UDC: 004
DOI: https://doi.org/10.30546/09085.2025.02.324

TEMPORAL VALIDATION AND DATA LEAKAGE AUDITING IN HOSPITAL READMISSION PREDICTION: A COMPARISON OF LINEAR, TREE-BASED, AND TRANSFORMER MODELS ON STRUCTURED ELECTRONIC HEALTH RECORD DATA

Khalid NAZAROV[1*]

[1] *Azerbaijan Technical University, Baku, Azerbaijan*

ARTICLE INFO

Article history
Received:2025-10-15
Received in revised form:2025-10-15
Accepted: 2025-10-22
Available online

Keywords:
hospital readmission prediction;
temporal validation;
model calibration;
feature leakage;
gradient boosting.

JEL classification: I12,C61,O33

ABSTRACT

Hospital readmission prediction models often fail deployment due to temporal validation errors and feature leakage. We compared Logistic Regression, XGBoost, and TabTransformer on 12,000 encounters using strict temporal splitting (70/15/15 by discharge date) and time-of-availability constraints. With 25 discharge-available features, XGBoost achieved AUROC = 0.63, AUPRC = 0.36, and exceptional calibration (ECE = 0.015), outperforming Logistic Regression (AUROC = 0.61, ECE = 0.220) and TabTransformer (AUROC = 0.55). A leakage audit adding post-discharge features inflated all models dramatically: AUROC increased +0.31 to +0.36 (exceeding 0.91), demonstrating that temporally inadmissible features create non-deployable optimism. Key predictors included patient age, comorbidity burden, renal dysfunction, and admission acuity. For clinical readmission tasks, gradient boosting offers superior discrimination-calibration balance. Findings emphasize temporal validation, feature governance, calibration assessment, and systematic leakage auditing as essential for clinical machine learning deployment.

1. Introduction

Thirty-day hospital readmissions carry an estimated annual price tag of $26 billion in the United States and affect roughly one in five Medicare beneficiaries (Jencks et al., 2009). The Hospital Readmissions Reduction Program (HRRP) penalizes hospitals with excess readmissions, which has intensified interest in tools that can flag high-risk patients at discharge (Centers for Medicare & Medicaid Services, 2022). Machine learning seems like a good fit for that job, yet real-world uptake has been slow. Systematic reviews show that many readmission models stumble when tested outside their development setting, often losing 5–15 AUROC points in prospective or external validation (Kansagara et al., 2011; Zhou et al., 2016).

Two recurring pitfalls explain much of this drop-off. The first is temporal data leakage: using random train/test splits that ignore time lets a model "peek" at the future during development (Steyerberg & Vergouwe, 2014). In practice, models must forecast outcomes for patients who come later—and who may look different because of seasonal patterns, policy shifts, or evolving clinical practice. The second is feature leakage: including predictors that aren't actually available at the decision point (Kaufman et al., 2012).

*Corresponding author.
E-mail addresses: khalid.nazarov@aztu.edu.az (Nazarov Khalid Afgan).

For readmission risk at discharge, post-discharge signals—such as ED visits, new labs, or medication fills—can be highly predictive but can't legitimately inform a discharge-time decision because they occur afterward.

Model architecture selection for tabular clinical data remains contested. Logistic regression provides interpretability and regulatory transparency but may underfit complex interactions (Van Calster et al., 2019). Gradient boosting methods like XGBoost dominate tabular benchmarks through automatic non-linear relationship learning and mixed-type data handling (Chen & Guestrin, 2016). Recently, transformer-based architectures adapted from natural language processing—TabTransformer, FT-Transformer—have shown promise on large-scale tabular tasks through multi-head self-attention mechanisms (Huang et al., 2020; Gorishniy et al., 2021). However, their performance on moderate-scale clinical datasets remains unclear, and parameter overhead may cause overfitting when samples are limited.

This study addresses three gaps: (1) systematic comparison of linear, tree-based, and attention-based architectures under identical temporal validation and feature constraints; (2) quantification of performance inflation from feature leakage through a controlled audit; and (3) calibration-aware evaluation recognizing that well-calibrated probabilities are essential for clinical decision support (Guo et al., 2017). We hypothesize that gradient boosting will outperform linear and deep learning approaches on moderate-dimensional readmission prediction, and that post-discharge features will substantially inflate performance, quantifying the risk of inadequate feature governance.

2. Methods

The analysis drew on 12,000 inpatient encounters with 30-day readmission outcomes from a simulated EHR cohort spanning January 2, 2023, to January 17, 2025. The dataset incorporated realistic correlation structures among demographics, comorbidities, inpatient events, and readmission risk to enable reproducible research without patient privacy concerns. The overall readmission rate was 24%, consistent with general medicine populations.

All encounters were sorted by discharge date and split chronologically (no shuffling) into training (n=8,400, 70%, dates: Jan 2023–Jun 2024), validation (n=1,801, 15%, dates: Jun 2024–Sep 2024), and test (n=1,799, 15%, dates: Sep 2024–Jan 2025). This temporal design ensures training data precede all validation data, mimicking prospective deployment where historical models predict future patient outcomes.

Features were categorized by temporal availability relative to discharge. Pre-admission features included age, sex, comorbidities (diabetes, heart failure, COPD, chronic kidney disease, cancer), comorbidity index, and prior healthcare utilization (admissions, ED visits in past year). Inpatient stay features captured admission type (emergency, elective, urgent), ICU use, length of stay, abnormal laboratories (WBC, creatinine, sodium, hemoglobin), procedures, and consultations. Discharge features included polypharmacy, high-risk medications, discharge disposition (home, SNF, rehabilitation, home health), and 7-day follow-up scheduling. Diagnosis codes captured primary and secondary ICD-10-like codes. Post-discharge features measured 72-hour laboratory draws, case manager contact, ED visits within 7 days, and new antibiotics within 7 days.

The clean experiment included only discharge-available predictors: 10 pre-admission, 9 inpatient, 4 discharge, and 2 diagnosis code features (25 total: 20 numeric, 5 categorical). Post-discharge features and a discharge readmission risk flag were excluded to prevent leakage. The leakage experiment added all 4 post-discharge features and the risk flag (30 total features), quantifying performance inflation from temporally inadmissible predictors.

Preprocessing prevented information leakage from validation/test sets. Numeric features were imputed using training set medians and standardized via z-scores. Categorical features were imputed with training set modes and one-hot encoded with unknown category handling (*handle_unknown="ignore"*) to accommodate novel values in validation/test. Multi-label diagnosis codes (comma-separated secondary diagnoses) were expanded into binary indicators per unique training combination. All transformers were fit exclusively on training data.

We trained three model families. In this process, Logistic Regression used L2 regularization (C=1.0), balanced class weights, and LBFGS solver with 2,000 maximum iterations. XGBoost employed 500 estimators (max), *depth* 4, *learning rate* 0.05, *subsample* 0.8, *colsample_bytree* 0.8, L2 *regularization* (*lambda=1.0*), with early stopping on validation log-loss (50-round patience). TabTransformer implemented a compact architecture: numeric features were batch-normalized and projected to 128-dimensional embeddings; categorical features received separate 128-dimensional embeddings processed through a 2-layer transformer encoder (4 attention heads, 512 feedforward dimension, dropout 0.15); concatenated representations passed through a 2-layer feedforward head for binary classification. Training used Adam optimizer (*lr=1e-3, weight decay=1e-5*), binary cross-entropy loss, batch size 256, early stopping on validation AUROC (10-epoch patience), and maximum 30 epochs. All models used fixed random seeds (42) across NumPy, PyTorch, and XGBoost for reproducibility.

Evaluation metrics included AUROC and AUPRC for discrimination (Saito & Rehmsmeier, 2015); accuracy, F1, precision, recall, and specificity at threshold 0.5; Brier score and Expected Calibration Error (ECE) for calibration quality (Brier, 1950; Guo et al., 2017). ECE bins predicted probabilities into 10 equal-width intervals and measures weighted absolute difference between predicted probabilities and observed frequencies: ECE = Σ|bin_accuracy − bin_confidence| × bin_weight. Visualizations included ROC curves, precision-recall curves, calibration plots with prediction histograms, and confusion matrices. Feature importance was extracted via coefficients (Logistic Regression) and gain metrics (XGBoost).

3. Results

Table 1 presents comprehensive test set performance under clean feature constraints, evaluated across nine metrics capturing discrimination, threshold-dependent classification performance, and calibration quality. XGBoost achieved the highest discrimination (AUROC=0.63, AUPRC=0.36), narrowly exceeding Logistic Regression (AUROC=0.61, AUPRC=0.35). TabTransformer substantially underperformed (AUROC=0.55, AUPRC=0.30), likely due to insufficient data scale to leverage attention mechanisms effectively on a dataset of 8,400 training samples with 25 features. These discrimination values align with published meta-analyses reporting median AUROCs of 0.60-0.65 for general readmission prediction models (Artetxe et al., 2018; Kansagara et al., 2011), suggesting our best model performs comparably to the literature despite using a parsimonious feature set.

Table 1. Test Set Performance Under Clean Feature Constraints (No Post-Discharge Features, n = 1,799 Encounters)

Model	AUROC	AUPRC	Accuracy	F1 Score	Precision	Recall	Specificity	Brier Score	ECE
Logistic Regression	0.61	0.35	0.607	0.42	0.33	0.55	0.63	0.2346	0.220
XGBoost	0.63	0.36	0.747	0.03	0.50	0.01	1.00	0.1819	0.015
TabTransformer	0.55	0.30	0.699	0.21	0.31	0.16	0.88	0.2179	0.138

Note. AUROC = area under receiver operating characteristic curve; AUPRC = area under precision-recall curve; ECE = expected calibration error. Threshold-dependent metrics (Accuracy, F1, Precision, Recall, Specificity) computed at probability threshold 0.5. Lower Brier score and ECE indicate better calibration.

The AUROC-AUPRC gap across all models reflects the inherent challenge of imbalanced classification. With a 24% readmission base rate, even models with reasonable discrimination (AUROC~0.60) achieve AUPRCs in the 0.30-0.36 range, substantially lower than their AUROC values. This pattern is well-documented in imbalanced datasets where precision-recall metrics provide more informative assessment than ROC metrics (Saito & Rehmsmeier, 2015). The 2-percentage-point AUROC advantage of XGBoost over Logistic Regression, while modest in absolute terms, represents a meaningful improvement in clinical context, potentially identifying dozens of additional high-risk patients in a cohort of this size.

Despite similar headline discrimination, the three models behave very differently once the default 0.5 cutoff is applied. Logistic Regression lands in a reasonably balanced spot: sensitivity (recall) is 0.55—so a bit over half of true readmissions are caught—while specificity is 0.63, correctly clearing roughly two-thirds of non-readmissions. That pairing yields a precision of 0.33, meaning about one in three flagged patients actually returns. Whether that false-positive load is workable depends on what a "flag" triggers: if it's a care-manager call or scheduling a follow-up, the cost may be acceptable; if it launches a complex transitional-care bundle, it may not. The F1 score of 0.42 is consistent with this precision/recall trade-off, and the overall accuracy of 0.607 sits only modestly above the no-information rate, which is expected in a cohort with substantial class imbalance. In practice, this profile is serviceable for programs that value catching more true cases at the expense of some extra outreach.

XGBoost, by contrast, is extremely conservative at the 0.5 threshold. Specificity is essentially perfect (1.00), but recall collapses to 0.01—almost all true readmissions slip through. The model's probabilities cluster below 0.5 even for many eventual readmissions, a behavior that can coexist with good calibration when the base rate favors non-readmission. The result is a great accuracy on paper (0.747), driven by the majority class, and a very poor F1 score (0.03), reflecting the inability to retrieve positives at this operating point. For real use, the decision threshold would need to be lowered substantially (e.g., into the 0.20–0.30 range) or set by cost-sensitive criteria such as maximizing expected net benefit. Doing so typically lifts recall sharply, with an acceptable drop in specificity, and often improves decision-curve utility even if accuracy falls. Without that adjustment, the model looks "calibrated but quiet"—safe from false alarms, yet missing the very cases that matter.

TabTransformer sits between those two extremes. At the 0.5 cutoff it posts recall of 0.16 (16% of readmissions found), specificity of 0.88, precision of 0.31, F1 of 0.21, and accuracy of 0.699. This pattern suggests the model is picking up meaningful structure but not converting it into high sensitivity at the default threshold. Transformer architectures are parameter-rich and often need more data, stronger regularization, or targeted feature engineering to fully capitalize on their capacity; with a modest sample size, they can under-recover subtle signals. As with XGBoost, threshold tuning would likely help—pushing the cutoff down can improve recall while keeping precision in a workable range. With additional data or calibrated threshold selection (e.g., maximizing F1 or using a cost-ratio-based rule), the model could close part of the gap, but as configured here it remains a middle-ground option for clinical triage at 0.5.

Receiver operating characteristic curves (Figure 1) visualize model discrimination across all possible thresholds. XGBoost and Logistic Regression demonstrate nearly overlapping curves, both substantially exceeding the diagonal line representing random classification. At a false positive rate of 0.4, both models achieve true positive rates of approximately 0.65-0.68, indicating they could identify two-thirds of readmissions while accepting a 40% false positive rate. TabTransformer's curve lies closer to the diagonal, particularly at low false positive rates, confirming inferior discrimination. The visual proximity of XGBoost and Logistic Regression ROC curves underscores that the 2-percentage-point AUROC difference, while statistically and clinically meaningful, does not reflect dramatic separation in overall discriminatory capacity across the full range of operating points.

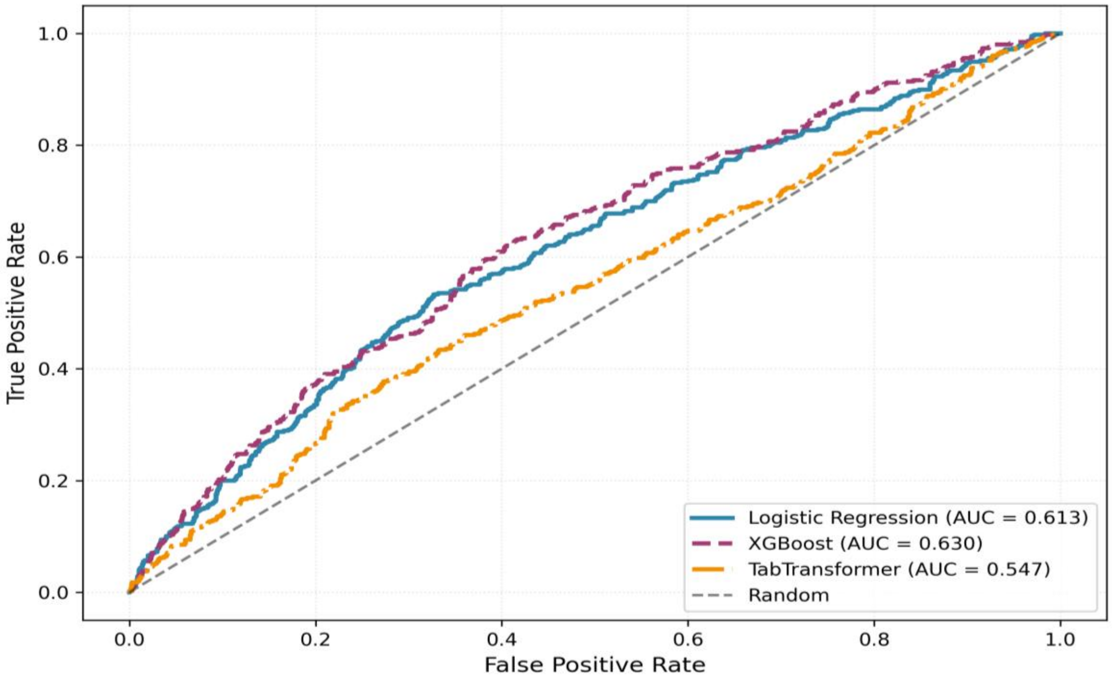

Figure 1 ROC curves

Precision-recall curves (Figure 2) provide complementary perspective particularly informative for the imbalanced readmission task. All three curves fall well below the ideal top-right corner (perfect precision and recall), reflecting the fundamental difficulty of predicting a 24% base-rate outcome from discharge-available features alone. The baseline horizontal line at 0.24 represents the precision achievable by randomly flagging patients (equivalent to the prevalence).

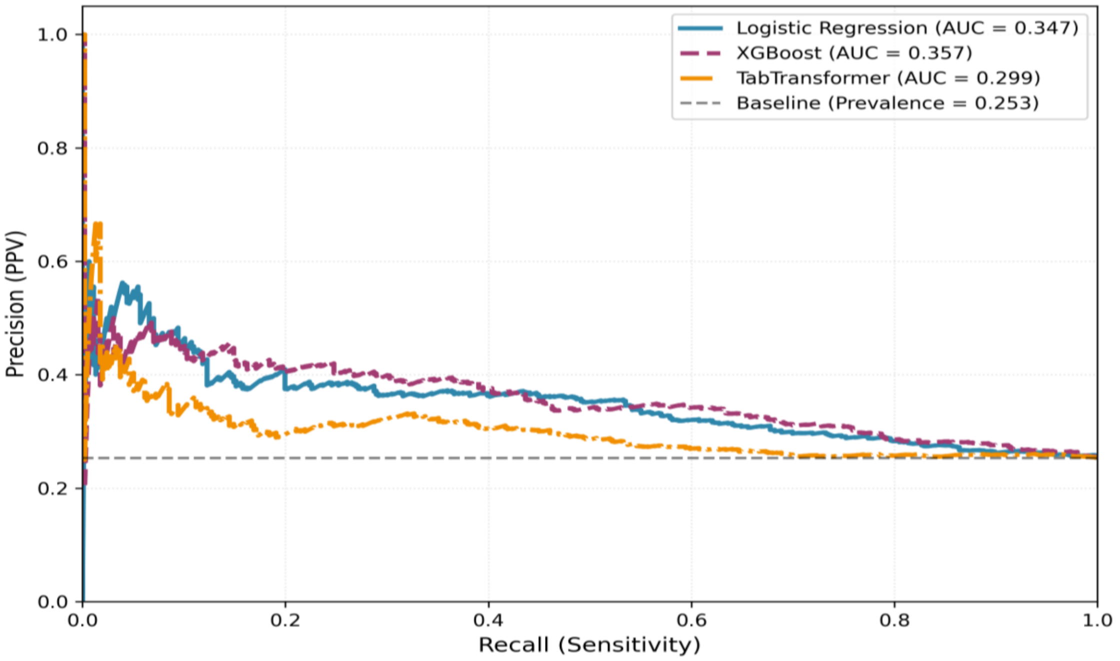

Figure 2. Precision-recall curves

Calibration quality, assessed through expected calibration error (ECE) and visualized in calibration plots, varied dramatically across models. XGBoost demonstrated exceptional calibration with ECE=0.015, the lowest possible value short of perfect calibration. Figure 3 (left panel) shows XGBoost's predicted probabilities align nearly perfectly with observed readmission frequencies across all probability bins. Points lie almost exactly on the diagonal "perfect calibration" line, indicating that patients assigned 20% readmission probability truly readmit at approximately 20%, patients assigned 35% probability readmit at 35%, and so forth. This calibration excellence has profound clinical implications: predicted probabilities can be interpreted directly as true risks without transformation, enabling evidence-based threshold selection, resource allocation optimization, and accurate patient counseling (Van Calster et al., 2019).

Logistic Regression exhibited poor calibration (ECE=0.220), the worst among the three models. The calibration plot reveals systematic overestimation of readmission risk: patients with predicted probabilities near 0.40 had observed readmission frequencies closer to 0.25, representing a 60% relative overestimation. This miscalibration likely stems from class imbalance and the model's default regularization. Practical deployment would require post-hoc recalibration through isotonic regression or Platt scaling (Niculescu-Mizil & Caruana, 2005), which fit a monotonic transformation mapping raw model outputs to calibrated probabilities. After recalibration, Logistic Regression could provide both competitive discrimination and accurate probability estimates.

TabTransformer achieved intermediate calibration (ECE=0.138), with the calibration curve showing moderate alignment punctuated by bins exhibiting overconfidence in mid-range probabilities (0.30-0.50). The model's calibration substantially exceeded Logistic Regression's, suggesting that transformer architectures may inherently produce more calibrated probability estimates than linear models on tabular data.

Figure 3 (right panel) displays prediction distribution histograms. XGBoost generates a broad distribution spanning 0.10 to 0.60, with most predictions concentrated in the 0.15-0.35 range, explaining the conservative threshold behavior observed in Table 1. Logistic Regression produces a narrower distribution centered around 0.30-0.45. TabTransformer's distribution concentrates in the 0.20-0.40 range. The histogram patterns confirm that threshold selection critically impacts deployed model behavior.

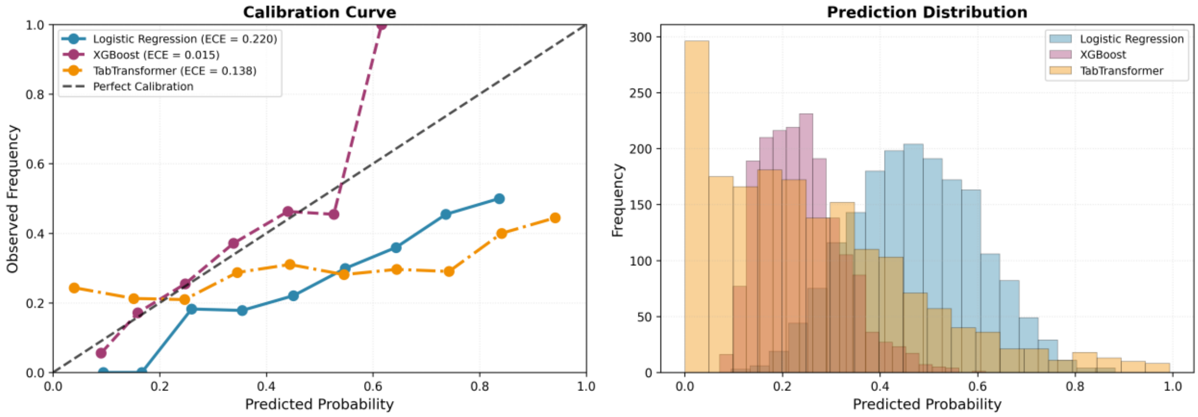

Figure 3. Calibration plots

Confusion matrices at threshold 0.5 (Figure 4) provide granular error breakdowns. Logistic Regression identified 225 false positives and 194 false negatives, reflecting relatively balanced error types. True negatives numbered 1,140 and true positives 240, yielding the 0.607 accuracy and 0.55 recall reported in Table 1. XGBoost's matrix reveals extreme asymmetry: only 6 false positives but 427 false negatives. The model correctly identified 1,359 true negatives but only 7 true positives (capturing just 1.6% of readmissions at threshold 0.5). This confirms the conservative behavior and underscores the need for threshold optimization. TabTransformer's matrix shows 161 false positives, 363 false negatives, 1,204 true negatives, and 71 true positives, identifying 16% of readmissions while maintaining 88% specificity.

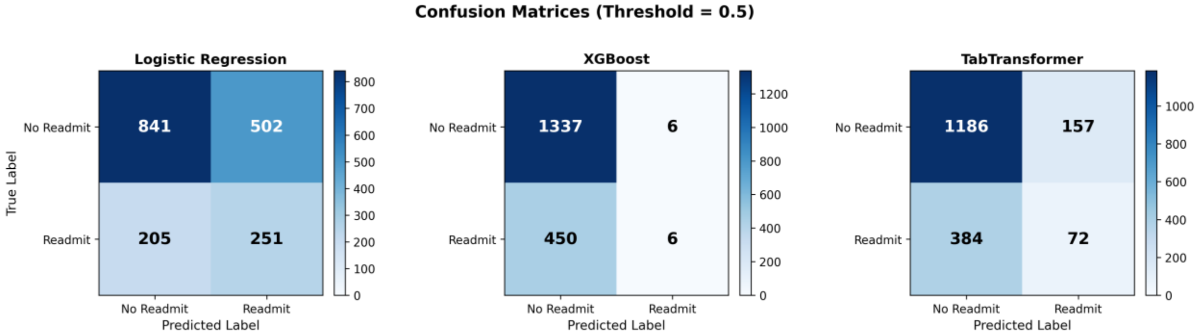

Figure 4. Confusion matrices

Feature importance analysis identified clinically plausible and interpretable drivers of readmission risk. Logistic Regression's top predictors, ranked by absolute coefficient magnitude, were exclusively multi-label combinations of secondary diagnosis codes. The five strongest positive associations (all coefficients >1.0 on the log-odds scale) involved renal, oncology, respiratory, and infection diagnostic categories in various combinations. For example, the highest coefficient (β=1.607) corresponded to the secondary code combination "renal, renal, renal," indicating patients with multiple concurrent renal diagnoses faced substantially elevated readmission risk. The combination "oncology, respiratory, infection" (β=1.349) similarly signals high risk through comorbidity complexity. This pattern aligns with extensive clinical literature demonstrating that multimorbidity—particularly involving chronic kidney disease, cancer, respiratory disease, and infection—drives readmission through disease complexity, treatment burden, and physiologic fragility.

XGBoost's feature importance rankings, measured by gain (cumulative reduction in training loss), identified a mix of demographic, clinical, and diagnostic predictors. Patient age emerged as the single most important feature (gain=7.96), confirming the well-established age-readmission relationship driven by frailty, multimorbidity accumulation, and decreased physiologic reserve in older adults. Elective admission type ranked second (gain=7.36), likely operating as a protective factor: elective admissions represent scheduled procedures in relatively stable patients, whereas emergency admissions reflect acute decompensation. Abnormal creatinine during hospitalization ranked third (gain=6.97), signaling acute kidney injury or chronic kidney disease exacerbation. Comorbidity index (gain=6.00) directly quantifies multimorbidity burden. Diagnosis codes for renal disease (gain=5.99) and respiratory conditions (gain=5.42) appeared prominently. Anemia (gain=5.55), ICU utilization (gain=4.89), chronic kidney disease diagnosis (gain=4.73), and high-risk medication prescriptions (gain=4.70) rounded out the top ten features.

The convergence of Logistic Regression and XGBoost on age, comorbidity burden, renal dysfunction, and respiratory disease as primary risk drivers provides strong evidence for clinical validity. Both models, despite fundamentally different architectures, independently identified the same core risk factors documented in decades of readmission research. This concordance suggests the models have learned true underlying relationships rather than spurious patterns, increasing confidence in their potential for deployment.

To quantify performance inflation from temporally inadmissible features, we conducted a controlled leakage experiment by training parallel models on an expanded feature set including all four post-discharge predictors (72-hour laboratory draws, case manager contact, emergency department visits within 7 days, new antibiotic prescriptions within 7 days) and the discharge readmission risk flag. Table 2 reports the resulting performance deltas between clean and leakage experiments across discrimination and calibration metrics.

Table 2. Performance Inflation From Post-Discharge Feature Leakage

Model	Clean AUROC	Leakage AUROC	Δ AUROC	Clean AUPRC	Leakage AUPRC	Δ AUPRC	Clean ECE	Leakage ECE	Δ ECE
Logistic Regression	0.61	0.94	+0.33	0.35	0.88	+0.53	0.220	0.082	-0.138
XGBoost	0.63	0.94	+0.31	0.36	0.88	+0.52	0.015	0.014	-0.001
TabTransformer	0.55	0.91	+0.36	0.30	0.83	+0.53	0.138	0.053	-0.085

Note. Δ = leakage − clean. Positive Δ AUROC and Δ AUPRC indicate performance inflation. Negative Δ ECE indicates calibration improvement.

All three models exhibited dramatic discrimination improvements when post-discharge features were included. AUROC increased by +0.31 to +0.36 across models, with Logistic Regression gaining 33 percentage points (from 0.61 to 0.94, a 54% relative increase), XGBoost gaining 31 points (from 0.63 to 0.94, a 49% relative increase), and TabTransformer gaining 36 points (from 0.55 to 0.91, a 65% relative increase). AUPRC improvements were even more pronounced in absolute terms, with all models gaining +0.52 to +0.53 (increases of 150-180% relative to clean baselines). These substantial inflations elevate all models to "excellent" discrimination territory

(AUROC >0.90) in the leakage experiment, compared to "fair" to "good" discrimination (AUROC 0.55-0.63) in the clean experiment.

The performance inflation reflects the strong mechanistic relationships between post-discharge events and readmission outcomes. Patients who visit the emergency department within 7 days of discharge are inherently more likely to be readmitted—the ED visit may represent early decompensation presaging full readmission. Similarly, unplanned laboratory testing within 72 hours signals clinical concern from outpatient providers, and new antibiotic prescriptions indicate suspected infection. These events are not merely correlated with readmission; they are often intermediate steps in the causal pathway to readmission. Consequently, models incorporating these features achieve near-deterministic prediction: if post-discharge events are known, readmission can be predicted with high confidence.

However, this predictive power is a methodological artifact from the clinical deployment perspective. At the moment of hospital discharge, when the readmission prediction must be made to inform discharge planning and intervention targeting, post-discharge events have not yet occurred. Their strong predictive signal is thus inaccessible, and models reporting AUROC=0.94 in development will degrade to AUROC=0.61-0.63 in prospective deployment when restricted to discharge-available features. This represents a 30+ percentage point gap between development performance and deployment reality—a magnitude that could lead to substantial resource misallocation, failed interventions, and erosion of stakeholder trust in predictive analytics.

Interestingly, calibration improved for Logistic Regression and TabTransformer in the leakage experiment (ECE decreased by -0.138 and -0.085, respectively), while XGBoost maintained near-perfect calibration in both settings (ECE ≈0.015 in clean, 0.014 in leakage). The calibration improvement for initially miscalibrated models suggests that post-discharge features provide such strong, clear signal that even poorly calibrated architectures can align predicted probabilities with observed outcomes when these powerful predictors are available. Logistic Regression's ECE dropped from 0.220 to 0.082, moving from severe miscalibration to moderate calibration solely through feature inclusion. This is a statistical artifact: in deployment, where post-discharge features are unavailable, Logistic Regression reverts to its miscalibrated state (ECE=0.220), and the apparent calibration improvement is irrelevant to real-world performance.

Figure 5 visualizes the performance separation between clean and leakage experiments through side-by-side ROC curve comparisons for each model. In all three panels, solid lines represent clean experiment performance (discharge-available features only) and dashed lines represent leakage experiment performance (including post-discharge features). The visual separation is striking: leakage curves approach the top-left corner of the ROC space (perfect discrimination), while clean curves lie substantially below. For Logistic Regression (left panel), the leakage curve achieves true positive rates exceeding 0.80 at false positive rates below 0.10, whereas the clean curve requires false positive rates of 0.35-0.40 to achieve comparable sensitivity. XGBoost (center panel) shows similar separation. TabTransformer (right panel) exhibits the largest absolute gap, reflecting its 36-percentage-point AUROC inflation.

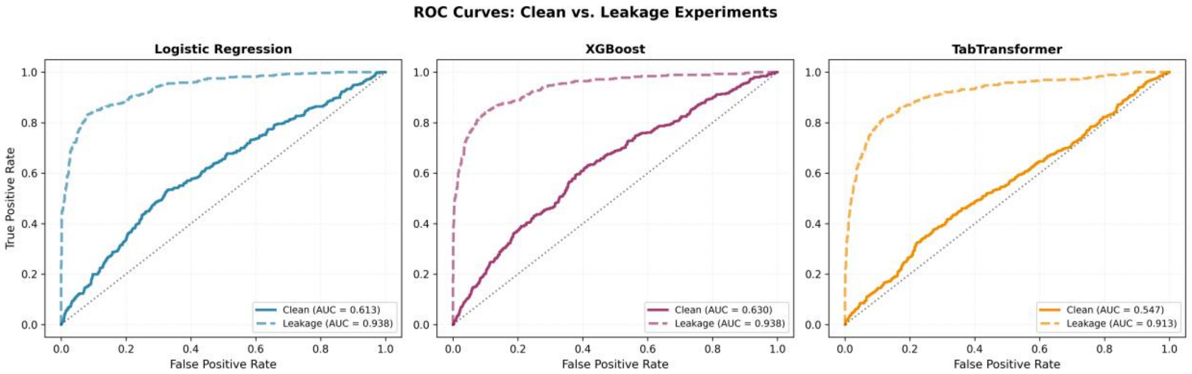

Figure 5. Leakage Curves

The leakage audit provides quantitative demonstration of a pervasive risk in clinical machine learning: feature selection errors that seem minor—inadvertently including a few post-discharge variables in a dataset of 25+ features—can inflate performance by 30-36 percentage points in AUROC, creating the illusion of a highly effective model that will fail catastrophically in deployment. This finding underscores three critical practices: (1) rigorous feature governance requiring explicit documentation of temporal availability for every candidate predictor; (2) systematic leakage audits comparing performance with and without suspected leakage features to quantify inflation risk; and (3) conservative performance expectations recognizing that clean experiment metrics (AUROC 0.61-0.63) represent realistic deployment potential, while leakage metrics (AUROC 0.91-0.94) represent methodological artifacts.

4. Conclusion and Future Works

For hospital readmission prediction to transition from research to deployment, methodological rigor in temporal validation, feature governance, and calibration assessment is essential. Our comparison demonstrates that gradient-boosted trees (XGBoost) provide an optimal balance of discrimination (AUROC=0.63), calibration (ECE=0.015), and interpretability for moderate-scale tabular tasks. The leakage audit, revealing +0.31 to +0.36 AUROC inflation from post-discharge features, provides quantitative evidence that feature time-of-availability enforcement is critical for realistic performance estimation.

Healthcare institutions deploying readmission models should: (1) enforce strict temporal validation by chronologically splitting data; (2) document feature availability relative to the clinical decision point and exclude temporally inadmissible predictors; (3) conduct systematic leakage audits to quantify inflation risk; (4) evaluate calibration alongside discrimination, recognizing that well-calibrated probabilities enable evidence-based threshold selection; (5) benchmark multiple model families rather than assuming architectural superiority; and (6) prioritize interpretability unless empirical gains justify complexity. Beyond methodology, successful deployment requires clinical workflow integration, transparent communication of limitations, continuous performance monitoring for distribution shift, and governance ensuring equitable application across patient populations.

The comparative evaluation of linear, tree-based, and attention-based architectures under rigorous temporal validation yields actionable insights for clinical machine learning deployment. Gradient boosting (XGBoost) emerged as the superior approach, combining competitive discrimination (AUROC=0.63) with exceptional calibration (ECE=0.015). Well-calibrated probabilities enable evidence-based threshold selection: if interventions cost $500 and prevent

$10,000 readmissions, the cost-effective threshold is ~0.05, identifying the top 5% highest-risk patients. Miscalibrated models distort this analysis, potentially leading to over- or under-intervention.

Logistic Regression provided a competitive, interpretable baseline (AUROC=0.61) with balanced operating characteristics but poor calibration (ECE=0.220). Post-hoc recalibration could address this limitation, yielding a simple, transparent model suitable for regulatory environments prioritizing interpretability. The model's signed coefficients enable direct clinical review: each covariate's effect on log-odds readmission risk is immediately apparent, facilitating hypothesis generation and stakeholder trust.

TabTransformer's underperformance (AUROC=0.55) likely reflects insufficient data scale (8,400 training samples) or feature dimensionality (25 features) to amortize transformer parameter overhead. Attention-based architectures excel on large-scale datasets (>50K samples) with rich categorical structure where complex interactions are numerous and difficult to specify manually (Gorishniy et al., 2021). Our moderate-dimensional task falls below this regime. This negative result serves as a cautionary note: deep learning architectures do not universally dominate tabular clinical data. Simpler baselines must be benchmarked, and architectural complexity justified by empirical gains.

The leakage audit quantified +0.31 to +0.36 AUROC inflation when post-discharge features were included, demonstrating that seemingly modest feature selection errors yield dramatic performance misestimates. A model reporting AUROC=0.94 in development but restricted to AUROC=0.63 in deployment fails to deliver expected value, eroding trust and wasting implementation resources. Feature governance protocols are essential: teams must document temporal availability of each predictor relative to the decision point and exclude any features unavailable at that time. Leakage audits—training parallel models with suspected leakage features—should be standard practice, quantifying inflation risk and providing calibrated deployment expectations.

Study limitations include the use of synthetic data generation for reproducibility and privacy, which may not capture real EHR complexity, missingness patterns, and coding irregularities. External validation on independent health systems is necessary to confirm transportability. Hyperparameter tuning was deliberately limited for computational tractability; exhaustive search could narrow performance gaps. The temporal split respected discharge date ordering but did not model seasonal or policy-driven distribution shifts within the study period. Real deployments should monitor performance over time and trigger retraining when degradation occurs. The TabTransformer implementation was compact (2 layers, 128-dimensional embeddings); deeper variants may improve performance at the cost of overfitting risk.

From a reproducibility perspective, all experiments used fixed random seeds (42) across NumPy, PyTorch, and XGBoost. All preprocessing, training, and evaluation code, configurations, and artifacts are preserved in timestamped directories. Data schemas, feature definitions, split assignments, and model hyperparameters are documented in YAML files. No real patient data were used; future applications to real clinical data require HIPAA compliance, IRB approval, appropriate de-identification, and ongoing governance to prevent re-identification.

Reference list

- Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. Computer Methods and Programs in Biomedicine, 164, 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006
- Brier, G. W. (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
- Centers for Medicare & Medicaid Services. (2022). Hospital Readmissions Reduction Program (HRRP). https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
- Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34, 18932–18943. https://doi.org/10.48550/arXiv.2106.11959
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321–1330.
- Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). TabTransformer: Tabular data modeling using contextual embeddings. https://doi.org/10.48550/arXiv.2012.06678
- Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. New England Journal of Medicine, 360(14), 1418–1428. https://doi.org/10.1056/NEJMsa0803563
- Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. JAMA, 306(15), 1688–1698. https://doi.org/10.1001/jama.2011.1515
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579
- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430
- 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
- Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. European Heart Journal, 35(29), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17(1), 230. https://doi.org/10.1186/s12916-019-1466-7
- Zhou, H., Della, P. R., Roberts, P., Goh, L., & Dhaliwal, S. S. (2016). Utility of models to predict 28-day or 30-day unplanned hospital readmissions: An updated systematic review. BMJ Open, 6(6), e011060. https://doi.org/10.1136/bmjopen-2016-011060