Development and Temporal Evaluation 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. Early identification of high-risk patients may improve opioid stewardship. We developed machine learning models to predict high opioid exposure during hospitalization using electronic health record data from MIMIC-IV. We conducted a retrospective study of 223,452 unique first hospital admissions in MIMIC-IV. The outcome was high opioid exposure, defined as the top decile among opioid-exposed admissions (MME/day ≥ 225), 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 interpretable bigram features. Models were trained using an 80/10/10 split and evaluated with temporal validation on the most recent 10% of admissions. Performance was assessed using ROC-AUC and PR-AUC with 95% confidence intervals. Among structured-only models, XGBoost achieved the best test performance (ROC-AUC 0.932 [0.924-0.940]; PR-AUC 0.223 [0.193-0.262]). The 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 evaluation showed similar discrimination (ROC-AUC 0.929; PR-AUC 0.223). High-risk bigrams included procedural terms such as “external fixation” and “cervical discectomy.” Integration of structured and text-derived features improved risk stratification compared to structured data alone. Interpretable bigram signals reflected procedural complexity and orthopedic pathology, reinforcing the clinical plausibility of model predictions. Multimodal EHR-based models accurately predict high inpatient opioid exposure and may support targeted opioid stewardship during hospitalization.

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, with negative side effects such as opioid addiction. Identifying patients at risk early during their hospital stay could help physicians make safer prescribing decisions. In this study, we used electronic health record data from over 220,000 hospital admissions to develop computer models that estimate the likelihood that a patient will receive high levels of opioids. We focused on information available within the first 24 hours of admission, including basic patient characteristics, laboratory testing patterns, and procedures. We also used information from clinical notes to capture additional context about patient care.

We found that these models were able to accurately identify patients at higher risk, and that combining structured data with information from clinical notes improved performance. Importantly, the patterns identified by the models, such as certain surgical procedures, were consistent with clinical expectations. Our findings suggest that routinely collected hospital data can be used to support earlier identification of patients at risk for high opioid exposure. This approach could help guide more targeted and cautious opioid prescribing practices in inpatient settings.

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 by approximately one quarter between March 2024 and March 2025, 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 and that a subset of hospitalizations may experience 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

88 opioid exposure during hospitalization, despite the hospital setting representing a critical and
89 potentially modifiable point of care. A predictive model capable of identifying hospitalizations
90 at risk for unusually high opioid dosing could enable targeted stewardship interventions, real-
91 time monitoring, clinical decision support, and individualized pain management strategies.

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

98 We additionally evaluated a notes-only model trained on clinical documentation to determine
99 whether unstructured text independently contains predictive signals associated with opioid
100 intensity. Furthermore, we constructed a multimodal model integrating structured features
101 with note-derived embeddings to assess whether combining quantitative utilization data with
102 clinical language improved discrimination and calibration. Finally, we conducted an exploratory
103 bigram analysis of discharge documentation to identify specific narrative phrases enriched
104 among high- and low-exposure admissions, providing interpretable context for the text-derived
105 signals captured by the models. By systematically comparing structured, text-based, and
106 multimodal approaches, we aimed and were able to characterize the predictive value of diverse
107 data modalities for identifying high-risk inpatient opioid exposure across heterogeneous clinical
108 contexts.

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 equivalent (MME ≥ 225 mg/day) among opioid-exposed admissions. 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, with a median of 54.7 (IQR 19.5-122.1), 99th percentile of 643.5, and a maximum of 4,259.1.

Admissions were randomly assigned at the admission level 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 (median 52 vs. 57 years, $p < 0.001$) and had higher laboratory utilization within 24 hours (median 36 vs. 33 labs, $p < 0.001$). Medication administrations within 24 hours were substantially higher in the high-risk group (median 49 vs. 30, $p < 0.001$), and surgical admissions were more frequent among high-risk cases (70% vs. 51%, $p < 0.001$). Significant differences were also observed across admission type, admission location, insurance, primary language, marital status, race, sex, and anchor year (a de-identified calendar year assigned after date shifting to preserve patient privacy) group) ($p < 0.001$).

130 **Missing Data and Feature Processing:**

131 The most frequently missing variables included medication event counts (53.6%) and
132 laboratory-derived features (29%). Indicator variables were used to capture absence of
133 laboratory testing within 24 hours (29.2%). After imputation and encoding, no residual missing
134 values remained in the analytic dataset.

135 **Model Development and Evaluation:**

136 Structured prediction models were developed using ridge logistic regression, random forest,
137 and extreme gradient boosting. Model development was performed in the training set,
138 hyperparameter tuning was conducted using the validation set, and final model performance
139 was evaluated in the independent test set. Model outputs were predicted probabilities of high-
140 risk opioid exposure, and classification thresholds were selected using the Youden index in the
141 validation set. Primary discrimination and calibration metrics are summarized in Table 1, while
142 threshold-dependent performance metrics are provided in Supplementary Table 2.

143 Table 1: Primary model performance

144

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

Model Performance: Structured-Only Models

In the validation cohort, ridge logistic regression achieved a ROC-AUC of 0.892 and PR-AUC of 0.166, random forest achieved ROC-AUC 0.909 and PR-AUC 0.201, and XGBoost achieved ROC-AUC 0.918 and PR-AUC 0.216, demonstrating the highest discrimination among candidate models.

In the independent test cohort, XGBoost achieved a ROC-AUC of 0.932 (95% CI: 0.924-0.940) and PR-AUC of 0.223 (95% CI: 0.193-0.262), with a Brier score of 0.0166 and favorable calibration (intercept 0.12; slope 1.06; ECE 0.0012) indicating minimal overfitting and

agreement between predicted and observed risk. Visual inspection of calibration curves demonstrated close alignment across risk deciles, with no substantial miscalibration observed in the highest-risk strata. Random forest achieved ROC-AUC 0.919 (95% CI: 0.909-0.928) and PR-AUC 0.204, while ridge logistic regression achieved ROC-AUC 0.915 (95% CI: 0.903-0.925) and PR-AUC 0.185. Bootstrap resampling (B = 2000) confirmed stability of performance estimates (Table 1) (Fig 1).

Model Interpretation: SHAP Analysis

SHAP analysis of the structured XGBoost model demonstrated that early inpatient utilization features were the strongest contributors to predicted risk of high opioid exposure (Fig 2). The number of medication events was the most influential variable, with higher counts strongly associated with increased predicted probability. Age also exhibited substantial contribution, with younger age values generally associated with higher model output. Surgical status and admission characteristics contributed moderately to risk stratification.

Notes-Only Model and Multimodal Integration:

Embeddings were available for 130,557 admissions (58.4% coverage), corresponding to 58.4% of training admissions, 58.9% of validation admissions, and 58.4% test admissions. In the validation subset with embeddings, the notes-only model achieved ROC-AUC 0.865 and PR-AUC 0.150. In the test subset with embeddings, the notes-only model achieved ROC-AUC 0.861 (95% CI: 0.840-0.881) and PR-AUC 0.168 (95% CI: 0.134-0.208), with Brier score 0.0189.

We then evaluated a combined model integrating structured predictors with discharge note embeddings within the subset with available text features. In validation, the combined model achieved ROC-AUC 0.933 and PR-AUC 0.262. In the test subset with embeddings (n = 13,051), the combined model achieved ROC-AUC 0.932 (95% CI: 0.920–0.943) and PR-AUC 0.276 (95% CI: 0.229–0.331), with Brier score 0.0173 and calibration slope 1.02 (Table 1) (Fig 3). Given the outcome prevalence of 2.65%, the baseline PR-AUC of a random classifier would be 0.0265. Thus, the combined model achieved more than a 10-fold improvement over random precision-recall performance, reflecting substantial enrichment of extreme opioid exposure among high-risk predictions. However, admissions with discharge notes may represent longer or more complex hospitalizations, and reliance on discharge documentation limits applicability to retrospective analysis rather than prospective early prediction.

Contrastive Text Analysis:

To characterize clinically interpretable text signals, we performed a bigram-based analysis and identified bigrams enriched among high-risk admissions, including terms related to orthopedic and spine interventions as well as substance use history (Fig 4). A complete ranked list of bigrams is provided in Supplementary Table 3.

Subgroup Performance:

We evaluated performance separately in surgical and non-surgical admissions to assess heterogeneity in discrimination and calibration. In the full test cohort, outcome prevalence was 2.96% among surgical admissions (n = 11,503; 340 events) and 0.87% among non-surgical

admissions (n = 10,842; 94 events). For XGBoost, ROC-AUC was 0.919 (95% CI 0.910-0.929) with PR-AUC 0.242 in the surgical subset, and ROC-AUC was 0.925 (95% CI 0.904-0.944) with PR-AUC 0.161 in the non-surgical subset. As precision-recall performance depends on outcome prevalence, these values should be interpreted relative to subgroup baseline risk (0.0296 for surgical and 0.0087 for non-surgical admissions). Thus, PR-AUC in the surgical group reflects approximately 8-fold enrichment over random classification, whereas the non-surgical group demonstrates nearly 18-fold enrichment despite its lower absolute PR-AUC. Calibration slopes remained close to 1.0 in both groups.

In the subset with available discharge note embeddings, the notes-only model achieved ROC-AUC 0.861 and PR-AUC 0.191 in surgical admissions (n = 7,768) and ROC-AUC 0.812 and PR-AUC 0.084 in non-surgical admissions (n = 5,283). In the same subset, the combined structured + notes model achieved ROC-AUC 0.934 and PR-AUC 0.316 in surgical admissions and ROC-AUC 0.908 and PR-AUC 0.099 in non-surgical admissions (Supplementary Table 4) (Fig 5).

Temporal Validation:

To assess temporal generalizability, we performed a time-based split using admission time with the most recent 10% of admissions reserved as a temporally held-out test cohort. Outcome prevalence remained stable across temporal splits.

In the temporally held-out test cohort, XGBoost achieved a ROC-AUC of 0.929 and PR-AUC of 0.223, with a Brier score of 0.0161. Random forest achieved ROC-AUC 0.920 and PR-AUC 0.231, while ridge logistic regression achieved ROC-AUC 0.900 and PR-AUC 0.171. Discrimination and

213 calibration in the temporal test cohort were comparable to those observed in the randomly
214 held-out test cohort (Table 1) (Fig 6).

215 **DISCUSSION**

216 In this study, we developed and internally validated machine learning models to identify
217 hospital admissions at risk for extreme inpatient opioid exposure. Rather than focusing solely
218 on discrimination metrics, our findings highlight how patterns embedded within routine
219 inpatient data reflect broader care complexity and resource intensity that are associated with
220 unusually high opioid dosing.

221 In clinical practice, hospitalized patients generate multimodal data streams. Structured data
222 quantify measurable aspects of care, such as laboratory frequency, procedural burden, and
223 medication administration. Unstructured notes capture physician interpretation, procedural
224 narratives, and contextual detail surrounding the patient's clinical course. These data
225 modalities represent complementary views of the same hospitalization. Our results suggest
226 that structured utilization patterns provide a strong foundation for identifying high-intensity
227 opioid exposure, while narrative documentation contributes additional context in certain
228 settings, particularly surgical admissions.

229 Extreme opioid exposure was uncommon but clinically meaningful, occurring in a small minority
230 of admissions. These cases were characterized by greater early laboratory utilization, higher
231 medication event counts, and more frequent surgical procedures. Such patterns likely reflect
232 more complex operative courses, trauma-related admissions, or high-acuity postoperative care

233 – contexts in which aggressive analgesia may be required. [16] Importantly, the model did not
234 rely on opioid-specific inputs alone but instead captured broader signals of clinical intensity.
235 This suggests that extreme dosing may be predictable from the overall trajectory of inpatient
236 care rather than from pain management variables in isolation.

237 Subgroup analyses further illustrate how context shapes predictive signals. Structured models
238 demonstrated stable discrimination across surgical and non-surgical admissions, indicating that
239 quantitative utilization features generalize across heterogeneous inpatient populations. In
240 contrast, the notes-only model performed more strongly in surgical contexts and less
241 consistently in non-surgical admissions. Surgical discharge summaries often contain detailed
242 procedural narratives, implant descriptions, and postoperative plans, which may encode
243 implicit signals of anticipated analgesic need. Medical admissions, by comparison, may contain
244 more templated or symptom-focused language, potentially limiting the strength of narrative-
245 based prediction. The improved performance observed when structured data were combined
246 with note embeddings suggests that quantitative measures and clinician documentation
247 capture partially distinct aspects of care complexity.

248 Bigram analysis of discharge notes supports the clinical plausibility of the learned narrative
249 signals. Terms enriched among high-risk admissions were predominantly related to orthopedic,
250 trauma, and spine procedures, such as “external fixation,” “laminectomy fusion,” and
251 “interbody fusion.” These are clinical scenarios commonly associated with substantial
252 postoperative pain and extended analgesic requirements and have been previously identified
253 as procedures likely to correlate with post-operative opioid prescription.[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 particular time period. Although prescribing behaviors and stewardship efforts continue to change 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 fluctuations.[18]

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, the model could be deployed at the 24-hour mark following admission, once initial opioid administration and early clinical data are available. This timepoint offers an opportunity to reassess analgesic strategies, optimize multimodal pain management, or involve specialty consultation. Given the observed prevalence (2.65%), risk thresholds could be selected to flag only the highest 5-10% of predicted risk scores, thereby limiting alert burden while focusing attention on patients most likely to experience extreme exposure. Integration into electronic clinical decision support systems could enable automated notification to primary teams, pharmacy services, or pain consultation programs. Importantly, predicted risk should inform 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 propagate underlying practice patterns, careful evaluation across demographic subgroups is essential prior to clinical deployment. Future work should assess subgroup discrimination, calibration, and risk 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, it was conducted within a single health system. External validation is necessary to assess generalizability across institutions with different prescribing practices and patient populations. Second, discharge note embeddings were available for only a subset of admissions, which may introduce documentation-related selection bias. Third, extreme opioid exposure was defined using a relative threshold within this dataset. Recalibration may be required in settings with different prescribing distributions. Fourth, contextual factors such as pain scores, intraoperative anesthetic regimens, and provider-level prescribing behaviors were not explicitly modeled. Finally, the low prevalence of extreme exposure limits positive predictive value, emphasizing that these models are best suited for risk stratification rather than deterministic decision-making.

MATERIALS AND METHODS

Data Source and Study Setting:

295 We conducted a retrospective cohort study using MIMIC-IV. MIMIC-IV is a publicly available, de-
296 identified electronic health record (EHR) database comprising detailed clinical data from
297 patients admitted to Beth Israel Deaconess Medical Center, a large tertiary academic medical
298 center in Boston, Massachusetts.[22-24] The MIMIC-IV hospital module contains structured
299 clinical data from 223,452 unique patients between 2008 and 2022, including demographic
300 characteristics, laboratory measurements, procedures, medication administrations, and
301 admission-level encounter data.

302 Free-text clinical documentation was obtained from the MIMIC-IV-Note module, which includes
303 331,793 de-identified discharge summaries from 145,915 patients during the same study
304 period.[25] Discharge summaries were linked to structured admission-level data using the
305 unique hospital admission identifier.

306 **Study Population:**

307 All inpatient hospital admissions recorded in the MIMIC-IV hospital module during the study
308 period were eligible for inclusion. To avoid correlated observations from repeat
309 hospitalizations, only the first hospital admission per patient was included in the analysis. No
310 age-based exclusions were applied. For model development and evaluation, the cohort was
311 temporally partitioned into development and held-out test sets to enable out-of-time
312 validation.

313 **Outcome Definition:**

314 The primary outcome was extreme inpatient opioid exposure within the first 24 hours of

315 hospital admission. Opioid exposure was quantified as the total morphine milligram equivalents
316 (MME) administered during the first 24 hours following admission. MME was calculated from
317 published guidelines converting individual opioid doses to a common morphine-equivalent
318 scale using standardized equianalgesic conversion factors.[26]
319
320 Among admissions receiving any opioid during the first 24 hours (MME > 0), total 24-hour MME
321 values were ranked, and the 90th percentile was used to define extreme exposure. This
322 percentile-based threshold was calculated using the full cohort prior to model development
323 and applied unchanged across all data partitions.
324
325 A relative threshold was selected rather than a fixed MME cutoff for several reasons. First,
326 inpatient opioid prescribing can vary substantially across institutions and clinical contexts, and
327 no standardized definition of high-dose opioid exposure exists for acute hospitalization.[27]
328 Second, percentile-based approaches allow identification of patients at the extreme tail of
329 institutional prescribing practices while maintaining adaptability across various health systems.
330 In our cohort, the 90th percentile corresponded to ≥225 MME within 24 hours, representing
331 10% of opioid-exposed admissions and 2.65% of all hospital admissions.
332 The 24-hour window was selected to capture early prescribing intensity during initial
333 hospitalization, a clinically actionable time frame when risk stratification and intervention may
334 be most feasible.
335

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. 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.

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

361 For remaining numeric predictors, missing values were imputed using median imputation.
362 Median values were calculated using the training dataset only and applied unchanged to the
363 validation and test datasets.

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

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

370 **Discharge Notes Preprocessing:**

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

374 A front-back truncation strategy was applied. For each note, two tokenized segments were
375 generated: 1) the first 256 tokens of the note (front segment) and 2) the last 256 tokens of the
376 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 embedding vectors were subsequently 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, while temporal validation was conducted as a secondary robustness analysis to assess model performance under 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. Similarly, 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 embedding features and to reduce overfitting.

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, parameters included a learning rate of 0.05, maximum tree depth of 6, minimum child weight of 1, subsampling rate of 0.8, and column subsampling rate of 0.8. Early stopping was applied based on validation set ROC-AUC, with training terminated if performance did not improve for 50 consecutive rounds; the optimal number of boosting rounds was selected accordingly.

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 intercept and slope were estimated by regressing observed outcomes on the logit of predicted probabilities.

416 Decision curve analysis was performed for the best-performing model to evaluate net clinical
417 benefit across probability thresholds ranging from 0 to 0.5

418 **Model Interpretation:**

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

422 **Subgroup Analysis:**

423 To evaluate model performance across clinically-distinct admission types with different
424 prevalences of opioid administration during hospital stays, analyses were conducted for global,
425 surgical, and non-surgical cohorts.

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

433 **Temporal Validation:**

To evaluate model robustness to temporal dataset shift, a secondary validation was performed using a time-based split of admissions. Hospital admissions were ordered chronologically based on admission timestamp. The earliest 80% of admissions were assigned to a temporal training set, the subsequent 10% to a temporal validation set, and the most recent 10% to a temporal test set.

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:

To identify clinically meaningful language patterns associated with high versus low opioid exposure, exploratory contrastive text analysis was performed on discharge summaries across the full cohort. 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

453 bigrams occurring fewer than 100 times overall. For each retained bigram, a log-odds ratio with
454 Dirichlet smoothing was computed to quantify enrichment in high- versus low-opioid groups.

455 **Acknowledgments:** The authors acknowledge the PhysioNet team and the investigators who
456 contributed to the MIMIC-IV database.[22-25] This study used the MIMIC-IV database (version
457 3.1), a publicly available, de-identified clinical dataset. The authors completed the required data
458 use training and certification for access.

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

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

463 **REFERENCES**

464 1. Humphreys K, Shover CL, Andrews CM, Bohnert ASB, Brandeau ML, Caulkins JP, et al.
465 Responding to the opioid crisis in North America and beyond: recommendations of the
466 Stanford–Lancet Commission. The Lancet. 2022 Feb 5;399(10324):555–604.
467 doi:10.1016/S0140-6736(21)02252-2.
468 2. Office USGA. Opioid Crisis: Status of Public Health Emergency Authorities | U.S. GAO
469 [Internet]. [cited 2026 Feb 18]. Available from: https://www.gao.gov/products/gao-18-
470 685r

471 3. Dowell D, Nataraj N, Rikard M, Park J, Zhang K, Baldwin G. Why have overdose deaths
472 decreased? Widespread fentanyl saturation and decreased drug use among key drivers.
473 The Lancet Regional Health – Americas. 2025 Nov 1;51. doi:10.1016/j.lana.2025.101226
474 4. Centers for Disease Control and Prevention. Provisional Drug Overdose Data [Internet].
475 CDC. 2024. Available from: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-
476 data.htm
477 5. National Center for Drug Abuse Statistics. Prescription Drug Abuse Statistics [Internet].
478 NCDAS. 2024. Available from: https://drugabusestatistics.org/prescription-drug-abuse-
479 statistics/
480 6. National Institute on Drug Abuse. Drug overdose death rates [Internet]. National
481 Institutes of Health; 2024. Available from: https://nida.nih.gov/research-topics/trends-
482 statistics/overdose-death-rates
483 7. Horn R, Hendrix JM, Kramer J. Postoperative pain control. In: *StatPearls*. Treasure
484 Island (FL): StatPearls Publishing; 2025. Available from: https://www-ncbi-nlm-nih-
485 gov.ezproxy.lib.utexas.edu/books/NBK544298/
486 8. Dowell D. CDC Clinical Practice Guideline for Prescribing Opioids for Pain — United
487 States, 2022. MMWR Recomm Rep. 2022;71. doi:10.15585/mmwr.rr7103a1.
488 9. Arwi GA, Schug SA. Potential for Harm Associated with Discharge Opioids After Hospital
489 Stay: A Systematic Review. Drugs. 2020 Apr;80(6):573–85. doi:10.1007/s40265-020-
490 01294-z.

491 10. Hapca S, Peet L, Gibson C, Harvey A, Forget P. Persistent post-discharge opioid use and
492 opioid stewardship: A single-centre retrospective cohort study. Br J Pain. 2024
493 Dec;18(6):497–503. doi:10.1177/20494637241276106.
494 11. Srivastava D, Hill S, Carty S, Rockett M, Bastable R, Knaggs R, et al. Surgery and opioids:
495 evidence-based expert consensus guidelines on the perioperative use of opioids in the
496 United Kingdom. British Journal of Anaesthesia. 2021 Jun 1;126(6):1208–16.
497 doi:10.1016/j.bja.2021.02.030
498 12. Lee S, Song W, Bates DW, Urman RD, Zhang P. The recent trend of twin epidemic in the
499 United States: a 10-year longitudinal cohort study of co-prescriptions of opioids and
500 stimulants. The Lancet Regional Health - Americas. 2025 Apr 1;44:101030.
501 doi:10.1016/j.lana.2025.101030
502 13. Leventhal EL, Nathanson LA, Landry AM. Variations in Opioid Prescribing Behavior by
503 Physician Training. West J Emerg Med. 2019 May;20(3):428–32.
504 doi:10.5811/westjem.2019.3.39311.
505 14. Substantial Variation in Opioid Prescribing Rates among ED Providers in the Same VA
506 Healthcare Facility – Publication Brief [Internet]. 2019 [cited 2026 Feb 18]. Available
507 from:
508 https://www.hsrd.research.va.gov/research/citations/pubbriefs/articles.cfm?RecordID=
509 975
510 15. National Academies of Sciences, Engineering, and Medicine, Health and Medicine
511 Division, Board on Health Sciences Policy, Committee on Pain Management and
512 Regulatory Strategies to Address Prescription Opioid Abuse. Pain Management and the

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

578 36. World Health Organization. The ICD-10 classification of mental and behavioural
579 disorders. Geneva: World Health Organization; 1993. Available from:
580 https://www.who.int/classifications/icd/en/
581

582 **FIGURE CAPTIONS**

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

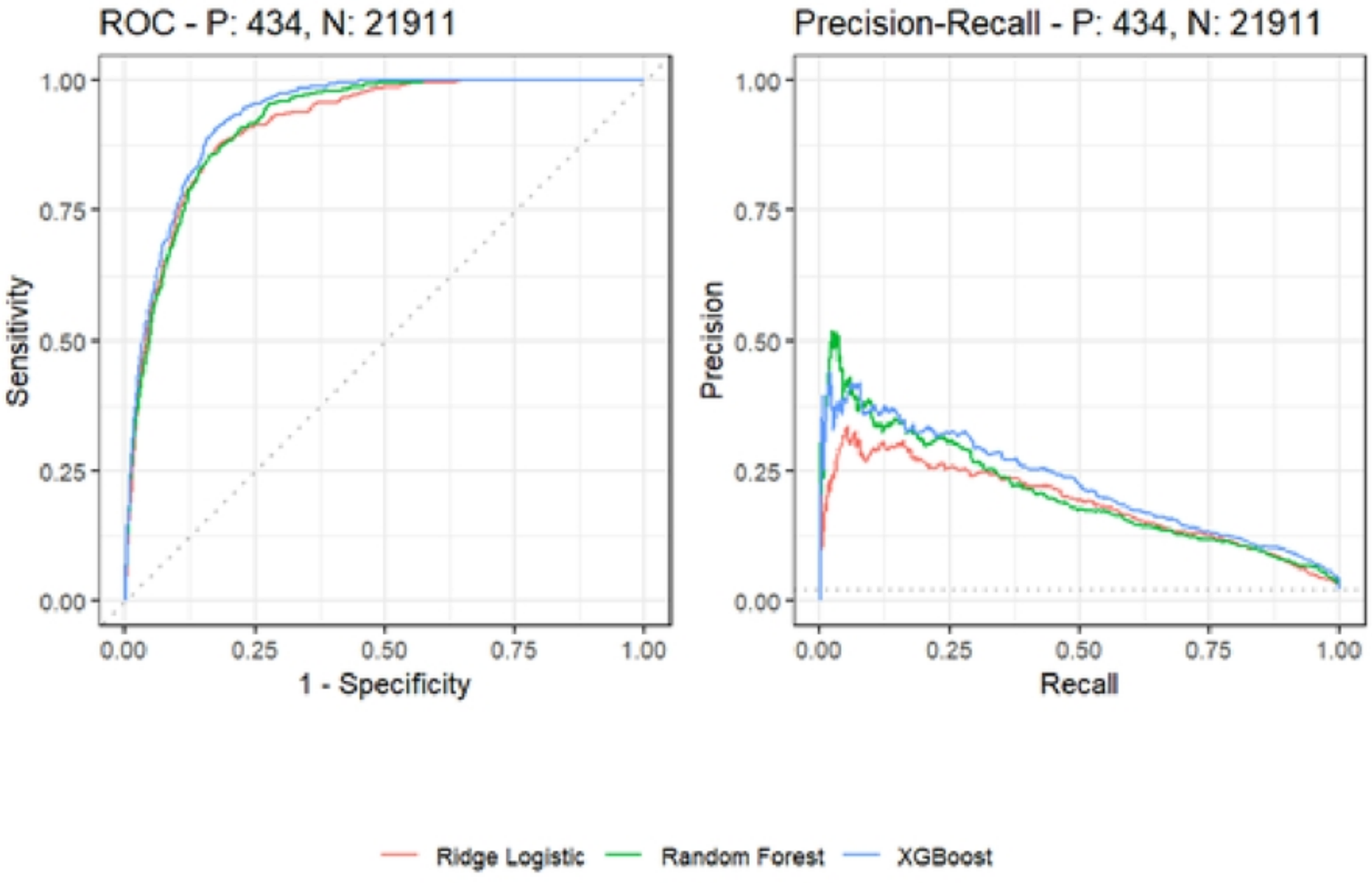
585 Fig 2. SHAP Summary plot for structured XGBoost model predicting high inpatient opioid
586 exposure

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

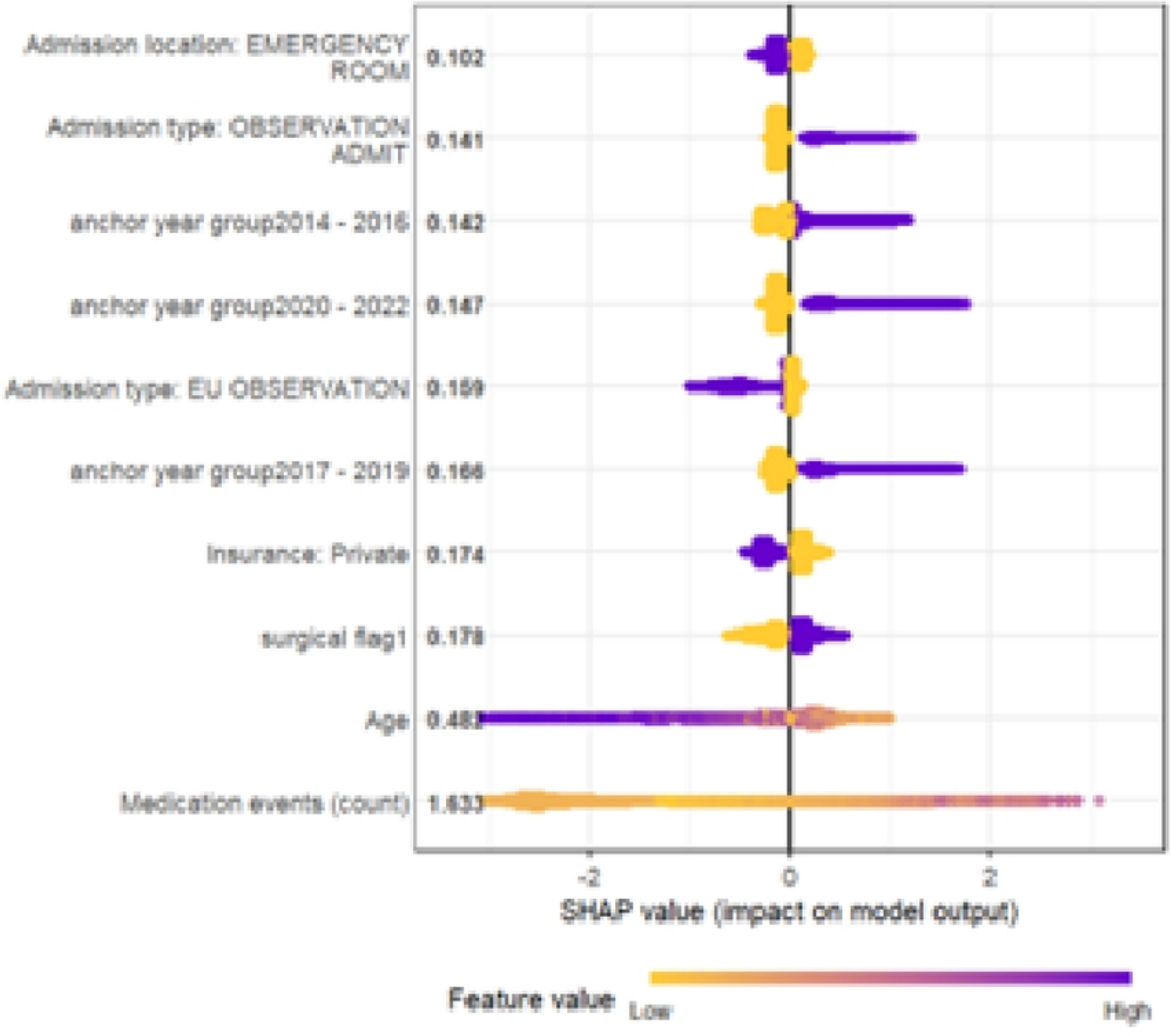
588 Fig 4. Contrastive bigrams in discharge notes associated with high versus low inpatient opioid
589 use

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

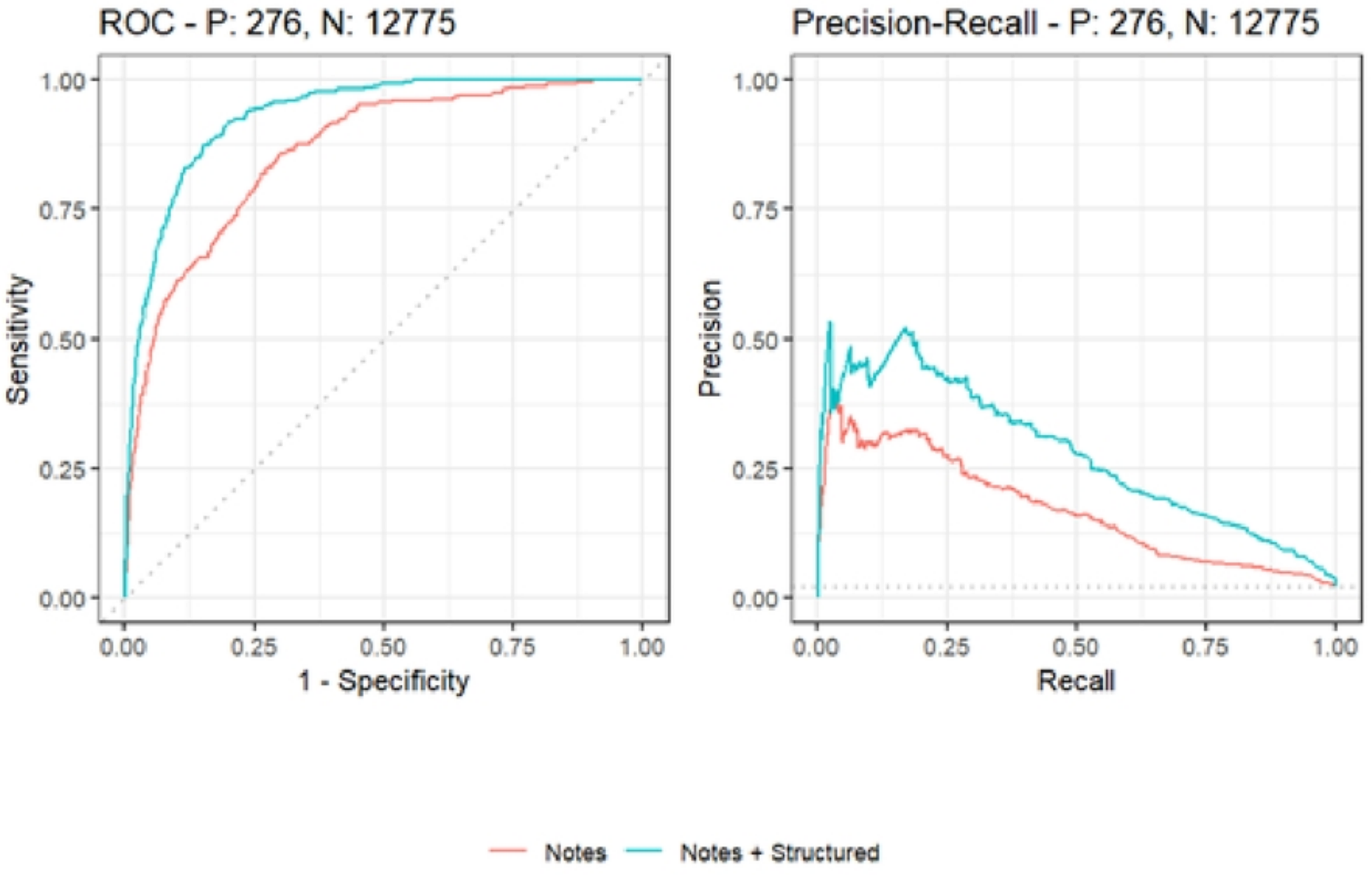
591 Fig 6. ROC and PR curves for ridge logistic regression, random forest, and XGBoost temporal
592 validation


ROC - P: 434, N: 21911
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall - P: 434, N: 21911
Precision
Recall
Ridge Logistic
Random Forest
XGBoost

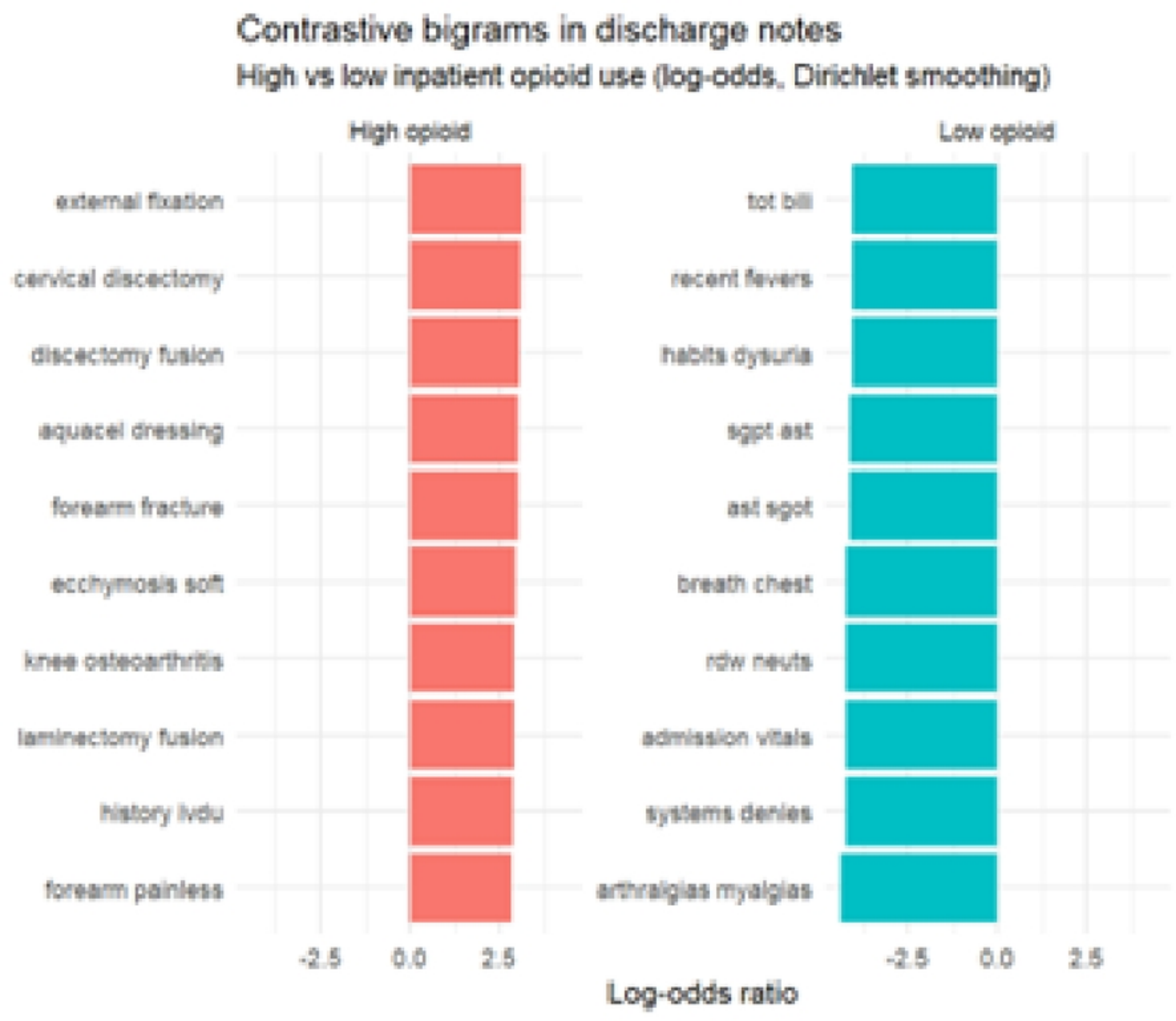
Figure


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
0.482
Medication events (count)
1.633
-2
0
2
SHAP value (impact on model output)
Feature value
Low
High

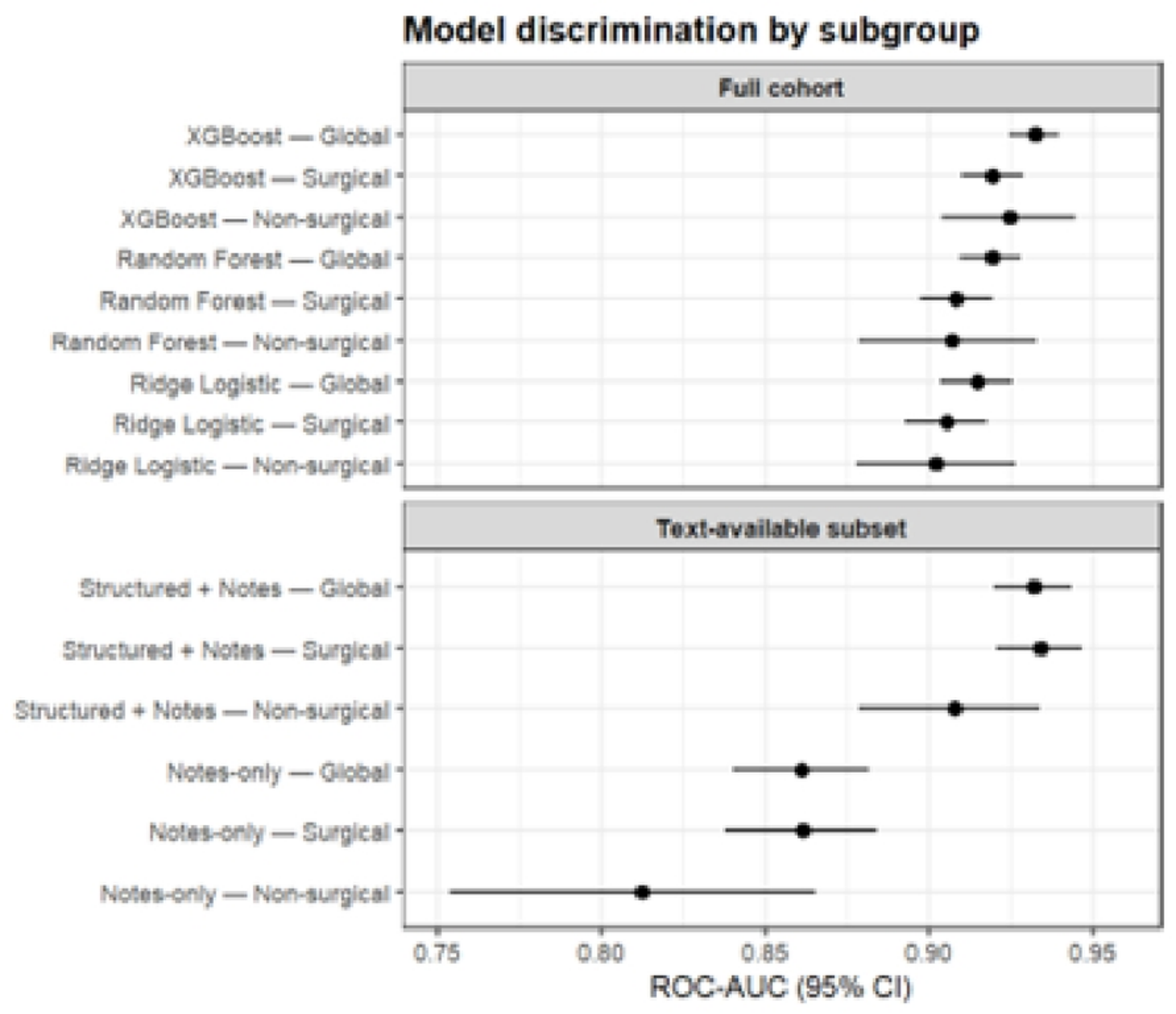
Figure


ROC - P: 276, N: 12775
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall - P: 276, N: 12775
Precision
Recall
Notes
Notes + Structured

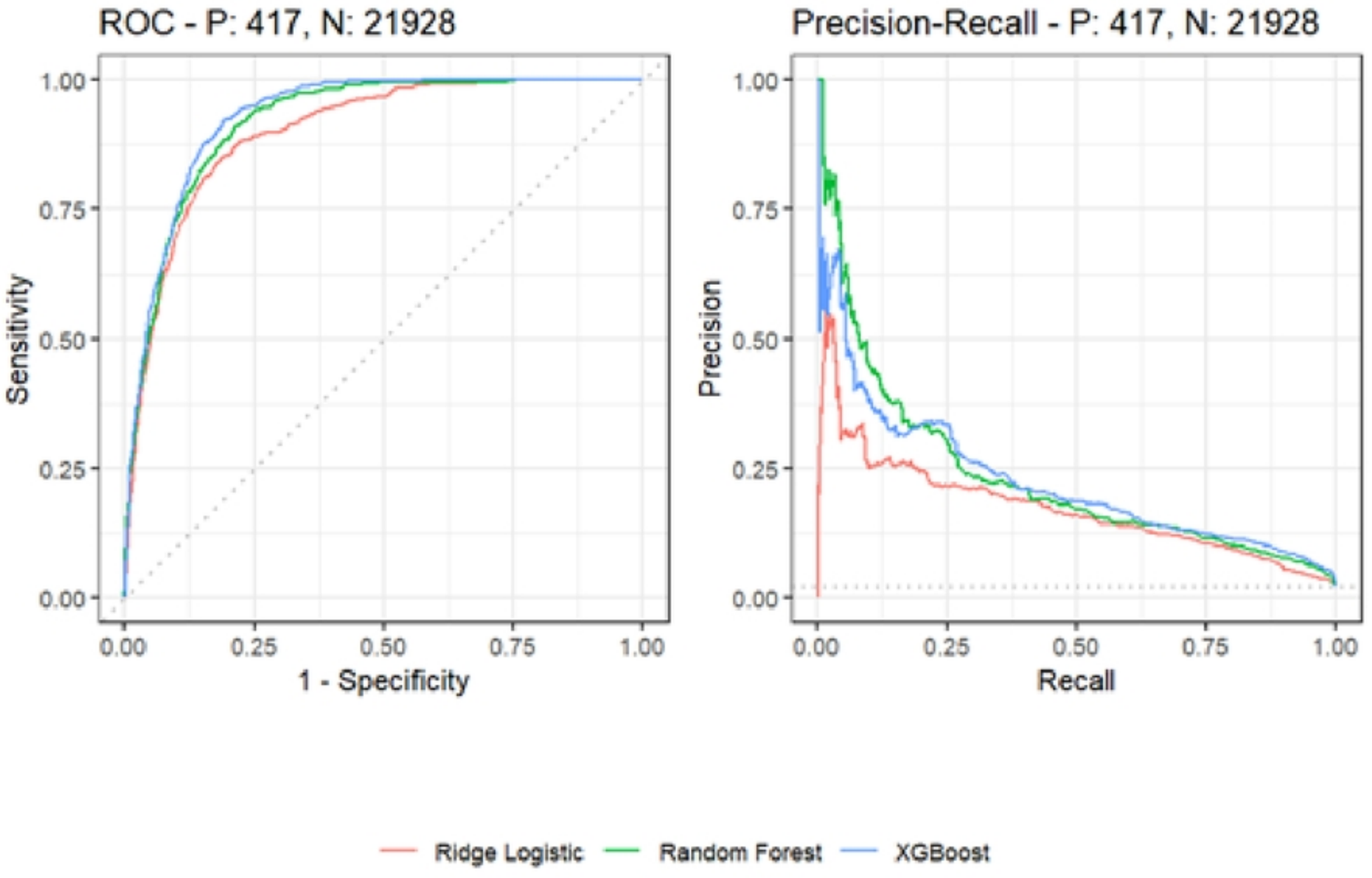
Figure


Contrastive bigrams in discharge notes
High vs low inpatient opioid use (log-odds, Dirichlet smoothing)
High opioid
external fixation
cervical discectomy
discectomy fusion
aquacel dressing
forearm fracture
ecchymosis soft
knee osteoarthritis
laminectomy fusion
history ivdu
forearm painless
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
-2.5
0.0
2.5
Log-odds ratio

Figure


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)

Figure


ROC - P: 417, N: 21928
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall - P: 417, N: 21928
Precision
Recall
Ridge Logistic
Random Forest
XGBoost

Figure