A Predictive Model for PICC-related Thrombosis in Sepsis Patients Using XGBoost Algorithm

Wei Hao
National Cancer Center, Chinese Academy of Medical Sciences & Peking Union Medical College
Tian-yu She
Xi'an Electric Power Central Hospital
Zhen-nan Yuan
National Cancer Center, Chinese Academy of Medical Sciences & Peking Union Medical College
Li-na Liu
LiulinaHB2H@163.com
The Second Hospital of Hebei Medical University

Article

Keywords: PICC, thrombosis, sepsis, XGBoost, predictive modeling

Posted Date: May 9th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6449820/v1

Additional Declarations: No competing interests reported.

Abstract

Background

Percutaneous insertion of central venous catheters (PICC) is critical for the management of sepsis patients requiring prolonged intravenous therapy; however, it poses significant complications, including thrombosis. Identifying risk factors for PICC-related thrombosis can enhance clinical management and patient outcomes. This study aimed to develop a predictive model for PICC-related thrombosis in sepsis patients using the XGBoost algorithm.

Methods

We analyzed data from 8,128 ICU patients diagnosed with sepsis and using PICC from the Medical Information Mart for Intensive Care IV version 3.1 (MIMIC-IV 3.1) database. Patients were divided into a training set (70%, n = 5,690) and a validation set (30%, n = 2,438). Variables included demographic, laboratory, and clinical factors potentially associated with PICC-related thrombosis. An XGBoost model was developed and validated, with performance assessed using the area under the receiver operating characteristic curve (AUC) and SHAP analysis for interpretability. Decision curve analysis confirmed the clinical utility of the model.

Results

The XGBoost model achieved an AUC of 0.761 (95% CI: 0.734–0.787) in the training set and 0.766 (95% CI: 0.731–0.801) in the validation set. The calibration curve demonstrated good calibration of the model, indicating that predicted probabilities of thrombosis closely aligned with observed outcome. Decision curve analysis confirmed clinical utility, yielding a net benefit of 0.31 at 20% risk threshold, outperforming treat-all/none strategies. Key predictors, including white blood cell count, hemoglobin levels, age, creatinine levels, and platelet count, were identified using SHAP analysis in the XGBoost predictive model for PICC-related thrombosis, with the top ten predictors significantly contributing to the model's performance.

Conclusions

The XGBoost model is an effective predictor of PICC-related thrombosis among sepsis patients, indicating its potential role in guiding clinical decision-making for the management of high-risk patients.

1 Background

Sepsis remains a leading cause of morbidity and mortality in critically ill patients[1, 2]. Central venous catheters, particularly peripherally inserted central catheters (PICC), are commonly used in the

management of sepsis patients requiring prolonged medication, nutrition, and fluid therapy[3]. However, the use of PICCs is not without risks, with PICC-related thrombosis being a serious complication[4–6] that can lead to decreased venous return, increased risk of infection, and prolonged ICU stays. Identifying patients at high risk for developing thrombosis is essential to implement preventive measures and optimize care. Numerous studies have identified various factors associated with PICC-related thrombosis[5, 7, 8], including patient demographics, laboratory values, and comorbid conditions. However, existing models often do not leverage advanced machine learning techniques capable of analyzing complex interactions between numerous predictive variables. The XGBoost algorithm has gained popularity due to its efficiency and high accuracy, making it suitable for developing robust predictive models in healthcare[9–11]. This study aims to construct a predictive model for identifying risk factors for PICC-related thrombosis in sepsis patients using the XGBoost algorithm, employing data from the Medical Information Mart for Intensive Care IV version 3.1 (MIMIC-IV 3.1).

2 Patients and Methods

2.1 Database and Study Population

This retrospective cohort study utilized de-identified data from the MIMIC-IV 3.1 database, which includes a comprehensive repository of clinical information on patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center from 2008 to 2022. The study included adult patients (age ≥ 18 years) diagnosed with sepsis who had a PICC catheter placed for more than 48 hours. Patients with substantial missing data (more than 20% for necessary variables), multiple ICU admissions, or hematologic tumors were excluded. A total of 8,128 patients were included, with a random split into a training set (70%) and validation set (30%). Flowchart illustrating the patient selection process in MIMIC IV 3.1 was described in Fig. 1. A total of 538 patients were diagnosed with PICC-related thrombosis through ultrasound.

2.2 Data Collection

Data were collected on various factors predicted to be correlated with PICC-related thrombosis. Variables included demographics (age, gender), laboratory results (e.g., hemoglobin, platelet count, creatinine), and comorbidities (such as myocardial infarction, congestive heart failure, renal disease). The primary outcome was the occurrence of PICC-related thrombosis during the ICU stay.

2.3 Model Development and Validation

The predictive model was developed using the XGBoost algorithm, incorporating hyperparameter tuning for optimal performance leveraging the caret package in R. The dataset was divided into two subsets, with 70% allocated to the training set and 30% to the validation set, which facilitated a rigorous internal validation process. The model's discriminatory ability was evaluated using the area under the ROC curve (AUC) on both training and validation sets, providing a quantitative measure of its predictive accuracy. To further interpret model predictions and identify the most significant features contributing to PICC-related

thrombosis, SHAP (SHapley Additive exPlanations) values were calculated using the shapviz R package. This analysis aids in understanding model behavior by attributing the contributions of individual features to the overall predictive output. SHAP values provide insights into how each feature affects the model's predictions, allowing us to discern which factors are most influential in determining the risk of thrombosis associated with PICC placement.

2.4 Statistical Analysis

Missing data (< 20% of variables) were handled via multiple imputation with the MICE package, generating five complete datasets using predictive mean matching, with pooled results adjusted for imputation uncertainty. To prevent data leakage, preprocessing steps (e.g., normalization, feature scaling) were applied exclusively to the training set, and transformations were later replicated in the validation set using saved parameters. Continuous variables were reported as mean ± standard deviation for normally distributed data or median with interquartile range (IQR) for skewed distributions, while categorical variables were expressed as counts and percentages. Group comparisons between training and validation cohorts utilized chi-square tests for categorical variables and independent samples t-tests or Mann-Whitney U tests for continuous variables, depending on distribution normality. To mitigate overfitting, the XGBoost model incorporated L2 regularization ($\lambda = 0.1$) and early stopping (50 rounds without validation AUC improvement). Hyperparameter optimization was conducted via a grid search over 200 combinations, including learning rate (0.01–0.3), max tree depth (2–6), and subsample ratio (0.6–1.0), with 5-fold cross-validation maximizing the F1 score. Model performance was evaluated using AUC (95% CI estimated by DeLong's method), precision, F1 score, Brier score (calibration error), and mean squared error (MSE). Calibration curves were generated via locally estimated scatterplot smoothing (LOESS), and decision curve analysis (DCA) quantified clinical utility by net benefit across threshold probabilities (10–50%). Statistical significance was set at $p < 0.05$ (two-tailed). All analyses were performed using R statistical software (version 4.3.0).

3 Results

3.1 Patient Characteristics and Data Preprocessing of enrolled population

The study cohort comprised a total of 8,128 sepsis patients who underwent the placement of PICC, and these patients were randomly divided into two distinct groups: a training set consisting of 5,690 (70%) individuals and a validation set comprising 2,438 (30%) individuals. Notably, the baseline demographics and clinical features of the patients were well-balanced between the two cohorts, as illustrated in Table 1, ensuring that the comparison between the training and validation sets would be both fair and scientifically valid.

3.2 Predictive Model Performance of the XGBoost model

The XGBoost model exhibited impressive capabilities in both discrimination and calibration, showcasing its effectiveness in predictive analytics. In the training set, the area under the curve (AUC) was recorded at 0.761, with a 95% confidence interval ranging from 0.734 to 0.787, as illustrated in Fig. 2A. This level of performance was not only commendable but also remained remarkably consistent when evaluated against the validation cohort, where the AUC was slightly higher at 0.766, accompanied by a 95% confidence interval of 0.731 to 0.801, as depicted in Fig. 2B. Furthermore, the model achieved a precision score of 0.75 and an F1 score of 0.76 within the training set, which underscores its high reliability and robustness in making positive predictions. The calibration curves further reinforced these findings, demonstrating minimal deviation from the ideal line, with a Brier score of just 0.14. This indicates a strong capability for accurate probability estimation, both in the training set, as shown in Fig. 2C, and in the testing set, represented in Fig. 2D.

3.3 Clinical Utility and Decision Curve Analysis of the XGBoost model

DCA demonstrated the model's superiority over default strategies (treat-all or treat-none) across threshold probabilities of 5–50% (Fig. 3). At a 10% risk threshold, the model provided a net benefit of 0.31, meaning 31 additional high-risk patients per 100 would receive appropriate interventions without unnecessary prophylaxis. Sensitivity analyses confirmed robustness to variations in imputation methods and hyperparameters.

3.4Key Predictors of PICC-Related Thrombosis based on SHAP analysis

SHAP analysis revealed the ten most significant predictors for PICC catheter-related thrombosis, ranked by their average absolute SHAP values: white blood count, platelet count, history of myocardial infarction, hemoglobin level, creatinine concentration, partial thromboplastin time (PTT), age, presence of mild liver disease, prothrombin time (PT), and diabetes without chronic complications (Fig. 4A). These parameters demonstrated varying degrees of influence on thrombosis risk within the predictive model. To elucidate individual prediction mechanisms, SHAP force plots were employed to visualize feature contributions in two representative cases (Fig. 4B, 4C). The color gradient illustrates directional impact: red features (leftward arrows) denote risk-reducing parameters with negative SHAP values, while yellow features (rightward arrows) indicate risk-enhancing parameters with positive SHAP values. Bar length quantitatively represents the magnitude of each feature's effect. The baseline reference value (E[f(x)]) corresponds to the model's mean predicted probability across the cohort. Notably, in true-positive cases, the model effectively stratified in-hospital mortality risk through this interpretable feature interaction framework.

4 Discussion

Sepsis remains a leading cause of morbidity and mortality worldwide, with a complex pathophysiology that includes inflammatory responses, immune dysregulation, and microcirculatory disturbances [12, 13]. One significant complication associated with sepsis is the increased risk of thrombosis, particularly in patients requiring PICC. The use of PICC has become commonplace in the management of sepsis due to their ability to facilitate long-term intravenous access for medication administration, nutrition, and fluid therapy. However, the risk of thrombus formation associated with PICC use is a growing concern, necessitating the development of predictive models to identify high-risk patients and optimize therapeutic strategies [14]. This study aims to explore the risk factors associated with PICC-related thrombosis in patients diagnosed with sepsis. Utilizing a retrospective cohort design, we analyzed data from the MIMIC-IV database, which provides a rich source of clinical information for understanding patient outcomes in critical care settings. By employing advanced machine learning techniques, we seek to develop a predictive model for identifying individuals at heightened risk for PICC-related thrombosis, thereby informing clinical practices and optimizing resource allocation in the management of septic patients [15].

The innovative aspect of this study lies in its comprehensive analysis of risk factors for PICC-related thrombosis in sepsis patients. Prior studies have primarily focused on isolated clinical parameters and their association with thrombosis risk[16–18], often neglecting the intricate interplay between these factors in the context of sepsis. By employing advanced machine learning techniques such as XGBoost and SHAP analysis, our research not only identifies key predictors but also elucidates their relative contributions to thrombosis risk, thereby enhancing model interpretability and clinical applicability. The clinical implications of our findings are significant. By identifying specific predictors of PICC-related thrombosis, healthcare providers can better tailor management strategies for septic patients, potentially reducing complications and improving outcomes. The predictive model, validated through rigorous statistical methodologies, demonstrates strong discriminative power, indicating its utility in clinical decision-making.

The SHAP analysis identified ten critical predictors of PICC-related thrombosis, including white blood cell count, platelet count, history of myocardial infarction, hemoglobin level, creatinine concentration, PTT, age, mild liver disease, PT, and diabetes without chronic complications. These findings align with prior studies emphasizing the role of inflammatory, coagulation, and comorbidity-related factors in thrombotic risk. Elevated white blood cell count, indicative of systemic inflammation, has been associated with endothelial dysfunction and hypercoagulability in sepsis patients[19]. Similarly, platelet count demonstrated a dual role, where both thrombocytopenia and elevated counts may reflect dysregulated coagulation pathways, as observed in models predicting liver-related events in NAFLD[20]. Coagulation parameters (PT and PTT) and mild liver disease highlighted the interplay between hepatic synthetic function and thrombosis. Prolonged PT/PTT, often linked to coagulopathy in critical illness, were also predictive in models for hepatitis B-related liver failure [21]. The interaction between the inflammatory response and coagulation mechanisms is a critical aspect of the pathophysiology of sepsis and its associated complications, including thrombosis. Mild liver disease may disrupt anticoagulant protein synthesis[22–24], further supporting its role in thrombogenesis. These predictors collectively underscore

the multifactorial nature of PICC-related thrombosis, integrating inflammation, coagulation dysregulation, and comorbid burden. The model's alignment with established risk factors across diverse clinical contexts [25–27] reinforces its robustness, while its machine learning framework enhances precision in sepsis-specific thrombosis prediction. Thrombosis is a multifactorial condition influenced by various risk factors, including the presence of underlying diseases and comorbidities in patients. Chronic conditions such as obesity, diabetes, hypertension, and cardiovascular diseases significantly increase the risk of thrombotic events. For instance, patients with diabetes have been shown to have a higher incidence of thrombosis due to factors such as increased platelet activation and altered coagulation profiles[28]. Additionally, conditions like chronic kidney disease (CKD) and liver dysfunction can exacerbate the risk of thrombosis by affecting hemostatic mechanisms, leading to a prothrombotic state[29]. The interplay between these underlying diseases and the coagulation cascade is complex, as inflammation associated with chronic diseases can lead to endothelial dysfunction, further promoting thrombosis[30].

Nevertheless, this study is not without its limitations. The retrospective design may introduce biases related to data collection and patient selection, and the reliance on a single database could limit the generalizability of findings. Moreover, while the predictive model shows promise, it requires validation in prospective studies to confirm its utility in varied clinical environments. Future research should also explore the long-term implications of identified risk factors and consider the integration of additional variables, such as treatment modalities and patient demographics, to refine the predictive capabilities of the model further[31, 32]. Addressing these limitations will be crucial for advancing our understanding of thrombosis risk in septic patients and improving clinical outcomes.

In conclusion, our study offers significant insights into the predictors of PICC-related thrombosis among septic patients, highlighting the potential utility of the XGBoost model in enhancing clinical decision-making. By identifying key risk factors, we provide a framework for tailored patient management that could ultimately improve outcomes in this vulnerable population. The strong performance metrics of the predictive model underscore its relevance in clinical practice, and decision curve analysis demonstrates its capacity to deliver meaningful benefits over traditional treatment strategies. As healthcare continues to evolve towards data-driven approaches, our findings lay the groundwork for the integration of predictive analytics in routine care, emphasizing the importance of early identification and intervention in high-risk patients to mitigate complications associated with sepsis.

Declarations

Ethics approval and consent to participate: The data in this study is based solely on publicly available, de-identified data from MIMIC-IV. After completing Collaborative Institutional Training Initiative (CITI program: 68902439), we got permission to access the database. This study involves no direct human participation occurred and that IRB exemption was granted due to the nature of the dataset.

Consent for publication: Not applicable.

Availability of data and materials: Data is provided within the manuscript.

Contributorship statement

(I) W.H. and TY.S. wrote the main manuscript text. ZN.Y. and LN.L. prepared figures and tables. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Wheeler, D. S. Oxidative Stress in Critically Ill Children with Sepsis. *Open. Inflamm. J.* **4** (s1), 74–81 (2011).
2. Mohammad, R. A. Use of granulocyte colony-stimulating factor in patients with severe sepsis or septic shock. *Am. J. Health Syst. Pharm.* **67** (15), 1238–1245 (2010).
3. Camara, D. Minimizing risks associated with peripherally inserted central catheters in the NICU. *MCN Am. J. Matern Child. Nurs.* **26** (1), 17–21 (2001). quiz 22.
4. McAuliffe, E., O'Shea, S. & Khan, M. I. PO-02 - Retrospective audit of the Peripherally Inserted Central Catheter (PICC) associated thrombosis in patients with haematological malignancies at Cork University Hospital. *Thromb. Res.* **140** (Suppl 1), S176 (2016).
5. Liu, Y. et al. Peripherally inserted central catheter thrombosis incidence and risk factors in cancer patients: a double-center prospective investigation. *Ther. Clin. Risk Manag.* **11**, 153–160 (2015).
6. Zochios, V., Umar, I., Simpson, N. & Jones, N. Peripherally inserted central catheter (PICC)-related thrombosis in critically ill patients. *J. Vasc Access.* **15** (5), 329–337 (2014).
7. Hao, N. et al. Nomogram predicted risk of peripherally inserted central catheter related thrombosis. *Sci. Rep.* **7** (1), 6344 (2017).
8. Pan, L., Zhao, Q. & Yang, X. Risk factors for venous thrombosis associated with peripherally inserted central venous catheters. *Int. J. Clin. Exp. Med.* **7** (12), 5814–5819 (2014).
9. Soares, F. M. et al. Design, construction, and validation of obstetric risk classification systems to predict intensive care unit admission. *Int. J. Gynaecol. Obstet.* **167** (3), 1243–1254 (2024).
10. Alabbad, D. A. et al. Machine learning model for predicting the length of stay in the intensive care unit for Covid-19 patients in the eastern province of Saudi Arabia. *Inf. Med. Unlocked.* **30**, 100937 (2022).
11. Cardoso, M. M. A. et al. Application of natural language processing to predict final recommendation of Brazilian health technology assessment reports. *Int. J. Technol. Assess. Health Care.* **40** (1), e19 (2024).
12. Li, G., Liu, W., Da, X., Li, Z. & Pu, J. The natural flavonoid pinocembrin shows antithrombotic activity and suppresses septic thrombosis. *Int. Immunopharmacol.* **142** (Pt B), 113237 (2024).

13. Rhodes, A. et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Med.* **43** (3), 304–377 (2017).
14. Chen, L., Lu, Y., Wang, L., Pan, Y. & Zhou, X. Construction of a nomogram risk prediction model for PICC-related venous thrombosis and its application. *Asian J. Surg.* **47** (1), 107–111 (2024).
15. Choi, D. W. & Ly, S. Y. Diagnostic recognition of Escherichia coli septicemia in in vivo vascular blood. *Med. Chem.* **11** (6), 590–594 (2015).
16. Wu, X. et al. Behcet's disease complicated with thrombosis: a report of 93 Chinese cases. *Med. (Baltim).* **93** (28), e263 (2014).
17. Wang, D. et al. Incidence of Thrombosis at Different Sites During the Follow-Up Period in Essential Thrombocythemia: A Systematic Review and Meta-Analysis. *Clin. Appl. Thromb. Hemost.* **29**, 10760296231181117 (2023).
18. Alvarez-Larran, A. et al. Application of IPSET-thrombosis in 1366 Patients Prospectively Followed From the Spanish Registry of Essential Thrombocythemia. *Hemasphere* **7** (8), e936 (2023).
19. Tang, C. Y. et al. Prediction models for COVID-19 disease outcomes. *Emerg. Microbes Infect.* **13** (1), 2361791 (2024).
20. Calzadilla-Bertot, L. et al. Predicting liver-related events in NAFLD: A predictive model. *Hepatology* **78** (4), 1240–1251 (2023).
21. Chen, X. et al. aCCI-HBV-ACLF: A Novel Predictive Model for Hepatitis B Virus-Related Acute-On-Chronic Liver Failure. *Aliment. Pharmacol. Ther.* **61** (2), 286–298 (2025).
22. Singhal, A. et al. Hypercoagulability in end-stage liver disease: prevalence and its correlation with severity of liver disease and portal vein thrombosis. *Clin. Appl. Thromb. Hemost.* **18** (6), 594–598 (2012).
23. Al Ghumlas, A. K. & AbdelGader, A. G. The liver and the haemeostatic system. *Saudi J. Gastroenterol.* **9** (2), 59–68 (2003).
24. Peck-Radosavljevic, M. Review article: coagulation disorders in chronic liver disease. *Aliment. Pharmacol. Ther.* **26** (Suppl 1), 21–28 (2007).
25. Alvarez-Mon, M. et al. A Predictive Model and Risk Factors for Case Fatality of COVID-19. *J. Pers. Med.* **11**(1). (2021).
26. Samadani, A., Wang, T., van Zon, K. & Celi, L. A. VAP risk index: Early prediction and hospital phenotyping of ventilator-associated pneumonia using machine learning. *Artif. Intell. Med.* **146**, 102715 (2023).
27. Jiang, Y. et al. Development and validation of a predictive model for acute kidney injury in patients with ureterolithiasis. *Ren. Fail.* **46** (2), 2394634 (2024).
28. Yu, S. C. et al. Comparison of Sepsis Definitions as Automated Criteria. *Crit. Care Med.* **49** (4), e433–e443 (2021).
29. Yu, D. et al. Correlation of clinical sepsis definitions with microbiological characteristics in patients admitted through a sepsis alert system; a prospective cohort study. *Ann. Clin. Microbiol. Antimicrob.*

21 (1), 7 (2022).

30. Lengquist, M. et al. Sepsis mimics among presumed sepsis patients at intensive care admission: a retrospective observational study. *Infection* **52** (3), 1041–1053 (2024).
31. Li, P., Wang, C. & Pang, S. The diagnostic accuracy of mid-regional pro-adrenomedullin for sepsis: a systematic review and meta-analysis. *Minerva Anestesiol.* **87** (10), 1117–1127 (2021).
32. Jiang, K. & Cao, T. Automated HIV Case Identification from the MIMIC-IV Database. *AMIA Jt. Summits Transl Sci. Proc.* **2024**, 555–564 (2024).

Tables

Table 1 is available in the Supplementary Files section.

Figures

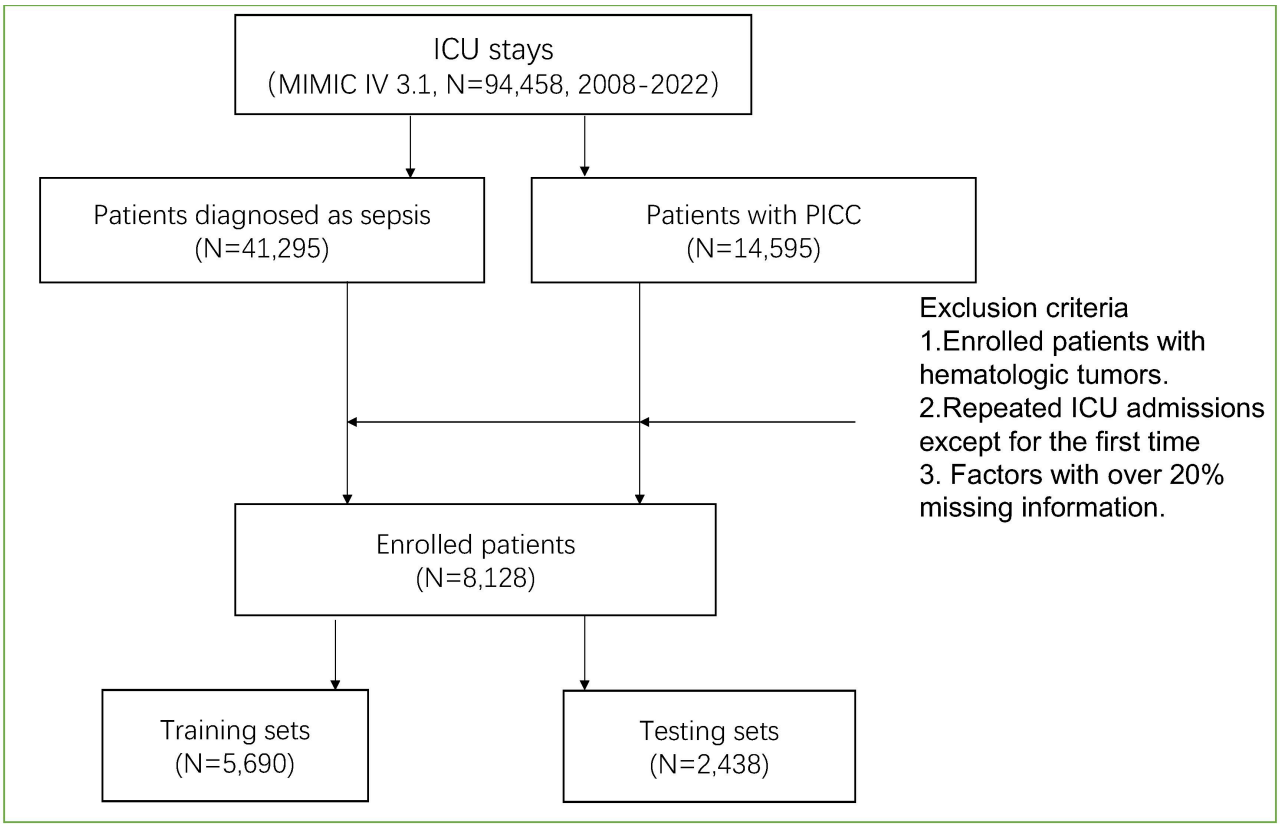

Figure 1

Flowchart illustrating the patient selection process in MIMIC IV 3.1. (MIMIC-IV, Medical Information Mart for Intensive Care).

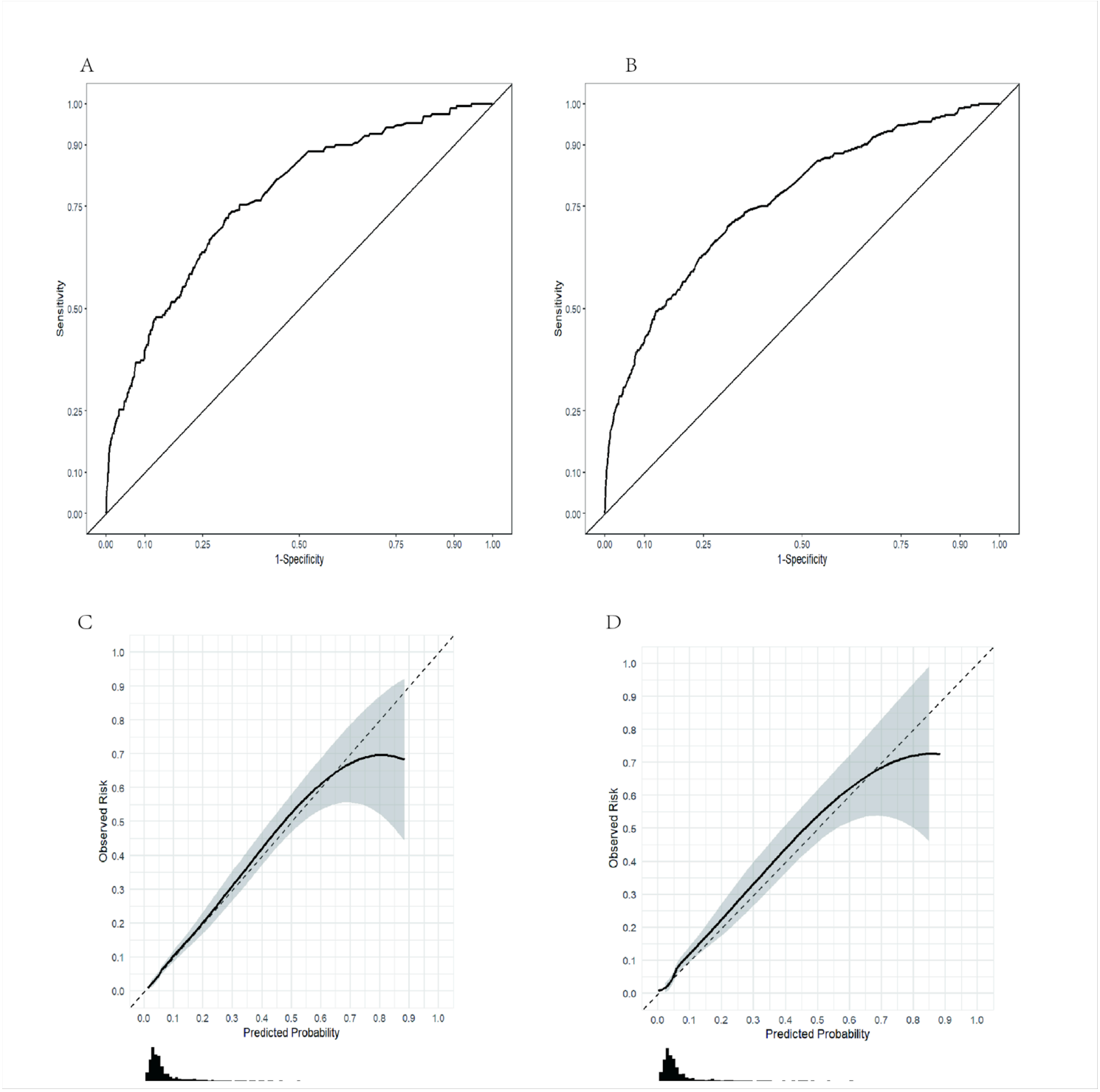

Figure 2

Performance metrics of the predictive model assessed within the training cohort (A) and validation cohort (B). Calibration plots of the predictive model for forecasting pressure injuries in both the training cohort (C) and validation cohort (D).

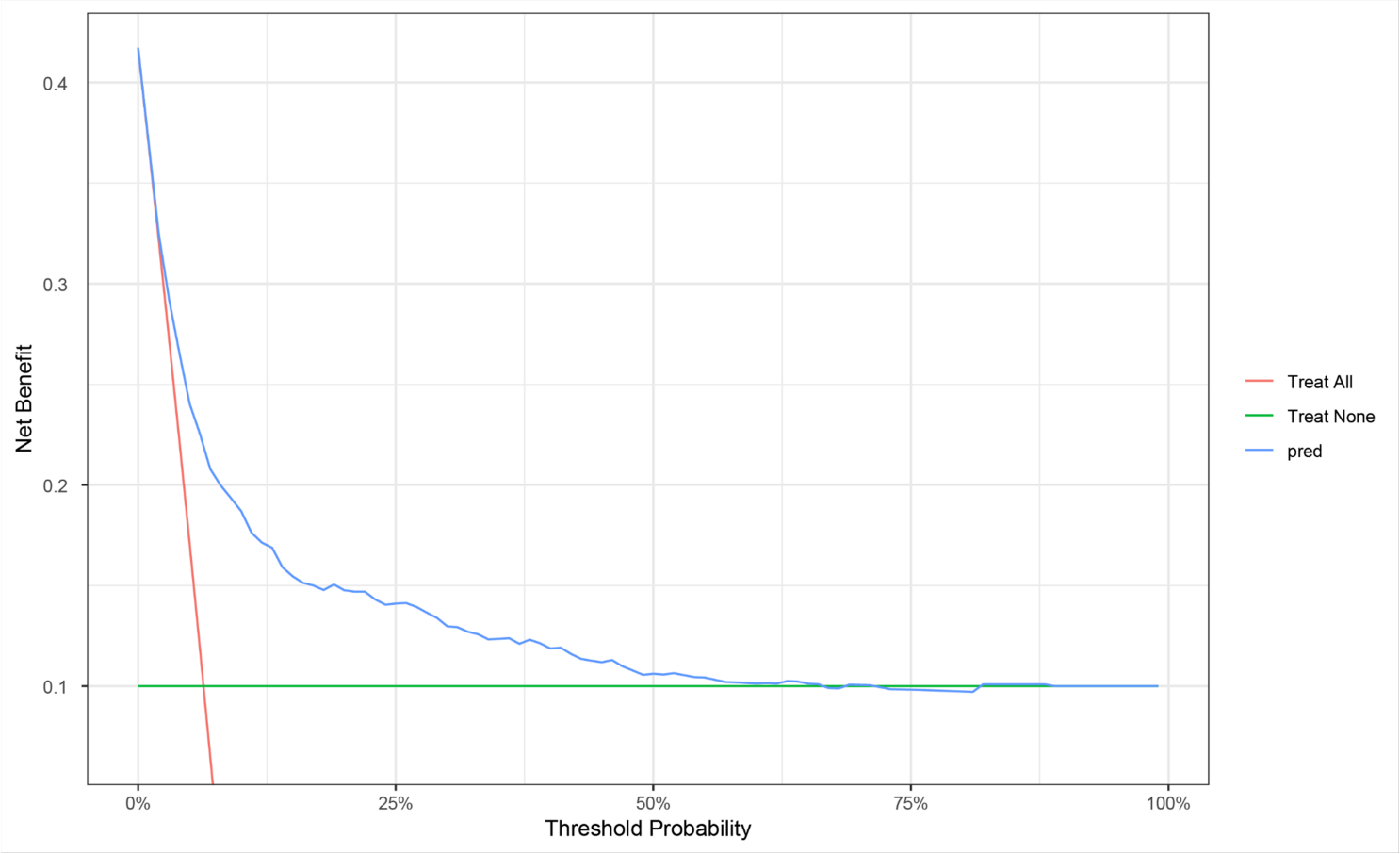

Figure 3

Decision-curve analysis for the predictive model within the validation cohort

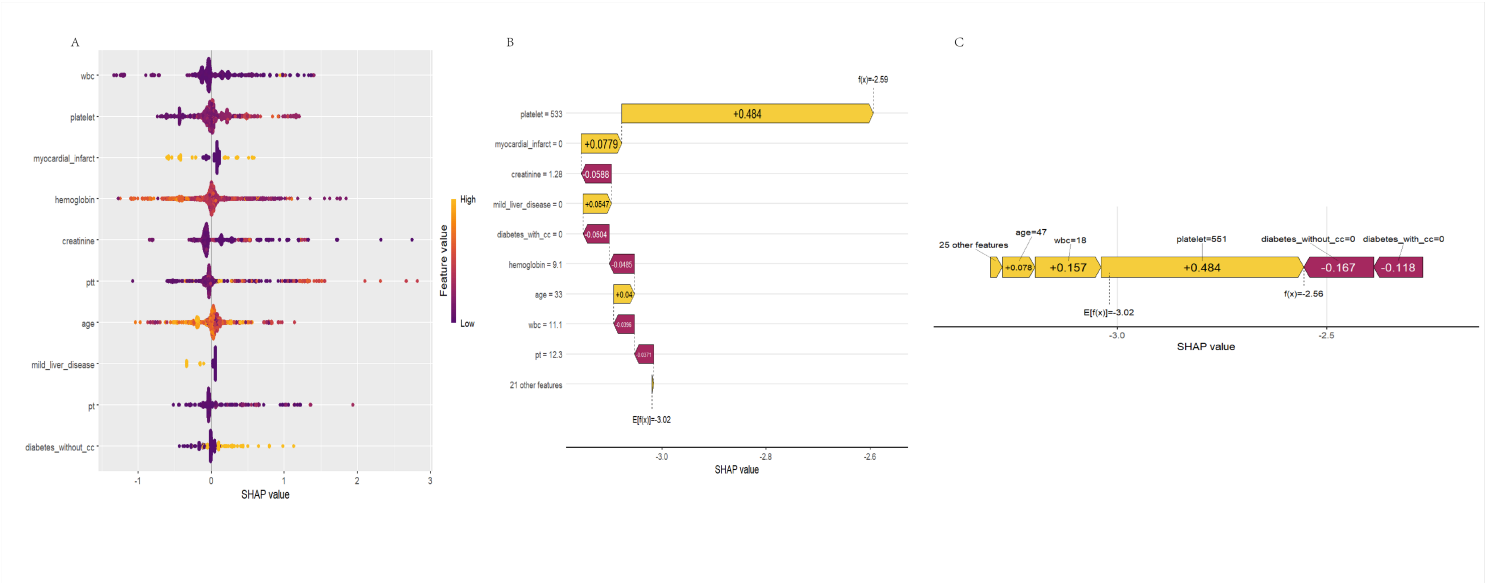

Figure 4

A. Importance chart of SHAP variables, with the selected features arranged by the average absolute SHAP value from highest to lowest. B, C. SHAP force plots for two specific cases: Color coding indicates

the contribution of each feature, where purple signifies a detrimental effect on predictions (arrow directed left, indicating a decrease in SHAP value), and yellow denotes a beneficial effect on predictions (arrow directed right, indicating an increase in SHAP value). The length of the color bar reflects the magnitude of the contribution, while E[f(x)] represents the SHAP reference value, corresponding to the mean predicted by the model. f(x) signifies the individual SHAP value.

Supplementary Files

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

- Table1.docx