scientific reports

OPEN

A predictive model for hospital death in cancer patients with acute pulmonary embolism using XGBoost machine learning and SHAP interpretation

Zhen-nan Yuan[1,4], Yu-juan Xue[2,4], Hai-jun Wang[1], Shi-ning Qu[1], Chu-lin Huang[1], Hao Wang[1], Hao Zhang[1], Min-ze Zhang[3,5]✉ & Xue-zhong Xing[1,5]✉

The prediction of in-hospital mortality in cancer patients with acute pulmonary embolism (APE) remains a significant clinical challenge. This study aimed to develop and validate a machine learning model using XGBoost to predict in-hospital mortality in this vulnerable population. A retrospective cohort study was conducted using the MIMIC-IV 2.2 database and external data from the intensive care unit of Cancer hospital, Chinese Academy of Medical Sciences, collected between May 1, 2021, and April 30, 2023. A total of 448 cancer patients with APE were included from the MIMIC-IV 2.2 database, divided into a training set (70%, n = 314) and an internal validation set (30%, n = 134). An external validation cohort consisted of 56 patients. An XGBoost model was trained and the SHAP (SHapley Additive Explanations) method was used to identify the top 10 predictors of in-hospital mortality. These predictors included Glasgow Coma Scale (GCS) score, albumin, platelet count, age, serum creatinine, hemoglobin, presence of metastasis, lactate, creatine kinase (CK), and types of cancer. The XGBoost model achieved an area under the ROC curve (AUC) of 0.806 (95% CI: 0.717–0.896) in the internal validation set and 0.724 (95% CI: 0.686–0.901) in the external validation set. Calibration curves indicated good model fit, and decision curve analysis (DCA) demonstrated a high clinical benefit across both the internal and external validation cohorts. The XGBoost model, leveraging SHAP for interpretation, effectively predicts in-hospital mortality in cancer patients with APE. This model provides valuable insights for clinical decision-making and has the potential to improve patient outcomes through early intervention and personalized treatment strategies. Further validation in diverse clinical settings is warranted to confirm its generalizability.

Keyword Machine learning, Acute pulmonary embolism, In-hospital mortality, Cancer

Acute pulmonary embolism (APE) is a potentially fatal complication that can occur in patients with malignancies, significantly affecting prognosis and clinical outcomes[1,2]. Cancer patients have a nine-fold higher venous thromboembolism risk than the general population[3] due to hypercoagulability associated with tumor presence, chemotherapy, and other oncologic treatments[4]. Despite advancements in oncology and critical care, predicting in-hospital mortality among cancer patients with APE remains challenging. Traditional risk assessment models, while valuable, often lack the precision required for individualized patient management, primarily due to the multifactorial and dynamic nature of the conditions involved[5]. The Wells Score is a common clinical tool used for assessing the risk of APE[6,7]. Despite its widespread use, the Wells Score has limitations in precision due to its reliance on linear relationships, which may not fully capture the complexity of patient conditions[8].

[1]Department of Intensive Care Unit, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100021, China. [2]Department of Pediatrics, Peking University People's Hospital, Peking University, Beijing, China. [3]People's Hospital of Lishui District, Nanjing, Lishui 211200, Jiangsu, China. [4]Zhen-nan Yuan and Yu-juan Xue contributed equally to this work and share first authorship. [5]Min-ze Zhang and Xue-zhong Xing contributed equally to this work and share corresponding authorship. ✉email: 18761825997@163.com; xxzncc@163.com

Machine learning (ML) techniques offer a promising alternative by leveraging vast datasets and uncovering complex patterns that may not be immediately apparent to human observers[9,10]. They can also effectively manage non-linear relationships and interactions among variables, thus enhancing the predictive capacity beyond that of traditional models, such as XGBoost[11,12]. However, ML models are often criticized for their difficult-to-explain "black box"trait[13]. SHapley Additive exPlanations (SHAP) is an advanced explainable ML framework designed to provide an in-depth explanation of the predictions of any machine learning model[14]. It enhances the transparency of ML models, thereby promoting the use and acceptance of artificial intelligence technology in clinic practice[15]. However, its application specifically for predicting in-hospital mortality in cancer patients with APE remains underexplored. This study aims to fill this gap by developing and validating a machine learning-based prediction model for in-hospital mortality among oncology patients diagnosed with APE. By integrating comprehensive clinical data, the proposed model seeks to enhance prognostic accuracy, guide clinical decision-making, and ultimately improve patients' outcomes.

Design and methods

Database

The model was developed and validated internally based on Medical Information Mart for Intensive Care IV (MIMIC IV, version 2.2) and validated externally based on department of Intensive Care Unit, Cancer Hospital, Chinese Academy of Medical Sciences. The MIMIC IV is a publicly available database of patients admitted to the Beth Israel Deaconess Medical Center, Boston, MA, between 2008 and 2019, which included over 76,000 ICU admissions. It contains detailed information for each admission, including laboratory examination, vital signs, administered medications and status on discharge. After completing Collaborative Institutional Training Initiative (CITI program), we got permission to access the database (Record ID: 36,067,767). The study was based on the suggestions of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.

Data collection and Participants

All eligible patients diagnosed with APE based on ICD codes (ICD-9: 41,511, 41,512, 41,519, 67,380–67,384; ICD-10 code: I26, I260, I2601, I2609, I269, I2690, I2693, I2699) from the MIMIC-IV database were included as participants in this study. The inclusion criteria were as follows: 1. Enrolled patients were diagnosed as malignancy and pulmonary embolism according to International Classification of Diseases, Ninth and Tenth Revision code (ICD-9 and ICD-10) 2. Patients' age were more than 18 years when admitted by ICU. The exclusion criteria included 1. enrolled patients with hematologic tumors. 2.repeated ICU admissions except for the first time and factors with over 50% missing information. Clinical data of cancer patients with pulmonary embolism consisted of the followings (Table 1): (1) demographic characteristics such as age and gender; (2) cancer stage and type; (3) chronic underlying diseases; (4) organ function, such as mean arterial pressure (MAP) and Glasgow score; (5) laboratory examination. The vital signs and laboratory results within the first 24 h after ICU admission were collected. The observed endpoints of this study were the hospital death or the safe discharge. R software was used to processed the raw data extracted by Navicat for Structure Query Language Server.

XGBoost machine learning

The model is constructed using the "XGBoost" package, utilizing the train () function to optimize parameters from the caret package and output the best parameter configuration. The XGBoost model is established by setting the learning rate eta to 0.1, the maximum depth (max_depth) to 2, and the number of iteration rounds (i.e., the number of boosting rounds) to 100. Cancer patients with APE from MIMIC IV database were designed for the training and internal validation of the model. 70% of the samples were allocated to the training set using the "caret"package, while the remaining 30% of the cases were assigned to the internal validation set. Cancer patients with APE, from cancer hospital, Chinese academy of medical sciences, were used for external validation.

SHAP

"Shapviz" is an R package for interpreting ML model predictions that provides visual explanations based on SHAP values. The SHAP value explains the extent to which each feature contributes to the model's predictions, either positively or negatively. The feature importance plot is used to display the features that have the most significant impact on the model prediction, and the feature importance is ranked based on the average absolute value of SHAP values.

Statistics analysis

All statistical analysis and data visualization were conducted using R 4.3.0. Continuous data is typically described using the mean ± standard deviation or median (interquartile range (IQR)), while categorical data is presented as frequency (percentage). Chi-square test was used for categorical variables, and independent samples t-test or nonparametric test was used for continuous variables to compare the differences between groups. The "tidyverse", "pROC", "CBCgrps", "rms", and "rmda"packages were used for data collation and visualization in this study. The ROC curve, calibration curve, and clinical decision curve analysis (DCA) were used to evaluate the model's performance. The predicted outcome probability was converted into a binary outcome using a threshold of 0.5. Calculate and output the accuracy, sensitivity, specificity, and other metrics for both the training and test set.

	Training Cohort	Internal testing Cohort	External testing Cohort	
Factors	N = 314	N = 134	N = 56	P
Age (mean, y)	66.2 ± 13.5	68.2 ± 13.1	65.3 ± 13.3	0.228
Female (N, %)	146(46.5)	55(41.0)	21(37.5)	0.328
COPD (N, %)	83(26.4)	30(22.4)	11(19.6)	0.435
Diabetes (N, %)	46(14.7)	29(21.6)	11(19.6)	0.170
Myocardial infarct	27(8.6)	9(6.7)	2(3.6)	0.387
Congestive heart failure	3.5(11.2)	16(11.9%)	6(10.7)	0.960
Peripheral vascular disease	17(5.4)	5(3.73)	1(1.79)	0.422
Cerebrovascular disease	29(9.2)	13(9.7)	6(10.7)	0.938
Dementia	5(1.59)	0(0)	0(0)	0.217
Peptic ulcer disease	4(1.27)	1(0.75)	0(0)	0.638
The presence of metastasis (N, %)	138(44.0)	60(44.8)	23(41.1)	0.894
Respiration (PaO2/FIO2)				0.029
≥ 400	251(79.9)	102(76.1)	41(73.2)	
300–400	5(1.6)	1 (0.8)	1(1.8)	
200–300	32(10.2)	18(13.4)	8(14.3)	
100–200&MV	20(6.4)	3(2.2)	1(1.8)	
< 100&MV	6(1.9)	10(7.5)	5(8.9)	
TIBL(umol/L)				0.098
< 20	232(73.9)	81(60.5)	34(60.7)	
20–32	46(14.7)	27(20.2)	14(25.0)	
33–101	25(8.0)	20(14.9)	7(12.5)	
102–204	6(1.9)	4(3.0)	0(0)	
> 204	5(1.6)	2(1.5)	1(1.8)	
MAP(mmHg)				0.150
≥ 70	65(20.7)	33(24.6)	21(37.5)	
< 70	213(67.8)	81(60.5)	30(53.6)	
DA < 5.0 (ug/kg.min)	1(0.3)	0(0)	0(0)	
DA:5–15 or NE ≤ 0.1(0.1ug/kg.min)	13(4.1)	14(3.0)	2(3.6)	
DA > 15 or NE > 0.1 (ug/kg.min)	22(7.0)	16(11.9)	3(5.4)	
GCS				0.011
15	152(48.4)	56(41.8)	24(42.9)	
13–14	110(35.0)	45(33.6)	25(26.8)	
10–12	16(5.1)	14(10.5)	6(10.7)	
6–9	16(5.1)	13(9.7)	10(17.9)	
< 6	20(6.4)	6(4.5)	1(1.8)	
Creatine(umol/L)				0.964
< 110	198(63.1)	80(59.7)	35(62.5)	
110–170	60(19.1)	35(18.7)	12(21.4)	
171–299	15(4.8)	9(6.7)	3(5.4)	
300–440	28(8.9)	12(9.0)	3(5.4)	
> 440	13(4.1)	8(6.0)	3(5.4)	
Lactate(mmol/L)	2.3 ± 1.4	2.5 ± 1.8	2.6 ± 1.8	0.617
Hemoglobin(g/L)	9.8 ± 2.1	9.8 ± 2.1	10.6 ± 2.2	0.024
Platelets count(10^6/L)	233.1 ± 136.8	195.5 ± 119.8	193.0 ± 110.7	0.004
Albumin(g/L)	2.9 ± 0.5	2.9 ± 0.5	3.1 ± 0.6	0.05
CK (iu/dl)	8.9 ± 3.9	7.2 ± 5.8	6.5 ± 2.9	0.857
Type of cancer (N, %)				0.010
Lung	117(37.3)	42(31.3)	22(39.3)	
Continued				

	Training Cohort	Internal testing Cohort	External testing Cohort	
Factors	N = 314	N = 134	N = 56	P
Gynecological	57(21.3)	17(12.7)	5(8.9)	
Digestive system	23(7.3)	23(17.2)	8(14.3)	
Urinary system	68(21.7)	29(21.6)	10(17.9)	
Others	39(12.4)	23(17.2)	11(19.6)	
Hospital death (N, %)	55(17.5)	26(19.4)	9(16.1)	0.833

Table 1. Characteristics of study participants from training set, internal validation set and external validation set. COPD, chronic obstructive pulmonary disease; GCS, Glasgow Coma Scale; TIBL, Total Bilirubin; CK, creatine kinase; MAP, mean arterial pressure. DA, Dopamine.

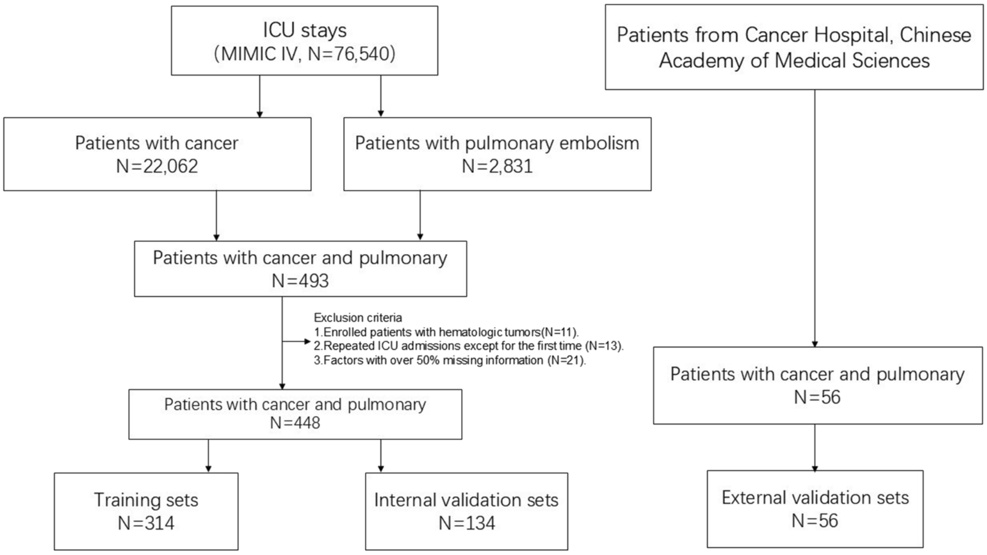
ICU stays
(MIMIC IV, N=76,540)
Patients from Cancer Hospital, Chinese Academy of Medical Sciences
Patients with cancer
N=22,062
Patients with pulmonary embolism
N=2,831
Patients with cancer and pulmonary
N=493
Exclusion criteria
1.Enrolled patients with hematologic tumors(N=11).
2.Repeated ICU admissions except for the first time (N=13).
3.Factors with over 50% missing information (N=21).
Patients with cancer and pulmonary
N=448
Patients with cancer and pulmonary
N=56
Training sets
N=314
Internal validation sets
N=134
External validation sets
N=56

Fig. 1. Flow diagram of the patient selection in MIMIC IV and ICU of cancer hospital, Chinese academy of medical sciences. (MIMIC-IV, Medical Information Mart for Intensive Care).

Results

Basic characteristics of patients

There are 76,540 ICU stays in MIMIC IV database. A flowchart of the patient selection process and study cohort is shown in Fig. 1. A total of 448 cancer patients with APE were included from the MIMIC-IV database, of which 314 patients were assigned to the training set and 134 patients to the internal validation set. Additionally, 56 patients from the ICU of cancer hospital of Chinese Academy of Medical Sciences Cancer Hospital served as the external validation cohort. The demographic and clinical characteristics of the patients were generally comparable between the training, internal validation, and external validation sets, ensuring the robustness of the model's performance across different datasets. In the training cohort, there were 55 (17.5%) cases of in-hospital mortality. In the internal validation cohort, 26 (19.4%) cases of in-hospital mortality were recorded. In the external validation cohort, there were 9 (16.1%) cases of in-hospital mortality. The baseline characteristics were described in Table 1. The percentage of missing data for each variable was described in the supplementary materials.

Development and validation of the predicted model

The XGBoost model was trained using the training set and evaluated using both internal and external validation cohorts. In the internal validation set, the model achieved an area under the receiver operating characteristic (ROC) curve (AUC) of 0.806 (95% CI: 0.717–0.896) (Fig. 2A). In the external validation set, the model's performance remained robust, with an AUC of 0.724 (95% CI: 0.686–0.901) (Fig. 2B).

Model calibration and clinical utility

Calibration curves for the internal (Fig. 2C) and external validation sets (Fig. 2D) demonstrated good agreement between predicted and observed mortality rates, indicating that the model is well-calibrated. The decision curve analysis further confirmed the clinical utility of the model, showing that the XGBoost model provides a net benefit across a wide range of threshold probabilities both in internal (Fig. 2E) and external sets (Fig. 2F).

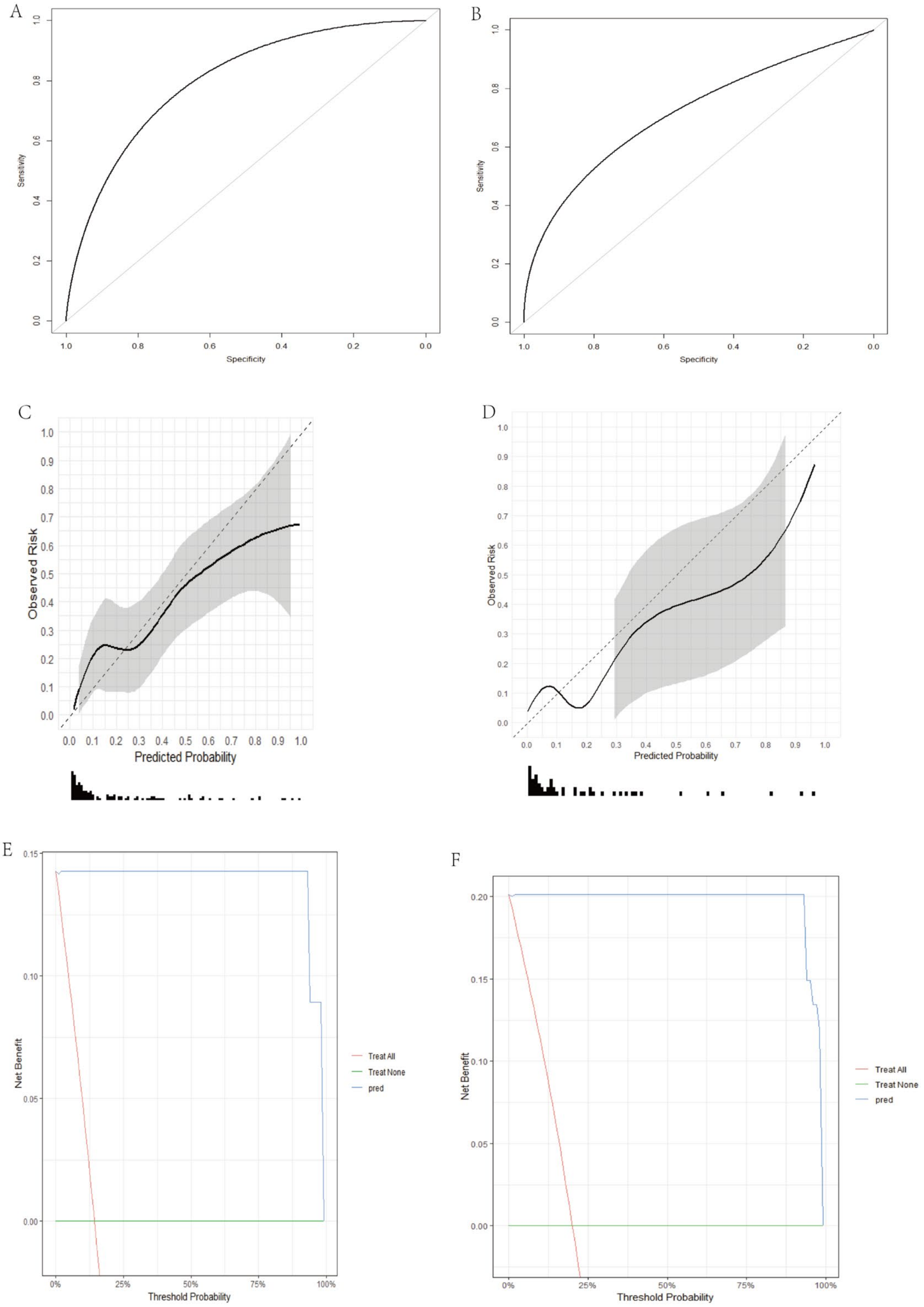

Fig. 2. The performance of the predicted model in the internal validation set (**A**) and external validation set (**B**). Calibration curves of the predicted model for predicting hospital morality both in the internal validation set (**C**) and external validation set (**D**). Decision-curve analysis of the predicted model in internal validation set (**E**) and external validation set (**F**).

Interpretation of the XGBoost ML model

The SHAP method was employed to interpret the XGBoost model and identify the most significant predictors of in-hospital mortality. The SHAP summary plot of the model shows the effect of features on the prediction model (Fig. 3A). The included features are sorted by average SHAP absolute value from highest to lowest. From high to

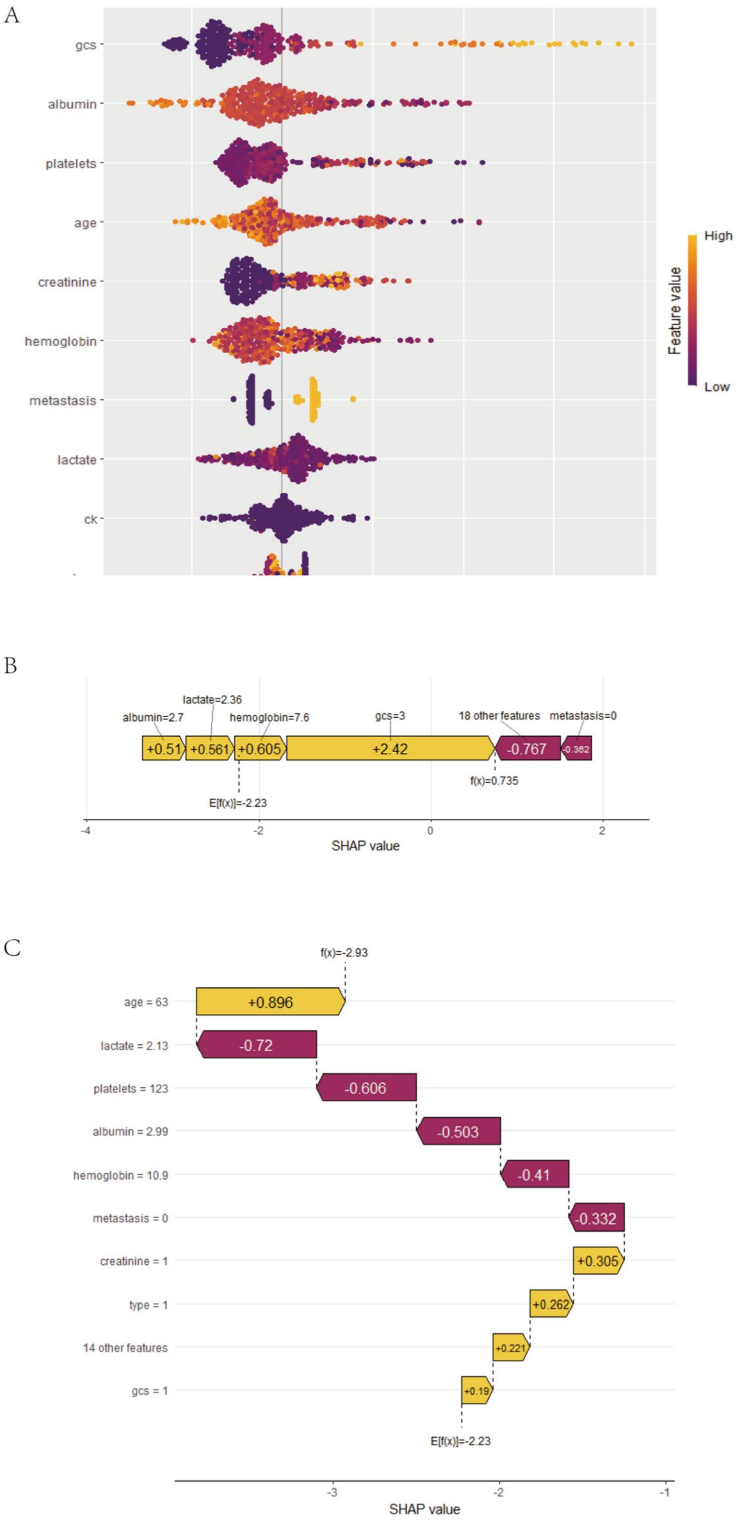

Fig. 3. (**A**) Importance chart of SHAP variables, with the included features sorted by the average absolute value of SHAP from highest to lowest. (**B**, **C**) SHAP force plot for two cases: Color indicates the contribution of each feature, purple indicates that the feature has a negative effect on the prediction (arrow to the left, SHAP value decreases), and yellow indicates that the feature has a positive effect on the prediction (arrow to the right, SHAP value increases). The length of the color bar indicates the strength of the contribution, and E[f(x)] indicates the SHAP reference value, which is the mean predicted by the model. f (x) represents the SHAP value of the individual.

low are: ‘GCS’, ‘Albumin’, ‘platelets’, ‘Age’, ’serum creatine’, ‘hemoglobin’, ‘the presence of metastasis’, ‘lactate’ , ‘CK’, ‘types of cancer’. According to the prediction model, the higher the SHAP value of the feature, the more likely it is to develop in-hospital mortality. The top 10 variables increasing in-hospital mortality outcomes were: low GCS score, low albumin levels, increased platelet count, increasing age, high serum creatinine, low hemoglobin levels, the presence of metastasis, high lactate levels, high creatine kinase (CK), and lung cancer. These predictors highlight critical physiological and pathological factors that are associated with an increased risk of mortality in cancer patients with APE.

Individual risk prediction using XGBoost and feature interaction

To illustrate the manner in which the XGBoost model assesses the contributions of specific patient characteristics, we employ SHAP force plots to analyze individual predictions for two distinct patients (Figs. 3B and C). The color coding represents the contribution of each characteristic, where red signifies a detrimental effect on the prediction (indicated by an arrow directed to the left, correlating with a decrease in SHAP value) and yellow denotes a beneficial effect (with an arrow oriented to the right, reflecting an increase in SHAP value). The length of the color gradient corresponds to the magnitude of the contribution, while E[f(x)] represents the SHAP reference value, which is the average prediction made by the model. For the cohort identified as “true positives,” the XGBoost model successfully predicted in-hospital mortality using SHAP values.

In our investigation, we employed SHAP interaction values to analyze the interrelations among features within our predictive framework. Figure 4 presents the interaction summary plot for the ten most significant interacting features, organized in descending order based on their interaction importance: GCS, platelets, age, albumin, creatinine, lactate, hemoglobin, metastasis, CK, and type. The interaction summary plot indicates that specific pairs of features exert a considerable combined influence on the predictions generated by the model. For example, the interactions between GCS and platelets, as well as GCS and age, exhibit marked SHAP interaction values, implying a strong interplay between these feature pairs in shaping the model’s outcomes. Furthermore, interactions involving albumin and creatinine also reveal a noteworthy effect, underscoring their collective significance in the predictive modeling process.

Discussion

In this study, we developed and validated an XGBoost machine learning model to predict in-hospital mortality among cancer patients with APE. The performance of the model was rigorously evaluated using both internal validation and external validation at our center. The model demonstrated robust predictive accuracy, underscoring the potential of advanced ML techniques in clinical decision-making for this high-risk population. Importantly, the use of SHAP values allowed us to interpret the model’s predictions, identifying the ten most influential factors associated with in-hospital mortality in this cohort: ‘GCS’, ‘Albumin’, ‘platelets count’, ‘Age’, ’serum creatine’, ‘hemoglobin’, ‘the presence of metastasis’, ‘lactate’ , ‘CK’, ‘types of cancer’. According to relevant studies, the Pulmonary Embolism Severity Index (PESI) score is the most commonly used method for predicting short-term mortality in patients with APE. However, the PESI score involves numerous variables with varying weights and is complex to calculate, making it less widely accepted among emergency medical personnel[16]. Hence, the simplified Pulmonary Embolism Severity Index (sPESI) has been developed to predict the risk of death in patients with APE[17]. Furthermore, the strengths of both the PESI and sPESI primarily lie in identifying patients with low- to moderate-risk acute PE, but their ability to identify patients with moderate- to high-risk APE is limited[16].

The identification of the top ten predictors of in-hospital mortality has significant clinical implications. The GCS score emerged as the most influential factor, suggesting that neurologic status is a critical determinant of outcome in this population. This finding aligns with existing literature, where altered mental status is frequently associated with worse outcomes in critically ill patients[18]. Our model emphasizes the need for close monitoring and potentially more aggressive intervention in cancer patients with APE who exhibit signs of neurologic compromise.

Low serum albumin levels, reflective of poor nutritional status and chronic illness[19], were also strongly associated with poor prognosis of patients with APE[20]. Albumin performs several crucial physiological functions, including exerting antithrombotic effects, inhibiting platelet aggregation, and promoting anticoagulation[21,22]. Hypoalbuminemia has been consistently linked to worse outcomes in cancer patients, as it may indicate both the severity of the underlying malignancy and the presence of systemic inflammation. This finding suggests that nutritional support and interventions aimed at improving the overall health status of patients could be important in reducing mortality.

Increasing age was the second most important predictor, which is consistent with the well-established relationship between advanced age and poorer outcomes in both cancer and thromboembolic diseases. This finding highlights the importance of age-stratified risk assessments in managing APE in cancer patients. In previous study, increasing age had limited impact on the prognosis in patients with active breast-, gastrointestinal- and lung cancer[23,24].

Hemoglobin and platelet levels, which reflect the patient’s hematologic status, were also important predictors. Anemia and thrombocytopenia are common in cancer patients and can result from the malignancy itself, chemotherapy, or other associated conditions. These factors are critical in assessing the overall health and resilience of the patient, and their inclusion in the top predictors highlights the need for comprehensive hematologic evaluation in this population. In previous study, Mean hemoglobin concentration (MCHC) is a standard indicator of anemia, Lower MCHC is an independent risk factor for increased 30-day mortality in patients with APE[25]. Studies have shown that increased platelet count is related to the risk of venous thrombosis and pulmonary thromboembolism in many patient groups[26,27]. Increasing platelet counts cause more inflammation, and this inflammation causes both an increase in thrombosis and embolism and worsening

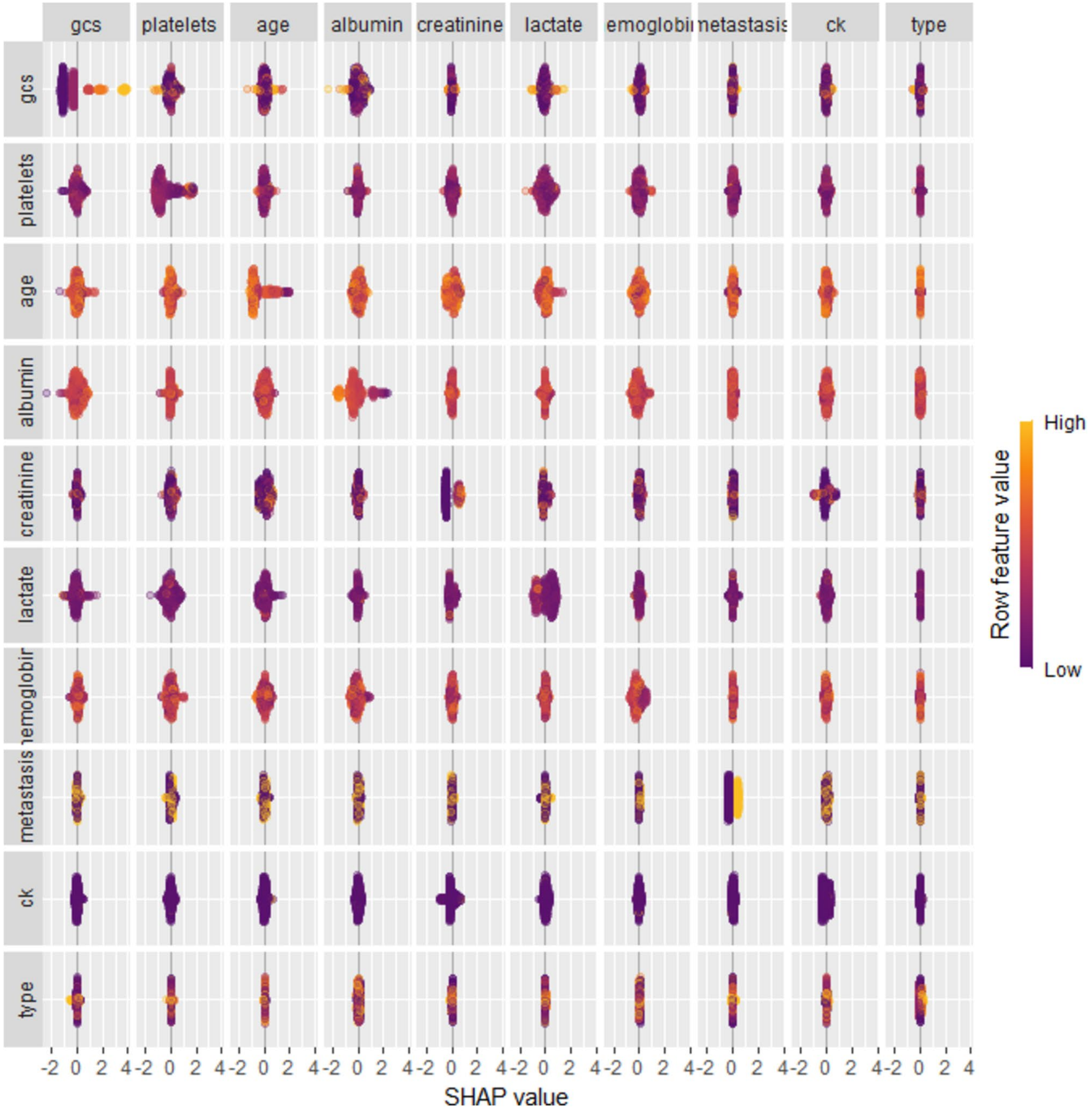

Fig. 4. **I**nteraction summary plot generated using SHAP values. This plot displays the top 10 most interacting features of the model. On the x-axis and y-axis, the features are listed according to their interaction importance, with the feature names ordered as follows: GCS, platelets, age, albumin, creatinine, lactate, hemoglobin, metastasis, CK, and type. Each point on the plot represents the SHAP interaction value for a specific feature interaction, highlighting how pairs of features together impact the model's predictions. (GCS, Glasgow Coma Scale; CK, creatine kinase.)

clinical outcomes by progressing[28]. Platelet-to-hemoglobin ratio (PHR) values predicted massive APE and were an independent predictor of mortality in APE[29].

Serum lactate, an indicator of tissue hypoxia and metabolic stress, was another significant predictor. Elevated lactate levels are a known marker of critical illness and are associated with increased mortality in various patient populations, including those with sepsis and APE[30]. The inclusion of lactate in the top predictors reinforces the need for its routine measurement and consideration in the management of cancer patients with APE, as it may help identify those at higher risk of adverse outcomes.

Renal function, as indicated by serum creatinine levels, was another critical predictor. Renal impairment is common in cancer patients, either as a consequence of the malignancy itself, its treatment, or associated comorbidities. Impaired renal function is a well-known risk factor for adverse outcomes in various medical conditions, including APE. Xing[31] conducted a study to explore the influence of impaired renal function on prognosis in APE patients and found that renal injury may be correlated with worse endpoints in APE patients. Our findings support the need for careful renal function monitoring and potentially adjusting therapeutic strategies in patients with renal impairment.

Creatine kinase, typically elevated in conditions involving muscle damage, was also among the top predictors. This might reflect the broader metabolic derangements in critically ill cancer patients or could be indicative of specific complications such as rhabdomyolysis or tumor lysis syndrome. Further research is warranted to explore the precise role of creatine kinase in this context and its potential as a prognostic marker. CK-MB was the strongest predictor of death from APE but its prevalence was low, thus limiting its value as a single prognostic indicator. The combination of high CK-MB, high cTnI, and RV dilatation tended to indicate the highest mortality[32].

The type of cancer emerged as a significant predictor, which is not surprising given the heterogeneous nature of malignancies and their varied impact on thrombotic risk and overall prognosis. Certain cancers, such as pancreatic and lung cancer, are known to have a higher thrombotic burden and poorer outcomes. The model's ability to account for cancer type underscores the importance of personalized approaches in managing APE in these patients. Cancer stage, as expected, was a significant determinant of in-hospital mortality. Advanced-stage cancer is associated with a higher thrombotic risk, greater physiological stress, and a diminished capacity to recover from acute illnesses like APE. This finding underscores the importance of considering the overall burden of cancer when managing these patients. The poor prognosis of venous thromboembolism in cancer patients is incompletely understood. The highest risk of venous thromboembolism is associated with aggressive types of cancer such as those originating from the pancreas, ovary, stomach, and lung[4,33]. An activated coagulation system can help tumors metastasize by evading the immune system[34]. Venous thromboembolism is also more frequent in patients with advanced disease than in those with localized disease[4]. The 1-year prognosis of APE was highly dependent on both cancer type and status in patients with breast-, gastrointestinal- or lung cancer[23].

Strengths and Limitations.

The primary strength of this study lies in the application of the XGBoost algorithm, a powerful and flexible machine learning tool, to predict in-hospital mortality in a specific and high-risk patient population. The use of SHAP values for model interpretation is another notable strength, as it allows for a transparent understanding of how individual predictors contribute to the risk of mortality, facilitating the translation of model insights into clinical practice.

However, there are limitations to our study. First, while the model demonstrated strong performance in both internal and external validation, its applicability to other settings and populations remains to be confirmed. External validation in other centers with different patient demographics and clinical practices is necessary to establish the generalizability of our findings. Second, although SHAP values provide insight into the contribution of individual variables, they do not capture potential interactions between variables. Future work could explore the incorporation of interaction effects to further refine the model's predictions. Additionally, the retrospective nature of the study introduces potential biases, such as selection bias and information bias. Specifically, both our internal and external validation datasets consist exclusively of ICU admissions, which may result in a cohort that is more representative of symptomatic pulmonary embolism (PE) patients, thereby introducing a potential selection bias. We acknowledge this limitation and will elaborate on it in the limitations section of the manuscript. Prospective validation studies are needed to confirm the model's predictive accuracy and clinical utility in real-time settings. Finally, while our model identifies key predictors of mortality, it does not directly inform interventions. Future research should focus on developing and testing intervention strategies tailored to the high-risk factors identified by our model.

Conclusion

Our study demonstrates that the XGBoost machine learning model, combined with SHAP-based interpretation, can effectively predict in-hospital mortality in cancer patients with APE. The identification of the ten most important predictors provides valuable insights into the factors driving mortality in this vulnerable population. These findings highlight the potential for ML models not only to enhance risk stratification but also to inform clinical decision-making and guide targeted interventions. Continued research and validation across diverse settings are essential to fully realize the potential of these models in improving outcomes for cancer patients with APE.

Data availability

The datasets used and/or analysed during the current study are available from corresponding author upon reasonable request.

Received: 29 October 2024; Accepted: 12 May 2025
Published online: 25 May 2025

References

1. Lee, A. Y. et al. Low-molecular-weight heparin versus a coumarin for the prevention of recurrent venous thromboembolism in patients with cancer. *N Engl. J. Med.* **349**(2), 146–153 (2003).
2. Sorensen, H. T., Mellemkjaer, L., Olsen, J. H. & Baron, J. A. Prognosis of cancers associated with venous thromboembolism. *N Engl. J. Med.* **343**(25), 1846–1850 (2000).
3. Lubetsky, A. Pulmonary embolism in cancer patients: A review. *ISR Med. Assoc. J.* **24**(3), 179–182 (2022).
4. Mulder, F. I. et al. Venous thromboembolism in cancer patients: A population-based cohort study. *Blood* **137**(14), 1959–1969 (2021).

5. Surov, A., Thormann, M., Bar, C., Wienke, A. & Borggrefe, J. Validation of clinical-radiological scores for prognosis of mortality in acute pulmonary embolism. *Respir Res* **24**(1), 195 (2023).
6. Posadas-Martinez, M. L. et al. Performance of the Wells score in patients with suspected pulmonary embolism during hospitalization: a delayed-type cross sectional study in a community hospital. *Thromb. Res.* **133**(2), 177–181 (2014).
7. Young, M. D. et al. Predicting pulmonary embolus in orthopedic trauma patients using the Wells score. *Orthopedics* **36**(5), e642-647 (2013).
8. Rosa-Jimenez, F. et al. Is time to search the Wells Score 4.0?. *Rev. Clin. Esp. (Barc)* **215**(5), 258–264 (2015).
9. Handelman, G. S. et al. eDoctor: Machine learning and the future of medicine. *J. Intern. Med.* **284**(6), 603–619 (2018).
10. Huang, Y., Li, J., Li, M. & Aparasu, R. R. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Med. Res. Methodol.* **23**(1), 268 (2023).
11. Xiong, S. et al. Compressive strength prediction of cemented backfill containing phosphate tailings using extreme gradient boosting optimized by whale optimization algorithm. *Materials (Basel)* **16**(1), 308 (2022).
12. Lv, C. X., An, S. Y., Qiao, B. J. & Wu, W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infect Dis* **21**(1), 839 (2021).
13. Zhang, Y. et al. Opening the black box: interpretable machine learning for predictor finding of metabolic syndrome. *BMC Endocr Disord* **22**(1), 214 (2022).
14. Ali, S. et al. The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Comput Biol Med* **166**, 107555 (2023).
15. Giacobbe, D. R., Zhang, Y. & de la Fuente, J. Explainable artificial intelligence and machine learning: novel approaches to face infectious diseases challenges. *Ann. Med.* **55**(2), 2286336 (2023).
16. Elias, A., Mallett, S., Daoud-Elias, M., Poggi, J. N. & Clarke, M. Prognostic models in acute pulmonary embolism: A systematic review and meta-analysis. *BMJ Open* **6**(4), e010324 (2016).
17. Venetz, C., Jimenez, D., Mean, M. & Aujesky, D. A comparison of the original and simplified pulmonary embolism severity index. *Thromb Haemost* **106**(3), 423–428 (2011).
18. Chen, H. et al. Glasgow coma scale as an indicator of patient prognosis: A retrospective study of 257 patients with heatstroke from 3 medical centers in Guangdong China. *Med. Sci. Monit.* **29**, e939118 (2023).
19. Quinlan, G. J., Martin, G. S. & Evans, T. W. Albumin: biochemical properties and therapeutic potential. *Hepatology* **41**(6), 1211–1219 (2005).
20. Hu, J. & Zhou, Y. The association between lactate dehydrogenase to serum albumin ratio and in-hospital mortality in patients with pulmonary embolism: A retrospective analysis of the MIMIC-IV database. *Front. Cardiovasc. Med.* **11**, 1398614 (2024).
21. Fanali, G. et al. Human serum albumin: From bench to bedside. *Mol. Aspects Med.* **33**(3), 209–290 (2012).
22. Folsom, A. R., Lutsey, P. L., Heckbert, S. R. & Cushman, M. Serum albumin and risk of venous thromboembolism. *Thromb. Haemost.* **104**(1), 100–104 (2010).
23. Nouhravesh, N. et al. Impact of breast-, gastrointestinal-, and lung cancer on prognosis in patients with first-time pulmonary embolism: A Danish nationwide cohort study. *Int. J. Cardiol.* **406**, 132001 (2024).
24. Junjun, L., Pei, W., Ying, Y. & Kui, S. Prognosis and risk factors in older patients with lung cancer and pulmonary embolism: a propensity score matching analysis. *Sci. Rep.* **10**(1), 1272 (2020).
25. Kong, W. et al. Mean corpuscular hemoglobin concentration correlates with prognosis of resected hepatocellular carcinoma. *Biomark. Med.* **14**(4), 259–270 (2020).
26. Simanek, R. et al. High platelet count associated with venous thromboembolism in cancer patients: Results from the Vienna Cancer and Thrombosis Study (CATS). *J. Thromb. Haemost.* **8**(1), 114–120 (2010).
27. Huang, C. B. et al. Risk factors for pulmonary embolism in ICU Patients: A retrospective cohort study from the MIMIC-III Database. *Clin. Appl. Thromb. Hemost.* **28**, 10760296211073924 (2022).
28. Chung, T. et al. Platelet activation in acute pulmonary embolism. *J. Thromb. Haemost.* **5**(5), 918–924 (2007).
29. Ozbeyaz, N. B. et al. Novel marker for predicting the severity and prognosis of acute pulmonary embolism: platelet-to-hemoglobin ratio. *Biomark Med* **16**(12), 915–924 (2022).
30. Wang, Y., Feng, Y., Yang, X. & Mao, H. Prognostic role of elevated lactate in acute pulmonary embolism: A systematic review and meta-analysis. *Phlebology* **37**(5), 338–347 (2022).
31. Xing, X. et al. Impact of renal function on the prognosis of acute pulmonary embolism patients: A systematic review and meta-analysis. *Expert Rev. Respir. Med.* **16**(1), 91–98 (2022).
32. Stein, P. D. et al. Prognosis based on creatine kinase isoenzyme MB, cardiac troponin I, and right ventricular size in stable patients with acute pulmonary embolism. *Am. J. Cardiol.* **107**(5), 774–777 (2011).
33. Sorensen, H. T., Pedersen, L., van Es, N., Buller, H. R. & Horvath-Puho, E. Impact of venous thromboembolism on the mortality in patients with cancer: A population-based cohort study. *Lancet Reg. Health Eur.* **34**, 100739 (2023).
34. Ward, M. P. et al. Platelets, immune cells and the coagulation cascade; friend or foe of the circulating tumour cell?. *Mol Cancer* **20**(1), 59 (2021).

Acknowledgements

None.

Author contributions

(I) Conception and design: ZN Yuan, XZ Xing and MZ Zhang; (II) Provision of study materials or patients: SN Qu, CL Huang, HJ Wang and H Wang; (III) Collection and assembly of data: YJ Xue and ZN Yuan; (IV) Data analysis and interpretation: ZN Yuan and H Zhang; (V) Final approval of manuscript: All authors.

Funding

None.

DeclarationsDeclarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The data in this study were from two public de-identified databases. After completing Collaborative Institutional Training Initiative (CITI program), we got permission to access the database (Record ID: 36,067,767).

Consent for publication

Not applicable.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-02072-1.

Correspondence and requests for materials should be addressed to M.-z.Z. or X.-z.X.

Reprints and permissions information is available at www.nature.com/reprints.