BMC Gastroenterology

RESEARCH

Open Access

Explainable machine learning model for predicting acute pancreatitis mortality in the intensive care unit

Meng Jiang[1*†], Xiao-peng Wu[2†], Xing-chen Lin[1] and Chang-li Li[3*]

Abstract

Background Current prediction models are suboptimal for determining mortality risk in patients with acute pancreatitis (AP); this might be improved by using a machine learning (ML) model. In this study, we aimed to construct an explainable ML model to calculate the risk of mortality in patients with AP admitted in intensive care unit (ICU) and compared it with existing scoring systems.

Methods A gradient-boosting ML (XGBoost) model was developed and externally validated based on two public databases: Medical Information Mart for Intensive Care (MIMIC, training cohort) and the eICU Collaborative Research Database (eICU-CRD, validation cohort). We compared the performance of the XGBoost model with validated clinical risk scoring systems (the APACHE IV, SOFA, and Bedside Index for Severity in Acute Pancreatitis [BISAP]) by area under receiver operating characteristic curve (AUC) analysis. SHAP (SHapley Additive exPlanations) method was applied to provide the explanation behind the prediction outcome.

Results The XGBoost model performed better than the clinical scoring systems in correctly predicting mortality risk of AP patients, achieving an AUC of 0.89 (95% CI: 0.84–0.94). When set the sensitivity at 100% for death prediction, the model had a specificity of 38%, much higher than the APACHE IV, SOFA and BISAP score, which had a specificity of 1%, 16% and 1% respectively.

Conclusions This model might increase identification of very low-risk patients who can be safely monitored in a general ward for management. By making the model explainable, physicians would be able to better understand the reasoning behind the prediction.

Keywords Acute pancreatitis, Prognostic factor, XGBoost, Mortality, Prediction

[†] Meng Jiang and Xiao-peng Wu contributed equally to this work.

*Correspondence:
Meng Jiang
jmhust@zju.edu.cn
Chang-li Li
lcl_hubei@126.com

[1] Emergency and Trauma Centre, The First Affiliated Hospital, Zhejiang University School of Medicine, #79 Qingchun Road, Hangzhou 310003, Zhejiang Province, P.R. China
[2] State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
[3] Department of FSTC Clinic, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, P.R. China

Introduction

Acute pancreatitis (AP) is a common gastrointestinal disease for acute hospitalization, with an annual incidence of 34.8/100,000 person-years globally [1]. Clinical outcomes vary broadly in AP from a self-limiting mild disease course (rarely die) in the majority of patients, to a severe course complicated by multiple organ failure in approximately 20% of patients (with a mortality of 36–50%) [2–4]. Risk stratification is important for triaging patients to the appropriate level of care and for deciding the aggressiveness of intervention. For this reason, approximately 20 scoring systems have been developed and evaluated since 1974 for AP severity assessment and mortality prediction [5], including the commonly used APACHE II [6], Ranson score [7], and Bedside Index for Severity in Acute Pancreatitis (BISAP) [8].

One major limitation of the scoring systems is that they are frequently cumbersome to calculate (such as APACHE II score) in clinical practice [5], and physicians continue to rely mainly on comorbidities, vital signs, and certain laboratory data to determine the severity for triaging and managing patients. This traditional strategy is experience depended and might lead to a significant number of AP patients unnecessarily admitted to the intensive care unit (ICU) for observation.

All current scoring systems use standard statistical methods to identify predictors and most allocate fixed weights according to the original dataset that used for model construction. Machine learning (ML) is a strategy that applies computational algorithms to learn from data, and the performance improves with experience (i.e., more training data) for executing a specific task. However, several factors have limited the studies of ML models in AP, such as small sample sizes, absence of external validation, or absence of head-to-head comparisons with existing risk tools [9–11].

Currently, the available evidence on risk assessment scores for patients with AP does not provide clear guidance on how to identify very low-risk patients who might be safely monitored in a general ward but not in ICU, thus to enhance the cost-effectiveness of care. The aim of this study was to develop and validate a ML model to identify very-low risk patients admitted to the ICU for AP, and perform a head-to-head comparison of its performance to existing clinical risk scores. Since one of the limitations for applying the ML model in clinical practice is the “black-box” pattern, we used explainable ML methods to provide a transparent interpretation on how a prediction was made.

Methods

Data source and outcome

Data were collected from two sizeable critical care databases: the MIMIC (MIMIC-III and MIMIC-IV, training set) and eICU-CRD (validation set) in accordance with the ethical standards of the institutional review board of the Massachusetts Institute of Technology (no. 0403000206) and with the 1964 Helsinki declaration and its later amendments. As the database is de-identified, we do not need to obtain informed consent from patients. The MIMIC-IV is the latest version of MIMIC database, which currently contains comprehensive and high-quality data of patients admitted to ICUs at the Beth Israel Deaconess Medical Center between 2008 and 2019. The MIMIC-IV covered 524,520 ICU admissions between 2008 and 2019 of 257,366 patients. In order to maximize the prediction power of the ML model, we also searched for eligible AP patients from the earlier version of MIMIC-III that was not included in the MIMIC-IV (2001–2008) for model construction. The eICU-CRD comprises 200,859 ICU admissions for 139,367 unique patients admitted between 2014 and 2015 at 208 hospitals located throughout the US. The outcome for this study was all-cause in hospital mortality of patients admitted to the ICU. One author (Dr. MJ) obtained access to both databases and was responsible for data extraction and analysis. The study was reported abide by the recommendations of the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement [12].

Selection of participants

The inclusion criterion for the study was adult patients (≥ 18 years old) admitted to ICU with AP. The diagnosis of AP (ICD-9: 5770) requires two of the following three features [4]: (1) abdominal pain (acute onset of a persistent epigastric pain, often radiating to the back); (2) serum lipase activity at least three times greater than the upper limit of normal; and (3) characteristic findings on contrast-enhanced computed tomography (CECT) or transabdominal ultrasonography. Patients who died within the first 24 h of their stay were excluded. Initial assessment of patients was performed when patients presented in ICU.

Data collection

We collected data from the electronic health records including demographical characteristics, vital signs, laboratory results, treatment and Glasgow Coma Scale (GCS) from the first 24 h in ICU, since: (1) the first 24 h of ICU admission are often considered a critical window for assessing the severity of AP and predicting patient outcomes; (2) many widely used clinical scoring systems, such as the APACHE II, SOFA, and BISAP scores, also rely on data collected within the first 24 h of ICU admission. This consistency allows for a more direct comparison between our ML model and these established scoring systems; (3) the first 24 h of ICU admission are when the most comprehensive data are typically available, ensuring that our model can be applied in real-world clinical settings.

The features were then analyzed for statistical and clinical significance with mortality, and a subset was selected as predictors for the ML model. The mean values in the first 24 h in the ICU were used for the vital sign: respiratory rate (RR), oxygen saturation (SpO2), heart rate (HR), systolic/diastolic blood pressure, mean arterial pressure (MAP), and temperature. For laboratory information, the maximum value in the first 24 h was selected for: serum creatinine, blood urea nitrogen (BUN), lipase, total bilirubin, lactate, prothrombin time (PT), international normalized ratio (INR), alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), white blood cells (WBC), potassium, sodium, hematocrit and glucose. The minimum value in first 24 h was chosen for serum bicarbonate, total calcium, albumin, platelets, and hemoglobin (Hb). Treatment information was included for the use of vasopressors, underwent invasive mechanical ventilation and renal replacement treatment (RRT) in the first 24 h. Comorbidities were recorded as binary variables for the presence of hepatic cirrhosis, metastatic cancer, diabetes mellitus, chronic obstructive pulmonary disease, congestive heart failure, and renal failure. Because missing data could create bias, variables with >20% missing values were excluded from further analysis. To reduce the impact of missing data on model construction, we used the KNNImputer (KNN) method to impute data missing less than 20%. We selected KNNImputer for its ability to preserve data structure, handle non-linear relationships, and maintain computational efficiency compare to other imputation techniques (e.g., mean imputation or multiple imputation by chained equations).

Study design

The MIMIC dataset was used to train a boosted ensemble model (XGBoost, using the open-source XGBoost library) and to conduct internal validation using tenfold cross validation. For internal validation, the MIMIC dataset was divided into 10 folds and each fold was used for internal validation with the remaining 90% folds as training set. Performing cross-validation is considered a robust strategy for a ML model assessment prior to external validation, which could maximize the potential prediction performance [13, 14]. External validation was performed on the eICU-CRD cohort.

For AP patients, using risk scores in clinical practice to identify very low-risk patients who might be safely monitored in a general ward or even managed in outpatient is rarely discussed. Achieving a high sensitivity for death prediction has obvious advantage, because false negatives need to be very rare so that patients who will become deteriorate or die can obtain the necessary intensive care. Therefore, to evaluate the clinical utility of the ML model, we planned to choose the low-risk cutoff value for the clinical risk tools and ML model by making sensitivity at 100% (or nearly 100% if none reached 100%) for death prediction in the external validation cohort, and to compare the specificity of different models [15, 16]. The higher specificity indicates that the tool would identify a greater proportion of patients presenting with AP who could be safely monitored in a step-down unit.

Prior to choosing and optimizing the XGBoost model, we performed rigorous exploratory analyses of logistic regression, random forest, and k-nearest neighbor (kNN). Separate models were developed with hyperparameter to optimize their respective performance, and all models underwent external validation. The preliminary findings suggested that the XGBoost model appeared to perform best on the eICU-CRD validation cohort. XGBoost is a recently generated gradient tree boosting algorithm, which is scalable and allows for faster computation [17]. Another advantage is that this algorithm can handle missing data effectively, whereas for other ML algorithms, we had to perform suitable data imputations, and the data imputation strategies might impact the comparisons between different models. Thus, the XGBoost was preferred over the other ML models in this task.

Model interpretation

To better interpret the prediction of the XGBoost model, we employed the SHapley Additive exPlanations (SHAP) method [18], which shows the summing effects of all variable attributions for approximating each patient's predicted outcome.

Statistical analysis

Continuous covariates are presented as median (interquartile range [IQR]) or mean (standard deviation [SD]) based on the normality of the data distribution, and analyzed using either the Student's t test or the Manne-Whitney U-test as appropriate. Categorical variables are reported as number and percentage, and analyzed by Chi-Squared test or Fisher's exact test. We used stepwise logistic regression model to select variables that were predictive of mortality. Both forward selection and backward elimination were applied, testing at each step for parameters to be included or excluded. The selection criteria to eliminate the predictors was Akaike Information Criterion (AIC).

For ML models, the XGBoost version 0.8 and scikit-learn version 2.1 packages were employed to develop models and tune hyperparameters in Python version 3.6 [17, 19]. Our primary metric was to compare the area under receiver operating characteristic curves (AUCs) for ML models with the SOFA, APACHE IV, and BISAP risk scores. The nonparametric DeLong test was used for comparing different AUCs [20]. The model's performance in classification that assigning patients into risk categories (e.g., low-risk vs. high-risk) was evaluated via sensitivity and specificity. Calibration of the model was assessed using the calibration curve and Hosmer-Lemeshow test as well as Brier score.

Results

Baseline characteristics

As shown in Figs. 1 and 1782 patients with AP in MIMIC (MIMIC-III: 942 and MIMIC-IV: 840) were included in the training cohort. A total of 250 patients died in hospital (14.0%, 95% CI: 12.4-15.7%). A cohort of 507 AP patients in eICU-CRD was included as external validation set.

The continuous and categorical variables (Table 1) were analyzed for association with mortality in both the MIMIC and eICU-CRD cohorts. Of the 1782 patients included in the training set, the median age was 67.1 years (IQR, 55.3–80.0]) in the non-survivors, which was significantly older than the survivors of 61.0 years (IQR, 48.0-77.2) ($P<0.001$). Most patients (46.1%) were admitted to the ICU from the emergency room, whereas 35.3% were transferred from the general ward. The remaining admissions were transferred from other hospitals.

The median age for the eICU-CRD group was 61.0 years (IQR, 48.0-77.2) for non-survivors compared with 52.0 years (IQR, 41.0–63.0) for survivors ($P=0.012$). Comparisons of the training and external validation cohorts are shown in Supplementary Table 1. The mortality rate was 6.7% (95% CI: 4.7%-9.25) in the eICU-CRD group.

Construction of the XGBoost model

For the training cohort, the variables with the greatest missingness include albumin ($N=537$), lactate ($N=1012$), and lipase ($N=1059$). They were excluded for further analysis. The features included for the ML model development were summarized in Table 2.

The hyperparameters used for training our XGBoost model were listed in Supplementary Table 2. With these hyperparameters, the training process is presented in Supplementary Fig. 1. As shown, the model performed

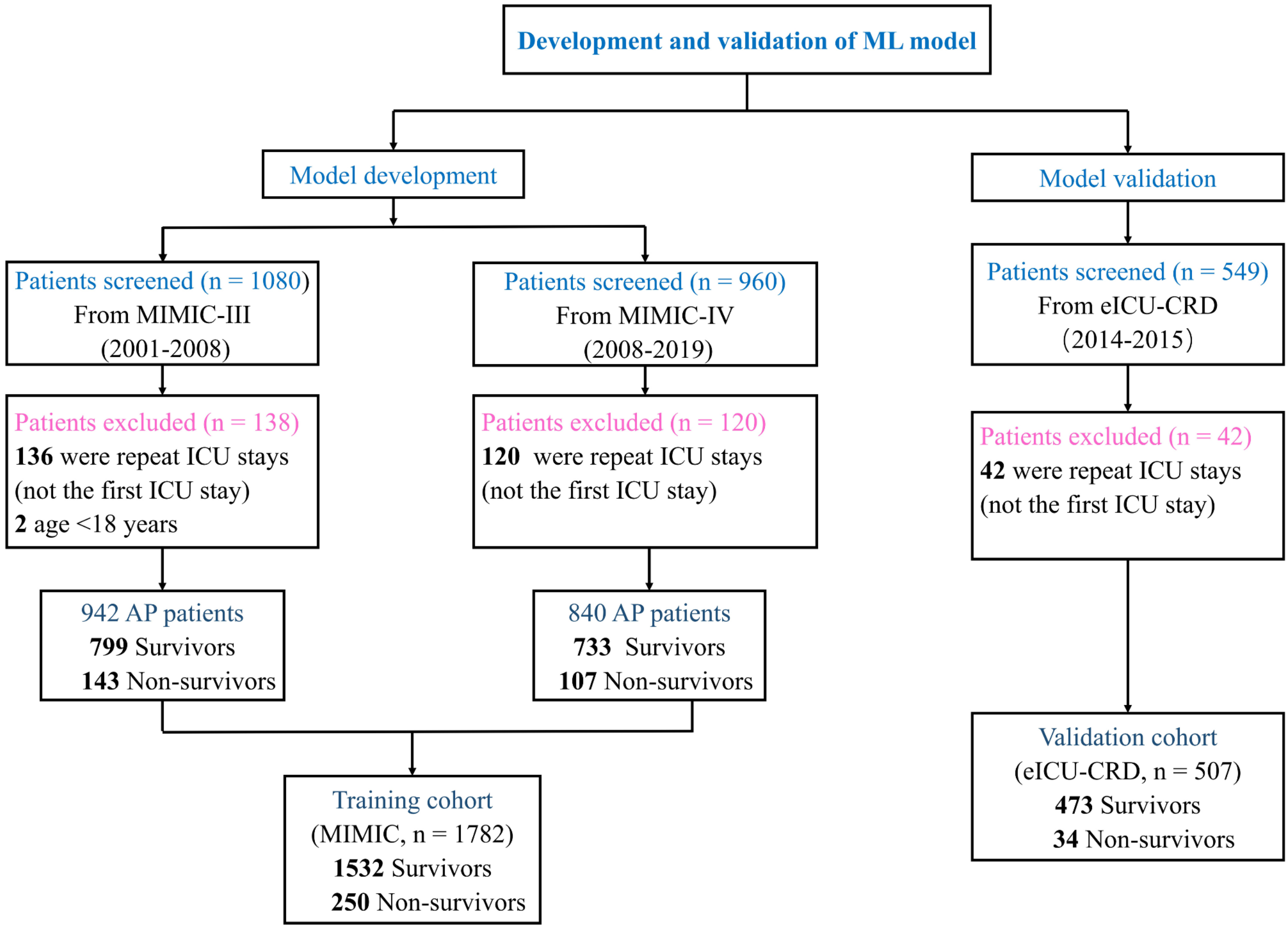

Fig. 1 Flow chart of patient selection

Table 1 Summary of demographic and clinical features for the training and validation cohort

	Training cohort (MIMIC dataset) (*n* = 1782)			**Validation cohort (eICU-CRD dataset) (*n* = 507)**		
	Non-survivors (*n* = 250)	**Survivors (*n* = 1532)**	***p***	**Non-survivors (*n* = 34)**	**Survivors (*n* = 473)**	***p***
Demographic						
Age (median [IQR])	67.1 [55.3, 80.0]	58.0 [45.5, 70.5]	< 0.001	61.0 [48.0, 77.2]	52.0 [41.0, 63.0]	0.012
Male (%)	140 (56.0)	883 (57.6)	0.677	20 (58.8)	272 (57.5)	1
Ethnicity, *n* (%)			0.001			0.152
African American	12 (4.8)	169 (11.0)		4 (11.8)	36 (7.6)	
Other	82 (32.8)	382 (24.9)		9 (26.5)	75 (15.9)	
White	156 (62.4)	981 (64.0)		21 (61.8)	362 (76.5)	
Comorbidities, *n* (%)						
Congestive heart failure	81 (32.4)	304 (19.8)	< 0.001	2 (5.9)	37 (7.8)	0.939
COPD	49 (19.6)	295 (19.3)	0.967	4 (11.8)	43 (9.1)	0.831
Renal failure	52 (20.8)	234 (15.3)	0.035	4 (11.8)	22 (4.7)	0.157
Metastatic cancer	21 (8.4)	35 (2.3)	< 0.001	0 (0.0)	2 (0.4)	1
Diabetes	74 (29.6)	458 (29.9)	0.984	14 (41.2)	156 (33.0)	0.43
Cirrhosis	15 (6.0)	31 (2.0)	0.001	2 (5.9)	19 (4.0)	0.935
Treatments, *n* (%)						
Renal replacement treatment	30 (12.0)	74 (4.8)	< 0.001	6 (17.6)	15 (3.2)	< 0.001
Invasive ventilation	89 (35.6)	344 (22.5)	< 0.001	21 (61.8)	71 (15.0)	< 0.001
Vasopressor use	129 (51.6)	329 (21.5)	< 0.001	18 (52.9)	41 (8.7)	< 0.001
Vitals (median [IQR])						
Heart rate, beats/min	95.8 [83.4, 110.5]	93.6 [80.5, 107.7]	0.035	100.6 [88.6, 112.1]	96.1 [83.3, 110.3]	0.193
Mean arterial pressure, mm Hg	72.6 [66.3, 82.1]	81.3 [73.5, 91.5]	< 0.001	68.3 [63.0, 71.1]	82.5 [73.0, 89.6]	< 0.001
Diastolic blood pressure, mm Hg	56.9 [50.9, 67.0]	66.7 [58.2, 75.5]	< 0.001	56.1 [51.0, 62.5]	70.6 [61.4, 79.6]	< 0.001
Systolic blood pressure, mm Hg	109.0 [100.8, 120.2]	121.8 [109.9, 136.7]	< 0.001	100.1 [95.5, 108.1]	124.1 [110.3, 138.7]	< 0.001
Respiratory rate, breaths/min	22.7 [20.0, 25.7]	19.8 [17.0, 23.1]	< 0.001	22.3 [19.0, 28.6]	19.9 [16.9, 23.6]	0.019
Oxygen saturation, %	96.0 [94.4, 97.5]	96.7 [95.2, 98.1]	< 0.001	95.7 [93.2, 97.6]	96.2 [94.7, 97.7]	0.299
Temperature, ℃	36.9 [36.3, 37.3]	37.0 [36.6, 37.4]	< 0.001	36.4 [36.1, 36.7]	36.6 [36.3, 36.9]	0.029
Metal status						
GCS score (median [IQR])	14.0 [10.0, 15.0]	15.0 [13.5, 15.0]	< 0.001	9.0 [3.8, 14.0]	15.0 [14.0, 15.0]	< 0.001
Laboratory results (median [IQR])						
Serum creatinine, mg/dL	1.9 [1.1, 3.2]	1.1 [0.8, 1.8]	< 0.001	2.4 [1.5, 4.6]	1.2 [0.8, 2.1]	< 0.001
Blood urea nitrogen, mg/dL	38.0 [23.0, 60.0]	20.0 [13.0, 35.0]	< 0.001	34.5 [21.2, 47.8]	19.0 [12.0, 34.0]	< 0.001
Total bilirubin, mg/dL	2.2 [0.9, 5.9]	1.1 [0.6, 2.8]	< 0.001	1.8 [0.8, 5.6]	1.2 [0.7, 2.4]	0.064
Prothrombin time, s	17.2 [14.4, 23.1]	14.9 [13.3, 19.8]	< 0.001	18.2 [15.5, 32.3]	14.3 [12.4, 16.9]	< 0.001
International normalized ratio	1.6 [1.3, 2.3]	1.3 [1.2, 1.6]	< 0.001	1.5 [1.3, 3.0]	1.2 [1.1, 1.4]	< 0.001
Bicarbonate, mmol/L	17.0 [13.0, 21.0]	21.0 [18.0, 24.0]	< 0.001	18.4 [11.0, 21.6]	20.0 [15.0, 23.0]	0.07
Platelets, K/µL	151.0 [75.0, 231.0]	181.0 [124.0, 261.0]	< 0.001	151.5 [110.5, 251.8]	165.5 [115.2, 233.8]	0.548
Hematocrit, %	34.5 [30.0, 40.3]	35.2 [31.0, 40.0]	0.268	37.0 [30.3, 45.4]	41.0 [35.7, 45.5]	0.052
Hemoglobin, g/dL	9.3 [7.8, 10.9]	10.3 [8.8, 11.9]	< 0.001	10.6 [9.0, 12.1]	12.0 [10.3, 13.6]	0.001
White blood cells, K/µL	16.7 [11.4, 23.2]	13.6 [9.5, 19.5]	< 0.001	15.6 [13.1, 23.8]	14.1 [9.7, 18.8]	0.032
Potassium, mmol/L	4.7 [4.2, 5.4]	4.3 [3.9, 4.9]	< 0.001	4.7 [4.3, 5.3]	4.3 [3.9, 4.8]	0.006
Sodium, mmol/L	141.0 [138.0, 144.0]	140.0 [137.0, 143.0]	0.286	140.0 [138.2, 143.8]	139.0 [137.0, 143.0]	0.21
Total calcium, mg/dL	7.4 [6.7, 8.0]	7.8 [7.2, 8.3]	< 0.001	7.4 [6.7, 8.0]	7.7 [7.0, 8.3]	0.26
Aminotransferase alanine, U/L	75.5 [29.2, 328.0]	62.0 [25.0, 256.0]	0.204	73.5 [20.8, 415.8]	59.0 [31.0, 138.0]	0.716
Aminotransferase aspartate, U/L	141.0 [48.2, 504.0]	84.0 [35.0, 316.2]	< 0.001	146.0 [39.5, 979.0]	84.0 [35.8, 213.5]	0.047
Alkaline phosphatase, U/L	128.5 [75.0, 270.2]	117.0 [76.0, 252.0]	0.731	115.0 [74.5, 169.5]	114.0 [82.0, 168.0]	0.93
Glucose, mg/dL	101.0 [82.0, 126.0]	102.0 [86.0, 123.0]	0.415	122.0 [99.2, 176.8]	113.0 [90.0, 147.0]	0.115
Clinical risk score (median [IQR])						
SOFA	9 [6, 13]	4 [2, 7]	< 0.001	10 [8, 13]	5 [2, 7]	< 0.001
APACHE IV score	NA	NA	NA	88 [76, 114]	48 [36, 66]	< 0.001
BISAP	NA	NA	NA	3 [3, 4]	3 [2, 3]	< 0.001

COPD, chronic obstructive pulmonary disease; IQR, interquartile range; GCS, Glasgow Coma Scale/Score; SOFA, Sequential Organ Failure Assessment; APACHE, Acute Physiology and Chronic Health Evaluation; BISAP, Bedside Index for Severity in Acute Pancreatitis; NA, not available

Table 2 Clinical features used for the machine learning model

Category	No. of variables	Variables
Demographic	1	Age
Comorbidities	4	Congestive heart failure
		Renal failure
		Metastatic cancer
		Cirrhosis
Treatments	3	Renal replacement treatment
		Invasive ventilation
		Vasopressor use
Vitals	7	Heart rate
		Mean arterial pressure
		Diastolic blood pressure
		Systolic blood pressure
		Respiratory rate
		Oxygen saturation
		Temperature
Metal status	1	Net Glasgow Coma Scale
Laboratory results	14	Serum creatinine
		Blood urea nitrogen
		Total bilirubin
		Prothrombin time
		International normalized ratio
		Bicarbonate
		Platelets
		Hemoglobin
		White blood cells
		Potassium
		Total calcium
		Aminotransferase aspartate
		Sodium
		Alkaline phosphatase

best at the training rounds of 24th, achieving the maximum AUC and minimum log-loss in the eICU-CRD validation cohort. To identify the variables that influenced the model the most, we depicted the SHAP summary plot (Fig. 2) and presented the top 25 variables of the prediction model. This plot depicts the associations between variables' values and SHAP values in the training dataset. According to the XGBoost model, the higher the SHAP value of a variable, the more likely for patients will die in hospital. The SHAP dependence plot (Supplementary Fig. 2) can also be applied to understand how a specific feature affects the prediction outcome. We could visualize how the variable's attributed importance altered as its values varied in the plot. SHAP values for a feature exceeding zero imply an increased risk of mortality.

Performance of the machine learning model

The ML model was tested on the eICU-CRD cohort. For predicting in hospital death, the XGBoost model performed better than other ML models, achieving an AUC of 0.89 (95% CI: 0.84–0.94) than the random forest AUC of 0.80 (95% CI: 0.73–0.87) and kNN AUC of 0.63 (95% CI: 0.53–0.72), with a P value of 0.01 and less than 0.001 respectively (Table 3a). Compared with APACHE IV and SOFA scores, the AUC of XGBoost was higher but the difference was not significant. The BISAP only achieved an AUC of 0.67 (95% CI: 0.58–0.76) for predicting mortality in this cohort. Figure 3 displays the ROCs for these models in the external validation cohort. The calibration curve (Supplementary Fig. 3) and Hosmer-Lemeshow test ($P = 0.342$) for the eICU-CRD validation cohort demonstrated a high accuracy of the ML model. Besides, the model achieved a Brier score of 0.08, further indicating the strong calibration (lower values denote better agreement between predicted and observed outcomes).

Identifying very low-risk patients

To identify very low-risk patients, we setted the sensitivity of the ML models and clinical risk scores at 100% for predicting death in the validation cohort. These models were then compared in view of the specificity. The XGBoost model performed better than the APACHE IV, SOFA and BISAP risk scores in correctly identifying patients who were at very low risk of death (true negatives). The XGBoost model achieved the specificity of 38% at a sensitivity of 100%, much higher than the APACHE IV, SOFA and BISAP score, which had a specificity of 1%, 16% and 1% respectively at the sensitivity of 100% (Table 3b).

Discussion

In patients with AP, a XGBoost model derived from a large cohort predicts the mortality better than the current commonly used clinical risk scores. The ML model relies on capturing important feature interactions to generate predictions, which could be used to improve the ability to identify very-low risk AP patients who can be safely monitored on a general medical ward. Importantly, the model increases the number of very-low risk AP patients who can be detected by more than two-fold as compared with the best performing clinical risk scale currently available (SOFA score).

AP is a potentially deadly disease; however, the clinical outcomes vary broadly among different cases. Thus, it is important to predict patient outcomes to facilitate clinicians to choose the appropriate level of care (i.e., transferred to a general ward or continue monitored in an ICU), to determine the aggressiveness of treatment, and to counsel patients' prognosis [5]. Although several guidelines recommend the application of certain scoring systems in clinical decision-making [21–23], it remains uncertain on how the results should guide therapy. Especially, no relevant studies were available to date that addressed the cost-effectiveness of these scoring systems. Currently, triaging AP patients is still strongly relying on the doctor's clinical assessment. This has raised concern

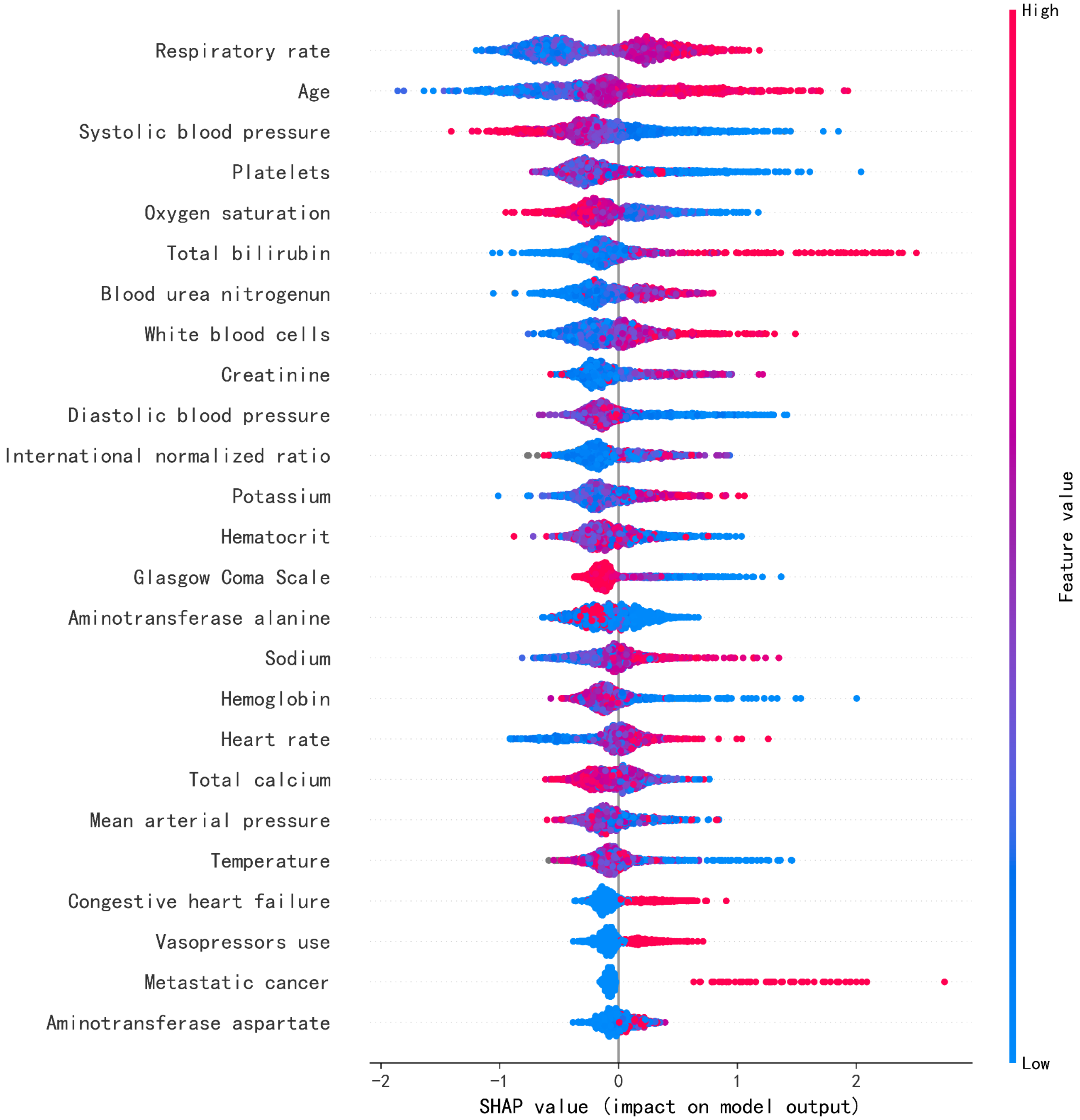

Fig. 2 SHAP summary plot of the top 25 features of the XGBoost model. The higher the SHAP value of a variable, the higher the probability of mortality. A dot is created for each feature attribution value for the model of each patient, and thus one patient is allocated one dot on the line for each variable. Dots are colored according to the values of variables for the respective patient and accumulate vertically to depict density. Red represents higher variable values, and blue represents lower variable values

about unnecessary ICU admissions and cost utilization. Our model's performance at identifying low-risk patients is promising, which could be integrated into clinical workflows (it would be integrated with electronic health record systems to provide real-time risk assessments) that if a patient with AP is predicted as at very low risk, she/he might be treated in a general ward but not in the ICU.

ICU provides an optimal environment for supportive care, has the ability to monitor and manage severe volume depletion and organ insufficiency. However, due to that intensive care is a limited resource, its valuable to screen for those very-low risk patients who could be

Table 3a Performance of machine learning models and clinical risk scores for mortality prediction in eICU-CRD validation cohort

Models	**AUC** **(95% CI)**	**Confidence Interval** **(2.5 – 97.5%)**		***p*-value**
XGBoost	0.89	0.84	0.94	reference
Random Forest	0.80	0.73	0.87	0.01
kNN	0.63	0.53	0.72	< 0.001
Logistic Regression	0.85	0.80	0.91	0.06
Clinical Risk Score				
APACHE IV	0.86	0.80	0.93	0.34
SOFA	0.83	0.75	0.91	0.09
BISAP	0.67	0.58	0.76	< 0.001

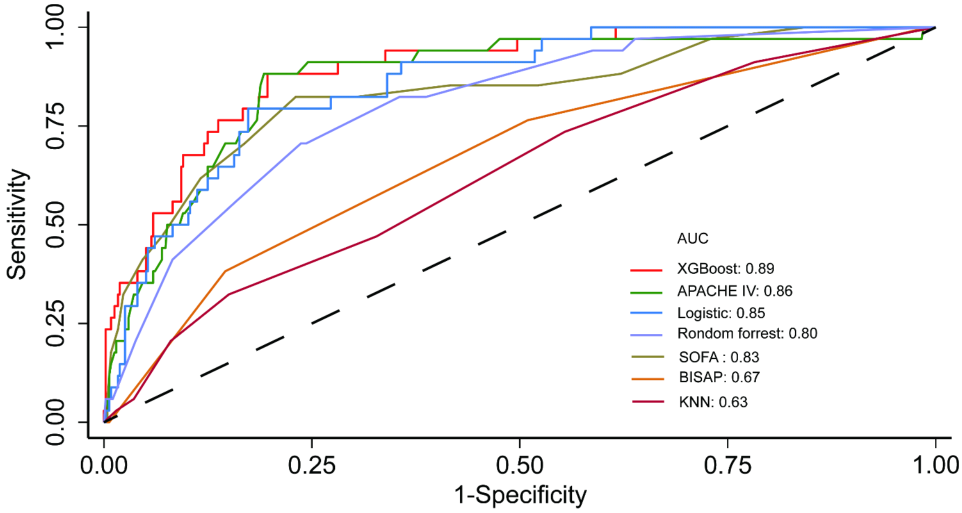

Fig. 3 Comparison of AUCs among machine learning models and clinical risk scores. XGBoost yielded the greatest AUC in the external validation cohort

Table 3b Performance characteristics for machine learning models and clinical risk scores for high sensitivity (100%) threshold in eICU-CRD validation cohort

Models	**Sensitivity**	**Specificity**	**PPV**	**NPV**
XGB	1.0	0.38	0.1	1.0
Random Forest	1.0	0.02	0.07	1.0
kNN	1.0	0	0	1.0
Logistic Regression	1.0	0.40	0.1	1.0
Clinical Risk Score				
APACHE IV	1.0	0.01	0.01	1.0
SOFA	1.0	0.16	0.08	1.0
BISAP	1.0	0.01	0.07	1.0

SOFA, Sequential Organ Failure Assessment; APACHE, Acute Physiology and Chronic Health Evaluation; BISAP, Bedside Index for Severity in Acute Pancreatitis; kNN, k-nearest neighbor; AUC, area under the receiver operator characteristic curve; PPV, positive prediction value; NPV, negative prediction value

safely managed in a general ward, so that the real potential severe cases would receive the appropriate intensive care. In this study, we compared our ML model's prediction performance with the APACHE IV, SOFA and BISAP score and found that the model was able to identify very low-risk cases (setting sensitivity at 100%) with 38% specificity compared with 1% specificity for the APACHE IV score, 16% specificity for the SOFA score, and 1% specificity for the BISAP score. This means that the model could screen for 38% of patients who are at very low risk of mortality and may not need to be admitted to the ICU for monitoring.

Explainability

The key factors in determining if the ML model predictions would be used by clinicians for clinical decision-making is if they can understand on how the prediction is made. The higher the interpretability for a model, the easier it is for clinicians to comprehend and thus make an appropriate clinical decision accordingly. In highly

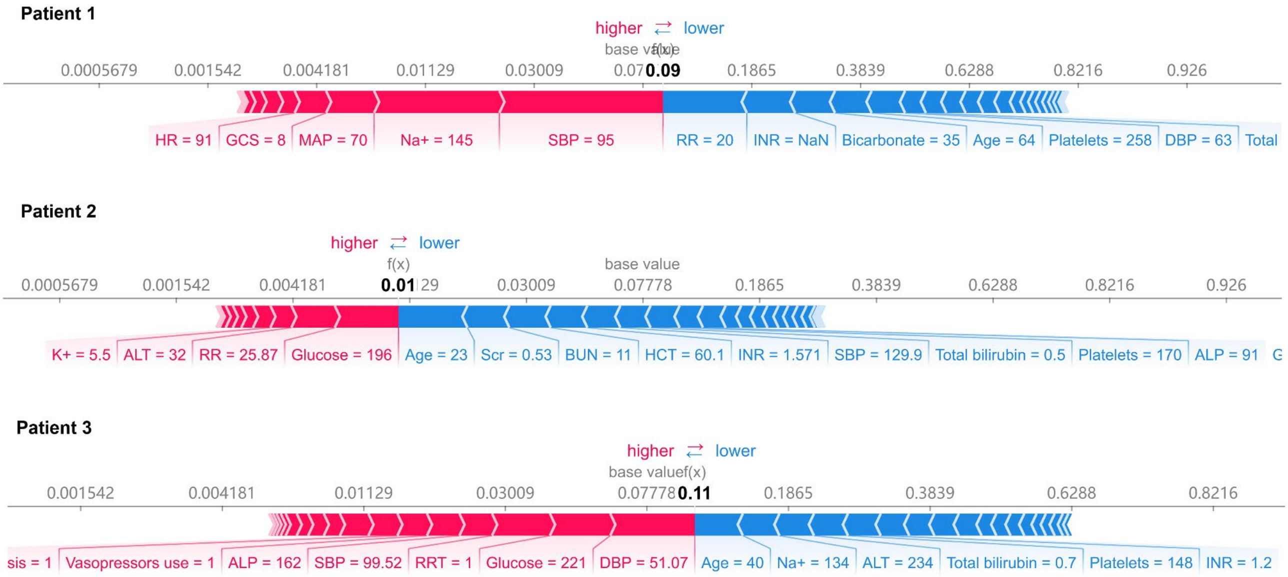

Fig. 4 SHAP explanation force plot for 3 patients from the eICU-CRD validation cohort of the ML model. ALT, aminotransferase alanine; BUN, blood urea nitrogen; GCS, Glasgow Coma Scale/Score; HR, heart rate; INR, international normalized ratio; MAP, mean arterial pressure; PT, prothrombin time; RR, respiratory rate; SPO2, oxygen saturation; WBC, white blood cell; Na+, sodium; K+, potassium; SBP, systolic blood pressure; RR, respiratory rate; INR, international normalized ratio; DBP, diastolic blood pressure; Scr, serum creatinine; HCT, hematocrit

interpretable models, such as sparse linear models, the coefficients of each feature indicates how much weight has been assigned to that factor; whereas in the black box models like XGBoost, it is unknown which variables are contributing to the decision for a specific case. In order to understand why the model came to a certain prediction, we used the SHAP method [24, 25] to explain individual predictions of three patients from the hold-out eICU-CRD cohort (Fig. 4).

ML explainability results for three patients

Patient 1.

This is a 64-year-old female patient who was admitted to the ICU for AP. This patient developed septic shock and ARDS while in the ICU and was intubated during this hospitalization. The predicted probability of death is 9% for her compared with the baseline of 7.8%. The factors identified by the ML model for predicting a higher mortality for this patient include decreased systolic blood pressure (95 mm Hg), low GCS score (8), borderline MAP (70 mmHg), sodium (145 mmol/L) and HR (91 beats/min). The predicted outcome of the model was death for this case, and the actual outcome was also death (true positive). However, the model predicted the mortality risk to be 9% based on information of the first 24 h after admission. The patient's length of stay was 118 h (approximately 5 days).

From a clinicians' point of view, ICU observation is justified for this patient, considering the low systolic blood pressure and GCS score around the time of admission in addition to development of septic shock later in the ICU. Supplementary Fig. 4 presented a visualization of the whole ICU stay for this patient. The patient's GCS score dropped to 3 after 28 h of admission (was intubated at this point), and the MAP fluctuated broadly during the ICU duration.

Patient 2.

This is a 23-year-old male patient with no comorbidities who presented with AP. This patient neither undergone intubation or RRT nor developed an infection during this hospitalization. The predicted probability for death by the ML model was 1% compared to the baseline of 7.8%. The model detected RR of 26 breaths/min, glucose of 196 mg/dL, ALT of 32 U/L, and potassium of 5.5 mmol/L as the risk factors for mortality, whereas the young age of 23 years, normal serum creatinine of 0.53 mg/dL, BUN of 7 mg/dL, and systolic blood pressure of 130 mm Hg decreased the risk of mortality. The model's predicted outcome was alive, in line with the final outcome of alive (true negative).

Because we aimed to use the ML model to screen for very low-risk patients, this is the group of interest for this study. This patient was observed in ICU for 57 h. However, from the clinical perspective of the whole ICU stay for this patient (Supplementary Fig. 5), this patient may not need an ICU level of care and might be safely dispositioned in a general unit.

Patient 3.

This is a 40-year-old diabetic woman with cirrhosis and a liver transplant history, taking immunosuppressive medications currently, who presented with AP and upper gastrointestinal bleeding. The predicted probability for death by the ML model was 11% compared to the baseline of 7.8%. Her mortality risk detected by the model including diastolic blood pressure of 51 mmHg, systolic blood pressure of 100 mmHg, glucose of 221 mg/dL, ALP of 162 U/L, need for RRT and vasopressors, as well as the comorbidities of liver cirrhosis. The patient's lower age of 40 years, total bilirubin of 0.7 mg/dL, and platelets of 148 K/μL partly offset the mortality risk predicted by the model. The model correctly predicted the outcome as death, and she died ten days after admission to the ICU (true positive). The visualization of the whole ICU stay of this patient was shown in Supplementary Fig. 6, the vital sign of HR and blood pressure varied broadly during the hospitalization in ICU.

The explainability in the ML model can help clinicians understanding why the model came to the conclusion. From a clinician's point of view, considering the blood pressure in the presence of vasopressors use, high glucose, need for RRT and cirrhosis, the patient should be monitored in the ICU.

In clinical practice, each patient would get a prediction outcome according to the ML algorithm just like Fig. 4, we can detect how the features that came to the conclusion and evaluate its reasonableness based on our clinical experiences.

Strengths

First, unlike prior reports [9, 11] on ML model in AP, we trained the model using a large cohort, and performed external validation for its generalization in an independent patient group. Second, we used SHAP method to show how the prediction was made by the model for specific instances. The reasoning behind the predictions makes it more acceptable by clinicians in clinical practice. Third, the XGBoost model outperformed existing clinical scoring systems in predicting mortality risk in AP patients, particularly in identifying low-risk patients who may not require ICU-level care. The model's strong calibration and consistent performance across patient subgroups (data not shown) further support its potential for clinical application.

Limitations

First, despite the improvement in performance for the specificity of 38% at the high sensitivity cutoff, it is less than optimal. This means that there are still several patients who will not deteriorate or die are not identified as low risk and are still unnecessarily monitored in ICU. By including additional variables or using other algorithms (ensemble methods) might further optimizing the model. However, the improvement in specificity for the XGBoost model compared with currently used clinical scores would potentially enhance the cost-effectiveness of ICU care for patients with AP. The trade-off between sensitivity and specificity should be considered flexibly based on the ICU resources' availability. Second, the etiology of AP was not available in the database. Besides, the data used was retrospectively collected and several features have some degree of missingness. This may somewhat affect the stability and precision of the prediction, leading to false positives or negatives that might affect patient care. Further improvement of the model by adding prospective high-quality data would be necessary for its application in clinical setting. Third, the MIMIC and eICU-CRD database are from independent cohort, however, would the ML model perform well in other populations (e.g., outside US) need to be validated in the future. Fourth, including additional factors (e.g., chronic kidney disease or obesity) or more granular comorbidity classifications could improve the model's predictive accuracy. However, limited by the database, we just included common comorbidities in the model construction. Finally, expanding the application of our XGBoost model to emergency departments and initial assessments might improve AP care across the healthcare continuum. By providing actionable risk stratification at the earliest stages, the model can enhance triage accuracy, reduce unnecessary ICU utilization, and ultimately save lives. However, the data availability in other clinical settings may impair our model's generalizability.

Conclusion

In summary, machine learning model has been proved to be superior to existing prediction scores for mortality prediction of AP. The use of most of previous ML models is limited in clinical practice, mainly due to the lack of explainability in the clinical setting. Our study used XGBoost model with SHAP method in a critically ill AP cohort to predict mortality, and identified variables that facilitated the model to make a reasonable prediction, thus making it more transparent and reliable. In the future prospective data are warranted to refine the model, thus to integrate it into clinical decision support systems.

Abbreviations

AP	Acute pancreatitis
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ALP	Alkaline phosphatase
BUN	Blood urea nitrogen
BISAP	Bedside Index for Severity in Acute Pancreatitis
eICU-CRD	eICU Collaborative Research Database
GCS	Glasgow Coma Scale
HR	Heart rate
INR	International normalized ratio
IQR	Interquartile range
INR	International normalized ratio
ICU	Intensive care unit
ML	Machine learning

MAP	Mean arterial pressure
MIMIC	Medical Information Mart for Intensive Care
PT	Prothrombin time
RR	Respiratory rate
RRT	Renal replacement therapy
SpO2	Oxygen saturation
SCr	Serum creatinine
SHAP	SHapley Additive exPlanations
WBC	White blood cells

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12876-025-03723-3.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

M.J. and C.L.L conceived the study idea and design, drafted the manuscript. M.J. obtained funding. M.J. performed acquisition of data. C.L.L and X.P.W performed analysis and interpretation of data. M.J and X.P.W performed critical revision of the manuscript for important intellectual content and statistical analysis. M.J. and C.L.L took administrative, technical, or material support tasks.

Funding

This work was supported by the National Natural Science Foundation of China (82000479), the Natural Science Foundation of Zhejiang Province (LMS25H180003), and the Fundamental Research Funds for the Central Universities (2022ZFJH003).

Data availability

Data were fully available at https://mimic.mit.edu and https://eicu-crd.mit.edu/. The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request. Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate

Since the study was an analysis of the third party anonymized publicly available database with pre-existing institutional review board (IRB) approval, IRB approval from The First Affiliated Hospital of Zhejiang university was waived.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 October 2024 / Accepted: 24 February 2025

Published online: 03 March 2025

References

1. Li C, Jiang M, Pan C, Li J, Xu L. The global, regional, and National burden of acute pancreatitis in 204 countries and territories, 1990–2019. BMC Gastroenterol. 2021;21(1):332.
2. Boxhoorn L, Voermans RP, Bouwense SA, Bruno MJ, Verdonk RC, Boermeester MA, et al. Acute pancreatitis. Lancet. 2020;396(10252):726–34.
3. Buter A, Imrie CW, Carter CR, Evans S, McKay CJ. Dynamic nature of early organ dysfunction determines outcome in acute pancreatitis. Brit J Surg. 2002;89(3):298–302.
4. Banks PA, Bollen TL, Dervenis C, Gooszen HG, Johnson CD, Sarr MG, et al. Classification of acute pancreatitis-2012: revision of the Atlanta classification and definitions by international consensus. Gut. 2013;62(1):102–11.
5. Di MY, Liu H, Yang ZY, Bonis PA, Tang JL, Lau J. Prediction models of mortality in acute pancreatitis in adults: A systematic review. Ann Intern Med. 2016;165(7):482–90.
6. Wilson C, Heath DI, Imrie CW. Prediction of outcome in acute pancreatitis: a comparative study of APACHE II, clinical assessment and multiple factor scoring systems. Brit J Surg. 1990;77(11):1260–4.
7. Ranson JH, Pasternack BS. Statistical methods for quantifying the severity of clinical acute pancreatitis. J Surg Res. 1977;22(2):79–91.
8. Wu BU, Johannes RS, Sun X, Tabak Y, Conwell DL, Banks PA. The early prediction of mortality in acute pancreatitis: a large population-based study. Gut. 2008;57(12):1698–703.
9. Xu F, Chen X, Li C, Liu J, Qiu Q, He M et al. (2021) Prediction of Multiple Organ Failure Complicated by Moderately Severe or Severe Acute Pancreatitis Based on Machine Learning: A Multicenter Cohort Study. Mediat Inflamm. 2021;5525118.
10. Qiu Q, Nian Y, Guo Y, Tang L, Lu N, Wen L, et al. Development and validation of three machine-learning models for predicting multiple organ failure in moderately severe and severe acute pancreatitis. BMC Gastroenterol. 2019;19(1):118.
11. Pearce CB, Gunn SR, Ahmed A, Johnson CD. Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C-Reactive protein. Pancreatology. 2006;6(1–2):123–31.
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
13. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J. 2017;38(7):500–7.
14. Rotondano G, Cipolletta L, Grossi E, Koch M, Intraligi M, Buscema M, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. Gastrointest Endosc. 2011;73(2):218–26.
15. Shung DL, Au B, Taylor RA, Tay JK, Laursen SB, Stanley AJ, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper Gastrointestinal bleeding. Gastroenterology. 2020;158(1):160–7.
16. Deshmukh F, Merchant SS. Explainable machine learning model for predicting GI bleed mortality in the intensive care unit. Am J Gastroenterol. 2020;115(10):1657–68.
17. Chen T, Guestrin C, XGBoost:. A scalable tree boosting system. KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
18. Lundberg S, Lee S. A Unified Approach to Interpreting Model Predictions. 2017.
19. Pedregosa F, Varoquaux G, Granfort A, et al. Scikit-learn: machine learning in Python. JMLR. 2011;12:2825–30.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
21. Forsmark CE, Baillie J. AGA Institute technical review on acute pancreatitis. Gastroenterology. 2007;132(5):2022–44.
22. Besselink M, Santvoort HV, Freeman M, et al. IAP/APA evidence-based guidelines for the management of acute pancreatitis. Pancreatology. 2013;13(4 Suppl 2):e1–15.
23. Leppäniemi A, Tolonen M, Tarasconi A, Segovia-Lohse H, Gamberini E, Kirkpatrick AW, et al. 2019 WSES guidelines for the management of severe acute pancreatitis. World J Emerg Surg. 2019;14:27.
24. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Proc Adv Neural Inf Process Syst. 2017;4768–77.
25. Strumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41:647–65.

Publisher's note