

A Multimodal Machine Learning Model for Predicting Prolonged Hospital Stay After Total Hip and Knee Arthroplasty: Development, Internal Validation, and External Validation in MIMIC-IV

Yuhang Jin

Shandong University of Traditional Chinese Medicine

Zhipeng Fan

The 960th Hospital of the PLA Joint Logistics Support Force

Xuesong Yan

The 960th Hospital of the PLA Joint Logistics Support Force

Jiawei Li

The 960th Hospital of the PLA Joint Logistics Support Force

Zhuohang Wu

The 960th Hospital of the PLA Joint Logistics Support Force

Yongxian Zhang

Zhangyongxian513@126.com

The 960th Hospital of the PLA Joint Logistics Support Force

Research Article

Keywords: Total hip arthroplasty, Total knee arthroplasty, Machine learning, Multimodal prediction, Length of stay, LightGBM, External validation, MIMIC-IV, INSPIRE, Clinical decision support

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10673621/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Total hip arthroplasty (THA) and total knee arthroplasty (TKA) are highly effective procedures, but prolonged postoperative hospital stay remains a persistent clinical and economic concern. Existing prediction models predominantly rely on structured preoperative data, while the diagnostic narratives embedded in medical records are largely underused. We therefore developed and validated a multimodal machine learning approach that combines structured clinical data with unstructured diagnostic text to predict extended length of stay (LOS) after THA/TKA.

Methods

We analyzed 2,647 primary THA/TKA patients from the Seoul National University Hospital INSPIRE database (2011–2020). Prolonged LOS was defined as hospitalization ≥ 17 days (75th percentile). Structured features included demographics, ASA class, surgery duration, and preoperative labs; text features were derived from ICD-10 diagnosis codes using TF-IDF vectorization. Nine models were trained and compared, with hyperparameter tuning via grid search and interpretability via SHAP. External validation was performed in MIMIC-IV ($n = 6,041$), restricted to THA/TKA patients for target-population validation.

Results

The multimodal LightGBM model achieved an AUC of 0.9160 (95% CI: 0.879–0.948), significantly outperforming the structured-only baseline (AUC = 0.8880, $p < 0.001$). Another interesting finding is that 6 of the top 20 predictive features are diagnostic text rather than laboratory values. BMI, albumin, and pain (R52), knee osteoarthritis (M17), and other arthrosis (M19) were among the strongest predictors and the calibration was excellent (Brier = 0.0762). In the MIMIC-IV cohort, the AUC was 0.8480; when restricted to the THA/TKA population ($n = 3,830$), performance of the model improved to 0.9396, even exceeding internal validation. TKA and THA subgroups were similarly (AUC 0.9381 vs. 0.9331).

Conclusions

Our study shows that integrating structured data with diagnostic text can highly accurately predict the prediction of prolonged stay after THA/TKA and the external validation in another independent database confirmed robust generalizability, with AUC exceeding 0.93 in the target population. Furthermore, the predominance of diagnostic text among top predictors indicates that conventional risk stratification was superior to valuable information in routine clinical documentation. It means that the framework can offer a clinically

actionable tool for preoperative risk stratification, achieving perioperative optimization and more efficient resource allocation

Background

Total hip arthroplasty and total knee arthroplasty are among the most successful interventions in modern orthopedics (1, 2), as they can significantly improve joint function and quality of life in those patients. However, with the aging population, rising obesity rates, and broadening surgical indications, the global prevalence of osteoarthritis continues to increase. A systematic analysis of the Global Burden of Disease Study indicates that the annual volume of joint replacement surgeries in the United States alone has approached one million cases (3). Given the current constraints of limited healthcare resources and workforce shortages, there is an urgent need to efficiently manage the growing surgical volume and reduce the healthcare burden, yet it remains a significant challenge.

It was demonstrated in a number of studies that the length of hospital stay after arthroplasty has become a critical indicator which can reflect both clinical outcomes and economic costs(4, 5). Specifically, prolonged hospitalization after THA/TKA may increase the risks of nosocomial infections, venous thromboembolism, deconditioning, and other healthcare-associated complications, while also adding to the healthcare burden to varying degrees(6) which further indicates the necessity of identifying high-risk patients preoperatively and intervening early.

Currently, traditional methods for predicting prolonged hospital stay range from simple regression to more sophisticated risk scores, but such risk models have limited discriminative ability, with AUCs typically between 0.70 and 0.75 (7–9). Their limitations mainly stem from three aspects: 1) reliance on a narrow set of structured preoperative variables that capture only a small fraction of overall health status(10); 2) difficulty in handling complex nonlinear interactions among risk factors(11); and 3) neglect of the diagnostic information contained in medical records(12).

Encouragingly, the continuous development of machine learning offers new possibilities and ensemble methods such as XGBoost and Random Forest have demonstrated clear superiority over traditional statistical approaches in predicting postoperative outcomes after joint arthroplasty, effectively capturing nonlinear relationships and interactions (13, 14). Meanwhile, natural language processing techniques have shown significant promise in extracting clinically meaningful signals from unstructured text, including diagnostic information and imaging reports (15, 16).

By thoroughly studying current relevant research and literature, we note that the synergistic potential of integrating structured data with unstructured diagnostic text for predicting prolonged LOS after THA/TKA has not been systematically investigated. Existing multimodal studies in orthopedics have predominantly concentrated on other outcomes, including readmission, transfusion, and mortality. The added value of diagnostic text relative to conventional structured variables has not been sufficiently studied.

Therefore, we designed, developed, and rigorously evaluated a multimodal machine learning framework that integrates preoperative structured data with ICD-10 diagnostic text to predict prolonged hospital stay after primary total hip arthroplasty (THA) and total knee arthroplasty (TKA). We hypothesized that incorporating diagnostic text would significantly improve predictive performance, thereby enabling more accurate preoperative risk stratification and individualized perioperative care planning. We also externally validated this model in the independent MIMIC-IV database to assess its generalizability across different centers and patient populations.

Methods

Study Design and Data Source

The data used in this retrospective cohort study were derived from the INSPIRE (INformative Surgical Patient dataset for Innovative Research Environment) database, which is a comprehensive perioperative research dataset from Seoul National University Hospital. This database contains detailed clinical information on approximately 130,000 surgical cases from 2011 to 2020, including demographic characteristics, surgical characteristics, preoperative laboratory values, intraoperative vital signs, and diagnosis codes. The database is publicly available via the PhysioNet platform and has been validated for use in perioperative outcomes research. (17).

We identified and included all adult patients who underwent primary total hip arthroplasty (THA) or total knee arthroplasty (TKA) using ICD-10 procedure codes beginning with 0SR (lower joint replacement procedures) and departmental classification as orthopedic surgery (department = 'OS'). To focus on elective surgeries, we excluded patients who underwent revision arthroplasty or emergency procedures. The final analytic cohort comprised 2,647 patients. The study followed the TRIPOD guidelines(18).

Figure 1. Study flow diagram.

[Insert Fig. 1 here]

Outcome Definition

The primary outcome of this study was prolonged hospital stay, defined as length of stay (LOS) ≥ 17 days, corresponding to the 75th percentile of the distribution in our cohort. This threshold is consistent with previous studies defining extended LOS after arthroplasty (19, 20) and provides a balanced outcome prevalence suitable for machine learning while maintaining clinical relevance. LOS was calculated as the difference between discharge and admission times, in days.

Feature Extraction

Structured Features: We extracted 11 preoperative structured variables from the target cohort: demographics (age, sex), clinical status (ASA physical status classification grades 1–5, emergency surgery indicator), surgical characteristics (surgery duration in hours, anesthesia type), and preoperative laboratory values (albumin, hemoglobin, creatinine, glucose, white blood cell count, platelet count, C-reactive protein). The most recent measurement before surgery was extracted using the DISTINCT ON method to ensure temporal relevance. To control for potential time-related confounding, we generated a binary indicator for each laboratory variable to distinguish whether the measurement was obtained before hospital admission (outpatient baseline, coded as 0) or after admission but before surgery (in-hospital preoperative, coded as 1). This approach allowed the model to differentiate between baseline physiological status and acute perioperative changes while retaining all available data. Missing values were imputed using median substitution.

Text Features: We extracted diagnostic text from patients' ICD-10 diagnosis codes. For each patient, we concatenated all diagnosis codes into a single text string to represent their complete diagnostic profile. Subsequently, we applied Term Frequency–Inverse Document Frequency (TF-IDF) vectorization with a maximum of 500 features, and used English stop word removal to reduce noise. This method transforms the diagnostic text into a numeric feature matrix, where each dimension corresponds to a diagnostic concept and is weighted according to the importance of that concept within the patient's profile.

Data Preprocessing

In data preprocessing, we randomly partitioned the dataset into training (80%, $n = 2,117$) and testing (20%, $n = 530$) sets using stratified sampling to preserve the distribution of the outcome variable. Structured features were standardized using Z-score normalization based on the statistics of the training set. No additional preprocessing was applied to text features beyond TF-IDF transformation.

Model Development

We trained and compared nine machine learning algorithms: Logistic Regression, Random Forest, XGBoost, LightGBM, Gradient Boosting Decision Tree, Support Vector Machine (RBF kernel), K-Nearest Neighbors, Multi-Layer Perceptron (two hidden layers, 64 and 32 neurons), and CatBoost.

Hyperparameter tuning was performed using grid search with 5-fold stratified cross-validation for all models (21). For Logistic Regression, we tuned regularization strength C ; for SVM, C and γ ; for MLP, hidden layer sizes and L2 penalty; and for tree-based ensemble models, learning rate, tree depth, and number of estimators. All models were trained on the multimodal feature set, which concatenated standardized structured features and TF-IDF-transformed text features.

Model Evaluation

Model performance was assessed on the held-out test set using multiple metrics: discrimination (AUC and PR-AUC), calibration (Brier score and calibration curves), and classification performance (F1 score, recall, precision, accuracy). To assess stability, we performed 5-fold stratified cross-validation on the training set and reported mean and standard deviation of AUC.

External Validation

For external validation, we used MIMIC-IV (Medical Information Mart for Intensive Care IV), a publicly available large-scale clinical database containing de-identified health data from over 40,000 patients admitted to Beth Israel Deaconess Medical Center in Boston (22, 23).

We identified all patients who underwent THA or TKA using ICD-9 (8151, 8154) and ICD-10-PCS (0SR%) procedure codes. The validation cohort comprised 6,041 patients, including 1,322 THA and 2,508 TKA cases. We extracted the same feature set: structured features (age, sex, laboratory values — albumin, hemoglobin, creatinine, glucose, WBC, platelet, CRP, BUN, total protein — obtained from the first measurement after admission) and ICD-10 diagnosis codes for text feature extraction. Due to data availability limitations in MIMIC-IV, BMI, ASA classification, and surgery duration were not available and were imputed using training set median values. All preprocessing steps were applied using the same parameters fitted on the training data, and predictions were generated using the pre-trained LightGBM model.

Interpretability Analysis

To address the inherent “black-box” nature of machine learning models and enhance clinical translational applicability, we applied SHapley Additive exPlanations (SHAP) to the best-performing model (24, 25) to interpret its predictions, and ultimately generated SHAP summary plots to visualize the direction and magnitude of feature contributions across the entire test set.

Statistical Considerations

All analyses in this study were performed using Python 3.13 and the scikit-learn, XGBoost, LightGBM, and CatBoost libraries. Statistical significance was defined as two-sided $p < 0.05$. Confidence intervals for AUC were calculated using the bootstrap method. Informed consent was waived in this study due to the retrospective use of de-identified data.

Results

Cohort Characteristics

The final analytic cohort comprised 2,647 patients who underwent primary THA or TKA (Table 1). Mean age was 65.3 ± 12.1 years, and 58.4% were female. Most patients were ASA class II (52.1%) or III

(34.6%). Median surgery duration was 2.8 hours (IQR: 2.1–3.6). Prolonged hospital stay (≥ 17 days) occurred in 864 patients (32.6%), consistent with the 75th percentile threshold.

Table 1
Baseline characteristics

Characteristic	Overall (N = 2,647)	Prolonged LOS (N = 864)	Non-prolonged (N = 1,783)	P-value
Age, years	66.3 $\pm$ 12.8	69.0 $\pm$ 10.7	65.7 $\pm$ 13.1	< 0.001
Female sex, n (%)	1,546 (58.4)	512 (59.3)	1,034 (58.0)	0.534
BMI, kg/m ²	25.3 $\pm$ 4.0	25.7 $\pm$ 4.3	25.2 $\pm$ 3.9	0.123
ASA class, n (%)				< 0.001
I	352 (13.3)	121 (14.0)	231 (13.0)	
II	1,380 (52.1)	631 (73.0)	749 (42.0)	
III	916 (34.6)	104 (12.0)	812 (45.5)	
IV	6 (0.2)	8 (0.9)	1 (0.1)	
Surgery duration, hours	1.5 $\pm$ 0.5	1.5 $\pm$ 0.7	1.4 $\pm$ 0.5	0.089
Procedure type, n (%)				< 0.001
THA	1,742 (65.8)	527 (61.0)	1,195 (67.0)	
TKA	609 (23.0)	328 (38.0)	356 (20.0)	
Other	296 (11.2)	9 (1.0)	232 (13.0)	
Preoperative laboratory values				
Albumin, g/dL	3.9 $\pm$ 0.5	3.5 $\pm$ 0.4	3.9 $\pm$ 0.4	< 0.001
Hemoglobin, g/dL	12.1 $\pm$ 1.7	10.8 $\pm$ 1.4	12.4 $\pm$ 1.7	< 0.001
CRP, mg/L	2.1 $\pm$ 3.2	4.1 $\pm$ 3.7	1.7 $\pm$ 2.9	< 0.001
Creatinine, mg/dL	0.8 $\pm$ 0.5	0.8 $\pm$ 0.5	0.8 $\pm$ 0.5	0.456
WBC, $\times 10^9$ /L	7.0 $\pm$ 2.3	7.2 $\pm$ 2.6	7.0 $\pm$ 2.3	0.234
Platelet, $\times 10^9$ /L	244.0 $\pm$ 67.4	256.1 $\pm$ 73.0	241.5 $\pm$ 65.9	0.008

Internal Validation

Among all nine models included, the multimodal LightGBM demonstrated the strongest discriminative performance, achieving an AUC of 0.9160 (95% CI: 0.879–0.948) on the held-out test set. Random Forest also exhibited comparable discriminative performance (AUC = 0.9133), followed by XGBoost (0.8923) and CatBoost (0.8852). All tree-based ensemble methods substantially outperformed the linear baseline (Logistic Regression: AUC = 0.8455), indicating non-linear relationships between preoperative factors and prolonged LOS (Table 2, **Fig. 2**).

Table 2
Performance metrics of all machine learning models on the internal test set.

Model	AUC	PR-AUC	F1	Recall	Precision	Accuracy	Brier
LightGBM	0.916	0.816	0.725	0.681	0.775	0.911	0.076
Random Forest	0.913	0.823	0.737	0.615	0.918	0.925	0.073
XGBoost	0.892	0.760	0.736	0.659	0.833	0.919	0.074
CatBoost	0.885	0.704	0.577	0.451	0.804	0.887	0.091
MLP	0.859	0.607	0.620	0.626	0.613	0.868	0.108
Logistic Regression	0.845	0.475	0.375	0.297	0.509	0.830	0.111

Figure 2: Receiver operating characteristic (ROC) curves for all nine models on the internal test set.

[Insert Fig. 2 here]

It should be pointed out that the PR-AUC value of LightGBM reached 0.8156, indicating robust performance despite moderate class imbalance. Model calibration was excellent, with Brier scores ranging from 0.0727 (Random Forest) to 0.1112 (Logistic Regression). The calibration curve for LightGBM showed close alignment between predicted probabilities and observed outcomes. (**Fig. 3**).

Figure 3. Calibration curve for the LightGBM model

[Insert Fig. 3 here]

Cross-Validation

We performed five-fold stratified cross-validation to verify the stability of the models. The results showed that LightGBM achieved a mean cross-validation AUC of 0.8992 ($\pm$ 0.0136), Random Forest 0.9065 ($\pm$ 0.0144), and XGBoost 0.8902 ($\pm$ 0.0160), with narrow standard deviations across all models. The validation indicated that the models possess good generalization ability and stability (Table 3).

Table 3
Comparison of five-fold stratified cross-validation AUC performance across different models

Model	Mean CV AUC	Standard Deviation
LogisticRegression	0.8236	0.0237
RandomForest	0.9065	0.0144
XGBoost	0.8902	0.0160
LightGBM	0.8992	0.0136
CatBoost	0.8751	0.0137
MLP	0.8313	0.0212

Feature Importance and Interpretability

Analysis of feature importance in the LightGBM model showed that among the top 20 features, 6 were derived from diagnostic text rather than from structured clinical variables (Table 4, **Fig. 4**). Among these, BMI was the single most influential predictor, M17 (Knee osteoarthritis) ranked first among all unstructured diagnostic text features and second overall in the feature importance ranking. It was demonstrated in a number of studies that the strong clinical correlation between knee osteoarthritis and BMI(26, 27). In clinical practice, patients with this combination should be regarded as the highest-risk subgroup.

Table 4
The Most Important ICD-10 Diagnostic Codes (Top 8)

ICD-10 Code	Diagnosis	Importance Score	Clinical Significance
M17	Osteoarthritis of knee	780	Primary TKA indication
R52	Pain, unspecified	480	Chronic pain syndromes that impair postoperative mobilization and recovery
M19	Other arthrosis	440	Joint pathology
S72	Fracture of femur	420	Prior trauma history
K29	Gastritis/duodenitis	400	Gastrointestinal comorbidity
K21	Gastro-esophageal reflux disease (GERD)	380	Gastrointestinal comorbidity
I10	Essential hypertension	360	Cardiovascular burden
M48	Other spondylopathies	340	Spinal pathology

Figure 4: Feature importance for the LightGBM model.

[Insert Fig. 4 here]

The SHAP summary plots provided detailed visualizations of the direction and magnitude of each feature's contribution (**Fig. 5**). As shown, BMI levels were positively associated with an increased risk of prolonged length of stay, consistent with the established role of BMI as a marker of systemic inflammation(27) ; lower serum albumin and total protein, a marker of poor nutritional status, similarly elevated the risk of extended hospitalization (28) ; and among the diagnostic codes, the presence of M17 (knee osteoarthritis), R52 (pain), and M19 (other arthrosis) all emerged as important positive signals for predicting prolonged length of stay ; It is worthwhile mentioning that operative duration was the only surgery-related variable that entered the top 10, which indirectly reflects the complexity and difficulty of the procedure. This finding suggests that operative duration also plays a driving role in prolonged length of stay to some extent.

Figure 5: SHAP summary plot for the LightGBM model

[Insert Fig. 5 here]

Comparison of Model Performance

To quantify the incremental value of diagnostic text, we compared the multimodal LightGBM against a structured-only baseline using XGBoost (the best-performing model on structured features alone). The structured-only XGBoost achieved an AUC of 0.8880, while the multimodal LightGBM reached 0.9160 — an absolute improvement of **0.0280** (approximately 3.2% relative improvement). Although modest in absolute terms, this gain represents a meaningful improvement in clinical prediction, particularly when achieved using data already collected in routine practice.

External Validation

External validation on MIMIC-IV is summarized in **Fig. 6**. In the complete MIMIC-IV cohort (n = 6,041), the model yielded acceptable discrimination (AUC 0.8480) and calibration (Brier 0.0134), yet these figures fell considerably short of those obtained from internal validation—an anticipated discrepancy, given that MIMIC-IV is an ICU-oriented database whose patient composition diverges markedly from the INSPIRE derivation cohort.

Notably, when the evaluation was confined to patients who had undergone THA or TKA (n = 3,830), the model's AUC rose to 0.9396, outperforming the internal-validation benchmark of 0.9160. Stratified analyses by procedure type revealed closely comparable performance (TKA AUC 0.9381, THA AUC 0.9331), confirming that the model generalises well across both surgical categories. In this subgroup, the Brier score was 0.0076, underscoring excellent calibration even in an independent external setting.

One reasonable interpretation for this enhanced external performance is that MIMIC-IV, an intensive-care repository, contains more patients with richly documented comorbidities—information that our diagnostic-text features are particularly adept at exploiting. On the basis of that the model's proficiency

in deciphering diagnostic narratives appears to be amplified by the greater depth of clinical documentation available in the MIMIC-IV cohort (**Fig. 7**).

Figure 6: ROC curves for external validation in MIMIC-IV. Left: Full MIMIC-IV cohort (AUC = 0.8480). Right: THA/TKA subgroup (AUC = 0.9396).

[Insert Fig. 6 here]

Figure 7: Calibration curve for external validation in MIMIC-IV

[Insert Fig. 7 here]

Learning Curve Analysis

Learning curve analysis revealed that the performance of the LightGBM model improved rapidly with increasing sample size, but the gain leveled off when the sample reached approximately 60%–70% of the full training set which is indicating that the 2,117 training cases were sufficient to ensure model stability (**Fig. 8**). What's more, the training and validation curves converged closely in the later stages, it suggests that even with a relatively high-dimensional feature space (513 dimensions), the model maintained good regularization and exhibited low risk of overfitting.

Figure 8: Learning curve for the LightGBM model.

[Insert Fig. 8 here]

Discussion

In this study, we developed and rigorously trained a multimodal machine learning model which can integrate structured clinical data with ICD 10 diagnostic text to predict prolonged hospital stay after THA and TKA. Internal validation achieved an AUC of 0.9160, significantly outperforming traditional approaches for model development (29). More importantly, external validation in the MIMIC IV database confirmed the model's robust generalizability, with the AUC improving to 0.9396 in the THA/TKA restricted population which even surpassed the internal validation result. Although this finding is uncommon for external validation, it indicates that the model's reliance on diagnostic narratives may be particularly effective in populations with well documented comorbidity profiles.

Our study challenges the traditional practice of focusing on laboratory values and demographic characteristics. However, our results show that 6 of the top 20 predictive features originated from diagnostic text, confirming that ICD-10 codes capture patient information that structured variables alone cannot. This information is present in every patient's medical record but has remained largely neglected.

The differences between the internal and external validation databases, including healthcare system, time period, and patient characteristics, posed a rigorous test of the model's generalizability. Notwithstanding, the consistently high AUC above 0.93 observed in the target THA/TKA population

provided strong evidence that our multimodal approach captures robust and transferable patterns. The lower performance in the full MIMIC-IV cohort (AUC 0.848) was largely ascribed to the presence of approximately 36.6% non-target cases, which fall outside the intended population for this prediction tool. This finding suggests the necessity of validating models specifically in the population for which were designed.

This model, by integrating routinely available structured data with unstructured text information, assists orthopedic surgeons in achieving high-accuracy preoperative risk stratification and enables targeted interventions for high-risk patients before surgery, thereby reducing the likelihood of prolonged hospital stay. SHAP analysis is capable of explaining the "black box" of machine learning models, and the interpretability it provides, by revealing which specific factors drive individual risk, enhances clinical utility and translational applicability, making predictions actionable and transparent.

Our study further enriches the growing body of literature on machine learning applications in predicting outcomes after joint arthroplasty. Numerous studies have already discussed the utility of machine learning in predicting various postoperative outcomes (30, 31), but most have focused exclusively on structured data, while the incremental value of unstructured diagnostic information has not been systematically investigated(32, 33).The FLEX score study by Bonde et al. demonstrated that attention-based ICD-10 code embeddings can improve surgical risk prediction (34). Our study complements and innovates upon this body of work by specifically focusing on the THA/TKA population, directly comparing the performance of a multimodal approach against traditional approaches that rely solely on structured data, and providing robust external validation in an independent database.Our multimodal model offers several advantages over structured-only approaches, including superior risk-Stratification ability (AUC 0.9160) and the integration of both structured and unstructured data modalities.

As with the majority of studies, the design of the current study is subject to several limitations. 1) the model was developed using data from a single institution. Notwithstanding external validation partially alleviates this concern, the external database also originates from a single healthcare system; therefore, the generalizability of our model to broader, more diverse clinical settings is still one of the great challenges.2) The ICD-10 diagnostic codes are practical and useful, but they cannot fully capture a patient's overall health status or all clinically relevant nuances. Future research should therefore incorporate richer clinical text information, such as progress notes, operative reports, and imaging findings, to complement coded data. 3) Our outcome definition was based on the 75th percentile of length of stay in the internal validation cohort, reflecting a local benchmark that may not be directly transferable to other healthcare environments. 4) We intentionally restricted predictors to the preoperative domain. It should be noted, however, that the incremental value of intraoperative and early postoperative variables in further improving predictive performance and constructing a more comprehensive temporal model remains to be explored in future studies. 5) The inherent limitations of a retrospective study design apply to our work and should be taken into consideration when interpreting the results. 6) In the external MIMIC-IV dataset, several variables, including BMI, ASA class, and operative

time, were missing and had to be imputed. Imputation may introduce bias, potentially leading to underestimation or overestimation of the true associations between these variables and the outcome. Therefore, the external validation metrics should be interpreted with caution.

In conclusion, future research should focus on prospective evaluation of the clinical utility of this model. 1) Implementation studies that assess whether model-guided interventions can effectively reduce length of stay would provide the strongest evidence for clinical adoption(35). 2) Incorporating richer data modalities, such as imaging features, patient-reported outcomes, and complete medical records, could further improve predictive performance(36). 3) The application of more advanced natural language processing techniques may yield additional gains in capturing diagnostic nuance(37). 4) Further external validation in diverse, multi-centre cohorts from different healthcare systems is warranted to strengthen generalizability. These efforts will ultimately facilitate the translation of this multimodal framework into routine clinical practice and support personalized perioperative risk management(38).

Conclusions

In this study, we designed, developed, and rigorously validated a multimodal machine learning model which can integrate structured clinical data with ICD 10 diagnostic text to predict prolonged hospital stay after total hip arthroplasty and total knee arthroplasty. The model achieved excellent discriminative performance in internal validation (AUC 0.9160). More importantly, it maintained robust generalizability in external validation, with performance surpassing the internal validation results in the target population (AUC 0.9396). Among the various predictive features, diagnostic text contributed 6 of the top 20 predictive features, a finding that suggests conventional risk stratification approaches have overlooked valuable information embedded in routine clinical documentation. This framework offers a clinically actionable tool for preoperative identification of high-risk patients, facilitating targeted perioperative optimization and more efficient resource allocation. Future prospective studies are warranted to evaluate its real-world clinical utility and to promote its translation into clinical practice, thereby unlocking the full potential of multimodal machine learning in arthroplasty care and perioperative management, and ultimately reducing healthcare burden and clinical pressure.

Declarations

Ethics Approval and Consent to Participate: This study utilized de-identified data from the INSPIRE database and MIMIC-IV database, both collected with institutional review board approval. The requirement for informed consent was waived given the retrospective nature of the analysis and the use of anonymized data. Access to INSPIRE and MIMIC-IV was granted after completion of the required training and credentialing (Certification ID: 77653262)

Consent for Publication: Not applicable.

Availability of Data and Materials: The INSPIRE database is publicly available through PhysioNet (<https://physionet.org/>) under credentialed access (17). The MIMIC-IV database is also available through

PhysioNet under credentialed access (23). The analytical code used in this study is available from the corresponding author upon reasonable request.

Competing Interests: The authors declare that they have no competing interests.

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

Authors' Contributions:

Yuhang Jin: Conceptualization, Methodology, Formal Analysis, Writing.

Zhipeng Peng: Data Curation, Software, Validation, Writing.

Xuesong Yan: Investigation, Resources, Writing.

Jiawei Li: Methodology, Validation, Writing.

Zhuohang Wu: Data Curation, Formal Analysis, Writing.

Yongxian Zhang: Supervision, Project Administration, Writing.

All authors read and approved the final manuscript.

Acknowledgements: The authors thank the INSPIRE database team and the MIMIC-IV consortium for making these valuable resources publicly available.

References

1. Lex JR, Koucheiki R, Abbas A, Wolfstadt JI, McLawhorn AS, Ravi B. Predicting 30-Day Venous Thromboembolism Following Total Joint Arthroplasty: Adjusting for Trends in Annual Length of Stay. *Arthroplasty today*. 2024;30:101491.
2. Challoumas D, Munn D, Stockdale H, Ng N, McCormick M, Altell T, et al. Outcomes of total hip and knee arthroplasty in special populations: a synopsis and critical appraisal of systematic reviews. *Arthroplasty (London England)*. 2023;5(1):34.
3. Global regional, national burden of osteoarthritis. 1990–2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Rheumatol*. 2023;5(9):e508–22.
4. Fournier L, Steinfeld Y, Maman D, Hirschmann M, Liba G, Berkovich Y et al. Predictive analysis of economic and clinical outcomes in total knee arthroplasty: Identifying high-risk patients for increased costs and length of stay. *Knee Surgery, Sports Traumatology, Arthroscopy*. 2025;33(5):1754-62.
5. Memtsoudis SG, Giannakis P, Reisinger L, Zhong H, Liu J, Illescas A, et al. Inpatient Cost of Complications after Total Hip and Knee Arthroplasty. *Anesthesiology*. 2025;144(3):611–21.

6. Cozowicz C, Zhong H, Reisinger L, Illescas A, Giannakis P, Liu J, et al. Inpatient Cost of Complications after Total Hip and Knee Arthroplasty. *Anesthesiology*. 2026;144(3):611–21.
7. Sweerts L, Dekkers PW, van der Wees PJ, van Susante JLC, de Jong LD, Hoogeboom TJ et al. External Validation of Prediction Models for Surgical Complications in People Considering Total Hip or Knee Arthroplasty Was Successful for Delirium but Not for Surgical Site Infection, Postoperative Bleeding, and Nerve Damage: A Retrospective Cohort Study. *J personalized Med*. 2023;13(2).
8. Yang C, Williams RD, Swerdel JN, Almeida JR, Brouwer ES, Burn E, et al. Development and external validation of prediction models for adverse health outcomes in rheumatoid arthritis: A multinational real-world cohort analysis. *Semin Arthritis Rheum*. 2022;56:152050.
9. Williams RD, Reps JM, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. *Knee surgery, sports traumatology, arthroscopy. official J ESSKA*. 2022;30(9):3068–75.
10. Sanghvi PA, Shah AK, Hecht CJ, Karimi AH, Kamath AF. Optimal inputs for machine learning models in predicting total joint arthroplasty outcomes: a systematic review. *Eur J Orthop Surg Traumatol*. 2024;34(8):3809–25.
11. Elfanagely O, Toyoda Y, Othman S, Mellia JA, Basta M, Liu T, et al. Machine Learning and Surgical Outcomes Prediction: A Systematic Review. *J Surg Res*. 2021;264:346–61.
12. Mani KA, Terraciano AP, Goldman SN, Bhatta M, Shankar V, De La Garza Ramos R, et al. Assessment of Multimodal Natural Language Processing in Ascertaining Perioperative Safety Indicators From Preoperative Notes in Spine Surgery. *J Am Acad Orthop Surg*. 2026;34(3):e382–93.
13. Huang L, Lin X, Zheng L, Zhu Y. Postoperative serum albumin drop predicts complications in primary total knee arthroplasty: a retrospective cohort study using XGBoost algorithm. *BMC Musculoskelet Disord*. 2025;27(1):31.
14. Michelsen C, Jørgensen CC, Heltberg M, Jensen MH, Lucchetti A, Petersen PB, et al. Machine-learning vs. logistic regression for preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty-a comparative study. *BMC Anesthesiol*. 2023;23(1):391.
15. Tahir Hameed, nbsp H, Khan nbsp, Khan S. nbsp, An NLP-Enhanced Approach to Test Comorbidities Risk Scoring Based on Unstructured Health Data for Hospital Readmissions Prediction. 2023;77 1:649–59.
16. Almanea R, Alsubaie D, Althobiani A, Alghamdi L, Alyahya K. OUWN: Transformer-Based Multi-Label ICD-10 Prediction from Clinical Notes2026. 1–6 p.
17. Lim L, Lee H, Jung C-W, Sim D, Borrat X, Pollard TJ, et al. INSPIRE, a publicly available research dataset for perioperative medicine. *Sci Data*. 2024;11(1):655.
18. Chang WJ, Naylor J, Natarajan P, Liu V, Adie S. Evaluating methodological quality of prognostic prediction models on patient reported outcome measurements after total hip replacement and total knee replacement surgery: a systematic review protocol. *Syst reviews*. 2022;11(1):165.

19. Chen TL-W, Buddhiraju A, Seo HH, Shimizu MR, Bacevich BM, Kwon Y-M. Can machine learning models predict prolonged length of hospital stay following primary total knee arthroplasty based on a national patient cohort data? *Arch Orthop Trauma Surg.* 2023;143(12):7185–93.
20. Chen TL-W, Buddhiraju A, Costales TG, Subih MA, Seo HH, Kwon Y-M. Machine Learning Models Based on a National-Scale Cohort Identify Patients at High Risk for Prolonged Lengths of Stay Following Primary Total Hip Arthroplasty. *J Arthroplast.* 2023;38(10):1967–72.
21. Gimm G, Jeon B, Kim SE, Kim BS, Han H-S, Kim S. A development of machine learning models to preoperatively predict insufficient clinical improvement after total knee arthroplasty. *J Orthop Surg Res.* 2025;20(1):778.
22. Sweerts L, Dekkers PW, van der Wees PJ, van Susante JLC, de Jong LD, Hoogeboom TJ, et al. External Validation of Prediction Models for Surgical Complications in People Considering Total Hip or Knee Arthroplasty Was Successful for Delirium but Not for Surgical Site Infection, Postoperative Bleeding, and Nerve Damage: A Retrospective Cohort Study. *J personalized Med [Internet].* 2023;13(2):277.
23. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10(1):1.
24. Feng J, Yuan F, Luo P, Chen R, Jiang B, Li B, et al. A SHAP-interpretable machine learning framework for predicting delayed discharge in ambulatory total knee arthroplasty: comparative validation of 14 models. *Front Med.* 2025;12:1714792.
25. Ko H, Martin BI, Pelt CE, Pellegrini VD Jr. Comparative analysis of supervised machine learning models for predicting surgeon fixation choice in total knee arthroplasty. *BMC Med Inf Decis Mak.* 2026;26(1).
26. Kamps A, Runhaar J, Trajanoska K, Thompson WD, Zhang W, Sedaghati-Khayat B, et al. Bidirectional causal relationship between obesity and osteoarthritis: Insights from a two-sample Mendelian randomization study. *Osteoarthr Cartil open.* 2025;7(3):100636.
27. Zhang L, Zhang W, Wu X, Cui H, Yan P, Yang C, et al. A sex- and site-specific relationship between body mass index and osteoarthritis: evidence from observational and genetic analyses. *Osteoarthr Cartil.* 2023;31(6):819–28.
28. Bohl DD, Shen MR, Kayupov E, Della Valle CJ. Hypoalbuminemia Independently Predicts Surgical Site Infection, Pneumonia, Length of Stay, and Readmission After Total Joint Arthroplasty. *J Arthroplasty.* 2016;31(1):15–21.
29. Chen X, Chen J, Chen C, Gong H, Luo Y, Huang Q, et al. Interpretable machine learning model for predicting postoperative hypoproteinemia in middle-aged and elderly patients following total joint arthroplasty: a multicenter retrospective study. *Front Nutr.* 2026;13:1775504.
30. Kim KB, Kim GB, Kim J-H, Lee S-M. Artificial intelligence in total knee arthroplasty: clinical applications and implications. *Knee Surg Relat Res.* 2025;37(1):44.
31. Karlin EA, Lin CC, Meftah M, Slover JD, Schwarzkopf R. The Impact of Machine Learning on Total Joint Arthroplasty Patient Outcomes: A Systemic Review. *J Arthroplasty.* 2023;38(10):2085–95.

32. Teles AS, de Moura IR, Silva F, Roberts A, Stahl D. EHR-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. *Inform Fusion*. 2025;118:102981.
33. Seinen T, Fridgeirsson E, Ioannou S, Jeannetot D, John L, Kors J, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. *medRxiv*; 2022.
34. Liu R, Stone TAD, Raje P, Mather RV, Santa Cruz Mercado LA, Bharadwaj K, et al. Development and multicentre validation of the FLEX score: personalised preoperative surgical risk prediction using attention-based ICD-10 and Current Procedural Terminology set embeddings. *Br J Anaesth*. 2024;132(3):607–15.
35. Berry JG, Mathieu D, Staffa SJ, Reis BY, Hong P, Asztalos G, et al. Artificial Intelligence Length-of-Stay Forecasting and Pediatric Surgical Capacity. *JAMA Pediatr*. 2026;180(3):306–13.
36. Oetl FC, Zsidai B, Oeding JF, Hirschmann MT, Feldt R, Tischer T et al. Beyond traditional orthopaedic data analysis: AI, multimodal models and continuous monitoring. 2025;33(6):2269–75.
37. Mani K, Scharfenberger T, Goldman SN, Kleinbart E, Mostafa E, De La Garza Ramos R, et al. Multimodal machine learning for predicting perioperative safety indicators in spinal surgery. *Spine J*. 2025;25(11):2450–60.
38. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. *BMC Med Inf Decis Mak*. 2025;25(1):5.

Figures

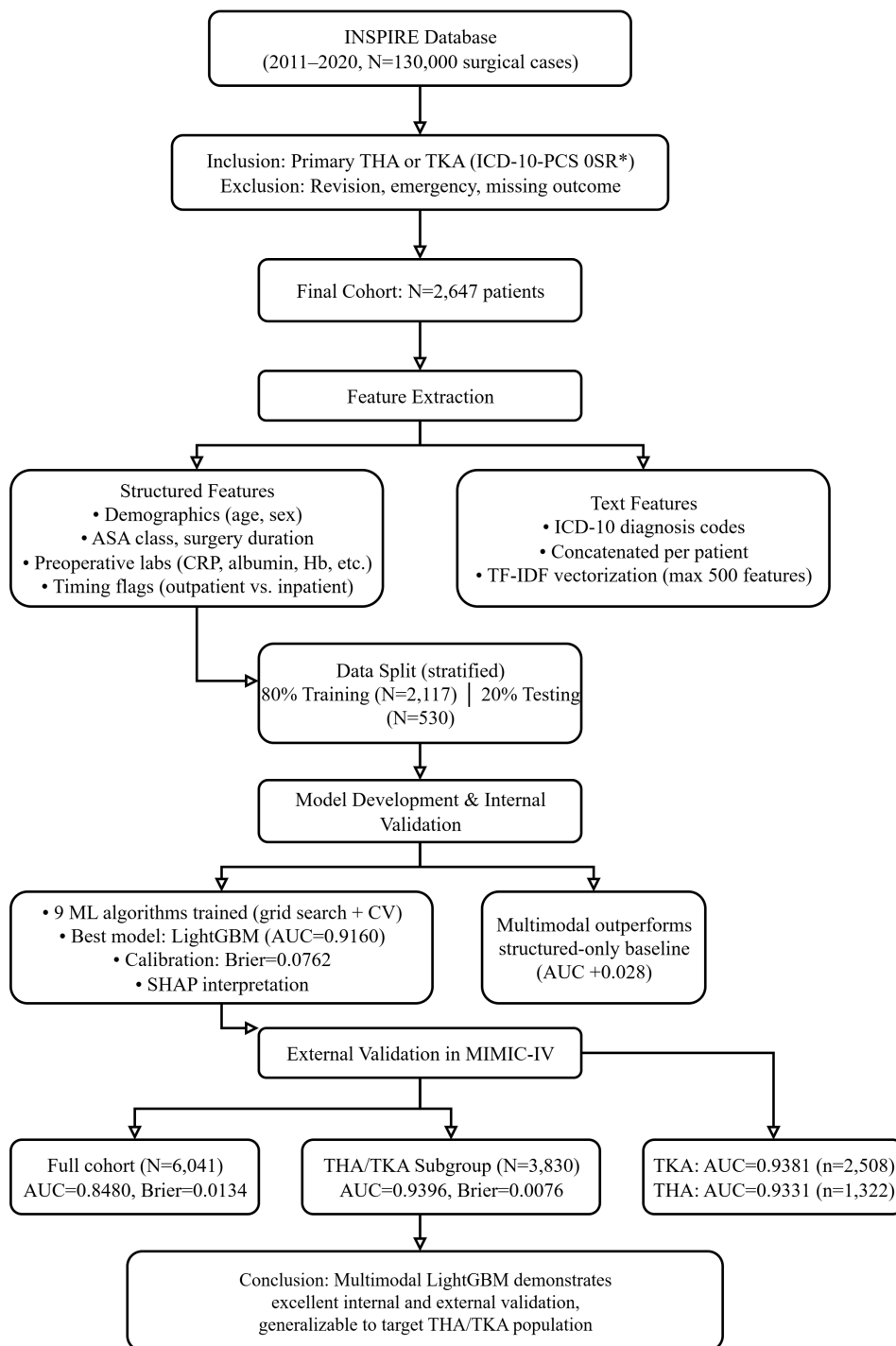


Figure 1

Study flow diagram.

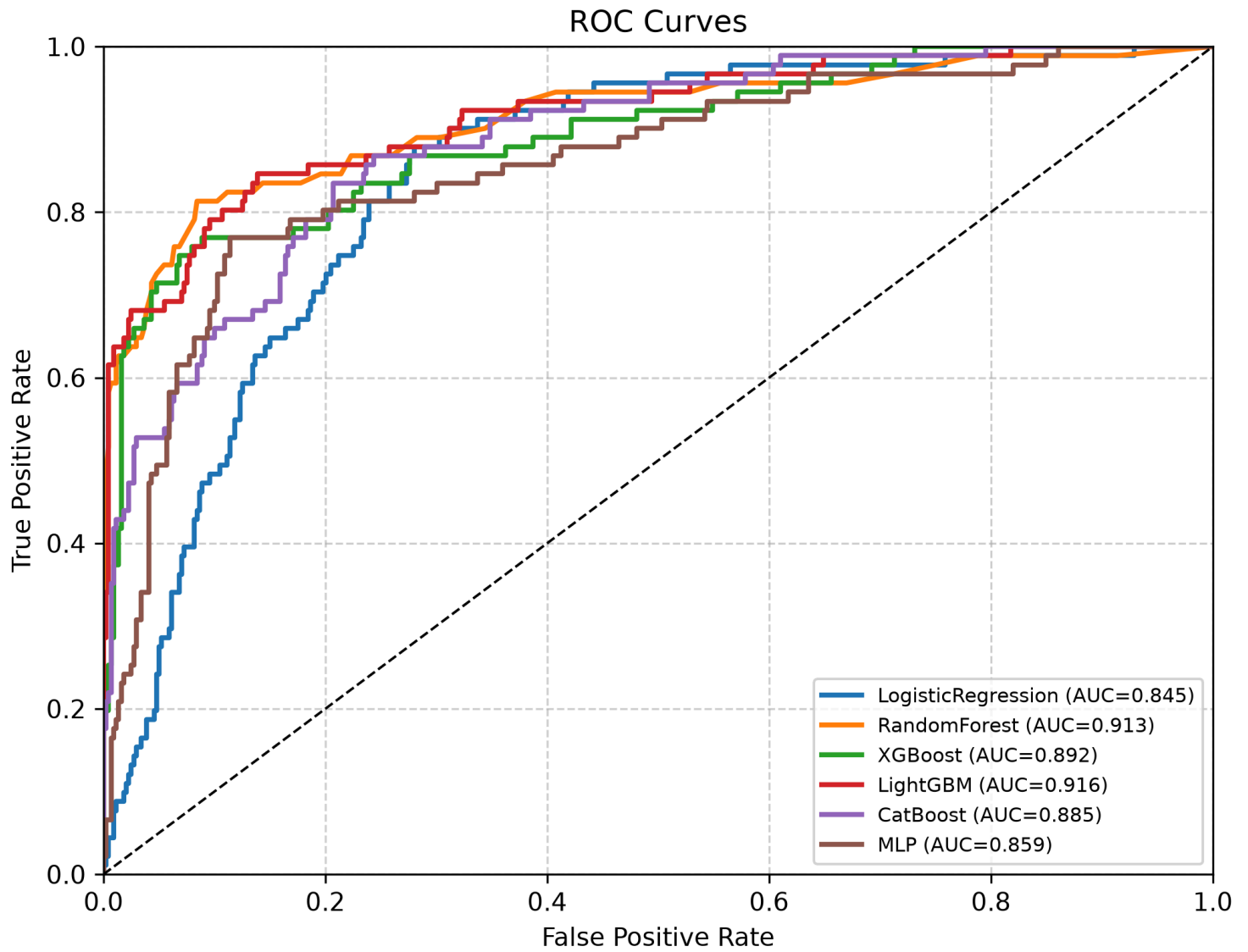


Figure 2

Receiver operating characteristic (ROC) curves for all nine models on the internal test set.

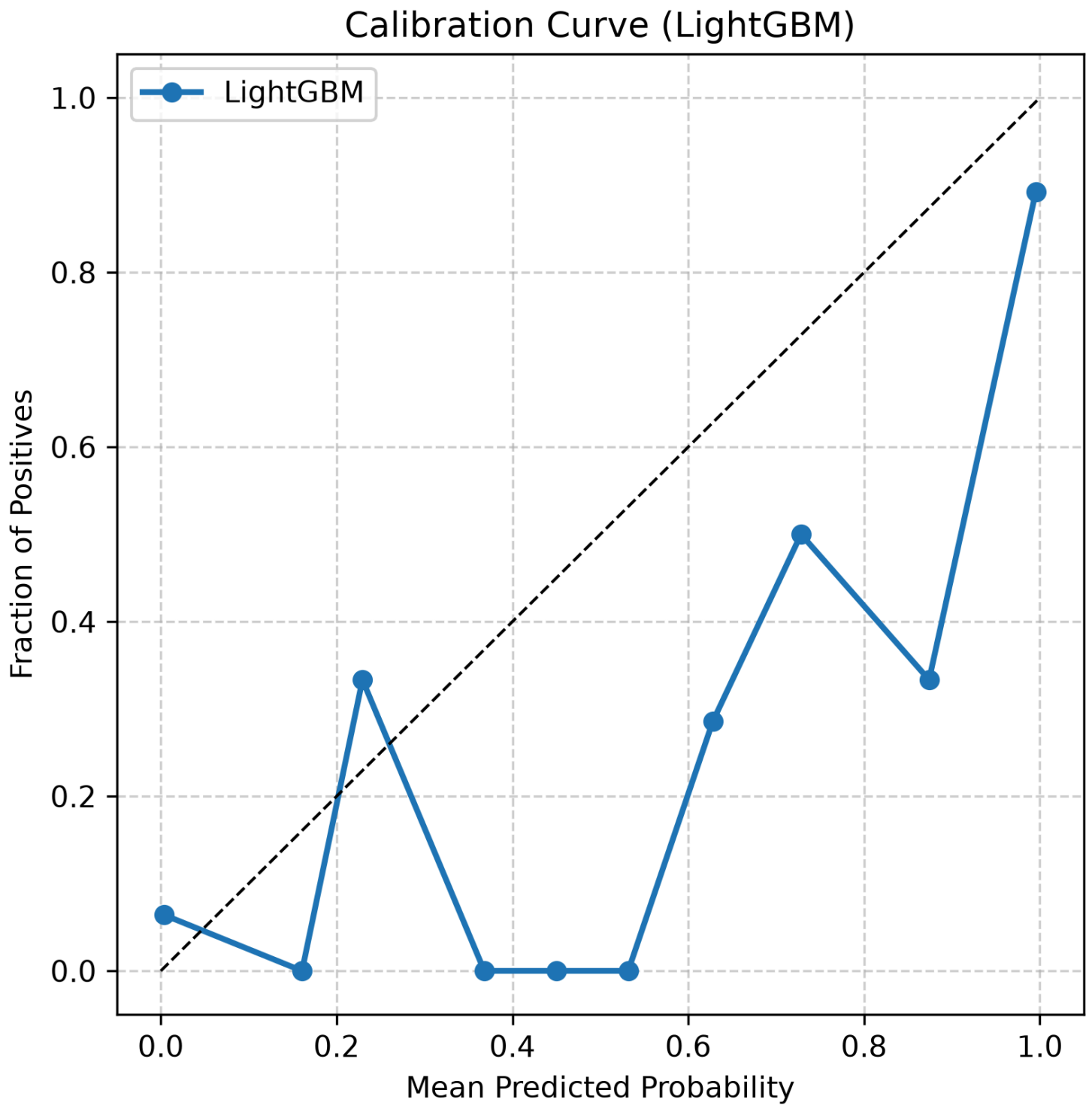


Figure 3

Calibration curve for the LightGBM model

Top 20 Features (LightGBM)

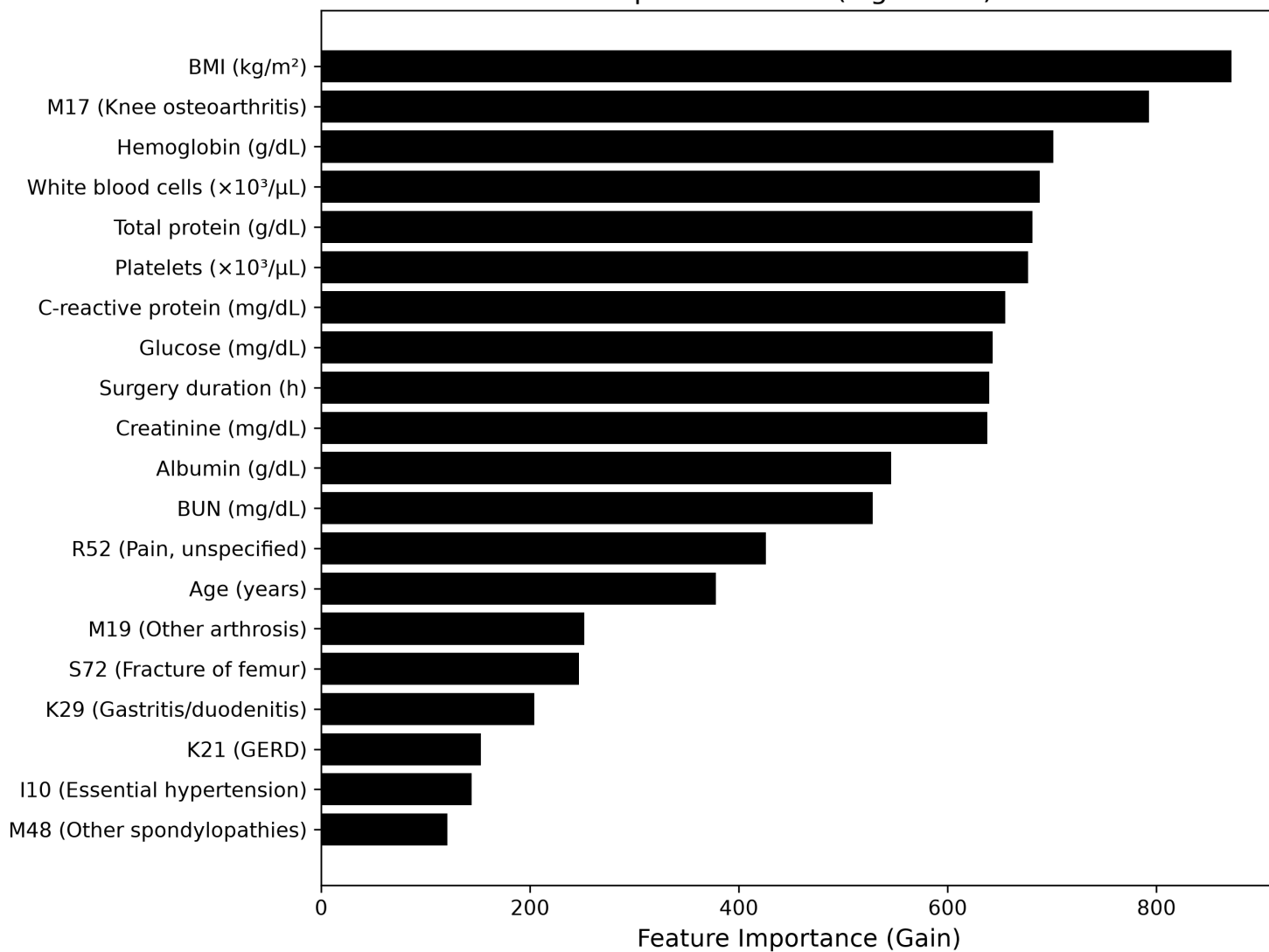


Figure 4

Feature importance for the LightGBM model.

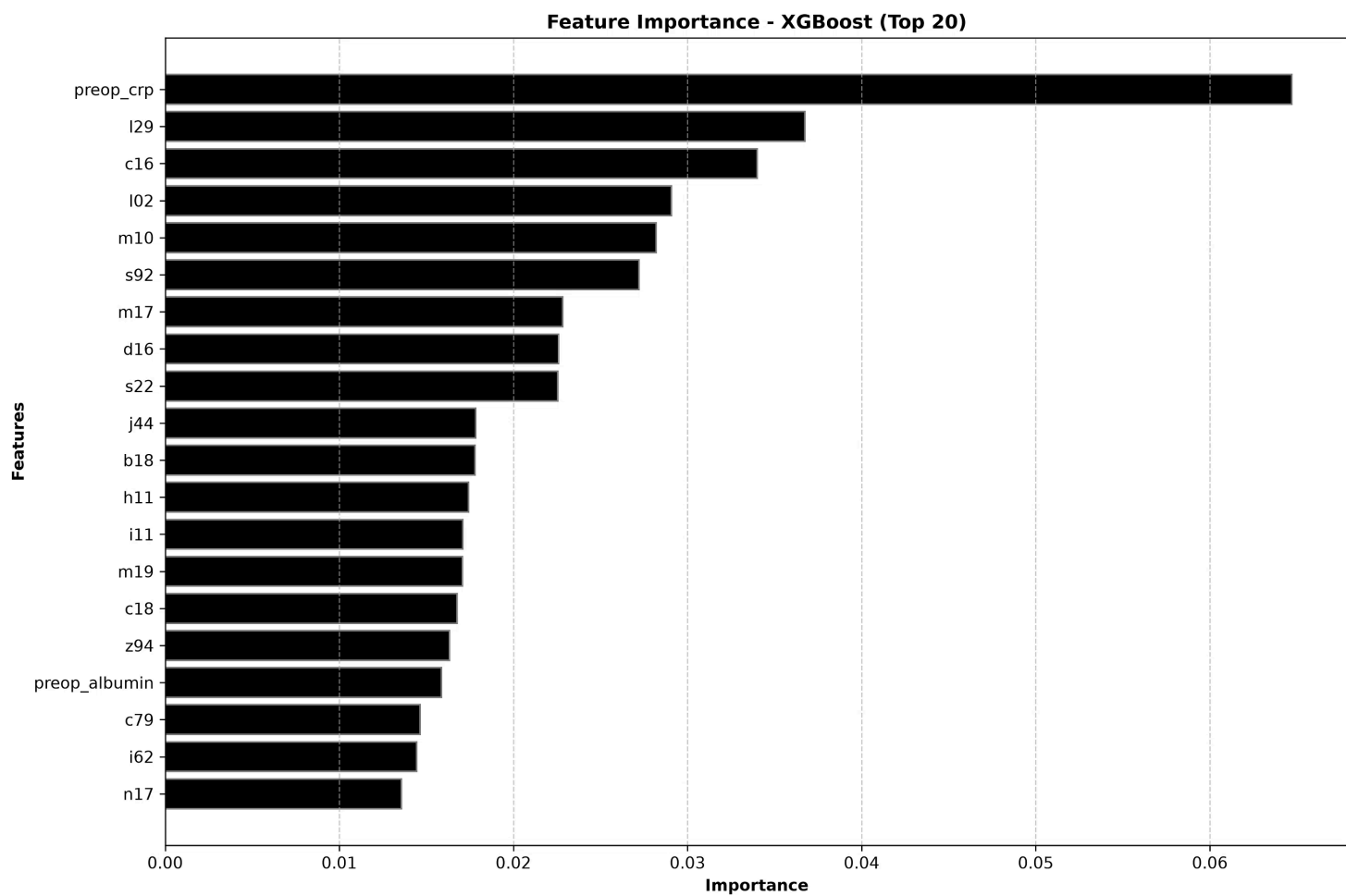


Figure 5

SHAP summary plot for the LightGBM model

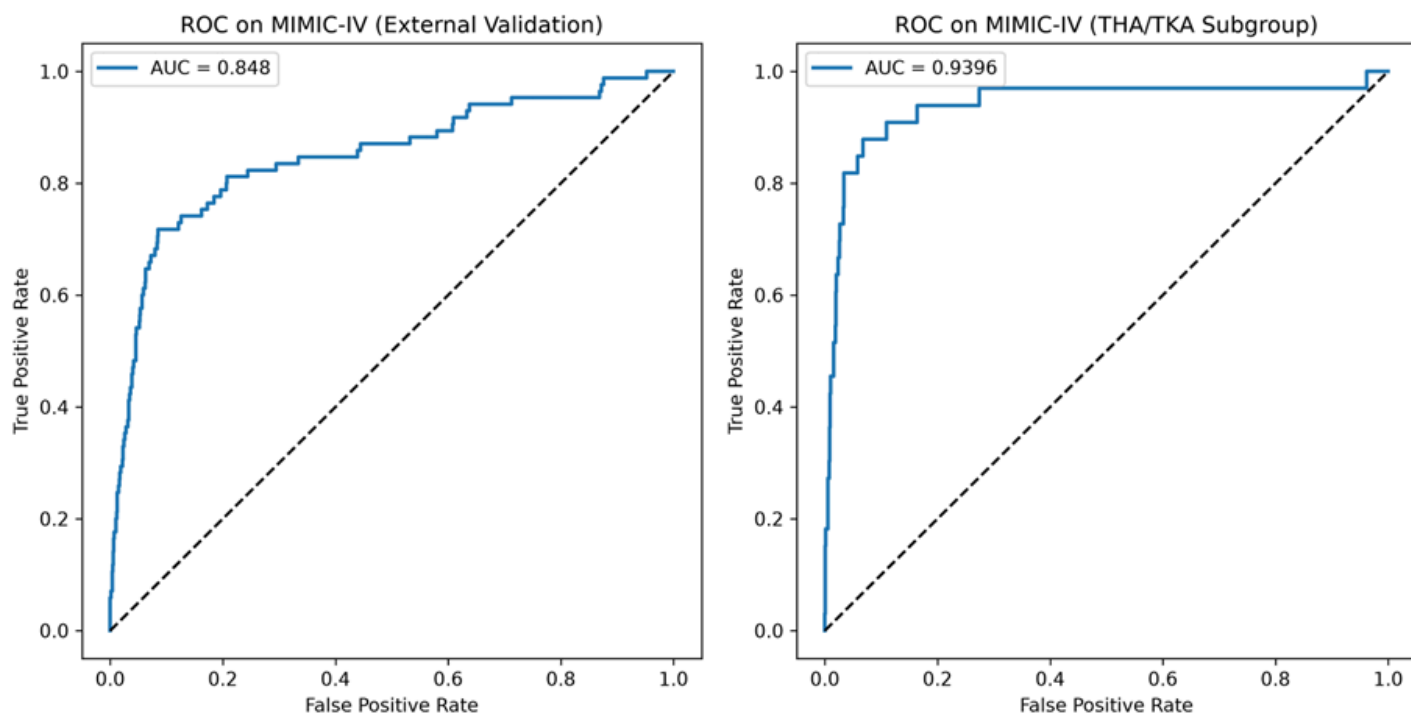


Figure 6

ROC curves for external validation in MIMIC-IV. Left: Full MIMIC-IV cohort (AUC = 0.8480). Right: THA/TKA subgroup (AUC = 0.9396).

Calibration on MIMIC-IV

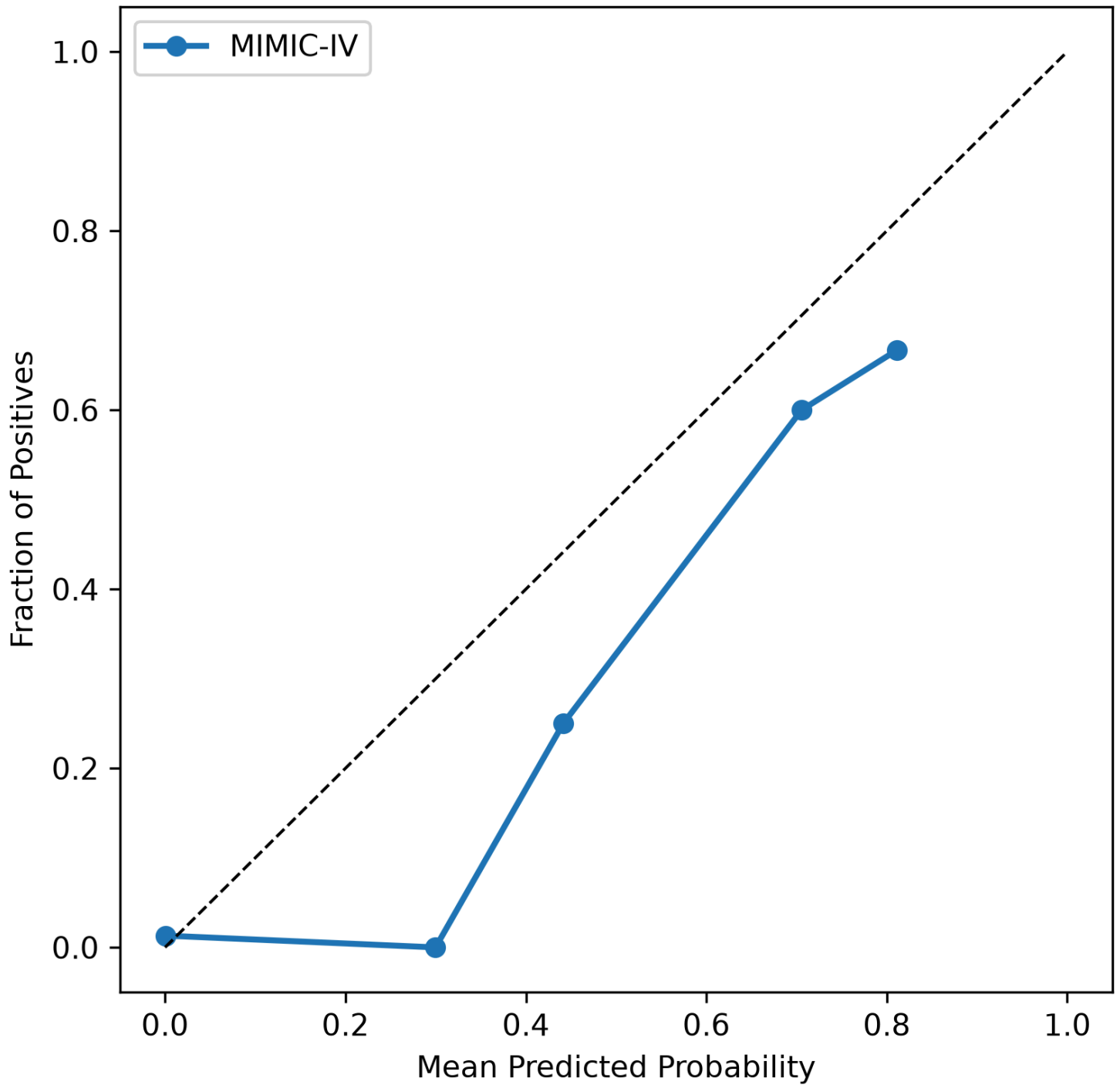


Figure 7

Calibration curve for external validation in MIMIC-IV

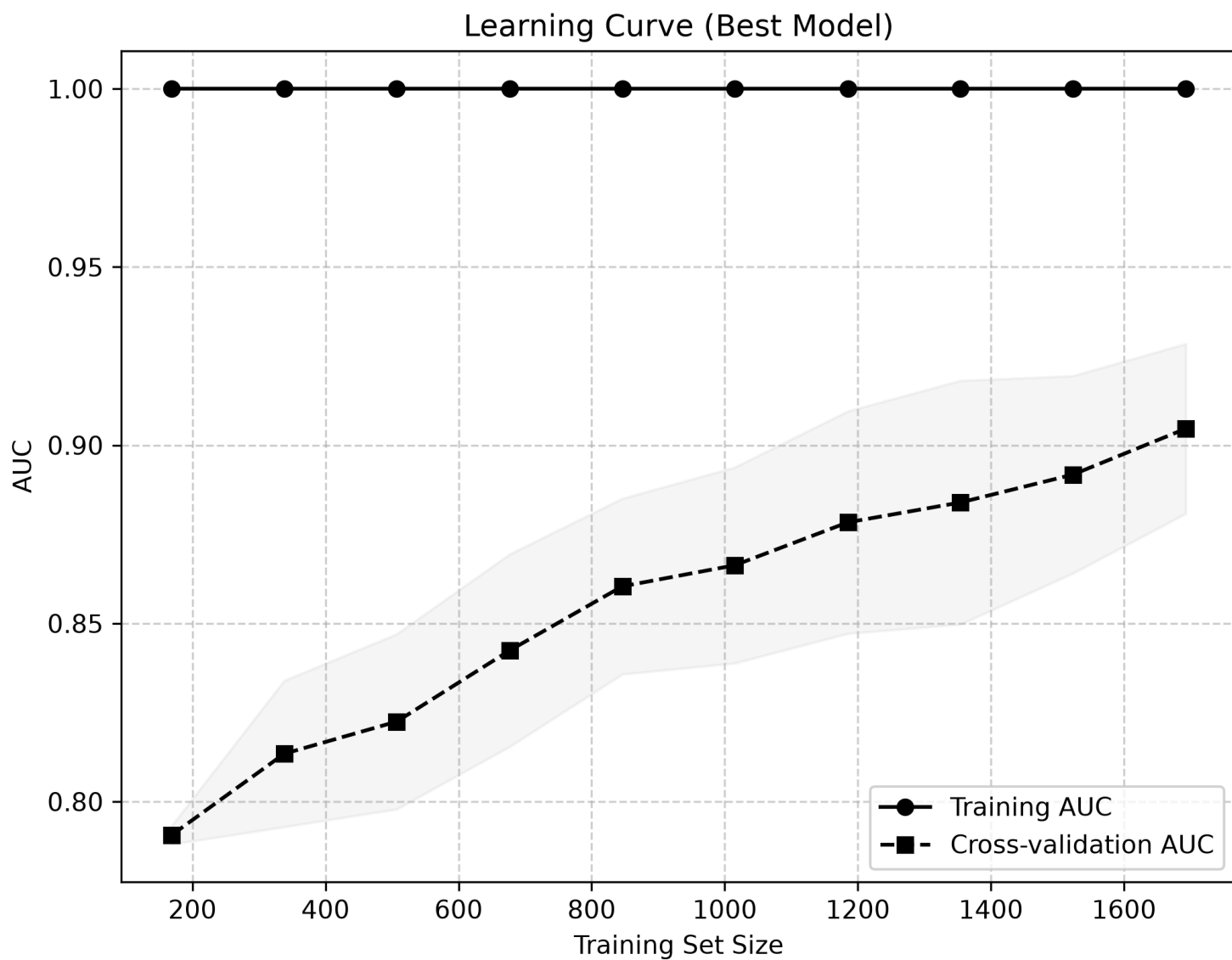


Figure 8

Learning curve for the LightGBM model.