Development and External Validation of Multimodal Machine Learning Models to Predict High Inpatient Opioid Exposure

Machine Learning Models for High Inpatient Opioid Exposure

Authors and Contributions:

Swara Kale (corresponding author) (B.S.)[a] - Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Writing - original draft

Dr. Devender Singh (Ph. D)[b] - Supervision; Validation; Writing - review & editing

Dr. Eeric Truumees (MD, attending physician)[b] - Supervision; Validation; Writing - review & editing

Dr. Matthew Geck (MD, attending physician)[b] - Supervision; Validation; Writing - review & editing

Dr. John Stokes, MD (MD, attending physician)[b] - Supervision; Validation; Writing - review & editing

Affiliations:

a - The University of Texas at Austin, 101 E 21st Street, Austin, Texas 78712, United States; swarak627@gmail.com

b - Ascension Texas Spine and Scoliosis, 1004 W 32nd St #200, Austin, TX 78705, United States;

ABSTRACT

High inpatient opioid exposure is associated with increased risk of persistent opioid use, yet early identification of high-risk patients during hospitalization remains limited. We developed and evaluated machine learning models to predict extreme opioid exposure using electronic health record data from MIMIC-IV. This retrospective cohort-based prediction modeling study included 223,452 unique first hospital admissions. The outcome was extreme opioid exposure, defined as the top decile of morphine milligram equivalents (MME) per day among opioid-exposed admissions (corresponding to ≥225 MME/day in the development cohort), representing 2.65% of all admissions.

Structured early-admission features included demographics, admission characteristics, laboratory utilization and abnormality summaries, and 24-hour procedural indicators. Discharge-note data were incorporated using ClinicalBERT embeddings and bigram features. Models were trained using an 80/10/10 split, with temporal validation performed on the most recent 10% of admissions. External validation was conducted using the MIMIC-III and eICU Collaborative Research Database cohorts. Performance was assessed using ROC-AUC and PR-AUC with 95% confidence intervals.

Among structured-only models, XGBoost achieved the best internal test performance (ROC-AUC 0.932 [0.924-0.940]; PR-AUC 0.223 [0.193-0.262]). A combined structured and notes model improved precision-recall performance (ROC-AUC 0.932 [0.920-0.943]; PR-AUC 0.276 [0.229-0.331]). Temporal validation showed similar discrimination (ROC-AUC 0.929; PR-AUC 0.223).

In external validation, performance decreased substantially. In MIMIC-III, the model achieved ROC-AUC 0.669 [0.659-0.680] and PR-AUC 0.018 [0.017-0.019], while in eICU performance was further attenuated (ROC-AUC 0.567 [0.556-0.576]; PR-AUC 0.018 [0.017-0.019]). Predicted probabilities were poorly calibrated in both external datasets, with limited correspondence between predicted and observed risk.

These findings demonstrate that while EHR-based machine learning models can achieve strong internal discrimination, their performance and calibration may degrade substantially across independent healthcare systems -- underscoring the need for dataset-specific validation and recalibration prior to clinical application.

AUTHOR SUMMARY

Opioid medications are commonly used in hospitals to treat pain, but some patients receive very high doses, which may increase their risk of long-term opioid use and dependence. Identifying these high-risk patients early during hospitalization could help doctors make safer prescribing decisions and improve pain management.

In this study, we analyzed electronic health record data from over 220,000 hospital admissions to develop machine learning models that estimate which patients are likely to receive high levels of opioids. We focused on information available within the first 24 hours of admission, including patient characteristics, laboratory testing patterns, and procedures. We also examined whether information from clinical notes could improve predictions.

We found that the models performed well within the original dataset, and that combining structured data with clinical notes improved performance. The patterns identified by the models -- such as links to surgical procedures and more intensive care -- were in line with expectations. However, when the models were tested in different hospital datasets, performance declined, and predicted risks no longer matched what actually happened to patients.

These findings highlight an important challenge in applying machine learning in healthcare: models that perform well in one setting may not work reliably in others. While routinely collected hospital data may help identify patients at risk early, predictive models must be carefully tested and adapted before they can be safely used in new clinical environments.

INTRODUCTION

Opioid use has been a critical public health concern in the United States for the past 30 years.[1, 2] Although overdose deaths have decreased in recent years, more than 68,000 individuals died from drug overdose between September 2024 and September 2025.[3, 4] Opioids may be obtained illicitly or through legitimate prescriptions. Among prescription medications, opioids remain the most frequently misused class, with more than 7.6 million prescription users transitioning to misuse annually.[5] In 2023 alone, 13,026 individuals died from prescription opioid overdose.[6]

Opioid administration during hospitalization is common, particularly following major surgical procedures and other high-intensity interventions.[7, 8] While short-term opioid therapy is

often clinically necessary for acute pain management, accumulating evidence indicates that inpatient opioid exposure may increase the likelihood of persistent opioid use after discharge and elevate the risk of long-term dependence or misuse.[9-11] Additionally, substantial variability exists in inpatient opioid prescribing intensity across geographic regions, institutions, and individual prescribers, even among patients with comparable clinical presentations.[12-14] This heterogeneity suggests that factors beyond clinical necessity contribute to high-dose exposure, with some hospitalizations experiencing disproportionately intense opioid administration.

Existing efforts to mitigate opioid-related harm have largely focused on outpatient prescribing practices, prescription drug monitoring programs, and community-level interventions.[15] Comparatively less attention has been directed toward identifying patients at risk for extreme opioid exposure during hospitalization, despite the hospital setting representing a critical and potentially modifiable point of care. A predictive model capable of identifying hospitalizations at risk for unusually high opioid dosing could enable targeted stewardship and earlier clinical review.

In this study, we developed, internally validated, and externally evaluated machine learning models to predict extreme inpatient opioid exposure using electronic health record (EHR) data. Extreme exposure was defined as the top decile of morphine milligram equivalents (MME) per hospital day among opioid-exposed admissions. Our primary structured model incorporated demographic, diagnostic, procedural, laboratory, and medication-derived features available early in hospitalization to estimate the probability of unusually high opioid intensity.

We additionally evaluated a notes-only model trained on clinical documentation to determine whether unstructured text independently contains predictive signals associated with opioid intensity. Furthermore, we constructed a multimodal model integrating structured features with note-derived embeddings to assess whether combining quantitative data with clinical language improved predictive performance. Exploratory analysis of clinical text was performed to provide interpretability of text-derived signals.

Finally, we evaluated model generalizability through external validation in independent datasets. Differences in patient populations, documentation practices, and data structure across healthcare systems may substantially impact model performance. By systematically comparing structured, text-based, and multimodal approaches, we aimed to examine how model discrimination and calibration change across heterogeneous clinical settings and to assess the extent to which predictive performance translates beyond the development dataset.

RESULTS

Participants and Cohort Characteristics:

A total of 546,028 hospital admission records were screened, of which 223,452 unique first admissions met eligibility criteria and were included in the final analytic cohort. Among these admissions, 4,627 (2.65%) met criteria for high-risk opioid exposure, defined as the top decile of daily morphine milligram equivalents (MME) among opioid-exposed admissions (corresponding to ≥225 mg/day in the development cohort). Overall, 59,441 admissions (26.5%) received any

opioid exposure during hospitalization. The distribution of MME per day among opioid-exposed admissions was right-skewed (median 54.7; IQR 19.5-122.1).

Admissions were randomly assigned into training (n = 178,762; 3,730 events), validation (n = 22,345; 465 events), and test (n = 22,345; 432 events) sets, with stable outcome prevalence across splits (training: 2.09%; validation: 2.08%; test: 1.94%).

Baseline characteristics stratified by outcome status are shown in Supplementary Table 1. Compared with non-top-decile admissions, high-risk admissions were younger and had higher laboratory utilization within 24 hours. Medication administrations within 24 hours were substantially higher in the high-risk group, and surgical admissions were more frequent among high-risk cases.

Missing Data and Feature Processing:

Medication and laboratory features had moderate missingness, which was addressed using indicator variables and imputation. No residual missing values remained after preprocessing.

Model Development and Evaluation:

Structured prediction models were developed using ridge logistic regression, random forest, and extreme gradient boosting. Models were trained on the training set, tuned using the validation set, and evaluated in the test set.

Model outputs were predicted probabilities of high-risk opioid exposure, and classification thresholds were selected using the Youden index in the validation set. Primary performance

metrics are summarized in Table 1. Threshold-dependent performance metrics are provided in Supplementary Table 2.

Table 1: Primary model performance

Dataset	Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Brier (95% CI)	Calibration intercept	Calibration slope
Validation	Ridge logistic	0.892	0.166	0.0187	—	—
Validation	Random forest	0.909	0.201	0.0181	—	—
Validation	XGBoost	0.918	0.216	0.0180	—	—
Validation	Notes-only	0.865	0.150	0.0196	—	—
Validation	Combined	0.933	0.262	0.0179	—	—
Test	Ridge logistic	0.915 (0.903-0.925)	0.185 (0.159-0.217)	0.0172 (0.0158-0.0188)	0.446	1.183
Test	Random forest	0.919 (0.909-0.928)	0.204 (0.174-0.240)	0.0169 (0.0155-0.0184)	0.083	1.094
Test	XGBoost	0.932 (0.924-0.940)	0.223 (0.193-0.262)	0.0166 (0.0152-0.0181)	0.123	1.063
Test	Notes-only	0.861 (0.840-0.881)	0.168 (0.134-0.208)	0.0189 (0.0169-0.0210)	-0.028	0.994
Test	Combined	0.932 (0.920-0.943)	0.276 (0.229-0.331)	0.0173 (0.0155-0.0191)	0.036	1.023

External (MIMIC-III)	XGBoost	0.669 (0.659-0.680)	0.018 (0.017-0.019)	0.0372 (0.0359-0.0388)	—	—
External (eICU)	XGBoost	0.567 (0.556-0.576)	0.018 (0.017-0.019)	0.0192 (0.0187-0.0197)	—	—

Model Performance: Structured-Only Models

In the independent test cohort, XGBoost achieved the best performance among structured models, with ROC-AUC 0.932 [0.924-0.940] and PR-AUC 0.223 [0.193-0.262], and demonstrated good internal calibration (slope 1.06; intercept 0.12). Random forest and ridge logistic regression showed slightly lower discrimination. Bootstrap resampling confirmed stability of performance estimates (Table 1) (Fig 1).

Model Interpretation: SHAP Analysis

SHAP analysis of the XGBoost model indicated that early inpatient utilization features were the strongest contributors to predicted risk, particularly medication event counts, age, and surgical status (Fig 2).

Notes-Only Model and Multimodal Integration:

Embeddings were available for 58.4% of admissions. The notes-only model achieved ROC-AUC 0.861 and PR-AUC 0.168 in the test subset. The combined structured + notes model improved performance (ROC-AUC 0.932; PR-AUC 0.276), representing substantial enrichment over baseline prevalence (Table 1) (Fig 3). However, because discharge summaries are generated

after hospitalization, these findings reflect retrospective signal capture rather than prospective prediction.

Contrastive Text Analysis:

Bigram analysis identified phrases enriched among high-risk admissions, including terms related to orthopedic and spine procedures (Supplementary Table 3) (Fig 4).

Subgroup Performance:

Model performance remained stable across surgical and non-surgical admissions, with differences in PR-AUC reflecting underlying outcome prevalence. Calibration remained stable across subgroups (Supplementary Table 4) (Fig 5).

Temporal Validation:

To assess temporal generalizability, we evaluated model performance in a temporally held-out cohort. In this cohort, XGBoost achieved ROC-AUC 0.929 and PR-AUC 0.223. Performance was similar to that observed in the primary test cohort (Table 1) (Fig 6).

External Validation:

In external validation, model performance decreased substantially compared with internal and temporal evaluation. In the MIMIC-III cohort (n = 46,476; outcome prevalence 2.6%), the model achieved a ROC-AUC of 0.669 [0.659-0.680] and PR-AUC of 0.018 [0.017-0.019]. In the eICU cohort (n = 200,859; outcome prevalence 1.6%), discrimination was further attenuated, with

ROC-AUC 0.567 [0.556-0.576] and PR-AUC 0.018 [0.017-0.019] approaching near-random discrimination (Table 1) (Fig 7). This represents a reduction in ROC-AUC of approximately 0.36 compared with internal test performance.

External calibration analysis demonstrated substantial degradation in predicted probability reliability in both external datasets. In MIMIC-III, predicted probabilities showed weak and non-monotonic association with observed outcomes, while in eICU, predicted risks exhibited minimal separation between high- and low-risk admissions. Calibration slopes were markedly reduced in both datasets, indicating severe distortion in probability scaling. These findings indicate that, although the model retained modest ranking ability, its probability estimates were not well calibrated in external settings (Fig 8).

DISCUSSION

In this study, we developed and validated machine learning models to identify hospital admissions at risk for extreme inpatient opioid exposure. Internal model performance demonstrated strong discrimination and suggested that routinely collected inpatient data captured broader signals of care complexity associated with high opioid exposure. However, this performance did not generalize to external datasets, where both discrimination and probability calibration declined substantially.

In clinical practice, hospitalized patients generate multimodal data streams. Structured data quantify measurable aspects of care, such as laboratory utilization, procedural burden, and medication administration, while unstructured notes capture physician interpretation,

procedural narratives, and contextual detail surrounding the patient's clinical course. These modalities represent complementary views of the same hospitalization. Our results suggest that structured utilization patterns provide a strong foundation for identifying high-intensity opioid exposure, while narrative documentation contributes additional context in specific settings, particularly surgical admissions.

Extreme opioid exposure was uncommon but clinically meaningful, occurring in a small minority of admissions. These cases were characterized by greater early laboratory utilization, higher medication event counts, and more frequent surgical procedures. Such patterns likely reflect more complex operative courses, trauma-related admissions, or high-acuity postoperative care -- contexts in which aggressive analgesia may be required.[16] Importantly, the model did not rely on opioid-specific inputs alone but instead captured broader signals of clinical intensity. This suggests that extreme dosing may be predictable from overall care trajectories rather than from pain management variables in isolation.

Subgroup analyses further illustrate how clinical context shapes predictive signals. Structured models demonstrated stable discrimination across surgical and non-surgical admissions, indicating that utilization-based features generalize across heterogeneous inpatient populations. In contrast, the notes-only model performed more strongly in surgical contexts and less consistently in non-surgical admissions. Surgical discharge summaries often contain detailed procedural narratives, implant descriptions, and postoperative plans, which may encode implicit signals of anticipated analgesic need. Medical admissions, by comparison, may contain more templated or symptom-focused language, potentially limiting the strength of

narrative-based prediction. The improved performance observed with multimodal integration suggests that quantitative measures and clinician documentation capture partially distinct aspects of care complexity.

Bigram analysis of discharge notes supports the clinical plausibility of the learned narrative signals. Terms enriched among high-risk admissions were predominantly related to orthopedic, trauma, and spine procedures, such as “external fixation” and “laminectomy fusion”. These clinical scenarios are commonly associated with substantial postoperative pain and extended analgesic requirements and have been previously linked to postoperative opioid prescribing.[17] Conversely, phrases enriched among low-risk admissions reflected general review-of-systems language or laboratory descriptors.

Temporal validation demonstrated that model performance remained stable when applied to more recent admissions, suggesting that the predictive relationships identified were not limited to a specific time period. Although prescribing behaviors and stewardship efforts continue to evolve nationally, the model’s relative stability across time supports the hypothesis that extreme opioid exposure reflects persistent underlying drivers of care intensity rather than transient prescribing variation.[18]

External validation revealed a marked and clinically meaningful decline in model performance, highlighting limited transportability across independent datasets. This degradation likely reflects multiple forms of dataset shift, including differences in patient case variety, clinical practice patterns, feature availability, and data representation. For example, variation in medication documentation, laboratory reference ranges, and procedural coding required

feature harmonization using proxy variables, potentially attenuating predictive signal. Additionally, differences in outcome prevalence and the use of dataset-specific percentile thresholds may have introduced label distribution shifts that further impacted performance.

Together, these results demonstrate limited transportability of the model across datasets and suggest that predictive performance is highly sensitive to differences in patient populations, feature representation, and prescribing practices. The relatively higher performance observed in MIMIC-III compared with eICU suggests partial generalizability across temporally distinct but structurally similar datasets, but poor generalization across more heterogeneous healthcare systems.

These findings suggest that a substantial portion of model performance in the development cohort may be driven by dataset-specific patterns rather than fully generalizable relationships. As a result, direct deployment of such models without recalibration may lead to inaccurate risk estimation in new clinical environments.

Methodologically, these results emphasize the importance of rigorous external validation in EHR-based machine learning studies. Models trained on a single dataset may overestimate performance due to implicit encoding of local practice patterns, documentation structures, and patient population characteristics. Future work should focus on domain adaptation, feature standardization, and recalibration strategies to improve generalizability across institutions.

From a clinical perspective, identifying admissions at risk for extreme opioid exposure has several potential applications. Because high inpatient dosing has been associated with

downstream opioid use, persistent exposure, and increased healthcare utilization, early identification may represent a modifiable intervention point.[19] Operationally, a model could be deployed at the 24-hour mark following admission, once early clinical data was available. This timepoint provides an opportunity to reassess analgesic strategies, optimize multimodal pain management, or involve specialty consultation. Given the low outcome prevalence, risk thresholds could be selected to flag only the highest-risk patients, thereby limiting alert burden while focusing attention on those most likely to experience extreme exposure. Potential future integration into clinical decision support systems could enable automated notification to care teams or pharmacy services. Importantly, predicted risk should support clinical judgment and proactive planning rather than mandate opioid restriction.

Additionally, opioid prescribing patterns vary across patient demographic characteristics.[20,21] Because predictive models may inadvertently reflect underlying practice patterns, careful evaluation across demographic subgroups is essential prior to clinical deployment. Future work should assess subgroup discrimination, calibration, and threshold performance to ensure equitable implementation and to avoid amplifying existing disparities in opioid prescribing, particularly among female patients and individuals from racial and ethnic minority backgrounds.

Limitations:

This study has several limitations. First, although external validation was performed using MIMIC-III and the eICU Collaborative Research Database, model performance declined

compared with internal and temporal validation, reflecting challenges in transportability across datasets with differing patient populations, feature availability, and prescribing practices. Second, discharge note embeddings were available for only a subset of admissions, introducing potential selection bias related to documentation availability and limiting generalizability of text-based models. Third, extreme opioid exposure was defined using a relative, dataset-specific threshold (top decile among opioid-exposed admissions), which may limit direct comparability across datasets and require recalibration in settings with different prescribing distributions. This approach also contributed to differences in outcome definition across external cohorts. Fourth, key clinical and contextual factors, including pain severity, intraoperative management, and provider-level prescribing behaviors, were not captured in the available data, potentially limiting model interpretability and performance. Finally, the low prevalence of extreme exposure limits positive predictive value, indicating that these models are best suited for risk stratification rather than definitive clinical decision-making.

MATERIALS AND METHODS

Data Source and Study Setting:

We conducted a retrospective cohort study using MIMIC-IV for the development and validation of prediction models. MIMIC-IV is a publicly available, de-identified electronic health record (EHR) database comprising detailed clinical data from patients admitted to Beth Israel Deaconess Medical Center, a large tertiary academic medical center in Boston, Massachusetts.[22-24] The MIMIC-IV hospital module contains structured clinical data from 223,452 unique patients between 2008 and 2022, including demographic characteristics,

laboratory measurements, procedures, medication administrations, and admission-level encounter data.

Free-text clinical documentation was obtained from the MIMIC-IV-Note module, which includes 331,793 de-identified discharge summaries from 145,915 patients during the same study period.[25] Discharge summaries were linked to structured admission-level data using the unique hospital admission identifier.

Study Population:

All inpatient hospital admissions recorded in the MIMIC-IV hospital module during the study period were eligible for inclusion. To avoid correlated observations from repeat hospitalizations, only the first hospital admission per patient was included in the analysis. No age-based exclusions were applied. The cohort was partitioned into training, validation, and test sets for model development, with temporal validation performed as a secondary analysis.

Outcome Definition:

The primary outcome was extreme inpatient opioid exposure within the first 24 hours of hospital admission. Opioid exposure was quantified as the total morphine milligram equivalents (MME) administered during this period, calculated using standardized equianalgesic conversion factors.[26]

Among admissions receiving any opioid within 24 hours (MME > 0), total MME values were ranked, and the 90th percentile was used to define extreme exposure. This threshold was calculated using the full cohort prior to model development and applied unchanged across all

data partitions. Because this approach yields dataset-specific absolute thresholds, direct comparability across external datasets may be limited.

Predictor Variables:

Predictor domains included demographic characteristics, laboratory utilization metrics, procedure indicators, and non-opioid medication administration counts. Demographic variables included age, sex, race, marital status, insurance type, primary language, admission type, admission location, calendar year of admission, and a categorical year grouping variable. Laboratory-related predictors included the total number of laboratory events within 24 hours, the number of laboratories with reference ranges available, counts of abnormal laboratory values (overall, low, and high), and the proportion of abnormal laboratory results. Procedure-related predictors included binary indicators for major procedural categories occurring within the first 24 hours (spine, orthopedic, neurologic, cardiac, abdominal, and vascular procedures), a composite indicator for any procedure, total procedure count, and a surgical admission flag. Medication-related predictors explicitly excluded opioid administrations and opioid-derived features. Variables directly representing opioid dose or opioid administration counts were excluded from predictors. In secondary analyses, free-text discharge summaries were used to construct text-based predictors. As discharge summaries are authored after hospitalization and may contain post-exposure information, text-based models were evaluated as retrospective comparators rather than real-time early prediction tools.

Data Preprocessing and Feature Engineering:

All preprocessing steps were performed after partitioning the dataset into training, validation, and test sets. Structured predictors were derived from data available within the first 24 hours of admission and aggregated at the admission level.

Laboratory-related variables were treated as structurally zero when no laboratory values were recorded during the 24-hour window. Indicator variables were created to denote admissions with no laboratory measurements and no laboratory results with reference ranges.

For remaining numeric predictors, missing values were imputed using median imputation. Median imputation values were derived from the training dataset and applied unchanged to validation and test sets.

Categorical predictors were encoded as factor variables. Missing or empty categorical values were assigned to a separate “Unknown” category. Factor levels were defined using the training dataset and applied consistently to validation and test datasets to ensure alignment across partitions.

All analyses were conducted in R (version 4.4.2) and Python (version 3.12.10) using ranger, glmnet, xgboost, and HuggingFace Transformers libraries.[28-32]

Discharge Notes Preprocessing:

Text representations were generated using a pretrained ClinicalBERT model, a domain-adapted transformer model trained on clinical text.[33] Embeddings for validation and test sets were generated without exposure to outcome labels and without updating model weights.

A front-back truncation strategy was applied. For each note, two tokenized segments were generated: 1) the first 256 tokens of the note (front segment) and 2) the last 256 tokens of the note (back segment).

Each segment was independently processed using ClinicalBERT, and the embedding corresponding to the [CLS] token was extracted as a 768-dimensional representation. Front and back embeddings were concatenated to produce a 1,536-dimensional vector for each admission-level note. Resulting embeddings were merged with structured admission-level predictors for downstream modeling.

Data Partitioning:

The cohort was randomly partitioned into training (80%), validation (10%), and test (10%) sets using stratified sampling to preserve outcome prevalence across partitions. The validation set was used for hyperparameter tuning and model selection. The test set was held out and used exclusively for final performance evaluation. The random stratified split served as the primary evaluation framework, with temporal validation conducted as a secondary analysis to assess robustness to distributional shift.

Models:

Three model classes were evaluated for structured predictors: random forest, penalized logistic regression, and gradient boosting. For text-only models, admission-level ClinicalBERT embeddings were used as predictors in a ridge-regularized logistic regression model. For combined models, structured predictors and text embeddings were concatenated and modeled

using ridge-regularized logistic regression to stabilize estimation in the presence of high-dimensional features and to reduce overfitting. The XGBoost model was selected as the primary model for external validation based on internal performance.

Hyperparameter Tuning:

Random forest models were trained using the ranger implementation with 500 trees. Class imbalance was addressed using inverse-prevalence class weights computed from the training data. The regularization parameter (λ) for ridge-regularized logistic regression models was selected using 5-fold cross-validation. For gradient boosting models, hyperparameters were tuned using the validation set, including learning rate, tree depth, subsampling, and column sampling, with early stopping based on validation ROC-AUC. For text-only and combined models, ridge regularization was selected via cross-validation on the training set.

Model Evaluation:

Model performance was evaluated on the held-out test set. Discrimination was assessed using the area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC), the latter included given the relatively low prevalence of the outcome. 95% confidence intervals were estimated using nonparametric bootstrapping with 2,000 resamples of the test set.

Overall prediction error was quantified using the Brier score. Calibration in external datasets was assessed using calibration curves and calibration slope and intercept.

Model Interpretation:

To characterize feature contributions, SHAP values were computed for the XGBoost model to quantify the relative importance and directionality of individual predictors in relation to high opioid exposure.[34]

Subgroup Analysis:

To evaluate performance across clinically distinct admission types, analyses were conducted for global, surgical, and non-surgical cohorts. A binary surgical admission flag was derived using International Classification of Diseases (ICD) procedure codes recorded for each hospital admission.[35, 36] Using the surgical flag, separate models were trained and evaluated within the surgical-only and non-surgical-only cohorts. The original global exposure threshold was retained for subgroup analyses. The same modeling pipeline, hyperparameter selection strategy, and evaluation framework described for the global models were applied to each subgroup. For each subgroup, discrimination, calibration, and clinical utility metrics were reported using the same procedures applied to the full cohort.

Temporal Validation:

To evaluate robustness to temporal dataset shift, a time-based split was performed using admission timestamp, with the most recent 10% of admissions reserved as a temporally held-out test cohort. Imputation parameters, categorical level definitions, and any high-cardinality category handling were derived from the temporal training set and applied unchanged to the temporal validation and test sets to prevent information leakage.

Random forest, ridge-regularized logistic regression, and XGBoost models were retrained using the temporal training set. Hyperparameter selection and early stopping procedures mirrored those used in the primary random split analysis. Final performance was evaluated on the temporal test set (latest 10% of admissions) using matching discrimination and calibration metrics.

Contrastive Text Analysis:

Exploratory contrastive text analysis was performed on discharge summaries to identify language patterns associated with high versus low opioid exposure. Discharge notes were tokenized and converted into bigrams. Tokens were restricted to alphabetic words 3-25 characters in length, with standard English stopwords and a curated set of clinical stopwords removed. Repeated-word bigrams were excluded, as were bigrams occurring fewer than 100 times overall. For each retained bigram, a log-odds ratio with Dirichlet smoothing was computed to quantify enrichment in high- versus low-opioid groups.

External Validation:

To evaluate model generalizability across independent datasets, external validation was performed using the MIMIC-III and eICU Collaborative Research Database cohorts.[37-40]

For MIMIC-III, an ICU-based cohort was constructed using first ICU stays per patient. The final cohort included 46,476 ICU admissions, with an outcome prevalence of 2.6% (n = 1,207). For eICU, we utilized the multicenter ICU cohort comprising 200,859 admissions across diverse hospital systems, with an outcome prevalence of 1.6% (n = 3,218).

Predictor variables for external validation were derived using data available within the first 24 hours of ICU admission, consistent with the feature definitions used in the MIMIC-IV development cohort. Due to differences in database structure and variable availability, predictors were harmonized across datasets using analogous data elements where direct equivalents were unavailable. Laboratory features were derived from 24-hour lab records as counts of measurements and abnormal values, with abnormality defined using available flags in MIMIC-III and approximated using total lab counts in eICU. Medication features were constructed using administration records and counts of opioid-related medication events within the first 24 hours were used as proxies for treatment intensity. Procedure-related features were derived from procedure and treatment tables, including counts of procedures and binary indicators of any procedure within 24 hours. A surgical admission flag was defined based on the presence of procedure codes in MIMIC-III and approximated in eICU using admission diagnosis text.

The primary outcome, extreme opioid exposure within 24 hours, was defined independently within each dataset using the same percentile-based approach as in the development cohort. Among admissions receiving any opioid within the first 24 hours, the top decile of total MME was used to define extreme exposure, ensuring consistency in relative outcome definition, while allowing absolute exposure thresholds to vary across datasets.

The final XGBoost model trained on the MIMIC-IV development cohort was applied directly to external datasets without retraining. No model recalibration was performed prior to external evaluation. All preprocessing steps, including median imputation values, categorical level

definitions, and feature encoding schemes, were derived from the MIMIC-IV training data and applied unchanged to MIMIC-III and eICU to prevent information leakage.

Model performance was evaluated using ROC-AUC and PR-AUC. Confidence intervals were estimated using nonparametric bootstrapping (B = 500). External validation results were compared with internal and temporal validation to assess generalizability across datasets.

Acknowledgments: The authors acknowledge the PhysioNet team and the investigators who contributed to the MIMIC-IV, MIMIC-III, and eICU databases.[22-25][37-40] Authors completed the required data use training and certification for access.

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Disclosures: The authors declare that they have no funding disclosures to report.

REFERENCES

1. Humphreys K, Shover CL, Andrews CM, Bohnert ASB, Brandeau ML, Caulkins JP, et al. Responding to the opioid crisis in North America and beyond: recommendations of the Stanford-Lancet Commission. The Lancet. 2022 Feb 5;399(10324):555-604. doi:10.1016/S0140-6736(21)02252-2.
2. Office USGA. Opioid Crisis: Status of Public Health Emergency Authorities | U.S. GAO [Internet]. [cited 2026 Feb 18]. Available from: https://www.gao.gov/products/gao-18-685r

3. Dowell D, Nataraj N, Rikard M, Park J, Zhang K, Baldwin G. Why have overdose deaths decreased? Widespread fentanyl saturation and decreased drug use among key drivers. The Lancet Regional Health - Americas. 2025 Nov 1;51. doi:10.1016/j.lana.2025.101226
4. Centers for Disease Control and Prevention. Provisional Drug Overdose Data [Internet]. CDC. 2024. Available from: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm
5. National Center for Drug Abuse Statistics. Prescription Drug Abuse Statistics [Internet]. NCDAS. 2024. Available from: https://drugabusestatistics.org/prescription-drug-abuse-statistics/
6. National Institute on Drug Abuse. Drug overdose death rates [Internet]. National Institutes of Health; 2024. Available from: https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates
7. Horn R, Hendrix JM, Kramer J. Postoperative pain control. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2025. Available from: https://www-ncbi-nlm-nih-gov.ezproxy.lib.utexas.edu/books/NBK544298/
8. Dowell D. CDC Clinical Practice Guideline for Prescribing Opioids for Pain — United States, 2022. MMWR Recomm Rep. 2022;71. doi:10.15585/mmwr.rr7103a1.
9. Arwi GA, Schug SA. Potential for Harm Associated with Discharge Opioids After Hospital Stay: A Systematic Review. Drugs. 2020 Apr;80(6):573-85. doi:10.1007/s40265-020-01294-z.

10. Hapca S, Peet L, Gibson C, Harvey A, Forget P. Persistent post-discharge opioid use and opioid stewardship: A single-centre retrospective cohort study. Br J Pain. 2024 Dec;18(6):497-503. doi:10.1177/20494637241276106.
11. Srivastava D, Hill S, Carty S, Rockett M, Bastable R, Knaggs R, et al. Surgery and opioids: evidence-based expert consensus guidelines on the perioperative use of opioids in the United Kingdom. British Journal of Anaesthesia. 2021 Jun 1;126(6):1208-16. doi:10.1016/j.bja.2021.02.030
12. Lee S, Song W, Bates DW, Urman RD, Zhang P. The recent trend of twin epidemic in the United States: a 10-year longitudinal cohort study of co-prescriptions of opioids and stimulants. The Lancet Regional Health - Americas. 2025 Apr 1;44:101030. doi:10.1016/j.lana.2025.101030
13. Leventhal EL, Nathanson LA, Landry AM. Variations in Opioid Prescribing Behavior by Physician Training. West J Emerg Med. 2019 May;20(3):428-32. doi:10.5811/westjem.2019.3.39311.
14. Substantial Variation in Opioid Prescribing Rates among ED Providers in the Same VA Healthcare Facility - Publication Brief [Internet]. 2019 [cited 2026 Feb 18]. Available from: https://www.hsrd.research.va.gov/research/citations/pubbriefs/articles.cfm?RecordID=975
15. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Sciences Policy, Committee on Pain Management and Regulatory Strategies to Address Prescription Opioid Abuse. Pain Management and the

526 Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription
527 Opioid Use [Internet]. Phillips JK, Ford MA, Bonnie RJ, editors. Washington (DC):
528 National Academies Press (US); 2017 [cited 2026 Mar 29]. doi: 10.17226/24781
529 16. Gan TJ. Poorly controlled postoperative pain: prevalence, consequences, and
530 prevention. J Pain Res. 2017 Sep 25;10:2287-98. doi:10.2147/JPR.S144066
531 17. Alessio-Bilowus D, Chua KP, Peahl A, Brummett CM, Gunaseelan V, Bicket MC, et al.
532 Epidemiology of Opioid Prescribing After Discharge From Surgical Procedures Among
533 Adults. JAMA Netw Open. 2024 Jun 26;7(6):e2417651.
534 doi:10.1001/jamanetworkopen.2024.17651
535 18. Luby AO, Alessio-Bilowus D, Hu HM, Brummett CM, Waljee JF, Bicket MC. Trends in
536 Opioid Prescribing and New Persistent Opioid Use After Surgery in the United States.
537 Annals of Surgery. 2025 Mar;281(3):347. doi:10.1097/SLA.0000000000006461
538 19. Gerlach EB, Plantz MA, Swiatek PR, Wu SA, Arpey N, Fei-Zhang D, et al. The Drivers of
539 Persistent Opioid Use and Its Impact on Healthcare Utilization After Elective Spine
540 Surgery. Global Spine Journal. 2024 Mar 1;14(2):370-9.
541 doi:10.1177/21925682221104731
542 20. Ly DP. Association of Patient Race and Ethnicity With Differences in Opioid Prescribing
543 by Primary Care Physicians for Older Adults With New Low Back Pain. JAMA Health
544 Forum. 2021 Sep 10;2(9):e212333. doi:10.1001/jamahealthforum.2021.2333
545 21. Yang N, Fang MC, Rambachan A. Sex Disparities in Opioid Prescription and
546 Administration on a Hospital Medicine Service. J Gen Intern Med. 2024
547 Nov;39(14):2679-88. doi:10.1007/s11606-024-08814-7

22. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi L A, Mark R. MIMIC-IV (version 3.1). PhysioNet. 2024. RRID:SCR_007345. Available from: https://doi.org/10.13026/kpb9-mt58
23. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023 Jan 3;10(1).
24. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13;101(23):E215-20. doi: 10.1161/01.cir.101.23.e215
25. Johnson A, Pollard T, Horng S, Celi L A, Mark R. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet. 2023. RRID:SCR_007345. doi: 10.13026/1n74-ne17
26. Adams MCB, Sward KA, Perkins ML, Hurley RW. Standardizing research methods for opioid dose comparison: the NIH HEAL morphine milligram equivalent calculator. Pain. 2025 Feb 3;166(8):1729-1737. doi: 10.1097/j.pain.0000000000003529
27. Cron DC, Hwang C, Hu HM, Lee JS, Dupree JM, Syrjamaki JD, Chung KC, Brummett CM, Englesbe MJ, Waljee JF. A statewide comparison of opioid prescribing in teaching versus nonteaching hospitals. Surgery. 2019 Apr;165(4):825-831. doi: 10.1016/j.surg.2018.10.005. Epub 2018 Nov 26.
28. R Core Team. *R: a language and environment for statistical computing*; 2025 [cited 2026 Mar 29]. Vienna: R Foundation for Statistical Computing [Internet]. Available from: https://www.R-project.org/

29. Python Software Foundation. *Python* (version 3.12); 2025 [cited 2026 Mar 29]. [Internet]. Available from: https://www.python.org

30. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*; 2017 Mar 31 [cited 2026 Mar 29];77(1):1-17. [Internet]. Available from: https://doi.org/10.18637/jss.v077.i01

31. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*; 2010 Feb 2 [cited 2026 Mar 29];33(1):1-22. [Internet]. Available from: https://doi.org/10.18637/jss.v033.i01

32. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J, Cortes D. *xgboost: extreme gradient boosting*; 2026 [cited 2026 Mar 29]. Version 3.3.0.0 [R package]. [Internet]. Available from: https://github.com/dmlc/xgboost

33. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; 2019 [cited 2026 Mar 29]; Minneapolis, MN, USA. p. 72-78. [Internet]. Available from: https://aclanthology.org/W19-1909/

34. Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4766-4777.

35. Slee VN. The International Classification of Diseases: ninth revision (ICD-9). Ann Intern Med. 1978 Mar;88(3):424-6. doi: 10.7326/0003-4819-88-3-424.

36. World Health Organization. The ICD-10 classification of mental and behavioural disorders. Geneva: World Health Organization; 1993. Available from: https://www.who.int/classifications/icd/en/

37. Johnson, A., Pollard, T., and Mark, R. (2016) 'MIMIC-III Clinical Database' (version 1.4), *PhysioNet*. RRID:SCR_007345. Available at: https://doi.org/10.13026/C2XW26

38. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

39. Pollard, T., Johnson, A., Raffa, J., Celi, L. A., Badawi, O., and Mark, R. (2019) 'eICU Collaborative Research Database' (version 2.0), *PhysioNet*. RRID:SCR_007345. Available at: https://doi.org/10.13026/C2WM1R

40. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. Scientific Data (2018). DOI: http://dx.doi.org/10.1038/sdata.2018.178.

FIGURE CAPTIONS

Fig 1. ROC and PR curves for ridge logistic regression, random forest, and XGBoost models (test set)

Fig 2. SHAP Summary plot for structured XGBoost model predicting high inpatient opioid exposure

Fig 3. ROC and PR curves for text-based models (test set)

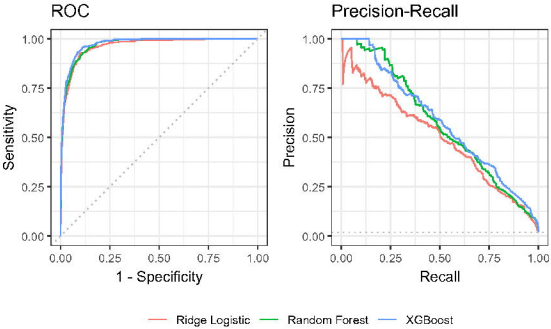
Fig 4. Contrastive bigrams in discharge notes associated with high versus low inpatient opioid use

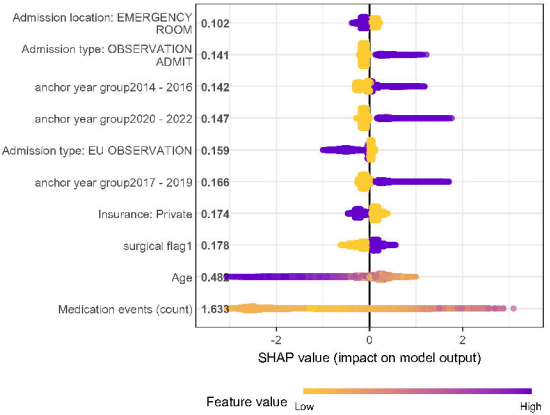
Fig 5. Model discrimination (ROC-AUC, 95% CI) stratified by surgical subgroup

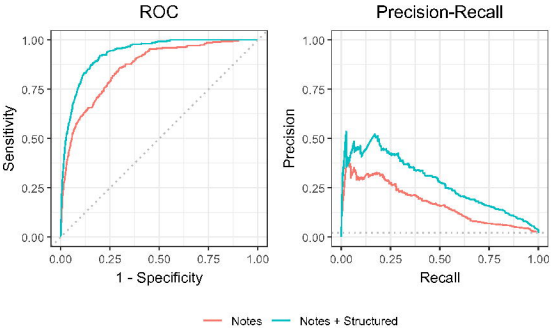
Fig 6. ROC and PR curves for ridge logistic regression, random forest, and XGBoost temporal validation

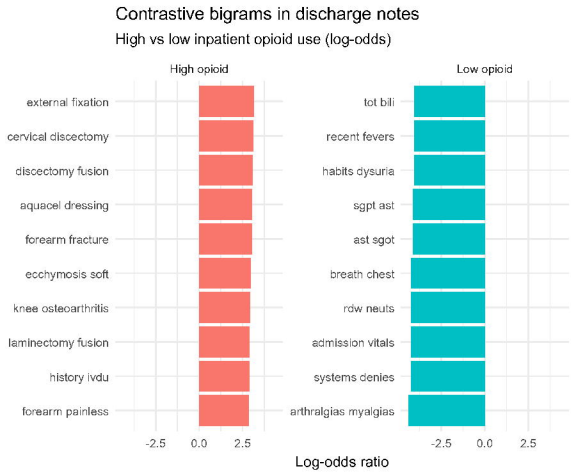
Fig 7. ROC and PR curves for MIMIC-III and eICU external validation XGBoost models

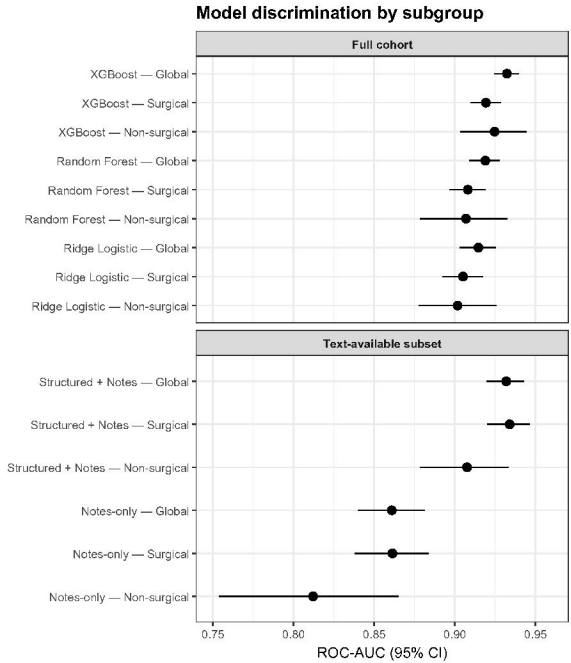
Fig 8. Calibration plots for MIMIC-III and eICU external validation

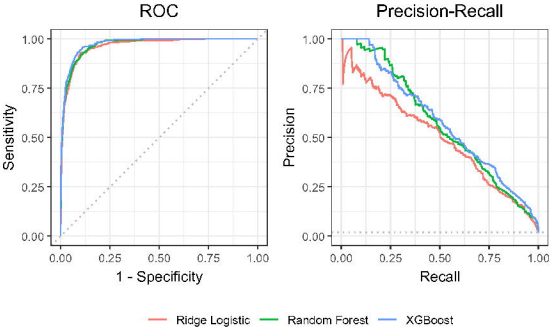

ROC
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall
Precision
Recall
Ridge Logistic
Random Forest
XGBoost

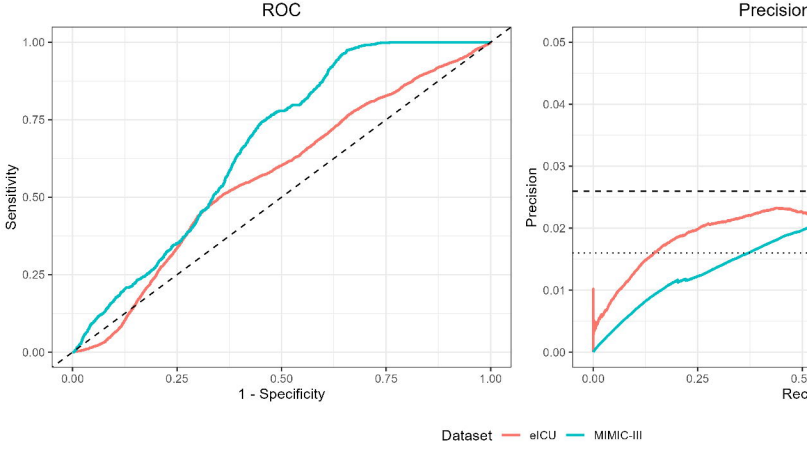

Admission location: EMERGENCY ROOM
0.102
Admission type: OBSERVATION ADMIT
0.141
anchor year group2014 - 2016
0.142
anchor year group2020 - 2022
0.147
Admission type: EU OBSERVATION
0.159
anchor year group2017 - 2019
0.166
Insurance: Private
0.174
surgical flag1
0.178
Age
0.489
Medication events (count)
1.633
-2
0
2
SHAP value (impact on model output)
Feature value
Low
High

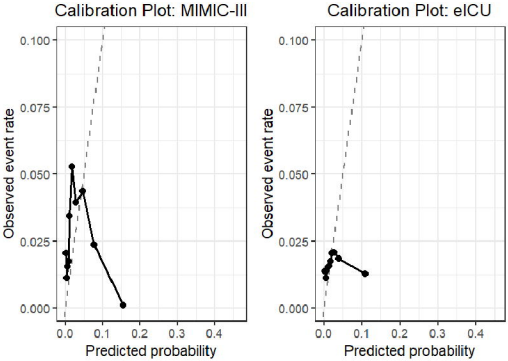

ROC
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall
Precision
Recall
Notes
Notes + Structured


Contrastive bigrams in discharge notes
High vs low inpatient opioid use (log-odds)
High opioid
external fixation
cervical discectomy
discectomy fusion
aquacel dressing
forearm fracture
ecchymosis soft
knee osteoarthritis
laminectomy fusion
history ivdu
forearm painless
-2.5
0.0
2.5
Low opioid
tot bili
recent fevers
habits dysuria
sgpt ast
ast sgot
breath chest
rdw neuts
admission vitals
systems denies
arthralgias myalgias
-2.5
0.0
2.5
Log-odds ratio


Model discrimination by subgroup
Full cohort
XGBoost — Global
XGBoost — Surgical
XGBoost — Non-surgical
Random Forest — Global
Random Forest — Surgical
Random Forest — Non-surgical
Ridge Logistic — Global
Ridge Logistic — Surgical
Ridge Logistic — Non-surgical
Text-available subset
Structured + Notes — Global
Structured + Notes — Surgical
Structured + Notes — Non-surgical
Notes-only — Global
Notes-only — Surgical
Notes-only — Non-surgical
0.75
0.80
0.85
0.90
0.95
ROC-AUC (95% CI)


ROC
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall
Precision
Recall
Ridge Logistic
Random Forest
XGBoost


ROC
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision
Precision
0.00
0.01
0.02
0.03
0.04
0.05
Dataset
eICU
MIMIC-III


Calibration Plot: MIMIC-III
Observed event rate
0.100
0.075
0.050
0.025
0.000
0.0
0.1
0.2
0.3
0.4
Predicted probability
Calibration Plot: eICU
Observed event rate
0.100
0.075
0.050
0.025
0.000
0.0
0.1
0.2
0.3
0.4
Predicted probability