SMOTE: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research 16:321–357. https://doi.org/10.1613/jair.953

[3] Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794, https://doi.org/10.1145/2939672.2939785

[4] Clore J, Cios K, DeShazo J, et al (2014) Diabetes 130-US Hospitals for Years 1999–2008. https://doi.org/10.24432/C5230J, URL https://doi.org/10.24432/C5230J, dataset

[5] Collins GS, Moons KGM, Dhiman P, et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378

[6] Emi-Johnson OG, Nkrumah KJ (2025) Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data. Cureus 17(4):e82437. https://doi.org/10.7759/cureus.82437

[7] Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330, URL https://proceedings.mlr.press/v70/guo17a.html

[8] Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18(17):1–5. URL https://www.jmlr.org/papers/v18/16-365.html

[9] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp 4765–4774

[10] 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):e0118432. https://doi.org/10.1371/journal.pone.0118432

[11] Strack B, DeShazo JP, Gennings C, et al (2014) Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Research International 2014:781670. https://doi.org/10.1155/2014/781670

[12] Vandewiele G, Dehaene I, Kovács G, et al (2021) Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artificial Intelligence in Medicine 111:101987. https://doi.org/10.1016/j.artmed.2020.101987

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip