scientific reports

Check for updates

OPEN

Machine learning model to predict sepsis in ICU patients with intracerebral hemorrhage

Lei Tang[1,2,3,4,5,6,7], Ye Li[1,2,3,4,5,6,7], Ji Zhang[1], Feng Zhang[1,2,3,4,5,6,7,8], Qiaoling Tang[1,2,3,4,5,6,7], Xiangbin Zhang[1,2,3,4,5,6,7], Sai Wang[1,2,3,4,5,6,7], Yupeng Zhang[1,2,3,4,5,6,7], Siyuan Ma[1,2,3,4,5,6,7], Ran Liu[1,2,3,4,5,6,7], Lei Chen[1,2,3,4,5,6,7], Junyi Ma[1,2,3,4,5,6,7], Xuelun Zou[1,2,3,4,5,6,7], Tianxing Yao[1,2,3,4,5,6,7], Rongmei Tang[1,2,3,4,5,6,7], Huifang Zhou[1,2,3,4,5,6,7], Lianxu Wu[1,2,3,4,5,6,7], Yexiang Yi[1,2,3,4,5,6,7], Yi Zeng[9], Duolao Wang[10]✉ & Le Zhang[1,2,3,4,5,6,7]✉

Patients with intracerebral hemorrhage (ICH) are highly susceptible to sepsis. This study evaluates the efficacy of machine learning (ML) models in predicting sepsis risk in intensive care units (ICUs) patients with ICH. We conducted a retrospective analysis on ICH patients using the MIMIC-IV database, randomly dividing them into training and validation cohorts. We identified sepsis prognostic factors using Least Absolute Shrinkage and Selection Operator (LASSO) and backward stepwise logistic regression. Several machine learning algorithms were developed and assessed for predictive accuracy, with external validation performed using the eICU Collaborative Research Database (eICU-CRD). We analyzed 2,214 patients, including 1,550 in the training set, 664 in the validation set, and 513 for external validation using the eICU-CRD. The Random Forest (RF) model outperformed others, achieving Area Under the Curves (AUCs) of 0.912 in training, 0.832 in internal validation, and 0.798 in external validation. Neural Network and Logistic Regression models recorded training AUCs of 0.840 and 0.804, respectively. ML models, especially the RF model, effectively predict sepsis in ICU patients with ICH, enabling early identification and management of high-risk cases.

Keywords Intracerebral hemorrhage, Sepsis, Machine learning, Prediction model

Intracerebral hemorrhage (ICH) represents about 15% of all stroke cases yet is associated with significant mortality, accounting for roughly 2.8 million deaths globally each year[1,2]. Approximately 11–31% of ICH cases lead to infections and long-term disabilities[3,4], often exacerbating into sepsis due to immunosuppression-induced systemic inflammations and metabolic disorders[5,6].

In the intensive care unit (ICU), approximately 25% of ICH patients who develop sepsis succumb within 28 days. Furthermore, sepsis is strongly associated with worsened long-term functional outcomes. These findings highlight the critical importance of early detection and effective management of sepsis in ICH patients[7]. A retrospective cohort study revealed that about 28% of ICH patients would develop sepsis, which poses significant challenges in their clinical management[8]. Sepsis is associated with an increased risk of a systemic infectious encephalopathy called sepsis-associated encephalopathy (SAE)[9], leading to in-hospital coma events and a higher risk of complications in ICH patients[10].

The initial clinical manifestations of sepsis are not specific, and the disease progresses rapidly[11,12]. Currently, there are no effective treatments, highlighting the significance of early detection and appropriate management to mitigate its impact[13–15]. A comprehensive understanding of the etiology of sepsis after ICH is essential for improving targeted prevention and treatment. However, the underlying cause of sepsis has not been fully

[1]Department of Neurology, Xiangya Hospital, Central South University, Jiangxi, Nanchang 330006, Jiangxi, China. [2]Department of Neurology, Xiangya Hospital, Central South University, Changsha 410008, Hunan, China. [3]Multi-Modal Monitoring Technology for Severe Cerebrovascular Disease of Human Engineering Research Center, Changsha, Hunan, China. [4]Brain Health Center of Hunan Province, Changsha, Hunan, China. [5]Human Brain Disease Biological Resources Platform of Hunan Province, Changsha, Hunan, China. [6]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, Hunan, China. [7]FuRong Laboratory, Changsha 410078, Hunan, China. [8]Department of Cardiovascular Medicine, Xiangya Hospital, Central South University, Changsha 410008, Hunan, China. [9]Department of Geriatrics, Second Xiangya Hospital, Central South University, Changsha 410011, Hunan, China. [10]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK. ✉email: duolao.wang@lstmed.ac.uk; zlzdzlzd@csu.edu.cn

elucidated to date[11]. This complicates the management of ICH-associated sepsis, posing challenges in reducing mortality rates and addressing cognitive complications. In this context, developing predictive models for early sepsis detection and risk factor identification is of great importance for improving early prevention strategies[16].

In recent years, advancements in statistical theory and computer technology have propelled machine learning (ML) into the forefront of medical research, garnering significant attention from clinicians. ML techniques have outperformed traditional methods like logistic regression and Cox regression in disease prediction, as evidenced by comparative studies[17,18]. Particularly, neural networks (NNs) have grown substantially in size and sophistication over the last decade, becoming leading tools in ML applications[19]. Among various algorithms, random forest (RF) and boosted trees with calibrated probabilities have demonstrated superior performance across multiple metrics[20]. Previous research has also shown consistent strong performance of RF algorithms on various biomedical data sets[21,22]. These algorithms offer several advantages, such as scalability to large data sets and greater robustness compared to other algorithm types[23]. We can find considerable effort in the application of ML algorithms in sepsis prediction[24–27]. Despite the demonstrated efficacy of ML in predicting various diseases, its application in predicting sepsis among patients with ICH remains underexplored, with limited data available. This study aims to address this gap by developing and validating multiple ML models to accurately predict the onset of sepsis in ICH patients, striving to determine the most effective model for clinical use.

Methods

Database

This study utilized clinical data from two sources: the Medical Information Mart for Intensive Care (MIMIC)-IV database (version 2.2), containing ICU patient records from Beth Israel Deaconess Medical Center between 2008 and 2019, and the eICU-CRD (Telehealth Intensive Care Unit Collaborative Research Database), a database of over 200,000 ICU admissions across 208 U.S. hospitals during 2014–2015, used for external validation. Data integrity and research compliance were ensured by author Y.L., who completed the National Institutes of Health's "Protecting Human Research Participants" program (certification number:53244021). The study is registered at Clinical Trials.gov (NCT06326385).

Study population and definitions

Data were extracted using Structured Query Language in PostgreSQL (version 14.6) and the study adhered to STROCSS criteria. Diagnoses from both databases were identified via the International Classification of Diseases, Ninth Revision (ICD-9) and Tenth Revision (ICD-10) codes[28]. Patients were included if they had a diagnosis of ICH and were admitted to the ICU. Exclusion criteria included pediatric patients (< 18 years), patients with sepsis diagnosed prior to ICU admission, and those with incomplete follow-up data. The primary outcome of this study was the occurrence of sepsis within 28 days after ICU admission, defined according to the Third International Consensus Definition (Sepsis-3) as suspected or confirmed infection and a Sequential Organ Failure Assessment (SOFA) score ≥ 2 points. For patients diagnosed with sepsis, the follow-up time was defined as the interval (in days) from ICU admission to the first documented diagnosis of sepsis. For non-sepsis patients, the follow-up time was defined as the duration from ICU admission to discharge.

Data collection

In this study, we extracted baseline patient data including age, sex, ethnicity, and Body Mass Index (BMI) from the database. ICU admission metrics such as Sequential Organ Failure Assessment (SOFA) and Glasgow Coma Scale (GCS) scores were collected, along with the use of mechanical ventilation (MV), continuous renal replacement therapy (CRRT), peripherally inserted central catheter (PICC), and intracranial pressure (ICP) monitoring. Patients were categorized into craniotomy, minimally invasive surgery (MIS), or non-surgical groups based on the type of cranial intervention received. Vital signs and comorbidities were also recorded at admission. Laboratory tests performed included White Blood Cell count (WBC), anion gap, creatinine, and more. Laboratory measurements taken after the diagnosis of sepsis were excluded from the analysis to avoid bias. The primary study outcome was the incidence of sepsis during the follow-up period.

Statistical analysis

The Shapiro–Wilk test was used to test the normality assumption. Continuous variables were summarized using median interquartile ranges and compared using the Kruskal-Wallis test. Categorical variables were summarized as numbers and percentages and compared using the chi-square test or Fisher's exact test. Missing values were addressed through multiple imputation (missForest R), with variables missing over 25% transformed into dummy variables to reduce bias. The number and percentage of missing values for each variable have been provided in Supplementary Table 1. To evaluate the impact of sepsis on prognosis, survival curves were plotted to compare the outcomes of ICH patients who developed sepsis and those who did not. The impact of sepsis on mortality was assessed using the log-rank test.

We divided ICH patients into training and testing sets to identify significant predictors. Using the glmnet package, we performed Least Absolute Shrinkage and Selection Operator (LASSO) regression to select non-zero coefficient features, followed by stepwise logistic regression to determine significant variables at $P < 0.05$. Patients were randomly split into a 7:3 training-to-testing ratio using the caTools package. Logistic regression was used as a baseline model to compare its performance with machine learning approaches (RF and NN). To account for nonlinear effects in continuous variables, restricted cubic splines (RCS) were incorporated into the model using the rms package in R. Knot placement followed standard recommendations, with four knots.

Our analysis included logistic regression, RF, and NN, with the model's performance assessed via the area under the receiver operating characteristics (ROC) curve using the pROC package. Decision Curve Analysis (DCA) was applied to establish the clinical utility of the models. The chosen model underwent five-fold cross-

validation and was evaluated based on metrics such as the area under the ROC curve (AUC), sensitivity, specificity, recall, accuracy, and F1 score[29], ensuring robust validation of its predictive capabilities. To further enhance the interpretability of the optimal model, SHapley Additive exPlanations (SHAP) analysis was conducted, providing insights into the contribution of each feature to the predictions.

Results

Demographic and clinical characteristics of ICH patients

We analyzed 2,214 patients who fulfilled the inclusion and exclusion criteria in MIMIC-IV database, with 1550 patients in the training cohort and 664 in the validation cohort. The patient selection process is illustrated in Supplemental Fig. 1. The baseline characteristics of patients upon admission are shown in Table 1. The study found that 53.5% of the participants, totaling 1184 patients, were men. In the group diagnosed with ICH, 813 patients (36.7%) also experienced complications due to sepsis. The prevalence of hypertension was 63.2% in the training cohort and 60.1% in the internal validation cohort. Additionally, the heart disease was present in 46.0% of the training cohort and 46.7% of the internal validation cohort. Diabetes was found in 25.1% and 22.4% of the training and internal validation cohorts, respectively. Acute pneumonia was observed in 10.9% of the training cohort and 9.6% of the validation cohort. The GCS score was consistent across both cohorts at 14 [IQR 11–15]. WBC counts (109/l) averaged 10.25 [IQR 8.40–12.10] in the training cohort and 10.16 [IQR 8.40–12.00] in the internal validation cohort. Average chloride levels (mmol/L) were noted at 103.40 [IQR 102.00-105.13] mmol/L in the training cohort and 103.00 [IQR 101.00-105.25] in the internal validation cohort. Additionally, baseline characteristics of the study cohort grouped by the presence or absence of sepsis are provided in Supplement Table 2. Furthermore, we investigated the impact of sepsis on the prognosis of patients with ICH. The results revealed that sepsis was associated with a significantly higher mortality rate. The survival curve (Supplemental Fig. 2) demonstrates a significant difference between the two groups, with ICH patients who developed sepsis having a markedly higher risk of mortality compared to those without sepsis (log-rank test, $p < 0.0001$).

Feature selection and nomogram construction

In our analysis of cohorts from the MIMIC-IV database, sepsis was identified in 813 of the total participants, representing 36.7% of the entire sample. This incidence rate was consistent across different study subsets, with 36.7% of both the training set (569 out of 1,551 patients) and the internal validation set (244 out of 664 patients) diagnosed with sepsis. To identify relevant variables, we used LASSO and backward stepwise logistic regression. The different mean square errors for different log(lambda) ranges are shown in Supplemental Fig. 3. Subsequently, RCS analysis was performed to explore the non-linear relationships of continuous variables in the model. The results demonstrated significant non-linear associations between sepsis and both WBC count (non-linear $p < 0.0001$) and chloride levels (non-linear $p = 0.018$), as shown in Supplemental Fig. 4.

Variables with non-zero coefficients from LASSO regression were further screened using stepwise logistic regression, which identified race, gender, acute pneumonia, fluid electrolyte disorders, heart diseases, liver diseases, renal failure, ICP monitoring, invasive ventilation, supplemental oxygen, GCS score, heart rate, chloride, and WBC as independent risk factors for sepsis in ICH patients. Patients undergoing ICP monitoring exhibited a significantly higher likelihood of developing sepsis, with odds 4.12 times greater than those not monitored (OR 4.12; 95% CI: 2.19, 8.00). Similarly, the presence of acute pneumonia in patients increases the odds of sepsis by 3.56 times compared to those without pneumonia (95% CI = 2.75, 4.63). Furthermore, patients with cerebral hemorrhage who also suffer from fluid and electrolyte disorders are 3.56 times more likely to develop sepsis than their counterparts without such disorders (95% CI = 2.75, 4.63). Detailed OR and 95% CI values are shown in Supplemental Table 3. These results were used to create a nomogram to estimate the odds of sepsis in ICU patients with ICH (Fig. 1). For instance, the male patient had liver disease, acute pneumonia, fluid electrolyte disorders, and underwent intracranial pressure monitoring, and invasive ventilation upon admission, with no supplemental oxygen upon admission, and a chloride level of 110 mmol/l. The sum of these points (68) is located on the total points axis. A line is then drawn downward on the axis to determine the probability of developing sepsis (> 90%). The feature importance ranking under the framework of RF algorithm is shown in Fig. 2.

Model performance comparisons and internal validation

We developed three ML models to predict the development of sepsis in patients with ICH after ICU admission. Figure 3 illustrates ROC curves measuring the discrimination of these models. In the training set, the RF model demonstrated the highest predictive performance for sepsis in ICH patients (AUC = 0.912, 95% CI: 0.898–0.927), followed by the NN model (AUC = 0.840, 95% CI: 0.820–0.861) and the LR model (AUC = 0.804, 95% CI: 0.781–0.827).

In internal validation, the RF model achieved the best performance with an AUC of 0.832 (95% CI: 0.801–0.864), outperforming the NN model (AUC = 0.811, 95% CI: 0.777–0.845) and the LR model (AUC = 0.799, 95% CI: 0.763–0.834) for predicting sepsis in patients with ICH. Detailed performance metrics for the three models are presented in Supplemental Table 4. Additionally, we incorporated RCS into the logistic regression model to account for potential nonlinear relationships between independent variables and sepsis risk. The inclusion of RCS terms improved the predictive performance of the LR model. As shown in Supplemental Fig. 5A, the AUC of the RCS-enhanced LR model increased to 0.812 (95% CI: 0.790–0.812) in the training set and 0.803 (95% CI: 0.768–0.803) in the validation set, compared to the original LR model without RCS terms (training set: AUC = 0.804, 95% CI: 0.781–0.827; validation set: AUC = 0.799, 95% CI: 0.763–0.834). Additionally, we incorporated interaction terms between variables significantly associated with the outcome to further enhance the LR model's predictive performance. As shown in Supplemental Fig. 5B, the inclusion of interaction terms improved the AUC of the LR model to 0.822 (95% CI: 0.800–0.822) in the training set and 0.809 (95% CI: 0.775–0.809) in the validation set.

Variable	*n* (%) or median (Interquartile range)			*P* value
	Total patients (*n* = 2214)	Training set (*n* = 1550)	Testing set (*n* = 664)	
Follow-up time until sepsis (days)	1.2 [0.3, 2.8]	1.2 [0.3, 2.8]	1.2 [0.4, 2.8]	> 0.900
Mortality	569.0 (25.7)	418.0 (27.0)	151.0 (22.7)	0.037
LOS	7.1 [3.7, 13.9]	7.5 [3.8, 14.0]	6.7 [3.7, 13.5]	0.300
Sepsis	813 (36.7)	569 (36.7)	244 (36.7)	1.000
Age, years				
<40	121 (5.5)	86 (5.5)	35 (5.3)	0.866
40–64	751 (33.9)	530 (34.2)	221 (33.3)	
≥65	1342 (60.6)	934 (60.3)	408 (61.4)	
Gender, Male	1184 (53.5)	830 (53.5)	354 (53.3)	0.956
Race				
Asian	80 (3.6)	57 (3.7)	23 (3.5)	0.290
Black	200 (9.0)	146 (9.4)	54 (8.1)	
Other	540 (24.4)	391 (25.2)	149 (22.4)	
White	1394 (63.0)	956 (61.7)	438 (66.0)	
BMI, kg/m²				
< 18.5	29 (1.3)	20 (1.3)	9 (1.4)	0.826
18.5–24.9	242 (10.9)	170 (11.0)	72 (10.8)	
25-29.9	284 (12.8)	194 (12.5)	90 (13.6)	
≥30	266 (12.0)	180 (11.6)	86 (13.0)	
Missing	1393 (62.9)	986 (63.6)	407 (61.3)	
Comorbidities				
Acute pneumonia	233 (10.5)	169 (10.9)	64 (9.6)	0.416
Chronic pulmonary	274 (12.4)	199 (12.8)	75 (11.3)	0.347
Diabetes	538 (24.3)	389 (25.1)	149 (22.4)	0.200
Fluid electrolyte disorders	753 (34.0)	520 (33.5)	233 (35.1)	0.514
Heart disease	1023 (46.2)	713 (46.0)	310 (46.7)	0.802
Hypertension	1378 (62.2)	979 (63.2)	399 (60.1)	0.188
Liver disease	120 (5.4)	93 (6.0)	27 (4.1)	0.082
Non-infectious colon disease	17 (0.8)	13 (0.8)	4 (0.6)	0.750
Obesity	133 (6.0)	90 (5.8)	43 (6.5)	0.610
Renal failure	465 (21.0)	320 (20.6)	145 (21.8)	0.566
Interventions				
CRRT	4 (0.2)	2 (0.1)	2 (0.3)	0.743
ICP monitoring	94 (4.2)	66 (4.3)	28 (4.2)	1.000
PICC	125 (5.6)	88 (5.7)	37 (5.6)	1.000
Invasive ventilation	532 (24.0)	374 (24.1)	158 (23.8)	0.909
Supplemental oxygen	735 (33.2)	509 (32.8)	226 (34.0)	0.618
Craniocerebral operations				
Craniotomy	24 (1.1)	16 (1.0)	8 (1.2)	0.414
MIS	41 (1.9)	25 (1.6)	16 (2.4)	
None	2149 (97.1)	1509 (97.4)	640 (96.4)	
First laboratory test				
Anion gap (mmol/l)	14.57 [13.00, 16.00]	14.55 [13.00, 16.00]	14.61 [13.00, 16.00]	0.646
Bicarbonate (mg/dl)	24.10 [22.00, 25.48]	24.09 [22.00, 25.36]	24.14 [22.00, 25.88]	0.861
BUN (mg/dl)	16.00 [13.00, 19.00]	16.00 [13.00, 19.00]	16.00 [13.00, 19.00]	0.773
Chloride (mmol/l)	103.24 [101.77, 105.16]	103.40 [102.00, 105.13]	103.00 [101.00, 105.25]	0.136
Creatinine (mg/dl)	0.89 [0.70, 1.00]	0.88 [0.70, 1.00]	0.89 [0.80, 1.00]	0.365
Hematocrit	36.46 [34.30, 39.50]	36.44 [34.20, 39.40]	36.50 [34.30, 39.60]	0.827
Hemoglobin (g/dl)	12.20 [11.40, 13.30]	12.20 [11.30, 13.30]	12.22 [11.50, 13.33]	0.356
Platelet ($\times 10^9$/l)	212.00 [177.00, 242.00]	213.00 [178.00, 243.00]	209.13 [176.75, 242.00]	0.319
Potassium (mmol/l)	3.90 [3.70, 4.20]	3.90 [3.70, 4.12]	3.90 [3.70, 4.30]	0.247
WBC ($\times 10^9$/l)	10.23 [8.40, 12.10]	10.25 [8.40, 12.10]	10.16 [8.40, 12.00]	0.534
Lymphocyte percentage				
Continued				

Variable	*n* (%) or median (Interquartile range)			*P* value
	Total patients (*n* = 2214)	Training set (*n* = 1550)	Testing set (*n* = 664)	
< 18	264 (11.9)	193 (12.5)	71 (10.7)	0.625
18–42	78 (3.5)	52 (3.4)	26 (3.9)	
> 42	4 (0.2)	3 (0.2)	1 (0.2)	
Missing	1868 (84.4)	1302 (84.0)	566 (85.2)	
Monocyte percentage				
< 2	17 (0.8)	15 (1.0)	2 (0.3)	0.257
2–11	306 (13.8)	219 (14.1)	87 (13.1)	
> 11	23 (1.0)	14 (0.9)	9 (1.4)	
Missing	1868 (84.4)	1302 (84.0)	566 (85.2)	
Neutrophil percentage				
< 50	11 (0.5)	10 (0.6)	1 (0.2)	0.302
50–70	64 (2.9)	42 (2.7)	22 (3.3)	
> 70	271 (12.2)	196 (12.6)	75 (11.3)	
Missing	1868 (84.4)	1302 (84.0)	566 (85.2)	
PT, sec				
< 10.4	27 (1.2)	17 (1.1)	10 (1.5)	0.852
10.4–13.4	982 (44.4)	685 (44.2)	297 (44.7)	
> 13.4	494 (22.3)	346 (22.3)	148 (22.3)	
Missing	711 (32.1)	502 (32.4)	209 (31.5)	
Sodium, mmol/l				
< 133	34 (1.5)	24 (1.5)	10 (1.5)	0.353
133–145	86 (3.9)	53 (3.4)	33 (5.0)	0.353
> 145	13 (0.6)	10 (0.6)	3 (0.5)	0.353
Missing	2081 (94.0)	1463 (94.4)	618 (93.1)	0.353
Heart rate (beats per minute)	80.00 [70.00, 91.75]	80.00 [70.00, 91.00]	80.00 [69.00, 92.25]	0.980
Respiratory rate (inspirations per minute)	18.00 [15.00, 21.00]	18.00 [15.00, 21.00]	18.00 [15.00, 21.00]	0.702
SBP, mmHg				
< 90	4 (0.2)	4 (0.3)	0 (0.0)	0.359
90–99	30 (1.4)	19 (1.2)	11 (1.7)	
100–109	94 (4.2)	63 (4.1)	31 (4.7)	
≥ 110	892 (40.3)	613 (39.5)	279 (42.0)	
Missing	1194 (53.9)	851 (54.9)	343 (51.7)	
DBP, mmHg				
< 140	1019 (46.0)	699 (45.1)	320 (48.2)	0.122
140–159	1 (0.0)	0 (0.0)	1 (0.2)	
Missing	1194 (53.9)	851 (54.9)	343 (51.7)	
GCS score	14.00 [11.00, 15.00]	14.00 [11.00, 15.00]	14.00 [11.00, 15.00]	0.892
SOFA score	1.00 [0.00, 2.00]	1.00 [0.00, 2.00]	0.50 [0.00, 1.00]	0.627

Table 1. Baseline characteristics of the study cohort (grouped by training and testing sets). Variables are initial values if not otherwise specified. *BMI* body mass index, *CRRT* continuous renal replacement therapy, *PICC* peripherally inserted central catheter, *ICP* intracranial pressure, *MIS* minimally invasive surgery, *BUN* blood urea nitrogen, *WBC* white blood cell, *PT* prothrombin time, *SBP* systolic blood pressure, *DBP* diastolic blood pressure, *GCS* Glasgow coma scale, *SOFA* sequential organ failure assessment, *LOS* length of stay, *MIMIC-IV* medical information mart for intensive care IV.

The RF model achieved the highest F1 scores in both the training set and the internal validation set (0.88 and 0.829, respectively), indicating it has the best discriminative ability. Sensitivity and specificity were also the most balanced. Taken together, the RF model is considered the optimal model. Finally, a DCA curve (Supplemental Fig. 6) demonstrated the RF model provided the highest net benefit and threshold probability, indicating its superior clinical utility compared with the other two models.

External validation

The RF model was validated in an external dataset from the eICU-CRD database using the same data extraction process as the derivation dataset. After screening, we ultimately selected 513 patients from the e-ICU for external validation. In the external validation dataset, the AUC was 0.798 (95% CI 0.606–0.99) (Fig. 4). This indicates

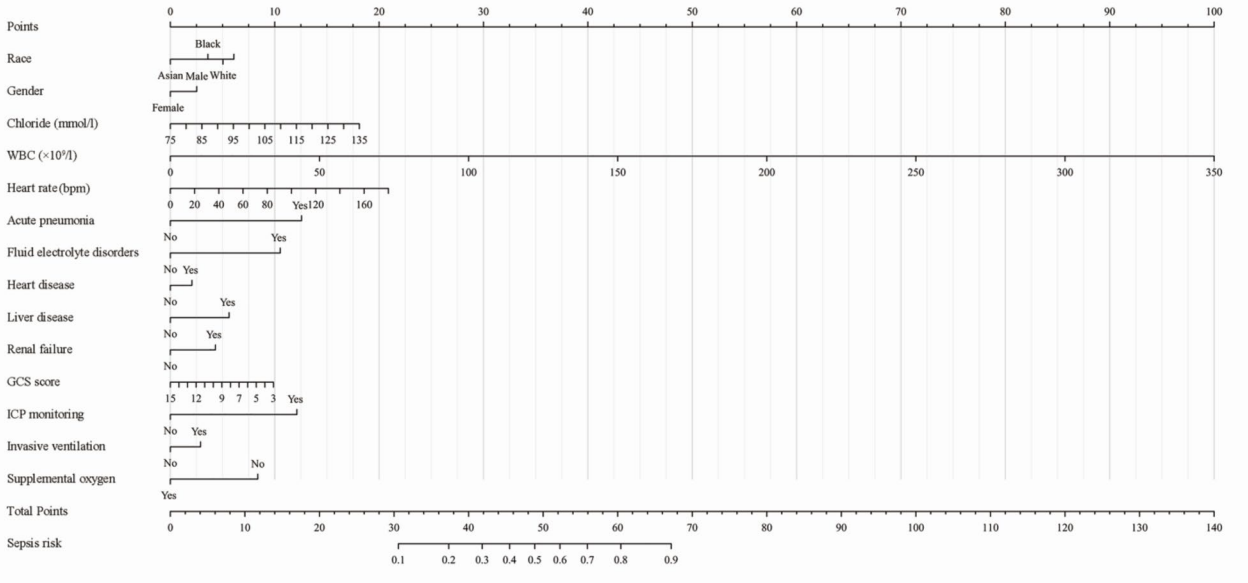

Fig. 1. Nomogram predicts the probability of sepsis in patients with ICH. Nomogram was established using variables including race, gender, acute pneumonia, fluid electrolyte disorders, heart diseases, liver diseases, renal failure, intracranial pressure (ICP) monitoring, invasive ventilation, supplemental oxygen, GCS score, heart rate, chloride, and WBC levels, for predicting the occurrence of sepsis after ICH. The total point was calculated as the sum of the individual values of the 14 variables included in the nomogram. Patients were scored for each variable and the total score was assigned according to the nomogram.

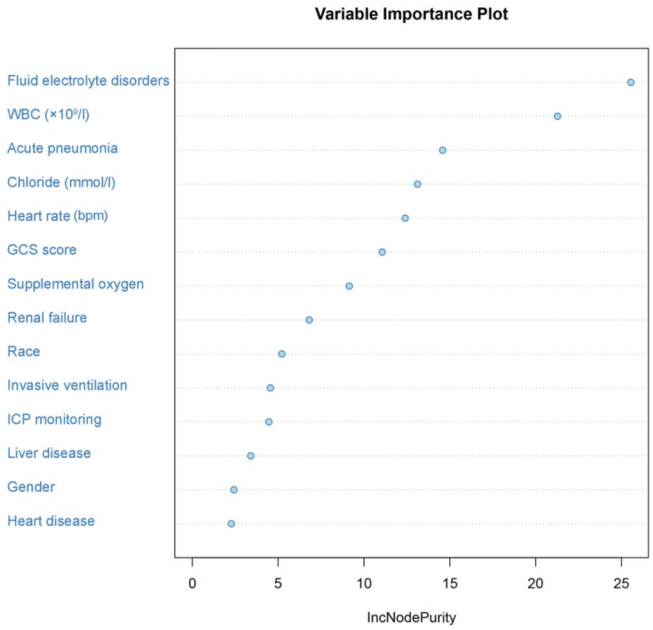

Fig. 2. The variable importance plot under the framework of random forest algorithm.

that the RF model demonstrated good predictive performance in independent external populations. However, further clinical evaluation is required to assess the robustness of the RF model across different populations.

SHAP analysis

To enhance the interpretability of the model, we conducted SHAP analysis on the optimal model, a RF classifier. The SHAP summary plot (Fig. 5) highlights the most influential features contributing to the prediction of sepsis risk in patients with ICH. The analysis identified fluid electrolyte disorders, WBC count, Supplemental oxygen use, GCS score, and renal failure as the top five contributors to the model's predictions. Among these, fluid

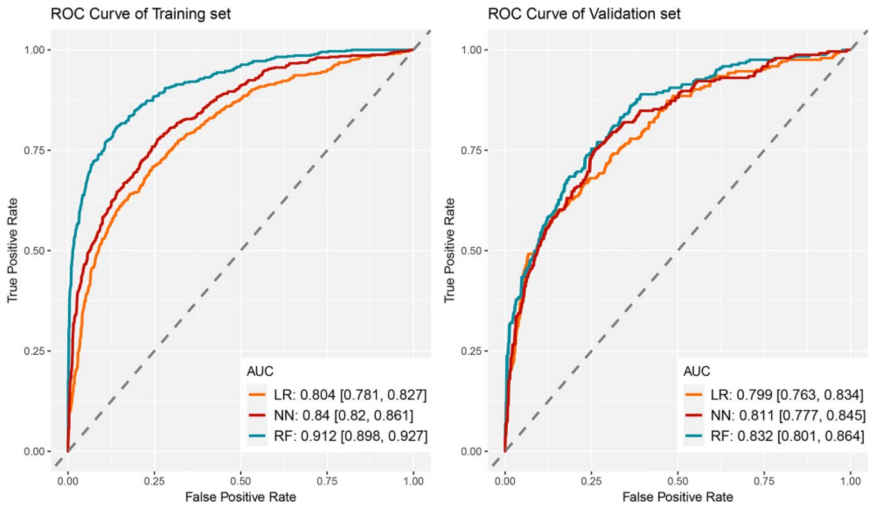

Fig. 3. ROC curves for sepsis on the (**A**) training and (**B**) validation sets. A greater AUC value indicated a higher predictive ability of the models. *ROC* receiver operating characteristic, *AUC* area under the curve, *LR* logistic regression without regularization, *RF* random forest, *NN* neural network.

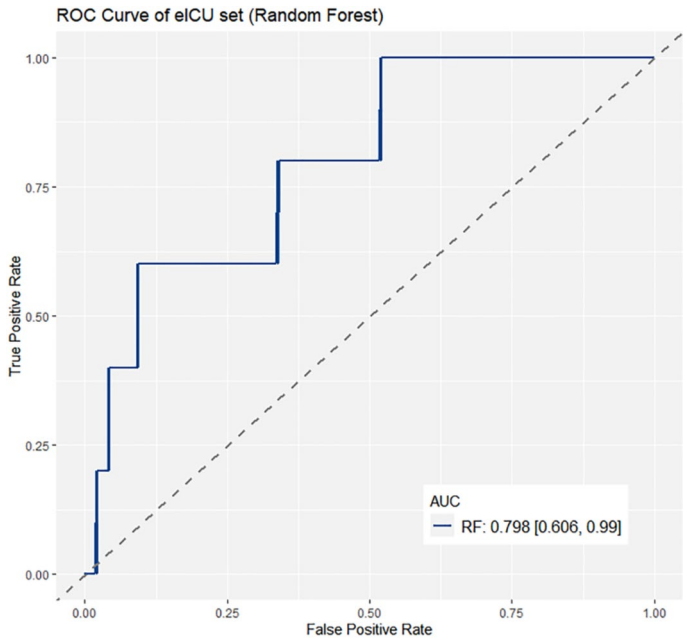

Fig. 4. Externally validated ROC curve of the RF model for predicting sepsis in ICH. *RF* random forest.

electrolyte disorders and elevated WBC counts demonstrated the strongest association with an increased risk of sepsis, as evidenced by their high SHAP values.

Discussion

To the best of our knowledge, this study represents a novel exploration of sepsis in ICH patients using the MIMIC-IV and eICU-CRD public databases. Our study demonstrates that machine learning models, particularly random forest models, exhibit high accuracy in predicting the onset of sepsis in patients with cerebral hemorrhage, showcasing superior clinical utility compared to alternative models. Compared to the LR algorithm which requires manual selection of independent variables, potentially introducing complex nonlinear relationships and interactions between independent variables into the error of the model, the RF model has several advantages. It efficiently handles missing data and creates effective predictive models by combining weak predictors. Due to its excellent accuracy and performance, the RF algorithm has received increasing attention as a competing alternative to LR analysis for predicting adverse clinical events.

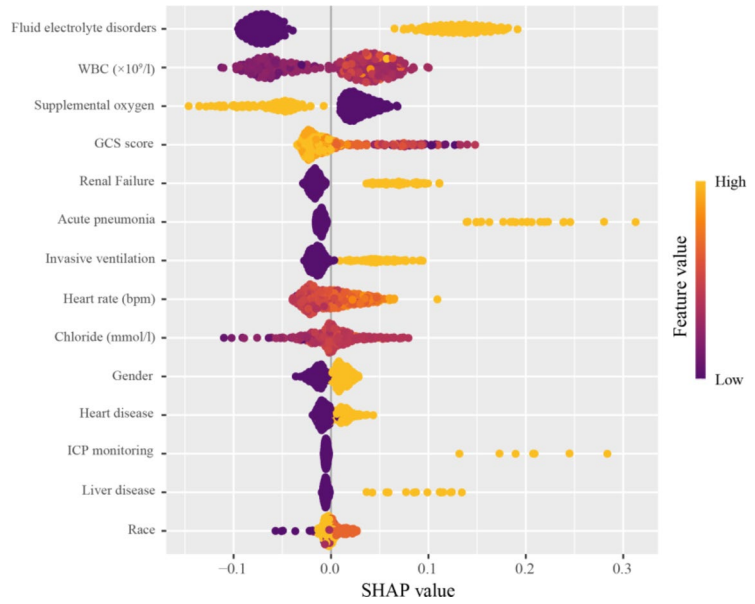

Fig. 5. Shapley value analysis for variable importance in the RF model. Each dot represents a Shapley value for an individual observation, indicating the contribution of a specific variable to the model's prediction. Variables are ordered by their average importance. The color gradient reflects the feature value, with yellow indicating higher values and purple indicating lower values. *RF* random forest.

Patients with ICH face a heightened susceptibility to sepsis due to immunosuppression and gut microbiota dysbiosis[16]. The combination of ICH and sepsis presents a significant challenge for clinicians, contributing to increased mortality rates and cognitive complications. Both conditions share similar pathophysiological mechanisms involving systemic inflammation and circulatory disturbance, resulting in high mortality and morbidity[30–34]. Timely risk assessment, prudent antibiotic therapy, and appropriate targeted therapy can reduce the incidence of sepsis[35], highlighting the importance of developing reliable predictive models for timely interventions and better management of high-risk patients.

Based on our study findings, we propose several early intervention strategies to mitigate sepsis risk in ICH patients. Integrating the RF model into ICU monitoring systems could enable real-time identification of high-risk patients, allowing for enhanced monitoring and timely interventions. Proactive infection control measures, such as early use of antibiotics, strict aseptic techniques during invasive procedures, and regular microbiological assessments, are crucial for reducing the risk of sepsis. Additionally, personalized interventions targeting key risk factors, including acute pneumonia, electrolyte imbalances, and renal failure, could further enhance patient care. Exploring immune-enhancing therapies to address immunosuppression in ICH patients also holds promise. These strategies underscore the critical importance of timely sepsis diagnosis, as early detection and intervention during the ICU stay can prevent the progression to severe sepsis or septic shock, ultimately optimizing patient care.

LASSO and multivariate logistic regression analysis identified nine clinical characteristic variables for assessing sepsis risk in ICH patients, including race, gender, acute pneumonia, fluid electrolyte disorders, heart disease, liver disease, renal failure, heart rate, GCS score, ICP monitoring, invasive ventilation, supplemental oxygen, chloride, and WBC. Liver disease has been consistently associated with an increased risk of infections, including sepsis, due to impaired synthetic function and detoxification processes[36,37]. This predisposition is particularly pronounced in ICU settings where patients are more susceptible to systemic infections[38]. Similarly, renal failure contributes to sepsis risk through mechanisms such as immune dysfunction and accumulation of uremic toxins, which impair host defenses[39–42]. Furthermore, heart disease impacts sepsis risk through multiple pathways, including reduced tissue perfusion and altered hemodynamic responses, which can precipitate organ dysfunction in critically ill patients[43–46]. In terms of invasive procedures, the use of invasive ventilation and ICP monitoring devices increases the risk of secondary infection, decreases hospital discharge rates, and raises mortality[47]. Aziz et al.[48] found that invasive mechanical ventilation is identified as an independent predictor of mortality in ICU-treated adults with sepsis. In addition, acute pneumonia is a significant independent predictor of sepsis in patients with ICH. Research indicates that around 43% of individuals with ICH develop acute pneumonia[49], which is linked to a four-fold increased risk of complications[50–52]. This aligns with the results of our study. Moreover, evidence has shown that elevated chloride levels independently correlate with mortality in ICU patients with ICH[53]. When patients have severe sepsis and hyperchloremia, this condition is typically observed upon admission to the ICU. Further, hyperchloremic patients admitted to the ICU experienced elevated chloride levels even after 72 h of admission, and the exacerbation of hyperchloremia was independently associated with all-cause in-hospital mortality[54,55].

The present study has certain limitations, including its retrospective and observational nature, which may introduce selection bias. The data from the MIMIC-IV database come from a single center in the United States,

potentially undermining the generalization to other populations. Therefore, external validation against different populations is required to accurately assess the model's performance. It is important to note that the model should be used as a medical reference only, as other complex clinical factors should also be considered in treatment decisions. Nevertheless, the established model can assist clinicians in the timely management of high-risk ICH patients with sepsis in the ICU.

Conclusion

In conclusion, ML models could be reliable tools for predicting sepsis in ICH patients. Among all prediction models, the RF model proved to be most effective in providing early identification and timely intervention for high-risk ICH patients with sepsis, potentially mitigating disease progression.

Data availability

The raw data for this study were sourced from the MIMIC-IV and eICU-CRD databases, both of which are accessible to the public. Detailed data pertinent to this study can be made available upon request by reaching out to the corresponding author.

Received: 15 October 2024; Accepted: 21 April 2025
Published online: 10 May 2025

References

1. Watson, N., Bonsack, F. & Sukumari-Ramesh, S. Intracerebral hemorrhage: the effects of aging on brain injury. *Front. Aging Neurosci.* **14**, 859067. https://doi.org/10.3389/fnagi.2022.859067 (2022).
2. Spontaneous Intracerebral Hemorrhage. *N. Engl. J. Med.* **388**, 1440. https://doi.org/10.1056/NEJMx230001 (2023).
3. Ali, M., Lyden, P., Sacco, R. L., Shuaib, A. & Lees, K. R. Natural history of complications after intracerebral haemorrhage. *Eur. J. Neurol.* **16**, 624–630. https://doi.org/10.1111/j.1468-1331.2009.02559.x (2009).
4. Lord, A. S. et al. Infection after intracerebral hemorrhage: risk factors and association with outcomes in the Ethnic/racial variations of intracerebral hemorrhage study. *Stroke* **45**, 3535–3542. https://doi.org/10.1161/strokeaha.114.006435 (2014).
5. Berger, B., Gumbinger, C., Steiner, T. & Sykora, M. Epidemiologic features, risk factors, and outcome of sepsis in stroke patients treated on a neurologic intensive care unit. *J. Crit. Care.* **29**, 241–248. https://doi.org/10.1016/j.jcrc.2013.11.001 (2014).
6. Cheng, Y. et al. Evaluation of intestinal injury, inflammatory response and oxidative stress following intracerebral hemorrhage in mice. *Int. J. Mol. Med.* **42**, 2120–2128. https://doi.org/10.3892/ijmm.2018.3755 (2018).
7. Lin, J. et al. Impact and risk factors of sepsis on long-term outcomes after spontaneous intracerebral hemorrhage. *Chin. Med. J. (Engl).* **135**, 1006–1008. https://doi.org/10.1097/cm9.0000000000001954 (2022).
8. Adam, N., Kandelman, S., Mantz, J., Chrétien, F. & Sharshar, T. Sepsis-induced brain dysfunction. *Expert Rev. Anti Infect. Ther.* **11**, 211–221. https://doi.org/10.1586/eri.12.159 (2013).
9. Gofton, T. E. & Young, G. B. Sepsis-associated encephalopathy. *Nat. Rev. Neurol.* **8**, 557–566. https://doi.org/10.1038/nrneurol.2012.183 (2012).
10. Wang, G. et al. [Sepsis associated encephalopathy is an independently risk factor for nosocomial coma in patients with supratentorial intracerebral hemorrhage: a retrospective cohort study of 261 patients]. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue.* **28**, 723–728. https://doi.org/10.3760/cma.j.issn.2095-4352.2016.08.011 (2016).
11. Singer, M. et al. The third international consensus definitions for Sepsis and septic shock (Sepsis-3). *Jama* **315**, 801–810. https://doi.org/10.1001/jama.2016.0287 (2016).
12. Cecconi, M., Evans, L., Levy, M. & Rhodes, A. Sepsis and septic shock. *Lancet* **392**, 75–87. https://doi.org/10.1016/s0140-6736(18)30696-2 (2018).
13. Angus, D. C. et al. Epidemiology of severe sepsis in the united States: analysis of incidence, outcome, and associated costs of care. *Crit. Care Med.* **29**, 1303–1310. https://doi.org/10.1097/00003246-200107000-00002 (2001).
14. Martin, G. S., Mannino, D. M., Eaton, S. & Moss, M. The epidemiology of sepsis in the united States from 1979 through 2000. *N. Engl. J. Med.* **348**, 1546–1554. https://doi.org/10.1056/NEJMoa022139 (2003).
15. Thompson, K., Venkatesh, B. & Finfer, S. Sepsis and septic shock: current approaches to management. *Intern. Med. J.* **49**, 160–170. https://doi.org/10.1111/imj.14199 (2019).
16. Lin, J., Tan, B., Li, Y., Feng, H. & Chen, Y. Sepsis-Exacerbated brain dysfunction after intracerebral hemorrhage. *Front. Cell. Neurosci.* **15**, 819182. https://doi.org/10.3389/fncel.2021.819182 (2021).
17. Hou, N. et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J. Transl. Med.* **18**, 462. https://doi.org/10.1186/s12967-020-02620-5 (2020).
18. Du, M., Haag, D. G., Lynch, J. W. & Mittinty, M. N. Comparison of the Tree-Based machine learning algorithms to Cox regression in predicting the survival of oral and pharyngeal cancers: analyses based on SEER database. *Cancers (Basel).* **12** https://doi.org/10.3390/cancers12102802 (2020).
19. Yu, J. R. et al. Energy efficiency of inference algorithms for clinical laboratory data sets: green artificial intelligence study. *J. Med. Internet Res.* **24**, e28036. https://doi.org/10.2196/28036 (2022).
20. R, C. & A, N.-M. in *23rd International Conference on Machine Learning*. Pittsburgh, PA, (2006).
21. Zhang, Y. et al. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *Biomed. Eng.* **16**, 125. https://doi.org/10.1186/s12938-017-0416-x (2017).
22. Harper, P. R. A review and comparison of classification algorithms for medical decision making. *Health Policy.* **71**, 315–331. https://doi.org/10.1016/j.healthpol.2004.05.002 (2005).
23. Murphy, K. P. *Machine Learning: a Probabilistic Perspective* (MIT Press, 2012).
24. Kam, H. J. & Kim, H. Y. Learning representations for the early detection of sepsis with deep neural networks. *Comput. Biol. Med.* **89**, 248–255. https://doi.org/10.1016/j.compbiomed.2017.08.015 (2017).
25. Scherpf, M., Gräßer, F., Malberg, H. & Zaunseder, S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput. Biol. Med.* **113**, 103395. https://doi.org/10.1016/j.compbiomed.2019.103395 (2019).
26. Van Steenkiste, T. et al. Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks. *Artif. Intell. Med.* **97**, 38–43. https://doi.org/10.1016/j.artmed.2018.10.008 (2019).
27. Aşuroğlu, T. & Oğul, H. A deep learning approach for sepsis monitoring via severity score Estimation. *Comput. Methods Progr. Biomed.* **198**, 105816. https://doi.org/10.1016/j.cmpb.2020.105816 (2021).
28. Mathew, G. et al. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. *Int. J. Surg.* **96**, 106165. https://doi.org/10.1016/j.ijsu.2021.106165 (2021).
29. Yacouby, R. & Axman, D. *Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models*. (2020).

30. Fu, Y., Liu, Q., Anrather, J. & Shi, F. D. Immune interventions in stroke. *Nat. Rev. Neurol.* **11**, 524–535. https://doi.org/10.1038/nrneurol.2015.144 (2015).
31. Westendorp, W. F., Nederkoorn, P. J., Vermeij, J. D., Dijkgraaf, M. G. & van de Beek, D. Post-stroke infection: a systematic review and meta-analysis. *BMC Neurol.* **11**, 110. https://doi.org/10.1186/1471-2377-11-110 (2011).
32. Singer, B. H. et al. Bacterial dissemination to the brain in Sepsis. *Am. J. Respir. Crit. Care Med.* **197**, 747–756. https://doi.org/10.1164/rccm.201708-1559OC (2018).
33. Kong, Y. & Le, Y. Toll-like receptors in inflammation of the central nervous system. *Int. Immunopharmacol.* **11**, 1407–1414. https://doi.org/10.1016/j.intimp.2011.04.025 (2011).
34. Kodali, M. C., Chen, H. & Liao, F. F. Temporal unsnarling of brain's acute neuroinflammatory transcriptional profiles reveals panendothelitis as the earliest event preceding microgliosis. *Mol. Psychiatry*. **26**, 3905–3919. https://doi.org/10.1038/s41380-020-00955-5 (2021).
35. Evans, L. et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensiv. Care Med.* **47**, 1181–1247. https://doi.org/10.1007/s00134-021-06506-y (2021).
36. Bellot, P., Francés, R. & Such, J. Pathological bacterial translocation in cirrhosis: pathophysiology, diagnosis and clinical implications. *Liver Int.* **33**, 31–39. https://doi.org/10.1111/liv.12021 (2013).
37. Bartoletti, M. et al. Epidemiology and outcomes of bloodstream infection in patients with cirrhosis. *J. Hepatol.* **61**, 51–58. https://doi.org/10.1016/j.jhep.2014.03.021 (2014).
38. Gustot, T. et al. Impact of infection on the prognosis of critically ill cirrhotic patients: results from a large worldwide study. *Liver Int.* **34**, 1496–1503. https://doi.org/10.1111/liv.12520 (2014).
39. Schrier, R. W. & Wang, W. Acute renal failure and sepsis. *N. Engl. J. Med.* **351**, 159–169. https://doi.org/10.1056/NEJMra032401 (2004).
40. Takasu, O. et al. Mechanisms of cardiac and renal dysfunction in patients dying of sepsis. *Am. J. Respir. Crit. Care Med.* **187**, 509–517. https://doi.org/10.1164/rccm.201211-1983OC (2013).
41. Ergin, B., Kapucu, A., Demirci-Tansel, C. & Ince, C. The renal microcirculation in sepsis. *Nephrol. Dial. Transpl.* **30**, 169–177. https://doi.org/10.1093/ndt/gfu105 (2015).
42. Schor, N. Acute renal failure and the sepsis syndrome. *Kidney Int.* **61**, 764–776. https://doi.org/10.1046/j.1523-1755.2002.00178.x (2002).
43. Laupland, K. B., Pasquill, K., Steele, L. & Parfitt, E. C. Burden of bloodstream infection in older persons: a population-based study. *BMC Geriatr.* **21**, 31. https://doi.org/10.1186/s12877-020-01984-z (2021).
44. Leng, Y. et al. Sepsis as an independent risk factor in atrial fibrillation and cardioembolic stroke. *Front. Endocrinol. (Lausanne)*. **14**, 1056274. https://doi.org/10.3389/fendo.2023.1056274 (2023).
45. Walker, A. M. N. et al. Prevalence and predictors of Sepsis death in patients with chronic heart failure and reduced left ventricular ejection fraction. *J. Am. Heart Assoc.* **7**, e009684. https://doi.org/10.1161/jaha.118.009684 (2018).
46. Shao, I. Y., Elkind, M. S. V. & Boehme, A. K. Risk factors for stroke in patients with Sepsis and bloodstream infections. *Stroke* **50**, 1046–1051. https://doi.org/10.1161/strokeaha.118.023443 (2019).
47. Zhang, H. et al. Risks and features of secondary infections in severe and critical ill COVID-19 patients. *Emerg. Microbes Infect.* **9**, 1958–1964. https://doi.org/10.1080/22221751.2020.1812437 (2020).
48. Mohamed, A. K. S., Mehta, A. A. & James, P. Predictors of mortality of severe sepsis among adult patients in the medical intensive care unit. *Lung India.* **34**, 330–335. https://doi.org/10.4103/lungindia.lungindia_54_16 (2017).
49. Hannawi, Y., Hannawi, B., Rao, C. P., Suarez, J. I. & Bershad, E. M. Stroke-associated pneumonia: major advances and Obstacles. *Cerebrovasc. Dis.* **35**, 430–443. https://doi.org/10.1159/000350199 (2013).
50. Koennecke, H. C. et al. Factors influencing in-hospital mortality and morbidity in patients treated on a stroke unit. *Neurology* **77**, 965–972. https://doi.org/10.1212/WNL.0b013e31822dc795 (2011).
51. Teh, W. H. et al. Impact of stroke-associated pneumonia on mortality, length of hospitalization, and functional outcome. *Acta Neurol. Scand.* **138**, 293–300. https://doi.org/10.1111/ane.12956 (2018).
52. Wilson, R. D. Mortality and cost of pneumonia after stroke for different risk groups. *J. Stroke Cerebrovasc. Dis.* **21**, 61–67. https://doi.org/10.1016/j.jstrokecerebrovasdis.2010.05.002 (2012).
53. Qureshi, A. I. et al. Early hyperchloremia is independently associated with death or disability in patients with intracerebral hemorrhage. *Neurocrit Care*. **37**, 487–496. https://doi.org/10.1007/s12028-022-01514-2 (2022).
54. Ditch, K. L., Flahive, J. M., West, A. M., Osgood, M. L. & Muehlschlegel, S. Hyperchloremia, not concomitant hypernatremia, independently predicts early mortality in critically ill Moderate-Severe traumatic brain injury patients. *Neurocrit. Care*. **33**, 533–541. https://doi.org/10.1007/s12028-020-00928-0 (2020).
55. Zhou, D. et al. Increase in chloride from baseline is independently associated with mortality in intracerebral hemorrhage patients admitted to intensive care unit: A retrospective study. *J. Intensive Med.* **2**, 274–281. https://doi.org/10.1016/j.jointm.2022.04.002 (2022).

Acknowledgements

The authors gratefully acknowledge the MIMIC-IV and eICU-CRD programs for open access to their database.

Author contributions

LT and LZ conceived and designed the study. YL, LT, JZ and QT collected the primary data. FZ, SM, and RL performed data analysis. XZ (Xiangbin Zhang) and SW developed analytical tools. YZ (Yupeng Zhang), LC, and JM contributed to data interpretation. XZ (Xuelun Zou), TY, and RT created visualizations and figures. YY and YZ (Yi Zeng) validated the methodology and findings. YL drafted the manuscript with input from all authors. LZ and DW supervised the study and critically revised the manuscript. All authors read and approved the final version of the manuscript.

Funding

This study was supported by The National Science & Technology Fundamental Resources Investigation Program of China to L.Z. (No.2018FY100900), Central South University Research Programme of Advanced Interdisciplinary Projects in Changsha Studies to L.Z. (No.2023QYJC011) , The Hunan Provincial Natural Science Foundation of China Grant to L.Z. (No.2023JJ60144) and Y.Z. (No.2021JJ30923), The Provincial Science and Technology Innovation Leading Talents Project to L.Z.(No.2021RC4014). Major Science and Technology Projects in Changsha to L.Z. (No.kq2301008), The Changsha Municipal Natural Science Foundation to L.T. (kq2202022), The Hunan Provincial Natural Science Foundation to L.T. (2023JJ60383), Hunan Provincial Health High-Level Talent Scientific Research Project to L.Z. (No.R2023069), The Major Basic Research Projects in Hunan Province to L.Z. (No.2024JC0004), The National Natural Science Foundation of China to L.Z. (No.82471364).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-99431-9.

Correspondence and requests for materials should be addressed to D.W. or L.Z.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.