A NEW TIME-WINDOW PREDICTION MODEL FOR TRAUMATIC HEMORRHAGIC SHOCK BASED ON INTERPRETABLE MACHINE LEARNING

Yuzhuo Zhao,[*] **Lijing Jia,**[*] **Ruiqi Jia,**[†] **Hui Han,**[*] **Cong Feng,**[*] **Xueyan Li,**[‡]
Zijian Wei,[§] **Hongxin Wang,**[||] **Heng Zhang,**[*] **Shuxiao Pan,**[¶] **Jiaming Wang,**[†]
Xin Guo,[**] **Zheyuan Yu,**[†] **Xiucheng Li,**[†] **Zhaohong Wang,**[†] **Wei Chen,**[††‡‡] **Jing Li,**[†]
and Tanshi Li[*]

[*]Department of Emergency, The First Medical Center of Chinese PLA General Hospital, Beijing, China;
[†]School of Economics and Management, Beijing Jiaotong University, Beijing, China; [‡]Management
School, Beijing Union University, Beijing, China; [§]Washington University in St. Louis; [||]Department of
Emergency, Armed Police Characteristic Medical Center, Tianjin, China; [¶]College of Computer Science
and Artificial Intelligence, Wenzhou University; [**]Beijing Friendship Hospital, Capital Medical University,
Beijing, China; [††]Department of Emergency, The Third Medical Center of Chinese PLA General Hospital,
Beijing, China; and [‡‡]Hainan Hospital of Chinese PLA General Hospital, Sanya, China

ABSTRACT—Early warning prediction of traumatic hemorrhagic shock (THS) can greatly reduce patient mortality and
morbidity. We aimed to develop and validate models with different stepped feature sets to predict THS in advance. From the
PLA General Hospital Emergency Rescue Database and Medical Information Mart for Intensive Care III, we identified 604
and 1,614 patients, respectively. Two popular machine learning algorithms (i.e., extreme gradient boosting [XGBoost] and
logistic regression) were applied. The area under the receiver operating characteristic curve (AUROC) was used to evaluate
the performance of the models. By analyzing the feature importance based on XGBoost, we found that features in vital signs
(VS), routine blood (RB), and blood gas analysis (BG) were the most relevant to THS (0.292, 0.249, and 0.225, respectively).
Thus, the stepped relationships existing in them were revealed. Furthermore, the three stepped feature sets (i.e., VS,
VS + RB, and VS + RB + sBG) were passed to the two machine learning algorithms to predict THS in the subsequent
T hours (where $T = 3, 2, 1,$ or 0.5), respectively. Results showed that the XGBoost model performance was significantly
better than the logistic regression. The model using vital signs alone achieved good performance at the half-hour time
window (AUROC = 0.935), and the performance was increased when laboratory results were added, especially when
the time window was 1 h (AUROC = 0.950 and 0.968, respectively). These good-performing interpretable models demon-
strated acceptable generalization ability in external validation, which could flexibly and rollingly predict THS T hours (where
$T = 0.5, 1$) prior to clinical recognition. A prospective study is necessary to determine the clinical utility of the proposed THS
prediction models.

KEYWORDS—Interpretability, machine learning, prediction window, shock index, time series, traumatic hemorrhagic shock

ABBREVIATIONS—95% CI—95% confidence interval; AUPRC—area under the precision-recall curve; AUROC—area
under the receiver operating characteristic curve; BE—base excess; BP—blood pressure; DBP—diastolic blood pressure;
EMR—electronic medical record; Hb—hemoglobin; Hct—hematocrit; HR—heart rate; IQR—interquartile range; Lac—
lactate; LOS—length of stay; MBP—mean blood pressure; MIMIC III—the Medical Information Mart for Intensive Care III;
$PaCO_2$—partial pressure of carbon dioxide; PaO_2—partial pressure of oxygen; PLAGH-ERD—the PLA General Hospital
Emergency Rescue Database; PLT—platelets; RESP—respiration rate; SBP—systolic blood pressure; SHAP—Shapley
additive explanation; SI—shock index; TCO_2—total carbon dioxide; TEMP—temperature; THS—traumatic hemorrhagic
shock; WBC—white blood cell count

INTRODUCTION

Traumatic hemorrhagic shock (THS) is a type of hypovolemic shock caused by severe trauma and is one of the main causes of death in severely injured patients (1–3). Compared with acute massive hemorrhage, the detection of THS is often delayed by occult bleeding. If THS could be detected at an early stage and even in advance, timely and effective interventions could be implemented, which could greatly reduce patient mortality and morbidity and improve the outcome of severe trauma (4).

Despite substantial advances in the management of hemorrhage (5), it remains the primary cause of preventable death (40% of trauma-related fatalities) (6–8). This is not only related to the characteristics of traumatic injury but also partly to the limited ability of humans to process information. Machine learning techniques excel in the analysis of complex signals in data-rich environments (9), and thus provide a powerful tool for early warning prediction of THS. A variety of signals are good indicators for THS, including increased shock index, decreased blood pressure, decreased hemoglobin, or blood transfusion in a short period of time, which provides a medical basis for the early warning prediction of THS. Currently, although there have been many clinical decision support studies on trauma patients worldwide (8), most of these investigations considered survival/death as the endpoint to make predictions (10–13), with little attention paid to the early warning prediction of THS (14, 15).

In this work, we developed a new approach to predict THS in advance on the basis of medical knowledge, data analysis, and start-of-the-art ML techniques, thus making it possible for clinicians to proactively prepare the necessary treatment resources. To improve the accuracy and speed of THS prediction, the electronic medical records of injured patients were analyzed to determine the important relationships between different types of indices and THS. In combination with the guidance of medical knowledge and experience, we analyzed the contributions of the indicators to prediction accuracy, and we grouped the indicator set into different indicator combinations to adapt to different scenarios.

METHODS

Data sources and study population

The study population was from the PLA General Hospital Emergency Rescue Database (PLAGH-ERD) (16) and the Medical Information Mart for Intensive Care III (MIMIC III) (17). The PLAGH-ERD has the depersonalized information of 22,941 patients from 2014 to 2018. This database was one of the first special databases in the field of first aid with independent intellectual property rights in China. Its advanced nature is representative of the first aid field in China and recognized by peers both in China and abroad. The MIMIC III is a public dataset, including non-private medical records of more than 50,000 patients at Beth Israel Deaconess Medical Center (BID) in Boston, Massachusetts, USA from 2001 to 2012 (18–20).

This research was aimed at patients with traumatic hemorrhagic shock. Studies have shown that the shock index and mean arterial pressure are commonly considered as good indicators to assess the severity of shock in a clinical setting (21–26). First, all adult patients aged 18 years or older who were admitted to the hospital due to trauma were included from the PLAGH-ERD and the MIMIC III. Second, shock was defined as simultaneous shock index ≥ 1 and mean arterial pressure ≤ 70. To accurately identify THS patients, traumatic hemorrhagic shock was defined as meeting at least one of the following conditions in addition to shock:

1) blood transfusion,
2) hemoglobin ≤ 90 g/L at admission (if no pre-existing cause of chronic anemia was present, including malignant tumors, hematological diseases, chronic kidney disease, and chronic liver disease), or
3) a hemoglobin decrease by 20% (baseline values were those at admission) (27–29).

Patients with one of the following conditions were excluded:

1) age < 18 years,
2) hospital admission not for trauma,
3) had one or more surgical treatment records,
4) suffered from septic shock, cardiogenic shock, or anaphylactic shock according to the ICD diagnosis codes, antibiotics, and blood cultures, or
5) died before shock or at discharge.

We developed Oracle SQL scripts to query the research cohort.

The study was approved by the Research Ethics Commission of the PLA General Hospital (S2020-129-01), and the requirement for informed consent was waived by the Ethics Commission.

Study variables and processing

Demographic characteristics such as age and sex were collected. Vital signs, such as blood pressure, heart rate, respiratory rate, and temperature, were included. In total, 39 laboratory measures, including routine blood, blood gas analysis, blood biochemistry, coagulation function, and routine urine, were collected (Supplemental Digital Content 1, http://links.lww.com/SHK/B357).

Continuous measurements were recorded every few seconds in PLAGH-ERD; therefore, the dataset contained more observations at a higher temporal resolution than MIMIC-III. The data was resampled to a 30-min resolution. If the index data contained multiple values within 0.5 h, the median value was taken. Cluster imputations were applied through Python to impute the missing data (Fig. 1B1). Considering the low resolution of laboratory measures, the cross-sectional data of the latest observations T hours ($T = 1, 2, 3$ h) before THS (or discharge) were collected. The missing data in laboratory measures were processed using multivariate imputation via chained equations implemented by the R *mice* package (30, 31). The features in this work with at least 50% data completeness were considered as predictors and were used for model establishment (32) (Fig. 1B2).

Outcome

The onset of THS during hospitalization was taken as the outcome. In the case of multiple times of occurrence, the first time was considered in the THS group. If THS did not occur, patients were classified as the non-THS group.

Statistical analysis

Differences in age, gender, body trauma sites, and lengths of hospital stay between THS and non-THS groups were analyzed using SPSS version 22.0. For continuous variables, the Mann–Whitney U test was utilized to compare the differences between the two groups. For binary variables, the chi-square test was employed for statistical analysis. All P values were two-sided, and values below 0.05 were considered significant.

Extreme gradient boosting (XGBoost) is derived from the gradient boosting decision tree and was proposed by Chen et al. (33). Several weak classifiers (i.e., decision tree) are transformed into a strong classifier to improve performance, which is achieved by an iterative computation of weak classifiers. Regularization is used to control the complexity of the tree to obtain a simpler model and avoid overfitting (20). Moreover, the algorithm ranks the importance of candidate predictors to reflect the contribution of each variable in classifying the THS versus the non-THS groups. Thus, individual predictions in XGBoost can be represented by decomposing a decision path into one component per feature. In this way, a decision can be tracked through the tree and used to explain a prediction through the contributions added in each decision node. In this work, a XGBoost model was employed to select variables predictive of THS using last cross-sectional data before shock. The feature importance was calculated and analyzed to determine the relationship between different feature types and THS, thereby providing a theoretical basis for the classification of stepped features to adapt to different scenarios (both pre-hospital and in-hospital). Furthermore, to enhance the
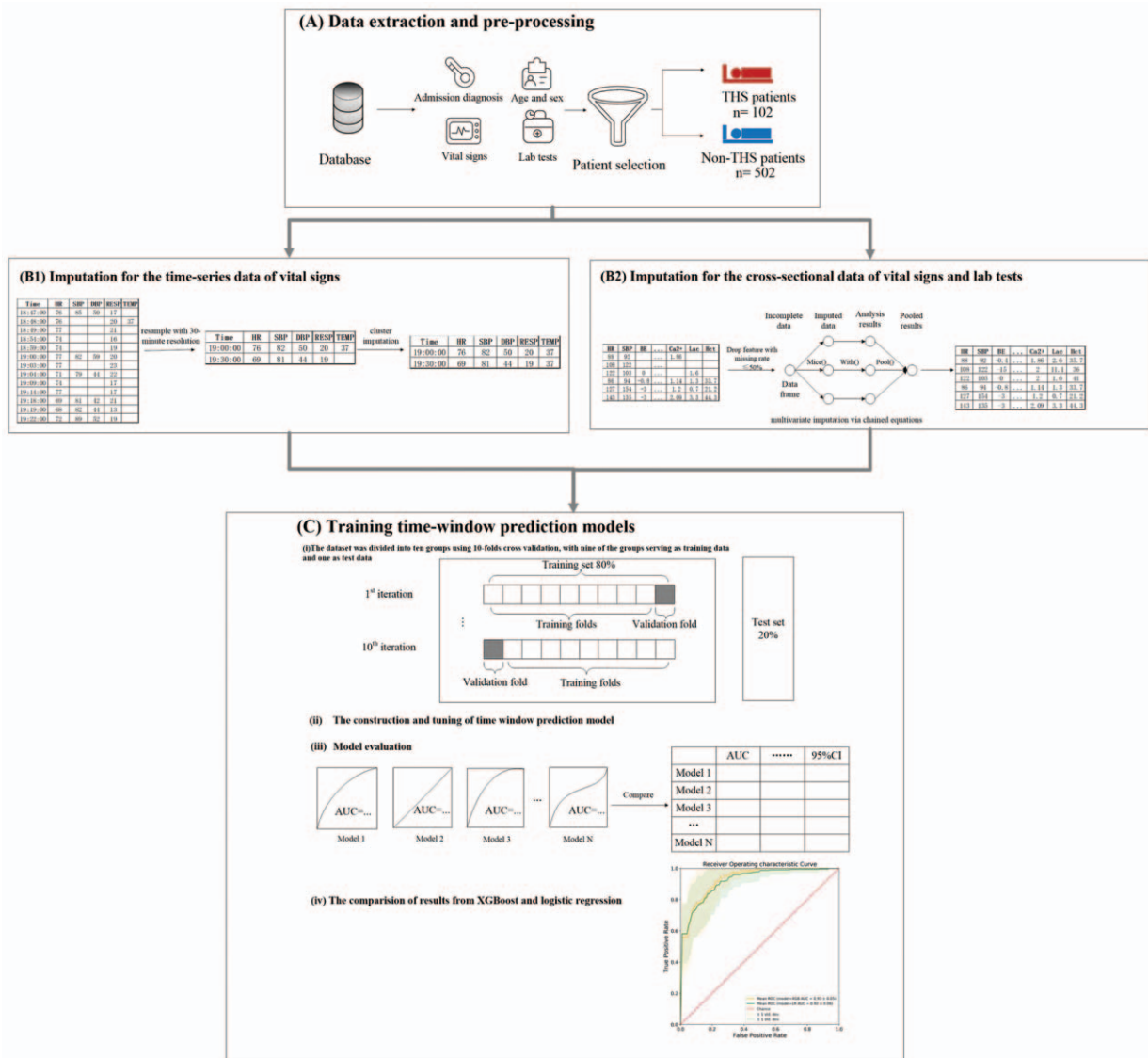
Fig. 1.   **Model development overview.** (A) Data extraction and processing. Data including admission diagnosis, demographic information (e.g., age and sex), vital signs, and laboratory results were extracted from PLAGH-ERD. Patients were divided into THS and non-THS groups. (B1) Imputation for the time-series data of vital signs based on cluster. (B2) Imputation for the time-series data of vital signs based on multivariate imputation via chained equations. Features with missing rates greater than 50% were removed. (C) Feature importance was calculated based on the average gain of XGBoost to analyze the relationship between the features and THS. (D) Training time-window prediction models. (i) The data set was divided into 10 groups using 10-fold cross-validation, with nine of the groups serving as training data and one as test data. (ii) The construction and tuning of time-window prediction. (iii) Evaluation. The AUROC, AUPRC, $F_{1.5}$, precision, recall, accuracy, and 95% confidence interval (CI) values were utilized to evaluate the performance of each model for different stepped feature sets and time windows. (iv) Comparison of results from XGBoost and logistic regression. AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; THS, traumatic hemorrhagic shock; XGBoost, extreme gradient boosting.

interpretability of the results, SHAP (SHapley Additive exPlanations) values were used to illustrate the effect of each feature on the classifier output: positive and negative SHAP values indicated an increase or decrease in the prediction score, respectively (34).

The PLAGH-ERD and MIMIC III were adopted to develop and validate the time-window prediction models, respectively. The negative samples were randomly partitioned into 10 equal-sized subsamples to overcome the problem of class imbalances. In the training set (80% randomly selected samples from the PLAGH-ERD), XGBoost, and logistic regression models with different time windows ($T = 0.5, 1, 2, 3\,h$) were implemented by the *xgboost* and *scikit-learn* packages in Python 3.6. For the XGBoost models, the best hyperparameters consisting of eta (step size shrinkage used in the update process to prevent overfitting), the maximum depth of a tree, and subsample (the proportion of the subsamples used for training the model to the whole sample set) were determined by grid search with 10-fold cross-validation. For LR models, Lasso regularization was applied to prevent overfitting. In the test set (the remaining 20% of the sample from the PLAGH-ERD), we computed the prediction

performance of each model derived above. Additionally, to validate the generalization ability of the model, MIMIC III was used for external validation. In terms of model evaluation, the following indicators were used: accuracy, recall, precision, $F_{1.5}$, the area under receiver operating characteristic (AUROC) curves, and precision-recall curves (AUPRC). We compared AUROCs using DeLong's test for the THS prediction models (35–37), which were performed using R package pROC, version 1.17.0.1. A two-sided 0.05 significance level was applied to general comparisons. All outcomes were compared and analyzed to select the models with the best performance (Fig. 1C).

RESULTS

Patient characteristics

In total, 604 patients from the PLAGH-ERD were obtained, 102 of whom developed THS (Fig. 2A). The proportion of male
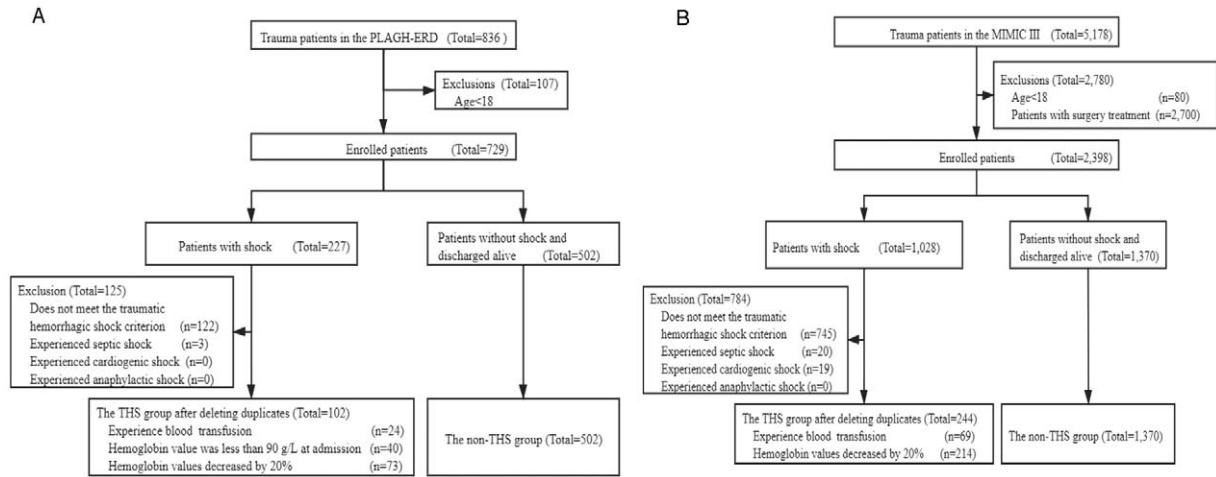
A

B

Fig. 2.   **Panel (A) shows the extraction process for the study cohort in the PLAGH-ERD.** Panel (B) shows the extraction process for the study cohort in the MIMIC III. MIMIC III, the Medical Information Mart for Intensive Care III.

trauma patients was greater than that of female patients (Table 1). The median time interval from admission to shock was 4.91 h. Compared with patients of non-THS, patients with THS had longer median lengths of hospital stay (1.03 vs. 0.38 days). The proportions of trauma sites in the abdomen (21.6% vs. 12.7%), pelvis (7.8% vs. 5.2%), and limbs (5.9% vs. 3.0%) were also higher. There were 1,614 patients in the MIMIC III that exhibited patterns similar to those of the PLAGH-ERD (Fig. 2B).

Relative sequence importance of features and analyses

After excluding the features with serious data loss, 27 features were obtained from both the PLAGH-ERD and the MIMIC III. The ranks of feature importance of these features were output (Supplemental Digital Content 2, http://links.lww.com/SHK/B357, Table 1). Vital signs accounted for the largest proportion of feature importance (0.292), followed by routine blood (0.249), blood gas analysis (0.225), blood biochemistry (0.147), and coagulation function (0.087).

Among them, the total feature importance of the top three types, which are easily obtained in a clinical setting, reached 76% (Supplemental Digital Content 2, http://links.lww.com/SHK/B357, Table 2).

Therefore, these three feature types were grouped into three stepped feature sets: vital signs alone, vital signs + routine blood, and vital signs + routine blood + blood gas analysis (Table 2). As vital signs are high-resolution indicators with strong timeliness and rapidity, and they are easy to obtain even in pre-hospital or other rough conditions, the time-series data of vital signs were used for short time-window prediction ($T = 0.5$, 1, 2 h). Considering the low resolution of routine blood and blood gas analysis, the cross-sectional data were used for relatively longer window prediction ($T = 1$, 2, 3 h).

To identify how a single feature influenced the outcome of a prediction model, we depicted the SHAP dependence plot of XGBoost (Fig. 3). The y-axis values indicated the SHAP values of the features, and the values of features for the x-axis were in the SHAP dependence plot. We visualized how the features'

Table 1.   **Baseline statistical characteristics of the study population**

Characteristics	The PLAGH-ERD		The MIMIC III	
	THS (n = 102)	Non-THS (n = 502)	THS (n = 244)	Non-THS (n = 1,370)
Age (years; median, IQR)	43.98 (31.92−58.87)	49.41* (35.56−63.01)	52.10 (36.13−72.16)	50.11* (31.61−73.36)
Gender, n (%)				
Female	20 (19.6)	110 (21.7)	82 (33.6)	438 (32.0)
Male	82 (80.4)	396 (78.3)	162 (66.4)	932 (68.0)
Injured body part, n (%)				
Head	32 (31.4)	239 (47.6)	156 (63.93)	897 (65.5)
Chest	12 (11.8)	63 (12.5)	72 (29.51)	241 (17.6)
Abdomen	22 (21.6)	64 (12.7)	55 (22.54)	254 (18.5)
Pelvis	8 (7.8)	26 (5.2)	38 (15.57)	78 (5.7)
Limbs	6 (5.9)	15 (3.0)	60 (24.59)	204 (14.9)
Other	64 (63.7)	164 (32.7)	19 (7.79)	152 (11.1)
Hospital LOS (days; median, IQR)	1.03 (0.44−1.90)	0.38* (0.18−0.82)	11.01 (5.77−20.18)	4.54* (2.58−7.75)
Time interval from admission to shock (h; median, IQR)	4.91 (1.97−12.14)	−	13.35 (4.66−50.84)	−

IQR, interquartile range; LOS, length of stay; THS, traumatic hemorrhagic shock.
*Statistically significant difference between the experimental and control groups.

TABLE 2. The stepped feature sets used for prediction

Forecast indicator dataset – 1 (vital signs)	Forecast indicator dataset – 2 (vital signs + routine blood)	Forecast indicator dataset – 3 (vital signs + routine blood + blood gas analysis)
HR	HR	HR
SBP	SBP	SBP
DBP	DBP	DBP
RESP	RESP	RESP
TEMP	TEMP	TEMP
	PLT	PLT
	WBC	WBC
	Hb	Hb
	RBC	RBC
	Hct	Hct
		BE
		Lac
		pH
		TCO$_2$
		PaCO$_2$
		PaO$_2$

BE, base excess; DBP, diastolic blood pressure; Hb, hemoglobin; Hct, hematocrit; HR, heart rate; Lac, lactate; PaO$_2$, partial pressure of oxygen; PLT, platelets; RESP, respiration rate; SBP, systolic blood pressure; TEMP, temperature; WBC, white blood cell count.

attributed importance changed as the values varied in the plot. SHAP values for specific features exceeding zero indicate an increased risk of THS.

Development of time-window prediction models of THS

We used the scikit-learn implementations of machine learning models to predict THS and optimized the parameters of each model by grid search with 10-fold cross-validation. The hyperparameters of XGBoost are listed in Supplemental Digital Content 2, http://links.lww.com/SHK/B357, Table 3. Each prediction model achieved good performance on different stepped feature sets and prediction windows, and the performance of XGBoost was significantly better than logistic regression (Table 4). For time-series data of vital signs when the timestep was 3, each model achieved good performance on each prediction window ($T = 0.5$, 1, 2 h) (see Table 3 and Fig. 4). When the time window was 0.5 h, the $F_{1.5}$ score of the THS prediction model based on XGBoost was up to 0.849, and the AUROC value was up to 0.935 (95% confidence interval [95% CI]: 0.911–0.959). Furthermore, the generalization ability of our model was verified by external validation. The $F_{1.5}$ score was 0.704, and the AUROC value was 0.785 (95% CI: 0.769–0.801).
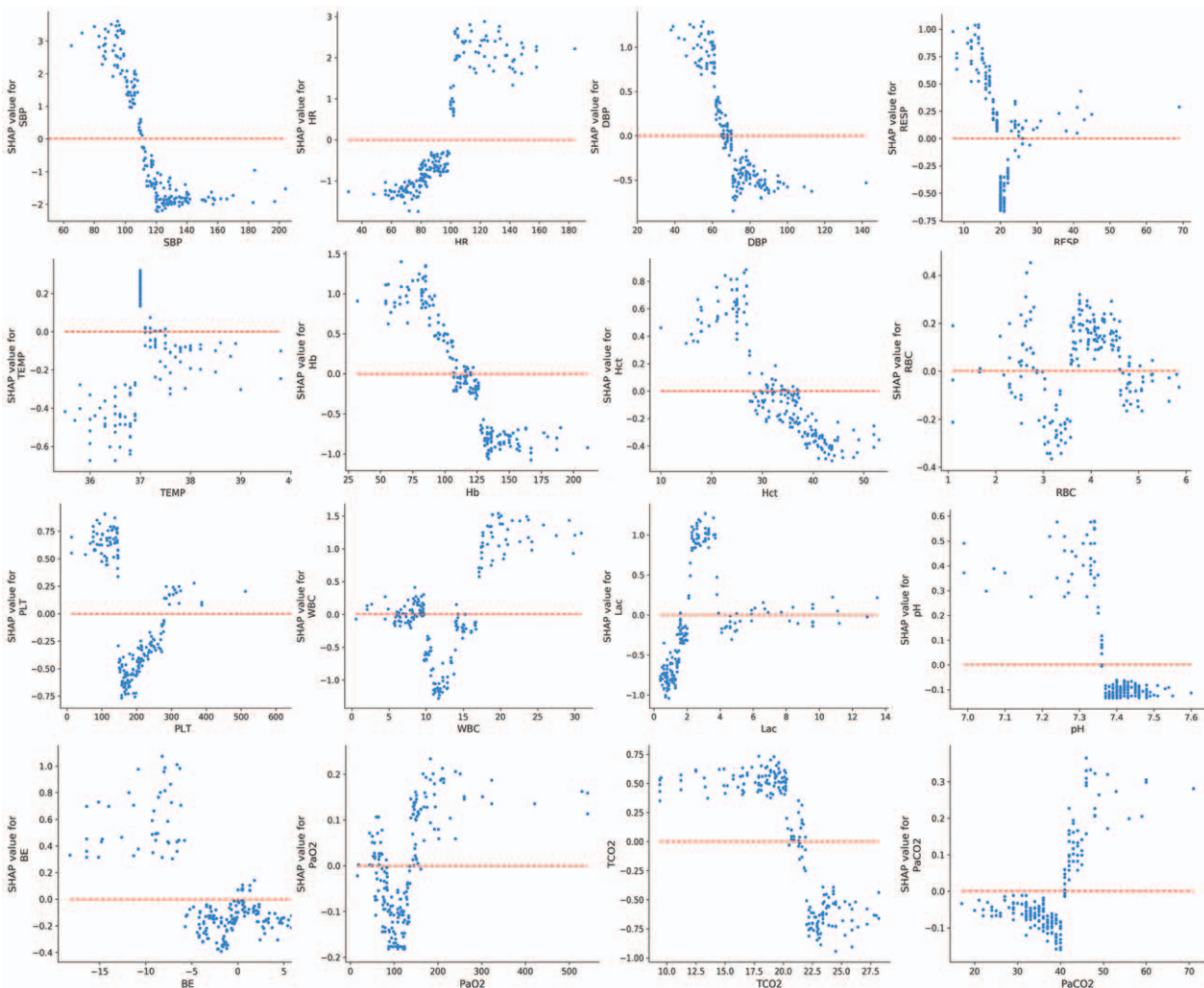

FIG. 3. **Partial SHAP dependence plots for features of vital signs, routine blood, and blood gas analysis.** SHAP, SHapley Additive exPlanations.

Table 3. **Validation of time-window prediction models for traumatic hemorrhagic shock**

Machine learning model	Prediction dataset	Internal validation						External validation					
		$F_{1.5}$	acc	pre	rec	AUROC	AUPRC	$F_{1.5}$	acc	pre	rec	AUROC	AUPRC
0.5 h in advance													
XGBoost	VS	0.849	0.865	0.866	0.847	0.935	0.943	0.704	0.665	0.571	0.804	0.785	0.720
LR	VS	0.794	0.835	0.853	0.778	0.875	0.887	0.701	0.661	0.558	0.801	0.797	0.773
1 h in advance													
XGBoost	VS	0.793	0.833	0.853	0.773	0.927	0.920	0.695	0.649	0.551	0.804	0.769	0.697
	VS + RB	0.866	0.883	0.889	0.860	0.950	0.943	0.834	0.841	0.851	0.827	0.913	0.915
	VS + RB + BG	0.900	0.903	0.898	0.903	0.968	0.962	0.804	0.822	0.847	0.787	0.901	0.908
LR	VS	0.775	0.819	0.836	0.755	0.878	0.897	0.703	0.652	0.549	0.815	0.792	0.761
	VS + RB	0.863	0.875	0.867	0.863	0.889	0.883	0.834	0.846	0.863	0.822	0.916	0.928
	VS + RB + BG	0.872	0.875	0.872	0.875	0.930	0.932	0.841	0.856	0.882	0.824	0.916	0.929
2 h in advance													
XGBoost	VS	0.781	0.859	0.873	0.760	0.937	0.905	0.679	0.653	0.558	0.777	0.772	0.699
	VS + RB	0.863	0.873	0.886	0.858	0.947	0.950	0.807	0.836	0.881	0.778	0.924	0.911
	VS + RB + BG	0.869	0.870	0.856	0.880	0.934	0.914	0.798	0.830	0.876	0.770	0.922	0.914
LR	VS	0.730	0.806	0.779	0.730	0.866	0.849	0.687	0.659	0.555	0.780	0.783	0.747
	VS + RB	0.835	0.847	0.875	0.828	0.905	0.898	0.835	0.846	0.862	0.823	0.909	0.924
	VS + RB + BG	0.860	0.847	0.813	0.887	0.912	0.891	0.832	0.847	0.870	0.817	0.913	0.926
3 h in advance													
XGBoost	VS + RB	0.857	0.888	0.935	0.828	0.946	0.959	0.703	0.766	0.842	0.658	0.842	0.840
	VS + RB + BG	0.863	0.873	0.886	0.858	0.957	0.950	0.807	0.836	0.881	0.778	0.924	0.911
LR	VS + RB	0.838	0.864	0.905	0.815	0.919	0.943	0.775	0.793	0.815	0.758	0.856	0.876
	VS + RB + BG	0.835	0.847	0.875	0.828	0.905	0.898	0.835	0.846	0.862	0.823	0.909	0.924

VS represents vital signs; VS + RB represents vital signs + routine blood; VS + RB + BG represents vital signs + routine blood + blood gas analysis. AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; BG, blood gas; RB, routine blood; VS, vital signs.

When controlling the same time window, routine blood and blood gas analysis data were added as prediction features along with vital signs, and the performance of the prediction models was improved. When the time window was 1 h and the cross-sectional data of the vital signs and routine blood were used, the $F_{1.5}$ score reached 0.866, and the AUROC was 0.950 (95% CI: 0.932–0.969). In external validation, the $F_{1.5}$ score was 0.834, and the AUROC value was 0.913 (95% CI: 0.905–0.922). Likewise, when using vital signs, routine blood, and blood gas analysis in internal validation, the $F_{1.5}$ score was up to 0.900, and the AUROC was 0.968 (95% CI: 0.956–0.980). In

external validation, the $F_{1.5}$ score was 0.804, and the AUROC value was 0.901 (95% CI: 0.887–0.915).

DISCUSSION

In this study, we constructed a series of good-performing prediction models based on XGBoost that could predict THS in advance using medical data combined with high-resolution time-series dynamics of vital signs and low-resolution laboratory indicators. Prediction performance was inversely proportional to the prediction window length and directly proportional to the number of features. $F_{1.5}$ scores for the XGBoost model decreased as the prediction window lengthened from 0.5 to 3 h, and the feature types decreased from 3 to 1; however, they still provided acceptable performance.

Both internal and external validations were utilized to test the reliability of the prediction models. Most early warning prediction models are only internally validated, which could result in an insufficient ability to provide new input data for practical applications (38–40). Damen pointed out that disease prediction models in the future should be externally validated (38). From the external validation results, we determined that the models established in this study had a certain generalization ability (Table 3).

This study made full use of the advantages of machine learning in the medical field. In the early 2000s, early warning prediction of diseases was pursued using entropy-based (e.g., approximate entropy) signal processing methods (41). The basic principle is to analyze the difference of nonlinear entropy between a pathological state and physiological state. In the past

Table 4. **The *P*-value of the DeLong's test compared the significant difference between performance of models with different stepped feature sets of the PLAGH-ERD and time windows**

	XGBoost	LR	*P* value
0.5 h in advance			
VS	0.935	0.875	<0.001
1 h in advance			
VS	0.927	0.878	0.028
VS + BR	0.95	0.889	<0.001
VS + BG + BR	0.968	0.93	<0.001
2 h in advance			
VS	0.937	0.866	<0.001
VS + BR	0.957	0.905	<0.001
VS + BG + BR	0.934	0.912	<0.001
3 h in advance			
VS + BR	0.946	0.919	0.016
VS + BG + BR	0.957	0.905	<0.001

BG, blood gas; PLAGH-ERD, the PLA General Hospital Emergency Rescue Database; VS, vital signs.
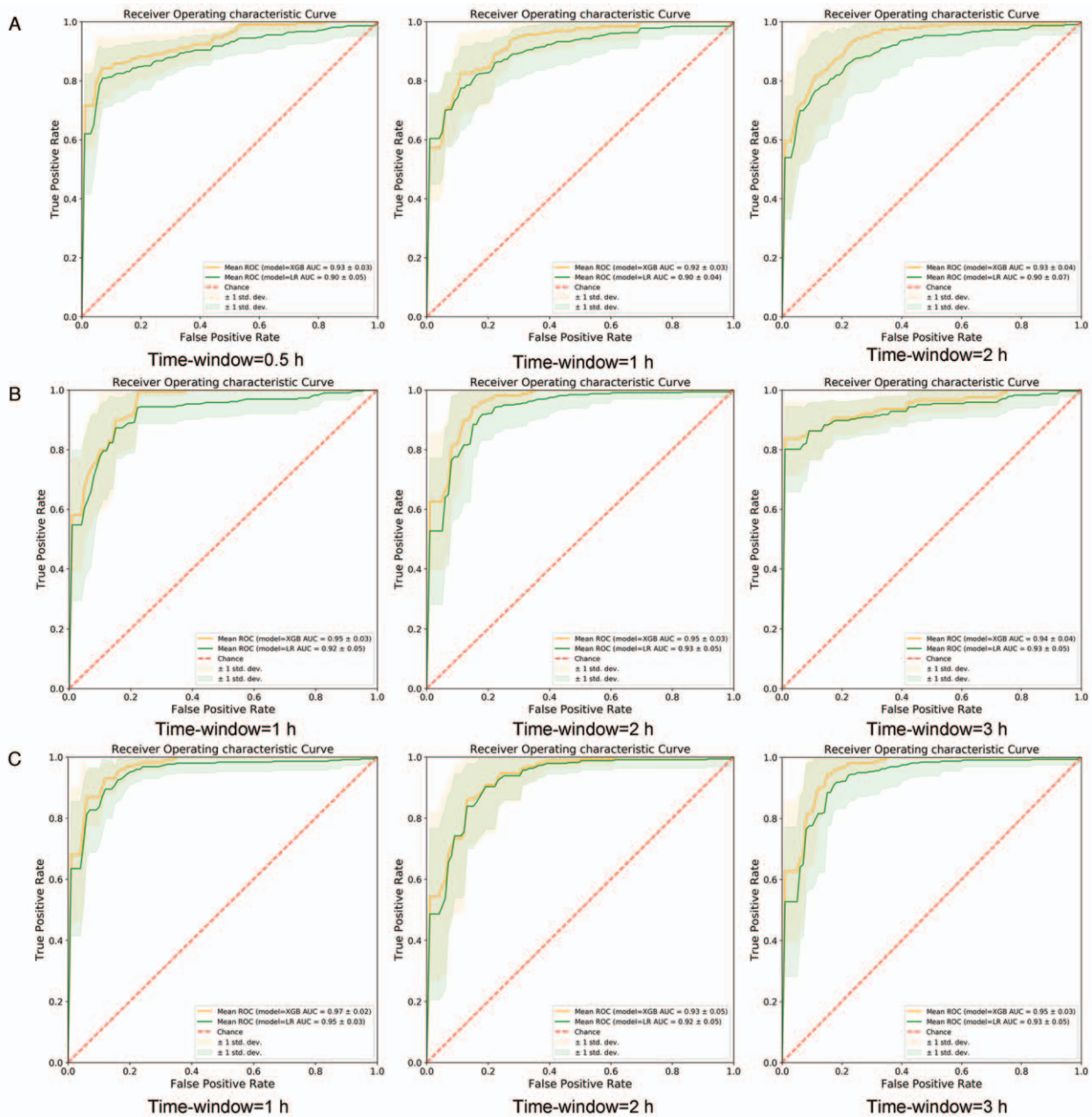
Fig. 4.   **(A) Receiver operating characteristic (ROC) curve of the prediction model for the vital signs dataset.** (B) ROC curve of the prediction model for the vital signs + routine blood dataset. (C) ROC curve of the prediction model for the vital signs + routine blood + blood gas analysis dataset.

10 years, ML techniques have made remarkable progress and can achieve unprecedented accuracy for classification tasks (42). With the help of machine learning, early warning prediction models have the potential to achieve better cross-individual recognition ability and generalization ability. Additionally, traditional statistical methods such as logistic regression are commonly employed in the medical field. Compared with traditional statistical methods, machine learning provides a powerful set of tools for describing relationships between features and the outcome(s) of interest (e.g., THS), particularly when they are nonlinear and complex (43). Moreover, machine learning has advantages in time-series prediction, especially with respect to the automated processing of time structures such

as autonomous learning and time-dependent trends. We used XGBoost, which is an interpretable algorithm, to analyze the importance of features in vital signs, routine blood, and blood gas analysis and then quantified their contribution to THS (0.292, 0.249, and 0.225, respectively). We were thus able to establish prediction models based on machine learning with good performance and generalization ability.

This study makes several significant contributions to the existing literature on THS prediction. Most similar studies used survival/death as outcomes, and few focused on the onset of THS, although early warning prediction detection analysis is common for sepsis (43–45). XGBoost, an interpretable machine learning algorithm, was applied to predict THS in

advance using vital signs. This is more effective for real-time prediction since clinicians can identify specific prediction windows in which THS could occur once time-series data have been input into the model.

Regarding input feature selection, we first introduced a concept known as "stepped features" classified in terms of practicability and timeliness. Due to the rapid progression of THS and the limitations of clinical conditions, the feasibility and timeliness of the model should be considered. At present, the vital signs acquisition equipment is developing toward miniaturization, and its performance is constantly optimized, allowing it to realize real-time dynamic acquisition in multiple scenes (e.g., in-hospital or an earlier stage). Therefore, the advantages of time-series data were fully utilized to achieve better prediction performance. Meanwhile, since a patient's vital signs are sensitive to external factors, clinicians usually run routine blood as well as blood gas analysis to supplement diagnosis. In this study, routine blood and blood gas analysis were regarded as stepped features. Additionally, our research confirmed that adding laboratory measures to vital signs increased the prediction ability of the models. Therefore, this study has flexible prediction ability of THS in-hospital even in earlier stages.

This study had some limitations. First, the research used retrospective electronic medical record data not originally collected for the analyses. Although the prediction model of THS exhibited good performance in the absence of data, the judgment of THS is a comprehensive process, and some features in the actual clinical work could not be recorded, such as pupil size, consciousness, and color of skin. Second, most of the current research based on machine learning has focused on the field of septic shock and utilized large sample sizes (43, 46, 47). In the field of THS, studies that utilized small sample sizes generally consisted of approximately 100 cases (14, 15, 48, 49). Third, our algorithm achieved good performance in both internal and external validation, and the results of the prediction models also provide certain decision support for clinical diagnosis, but prospective clinical validation is still required to determine whether the model can accurately identify THS in actual clinical scenarios. In addition, some aspects, such as data acquisition devices, data acquisition methods, and data transmission methods, need further enhancement for continuous data quality improvement and model optimization.

CONCLUSIONS

In this two-center retrospective study, we revealed that features in vital signs, routine blood, and blood gas analysis were the most relevant to THS. Thus, the stepped relationships existing in them were discovered. We confirmed that it is feasible to construct three models that can flexibly and rollingly predict THS prior to its occurrence by adopting different types of stepped features and time windows that can adapt to the scenarios of in-hospital even the earlier stage, and provide predictive accuracy and speed for actual clinical scenarios. In summary, these findings could reduce mortality, improve prognosis, and optimize the clinical treatment of severe trauma patients. The protocol used in our research also has reference value for other clinical syndromes and disease processes.

REFERENCES

1. Heckbert SR, Vedder NB, Hoffman W, Winn RK, Hudson LD, Jurkovich GJ, Copass MK, Harlan JM, Rice CL, Maier RV: Outcome after hemorrhagic shock in trauma patients. *J Trauma Acute Care* 45(3):545–549, 1998.
2. Sauaia A, Moore FA, Moore EE, Moser KS, Brennan R, Read RA, Pons PT: Epidemiology of trauma deaths: a reassessment. *J Trauma Acute Care* 38(2):185–193, 1995.
3. Kauvar DS, Lefering R, Wade CE: Impact of hemorrhage on trauma outcome: an overview of epidemiology, clinical presentations, and therapeutic considerations. *J Trauma Acute Care* 60(6):S3–S11, 2006.
4. Kai-yuan LI, Cong F, Chen LI, Fei P, Yu-zhuo Z, Tan-shi LI: A description of Combat Casualty Care statistics of U.S army and its illumination for our army. *China J Emerg Resuscit Disaster Med* 13(3):210–212, 2018.
5. Gruen RL, Brohi K, Schreiber M, Balogh ZJ, Pitt V, Narayan M, Maier RV: Haemorrhage control in severely injured patients. *Lancet* 380(9847):1099–1108, 2012.
6. Brohi K, Cohen MJ, Ganter MT, Schultz MJ, Levi M, Mackersie RC, Pittet J: Acute coagulopathy of trauma: hypoperfusion induces systemic anticoagulation and hyperfibrinolysis. *J Trauma Acute Care* 64(5):1211–1217, 2008.
7. Tremoleda JL, Watts SA, Reynolds PS, Thiemermann C, Brohi K: Modeling acute traumatic hemorrhagic shock injury: challenges and guidelines for preclinical studies. *Shock* 48(6):610–623, 2017.
8. Bing Y, Katrin B, Christian K, Borna R, Claudia N: Traumatic injury pattern is of equal relevance as injury severity for experimental (poly) trauma modeling. *Sci Rep-UK* 9(1):1–12, 2019.
9. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, Bock C, Horn M, Moor M, Rieck B, et al.: Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* 26(3):364–373, 2020.
10. Rau C, Wu S, Chien P, Kuo P, Chen Y, Hsieh H, Hsieh C: Prediction of mortality in patients with isolated traumatic subarachnoid hemorrhage using a decision tree classifier: a retrospective analysis based on a trauma registry system. *Int J Env Res Pub He* 14(11):1420, 2017.
11. Cannon CM, Braxton CC, Kling-Smith M, Mahnken JD, Carlton E, Moncure M: Utility of the shock index in predicting mortality in traumatically injured patients. *J Trauma Acute Care* 67(6):1426–1430, 2009.
12. Choi JY, Kim SK, Lee WH, Yoo TK, Kim DW: A survival prediction model of rats in hemorrhagic shock using the random forest classifier. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. 2012;5570–5573, 2012.
13. Liu NT, Salinas J: Machine learning for predicting outcomes in trauma. *Shock: Injury Inflammation and Sepsis: Laboratory and Clinical Approaches* 48(5):504–510, 2017.
14. Chen L, Reisner AT, McKenna TM, Gribok A, Reifman J: Diagnosis of hemorrhage in a prehospital trauma population using linear and nonlinear multiparameter analysis of vital signs. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. 2007;3748–3751, 2007.
15. Chen L, McKenna TM, Reisner AT, Gribok A, Reifman J: Decision tool for the early diagnosis of trauma patient hypovolemia. *J Biomed Inform* 41(3):469–478, 2008.
16. Yuzhuo Z, Junmei W, Fei P, Peiyao L, Lijing J, Kaiyuan L, Cong F, Tongbo L, Zhengbo Z, Desen C, et al.: Pilot research: construction of emergency rescue database. *Chin Crit Care Med* 30(6):609–612, 2018.
17. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG: MIMIC-III, a freely accessible critical care database. *Sci Data* 3(1):1–9, 2016.
18. Marshall DC, Hatch RA, Gerry S, Young JD, Watkinson P: Conditional survival with increasing duration of ICU admission: an observational study of three intensive care databases. *Crit Care Med* 48(1):91–97, 2020.
19. Gao Y, Wang Q, Li J, Zhang J, Li R, Sun L, Guo Q, Xia Y, Fang B, Wang G: Impact of mean arterial pressure fluctuation on mortality in critically ill patients. *Crit Care Med* 46(12):e1167–e1174, 2018.
20. Liu L, Yu Y, Fei Z, Li M, Wu F, Li H, Pan Y, Wang J: An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC Syst Biol* 12(6):29–38, 2018.
21. Liu YC: Modified shock index and mortality rate of emergency patients. *J World J Emerg Med* 3(2):114, 2012.
22. Sohn CH, Kim WY, Kim SR, Seo DW, Ryoo SM, Lee YS, Lee JH, Oh BJ, Won HS, Shim JY: An increase in initial shock index is associated with the

requirement for massive transfusion in emergency department patients with primary postpartum hemorrhage. *J Shock* 40(2):101–105, 2013.

23. Li T, Zhu Y, Hu Y, Li L, Diao Y, Jing T, Liu L: Ideal permissive hypotension to resuscitate uncontrolled hemorrhagic shock and the tolerance time in rats. *J Anesthesiology* 114(1):111–119, 2011.

24. Gutierrez G, Reines HD, Wulf-Gutierrez ME: Clinical review: hemorrhagic shock. *J Crit Care (London England)* 8(5):373–381, 2004.

25. Douzinas EE, Livaditi O, Andrianakis I, Prigouris P, Paneris P, Villiotou V, Betrosian AP: The effect of hypoxemic resuscitation from hemorrhagic shock on blood pressure restoration and on oxidative and inflammatory responses. *J Intensive Care Med* 34(6):1133–1141, 2008.

26. Peterson KL, Hardy BT, Hall K: Assessment of shock index in healthy dogs and dogs in hemorrhagic shock. *J Vet Emerg Crit Car* 23(5):545–550, 2013.

27. Cannon, Jeremy W: Hemorrhagic shock REPLY. *J N Engl J Med* 378(4):370–379, 2018.

28. Jackson Chornenki NL, James TE, Barty R, Liu Y, Rochwerg B, Heddle NM, Siegal DM: Blood loss from laboratory testing, anemia, and red blood cell transfusion in the intensive care unit: a retrospective study. *Transfusion* 60(2):256–261, 2020.

29. Kherad O, Restellini S, Martel M, Sey M, Murphy MF, Oakland K, Barkun A, Jairath V: Outcomes following restrictive or liberal red blood cell transfusion in patients with lower gastrointestinal bleeding. *Aliment Pharm Ther* 49(7):919–925, 2019.

30. Buuren SV, Groothuis-Oudshoorn K: mice: multivariate imputation by chained equations in R. *J Stat Softw* 45:1–68, 2010.

31. Liang W, Yao J, Chen A, Lv Q, Zanin M, Liu J, Wong S, Li Y, Lu J, Liang H, et al.: Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* 11(1):1–7, 2020.

32. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, Liang H: Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. *Sci Rep-UK* 7(1):1–9, 2017.

33. Chen T, Guestrin C: Xgboost a scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining; 785–794, 2016.

34. Lundberg SM, Lee S: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 4765–4774, 2017.

35. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837–845, 1988.

36. Mackenzie CF, Gao C, Hu PF, Amechi A, Hegang C, Theresa D, Cristina IP, Lauren H, Christopher S, Jay M, et al.: Comparison of decision-assist and clinical judgment of experts for prediction of lifesaving interventions. *Shock* 43(3):238–243, 2015.

37. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D: Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Netw Open* 1:e181018, 2018.

38. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu L, Moons KG, et al.: External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 14(1):1–11, 2014.

39. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, et al.: Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 353:i2416, 2016.

40. Goldstein BA, Navar AM, Pencina MJ, Ioannidis J: Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assn* 24(1):198–208, 2017.

41. Boshra R, Ruiter K, Reilly JP, Connolly JF: Machine learning based framework for EEG/ERP analysis. *J Int J Psychophysiol* 108:105, 2016.

42. Yun L, Mingjiang W, Qiquan Z, Yufei H: Identification of auditory object-specific attention from single-trial electroencephalogram signals via entropy measures and machine learning. *J Entropy* 20(5):386, 2018.

43. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG: An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 46(4):547–553, 2018.

44. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A: Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med* 48(2):210–217, 2020.

45. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, Calvert J, Das R: Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput Biol Med* 109:79–84, 2019.

46. Rhee C, Wang R, Zhang Z, Fram D, Kadri SS, Klompas M, Program CPE: Epidemiology of hospital-onset versus community-onset sepsis in US hospitals and association with mortality: a retrospective analysis using electronic clinical data. *Crit Care Med* 47(9):1169–1176, 2019.

47. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, Fuchs BD, Meadows L, Lynch M, Donnelly PJ, et al.: A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med* 47(11):1485–1492, 2019.

48. Zhu CS, Cobb D, Jonas RB, Pokorny D, Rani M, Cotner-Pouncy T, Oliver J, Cap A, Cestero R, Nicholson SE, et al.: Shock index and pulse pressure as triggers for massive transfusion. *J Trauma Acute Care* 87(1S):S159–S164, 2019.

49. Gelbard RB, Hensman H, Schobel S, Khatri V, Tracy BM, Dente CJ, Buchman T, Kirk A, Elster E: Random forest modeling can predict infectious complications following trauma laparotomy. *J Trauma Acute Care* 87(5):1125–1132, 2019.