Kong *et al. BMC Medical Informatics and Decision Making* (2020) 20:251
https://doi.org/10.1186/s12911-020-01271-2

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU

Guilan Kong[1,2*], Ke Lin[1,2] and Yonghua Hu[3,4]

Abstract

Background: Early and accurate identification of sepsis patients with high risk of in-hospital death can help physicians in intensive care units (ICUs) make optimal clinical decisions. This study aimed to develop machine learning-based tools to predict the risk of hospital death of patients with sepsis in ICUs.

Methods: The source database used for model development and validation is the medical information mart for intensive care (MIMIC) III. We identified adult sepsis patients using the new sepsis definition Sepsis-3. A total of 86 predictor variables consisting of demographics, laboratory tests and comorbidities were used. We employed the least absolute shrinkage and selection operator (LASSO), random forest (RF), gradient boosting machine (GBM) and the traditional logistic regression (LR) method to develop prediction models. In addition, the prediction performance of the four developed models was evaluated and compared with that of an existent scoring tool – simplified acute physiology score (SAPS) II – using five different performance measures: the area under the receiver operating characteristic curve (AUROC), Brier score, sensitivity, specificity and calibration plot.

Results: The records of 16,688 sepsis patients in MIMIC III were used for model training and test. Amongst them, 2949 (17.7%) patients had in-hospital death. The average AUROCs of the LASSO, RF, GBM, LR and SAPS II models were 0.829, 0.829, 0.845, 0.833 and 0.77, respectively. The Brier scores of the LASSO, RF, GBM, LR and SAPS II models were 0.108, 0.109, 0.104, 0.107 and 0.146, respectively. The calibration plots showed that the GBM, LASSO and LR models had good calibration; the RF model underestimated high-risk patients; and SAPS II had the poorest calibration.

Conclusion: The machine learning-based models developed in this study had good prediction performance. Amongst them, the GBM model showed the best performance in predicting the risk of in-hospital death. It has the potential to assist physicians in the ICU to perform appropriate clinical interventions for critically ill sepsis patients and thus may help improve the prognoses of sepsis patients in the ICU.

Keywords: Intensive care unit, Sepsis, Prediction model, Machine learning, In-hospital mortality

Background

Sepsis is a life-threatening illness that occurs when the body's response to infection is out of balance [1]. It can trigger body changes that may damage multiple organ systems and lead to death [2]. Sepsis has become a major contributor of public health and economic burden [3, 4]. It is associated with high risk of complications and in-hospital death, longer hospital stays and higher medical costs. Today, sepsis has become a major cause of in-hospital death for intensive care unit (ICU) patients. In the US, 10% of patients admitted to the ICU have sepsis, and around 25% of ICU beds are occupied by sepsis

* Correspondence: guilan.kong@hsc.pku.edu.cn
[1]National Institute of Health Data Science, Peking University, Beijing, China
[2]Center for Data Science in Health and Medicine, Peking University, Beijing, China
Full list of author information is available at the end of the article

patients [5–7]. Given the high mortality of sepsis patients in the ICU, the risk of in-hospital death of patients with sepsis needs to be discovered the earlier the better. Early and accurate identification of sepsis patients with high risk of in-hospital death can help ICU physicians make optimal clinical decisions, which can, in turn, improve their clinical outcomes [8].

The initial sepsis definition was proposed in 1991, referred as Sepsis-1 [9]. It was defined as infected patients meeting two or more of the systemic inflammatory response syndrome (SIRS) criteria, including 1) temperature > 38 °C or < 36 °C, 2) heart rate > 90/min, 3) respiratory rate > 20/min or partial pressure of carbon dioxide ($PaCO_2$) < 32 mmHg (4.3 kPa), and 4) white blood cell count (WBC) > 12,000/mm^3 or < 4000/mm^3 or > 10% immature (band) forms. Sepsis-1 was updated to Sepsis-2 [10] in 2001 by expanding the list of diagnostic criteria but did not offer alternatives. In effect, the definition of sepsis has remained largely unchanged for more than two decades. In 2016, a new sepsis definition, Sepsis-3 [1], was proposed. In Sepsis-3, instead of checking the SIRS criteria as in previous definitions, infected patients who have a sequential organ failure assessment (SOFA) score [11] higher than 2 are defined as having sepsis.

Some scoring tools have been developed to assess the illness severity of patients with sepsis. Amongst them, the simplified acute physiology score (SAPS) II [12]; the acute physiology and chronic health evaluation (APACHE) II, III and IV scores [13–15]; and the SOFA score are frequently used severity assessment tools in the ICU. Patient vital signs, laboratory results and demographic statistics are risk factors used in these scoring systems for severity assessment. Most of these severity scores were developed decades ago. They may perform well in the population or clinical settings at the time when they were developed. However, the performance of these severity scores has declined due to the fact that the population and clinical settings have changed over the time. Moreover, some studies [16, 17] showed that the calibration and discrimination capabilities of these severity scores are poor in predicting the risk of in-hospital death of sepsis patients. In addition, there is a specific severity score particularly for sepsis patients: Sepsis Severity Score (SSS) [18]. The SSS was developed as a specific severity score for predicting in-hospital mortality of sepsis patients. However, a study [17] showed that the discrimination performance of SSS is not as good as that of APACHE IV in predicting the risk of in-hospital death of patients with sepsis, and the calibration capability of SSS is poor.

Given the poor performance of existing severity scores, some new models have been developed for predicting the risk of in-hospital death amongst ICU patients with sepsis [19–21]. Fang et al. [22] developed and validated a scoring system for predicting 28-day mortality risk of patients with sepsis. The corresponding area under the receiver operating characteristic curve (AUROC) generated by the scoring system is 0.789. Xie et al. [23] used clinical features and biomarkers as predictors to develop a model based on traditional logistic regression (LR) algorithm for predicting the mortality risk of sepsis patients. The AUROC of their model is 0.778.

With the accumulation of big data and the development of techniques for data storage, machine learning methods have attracted considerable research attention [24–26]. Several innovative and pragmatic machine learning methods such as random forest (RF) [27], gradient boosting machine (GBM) [28] and the least absolute shrinkage and selection operator (LASSO) [29] which is a type of linear regression using shrinkage, have been proposed, and these models have good prediction performance in medicine. Some machine learning-based models have been developed for ICU mortality prediction in the literature [30, 31]. In sepsis area, Zhang et al. [32] used the LASSO method to build a tool for predicting the mortality risk of sepsis patients based on the medical information mart for intensive care (MIMIC) III dataset [33], and their research results showed that the LASSO-based prediction model was superior to SOFA score in discrimination. Taylor et al. [34] employed machine learning methods to build mortality prediction models for patients with sepsis, and their research results showed that the RF model performed better than the LR model in discrimination. Relevant studies about mortality prediction for sepsis patients are listed in Table 1, where the dataset, methodology, predictors, outcome and sepsis definition used in each study were presented.

From Table 1, we can find that most of the existing relevant studies have limitations such as limited sample sizes, limited predictor variables and old sepsis definitions, and most of them used traditional analytic methods such as LR to build models. LR models assume that the dependent variable has a linear functional relationship with predictor variables after a logit transformation, and this assumption may affect the model's discrimination power as the relationship may be nonlinear [36]. A previous study [37] recommended using the GBM model to predict the in-hospital mortality for sepsis as it often outperforms the RF model. However, few studies have employed the GBM model to predict the mortality of ICU sepsis patients.

Therefore, driven by the need of using big data and the latest sepsis definition to develop in-hospital mortality prediction models for sepsis patients in the ICU, we used the GBM, RF, LASSO and LR methods together with the latest sepsis definition criteria to build models for predicting the risk of in-hospital death of patients with sepsis in the ICU and compared their prediction performance with the existent scoring tool SAPS II.

Table 1 Relevant studies about mortality prediction for sepsis patients

Authors	Title	Dataset	Methodology	Predictors	Outcome	Sepsis definition
Masson, S. et al. [19]	Presepsin (soluble CD14 subtype) and procalcitonin levels for mortality prediction in sepsis: data from the Albumin Italian Outcome Sepsis trial	A multicentre, randomised Albumin Italian Outcome Sepsis trial, 100 patients	Cox regression model	Presepsin level, procalcitonin level and some covariates	28-day/ICU/90-day mortality	Sepsis-2
Adrie C. et al. [20]	Model for predicting short-term mortality of severe sepsis	A multicentre database including data from 12 ICUs, 2268 patients	Generalised linear model	SAPS II and LOD scores at ICU admission, septic shock, multiple organ failure, comorbidities, procedures, agents, bacteraemia and sources of infection	14-day mortality within ICU stay	Sepsis-2
Ripoll, V.J.R. et al. [21]	Sepsis mortality prediction with the Quotient Basis Kernel	MIMIC II	Support vector machines (SVMs), LR, SAPS	SOFA and SAPS scores at ICU admission	ICU mortality	Sepsis-2
Fang W-F et al. [22]	Development and validation of immune dysfunction score to predict 28-day mortality of sepsis patients	Sepsis patients admitted to ICU at a hospital in Taiwan, 151 patients	LR	Monocyte HLA-DR* expression, plasma G-CSF* level, plasma IL*-10 level, and serum SeMo* ratio	28-day mortality	Sepsis-3
Xie, Y. et al. [23]	Using clinical features and biomarkers to predict 60-day mortality of sepsis patients	Protocol-based care in early septic shock trial, around 530 patients	LR	Clinical features and biomarkers obtained during the first 24 h of hospital admission	60-day mortality	Not mentioned
Poucke, S.V. et al. [31]	Scalable predictive analysis in critically ill patients using a visual open data analysis platform	MIMIC II	Naïve Bayes, LR, RF, AdaBoost, Bagging, Stacking, SVM	Demographics, comorbidities, types of care unit, platelet count	ICU mortality	NA
Zhang, Z. & Hong, Y [32].	Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: the use of electronic healthcare records with LASSO regression	MIMIC III	LASSO, LR	Demographics, clinical and laboratory variables recorded during the first 24 h in ICU	Hospital mortality	Sepsis-2
Taylor, R.A. et al. [34]	Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach	Adult ED* visits over 12 months, 4676 patients	RF, CART, LR	Demographics, previous health status, ED health status, ED services rendered and operational details	Hospital mortality	Sepsis-2
Pregernig, A. et al. [35]	Prediction of mortality in adult patients with sepsis using six biomarkers: a systematic review and meta-analysis	44 articles in English	Qualitative analysis, meta-analysis	Angiopoietin 1 and 2, HMGB1*, sRAGE*, sTREM*-1, suPAR*	28-day/30-day/ICU/hospital/90-day mortality	Sepsis-1/Sepsis-2/Sepsis-3

**Abbreviations*: *HLA-DR* Human leukocyte antigen D-related, **G-CSF* Granulocyte-colony stimulating factor, *IL* Interleukin, *SeMo* Segmented neutrophil-to-monocyte, *ED* Emergency department, *HMGB1* High mobility group box 1 protein, *sRAGE* soluble receptor for advanced glycation endproducts, *sTREM* soluble triggering receptor expressed on myeloid cells 1, *suPAR* soluble urokinase-type plasminogen activator receptor

Methods

Dataset and subjects

MIMIC III, an ICU database from the Beth Israel Deaconess Medical Center (BIDMC) [33], was employed for model derivation and validation. As a database accessible to researchers worldwide, MIMIC III contains over 40, 000 records of patients receiving critical care in the ICU at BIDMC between 2001 and 2012. The diagnostic codes, vital signs, laboratory tests, demographics and some other clinical characteristics of each patient were included in MIMIC III. The Institutional Review Board of the BIDMC and Massachusetts Institute of Technology approved the research use of MIMIC III for researchers having attended their training course.

Adult patients between 18 and 90 years old were included in this study. We identified adult sepsis patients using the latest Sepsis-3. According to the Sepsis-3, infected patients who have a SOFA score higher than 2 are defined as having sepsis. First of all, we need to identify infected patients. We employed the International Classification of Diseases, ninth revision, Clinical Modification (ICD-9-CM) diagnosis codes provided by Angus et al. [38] to identify infected patients from the MIMIC-III database. Furthermore, we used the SOFA score as another criterion to identify sepsis patients from the infected. For those patients with twice or more ICU admissions during one hospitalisation, we included only the patient's first ICU admission. Patients whose records have a predictor variable missing rate higher than 30% were excluded.

Predictor variables and the primary outcome

The primary outcome of this study is in-hospital mortality of sepsis patients in the ICU.

In the process of predictor selection, we made reference to the established scoring tools, SAPS II and APAC HE III, and considered variables such as obesity, serum lactate and international normalised ratio (INR), which have been found relevant to mortality, as predictors to construct the mortality prediction models [32, 38].

To ensure the availability of all predictor variables in prediction model development, we excluded variables with data missing rate higher than 30%. Finally, a total of 86 predictor variables consisting of demographics, laboratory tests and comorbidities were used as independent predictor variables for model development.

MIMIC-III has some time-stamped physiological data. For example, blood pressure and heart rate are measured hourly. In model construction, as variations and sudden changes could be more informative than average values of time-stamped measurements, the minimum and maximum values during the first 24 h of these variables for each ICU stay were used as parallel inputs for prediction models. All predictor variables are listed in Table 2. All predictor variables were extracted or calculated from the patient data recorded in the MIMIC III database.

Table 2 Predictor variables used in this study

Predictors	
Acute physiology (first 24 h in the ICU)	**Chronic health status**
Heart rate*	Elixhauser comorbidity index
Systolic blood pressure*	Congestive heart failure
Diastolic blood pressure*	Cardiac arrhythmias
Mean blood pressure*	Valvular heart disease
Respiratory rate*	Pulmonary circulation
Temperature*	Peripheral vascular
SpO2* (blood oxygen saturation)	Hypertension
Total CO_2*	Other neurological diseases
pCO_2* (partial pressure of CO_2)	Chronic obstructive pulmonary disease
pH* (acidity in the blood)	Diabetes without complications
Urine output	Diabetes with complications
Glasgow Coma Score (GCS)	Hypothyroidism
GCS (eye)	Renal failure
GCS (motor)	Liver disease
GCS (verbal)	Metastatic cancer
Anion gap*	Coagulopathy
Bicarbonate*	Obesity
Creatinine*	Fluid electrolyte
Chloride*	Alcohol abuse
Glucose*	Depression
Haematocrit*	Renal replacement therapy
Haemoglobin*	**Other**
Lactate*	Gender
Platelet*	Weight loss
Potassium*	Ventilation
Partial thromboplastin time*	Age
INR*	Weight
Prothrombin time*	SAPS II score (first 24 h in the ICU)
Sodium*	SOFA score (first 24 h in the ICU)
Blood urea nitrogen (BUN)*	
WBC*	
Acute kidney injury	

*: each predictor marked with * means that it is a time-stamped variable, and its corresponding minimum and maximum values within the first 24 h in the ICU were used as inputs in model development

Prediction models

In this study, three machine learning methods, namely, LASSO, RF and GBM, together with the frequently used LR method were used to develop in-hospital mortality prediction models. In addition, the established scoring tool, SAPS II, was employed for the prediction. A brief introduction of the SAPS II, LR, LASSO, RF and GBM models is provided as follows.

SAPS II

SAPS II is a scoring tool developed to measure the disease severity of patients (≥15) admitted to ICUs. SAPS II was developed based on a large international sample of patients. It can provide an estimate of the risk of death without having to specify a primary diagnosis. Detailed variables and score assignment in SAPS II are shown in Table 3.

The probability of in-hospital death is estimated from the SAPS II score as follows:

$$\begin{aligned}\text{logit} = & -7.763 + (0.0737 + \text{SAPS II score}) \\ & + (0.9971\times\ \ln(\text{SAPS II score} + 1)\end{aligned} \quad (1)$$

$$p = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}} \quad (2)$$

Table 3 Variables and score assignment in SAPS II

Variables		**Maximum scores**
Acute physiology	Temperature	3
	Heart rate	11
	Systolic blood pressure	13
	WBC	12
	Bilirubin	9
	Serum sodium	5
	Serum potassium	3
	Serum bicarbonate	6
	BUN	10
	Urine output	11
	PaO_2[a]or FiO_2[a]	11
	GCS	26
Chronic health status	AIDS[a]	17
	Haematologic malignancy	10
	Metastatic cancer	9
Other	Age	18
	Type of admission	8
Overall score		182

[a]*Abbreviations*: *AIDS* Acquired immunodeficiency syndrome, *PaO_2* Partial pressure of oxygen, *FiO_2* Fraction of inspired oxygen

LR

The LR method has been widely used in medical research. The mathematical function between predictors and the risk of in-hospital death can be described as follows:

$$\text{logit} = \ log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_N x_N \quad (3)$$

where p denotes the probability of in-hospital death, $x_i(i = 1, 2, \ldots, N)$ represents independent predictors, $\beta_i(i = 1, 2, \ldots, N)$ are the coefficients associated with predictors, and N is the number of all predictors.

LASSO

The LASSO model was developed in 1996. It is similar to linear regression except that it shrinks the coefficients of some variables with multicollinearity towards zero via regularisation [39]. As a result, the coefficients of some correlated predictors will be zero after model training, and a so-called sparse model with only important predictors can be generated.

In LASSO regression, the relationship between all potential predictor variables $x_{ij}(i = 1, 2, \ldots, N; j = 1, 2, \ldots, P)$ and the outcome $y_i(i = 1, 2, \ldots, N)$ is assumed to be similar to that in linear regression. Here, N denotes the number of all cases in the training dataset, and P denotes the number of all predictor variables. Furthermore, the LASSO regression is different from linear regression in that it uses coefficient shrinkage. The coefficients in the LASSO regression are generated through model training by solving the following objective function:

$$\min_{\beta}\left\{\sum\nolimits_{i=1}^{N}\left(y_i - \beta_0 - \sum\nolimits_{j=1}^{P} x_{ij}\beta_j\right)^2 + \lambda \sum\nolimits_{j=1}^{P}\left|\beta_j\right|\right\} \quad (4)$$

The first half of the objective function is about the training loss, which is the difference between model generated results and the observed outcomes, measuring how well the model fits the training dataset. The second half of the objective function is for regularisation, the penalty on the coefficients, which measures the complexity of the model. It is through the regularisation that coefficients of correlated variables can be shrinked to 0 in LASSO regression. In the function, the parameter $\lambda \geq 0$ is for controlling the amount of shrinkage, and a larger λ means a greater shrinkage amount.

We used the glmnet package in R (https://www.r-project.org) software to train the LASSO model. In the LASSO model training, the penalty on the β-coefficients is controlled by the tuning parameter λ, and the optimal λ was found via cross-validation (CV) in the glmnet package.

RF

The RF model is a type of ensemble-learning model that uses multiple decision trees as its base models, and a majority voting system is used as the final aggregation method to synthesise the classification results of all the base models [27]. The training of decision trees in a RF model utilises the same learning algorithm, which uses Gini index [40] as the criterion for selecting appropriate variables for different nodes in tree growing. For more detailed introduction of the RF model, readers can refer to article [41].

We used the randomForest package in R to fit the RF model. In the RF model training, we set the total number of variables in each decision tree as the default value in R. We set the candidate numbers of all decision trees in the RF model to 500, 1000, 1500 and 2000, and the number that resulted in the highest AUROC was selected. We finally selected 1000 as the number of decision trees in our RF model.

GBM

While the RF model is an ensemble of parallel decision trees, the GBM model builds an ensemble of decision trees in a sequential way, in which the training purpose of each decision tree is to minimise the discrepancies between the observed and predicted outcomes made by all its preceding tress [28, 42]. In a GBM model, the ensemble of decision trees is trained sequentially, where each decision tree gradually corrects for the residuals of all its preceding trees via the gradient descent method. The training procedure of a GBM model consecutively and iteratively fits new decision trees to generate estimates closer to the observed outcomes. The training of the GBM model is illustrated as follows.

Let M be the number of all iterations or the number of all decision trees in a GBM model. $F_{m-1}(x)(m = 2, 3\ldots, M)$ is used to represent the GBM model in current status, which contains m-1 decision trees. Then, $h_m(x)(m = 1, 2, \ldots, M)$ is used to represent each decision tree that the GBM model attempts to find and include. Each single decision tree $h_m(x)$ is trained using the gradient boosting method, where the negative gradient of the loss function defined for the GBM model is used for model fitting. $F_m(x)$ is used to represent the GBM model after including the mth decision tree $h_m(x)$. The iterative updating process of the GBM model, $F_m(x)$, can be written as follows:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x) \qquad (5)$$

In the above formula, the coefficient γ_m is calculated in the process of minimising the loss function of $F_m(x)$ by a line search strategy. Overfitting is avoided by using shrinkage as a regularisation method, where $h_m(x)$ is multiplied by a small learning rate ν in each training iteration. $F_M(x)$ is the final model in which m reaches M.

We used the gbm package in R software to fit the GBM model. The grid search, an exhaustive searching strategy which attempts to find optimal parameters through a subset of manually specified values, was employed in the GBM model training. We set the candidate numbers of decision trees in the ensemble GBM model to 500, 1000, 1500 and 2000 and the candidate learning rates to 0.01 and 0.001. These candidate values were selected in accordance with a previous study by Hastie et al. [43]. We finally selected 2000 as the number of trees in the ensemble model and 0.01 as the learning rate.

The hypothesis of this study was that machine learning-based prediction models perform better than traditional regression models and existing scoring tools, and ensemble-learning models perform better than single-learning models.

Statistical analysis

After patient and variable exclusion, the mean value of each variable was utilised to compensate for the missing values of the corresponding variable for final analysis. We employed R 3.40 to perform calculations and statistical analysis. We employed mean ± standard deviation, percentages or actual numbers to describe patient data characteristics. We employed five-fold CV to find optimal model parameters and the model with best prediction performance amongst the developed models. By using the five-fold CV method, the whole dataset was divided into five folds, and five rounds of model training and test were conducted. In each round, four folds were used as the training dataset, and the last one was used as the test dataset. The test fold was different in each round. In this study, the model with the best average prediction performance on five test datasets was considered the optimal model.

We compared the developed models using five different performance measures: Brier score [44], AUROC [45], sensitivity, specificity and calibration plot. Brier score measures the overall prediction performance, which denotes the accuracy of a prediction, of a model. The best possible score is 0 for a totally accurate prediction, whereas the lowest possible score is 1 for a wholly inaccurate prediction. AUROC is an overall measure of discrimination of a model considering sensitivity and specificity, and it denotes the capability to distinguish the survivors from the deceased. Calibration plot is an approach to illustrate the calibration capability of a model, representing the consistency between the predictions and observed outcomes.

Results

A total of 16,688 patients with sepsis were included in model derivation, amongst which 2949 (17.7%) died. The proportions of males in all patients, in survivors and in decedents were 54.5, 54.1 and 56.05%, respectively. The average age of all patients, survivors and decedents were 65.61+/− 15.01, 65.00+/− 15.11 and 68.46+/− 14.19 years old, respectively. The average SOFA scores in all patients, the survivors and decedents were 5.57+/− 3.06, 5.10+/− 2.67 and 7.77+/− 3.72, respectively. Table 4 describes the characteristics of sepsis patients included in this study. Compared with the survivors, the decedents were older and had higher SOFA and SAPS II scores, longer ICU stays and shorter hospitals stays.

Table 5 presents the average AUROC value, Brier score, the sensitivity and specificity corresponding to the optimal cut-off point of each model in fivefold CV. The average AUROC values of the LASSO, GBM, RF, LR and SAPS II models were 0.829 (95% confidence interval (CI): 0.827–0.831), 0.845 (95%CI: 0.837–0.853), 0.829 (95%CI: 0.823–0.834), 0.833 (95%CI: 0.830–0.838) and 0.77 (95%CI: 0.760–0.780), respectively. The average Brier scores of the LASSO, GBM, RF, LR and SAPS II models were 0.108 (95%CI: 0.107–0.109), 0.104 (95%CI: 0.102–0.105), 0.109 (95%CI: 0.108–0.109), 0.107 (95%CI: 0.105–0.108) and 0.146 (95%CI: 0.142–0.150), respectively.

Figure 1 illustrates the calibration plots of the five models. It shows that the LASSO, LR and GBM models

Table 4 Characteristics of included sepsis patients

Items	Statistics		
	All	Survivors	Decedents
Total number	16,688	13,739	2949
Age (year)	65.61 ± 15.01	65.00 ± 15.11	68.46 ± 14.19
Gender			
Male	9087 (54.5%)	7434 (54.1%)	1653 (56.05%)
Female	7601 (45.5%)	6305 (45.9%)	1296 (43.95%)
SOFA score, 1st day of ICU admission	5.57 ± 3.06	5.10 ± 2.67	7.77 ± 3.72
SAPS II score, 1st day of ICU admission	40.87 ± 13.51	38.55 ± 12.18	51.68 ± 14.13
Hospital length of stay (day)	14.42 ± 14.77	14.55 ± 14.26	13.82 ± 16.96
ICU length of stay (day)	6.41 ± 8.52	6.07 ± 8.30	7.99 ± 9.34
Number of deaths	2949 (17.7%)	NA	NA

had good calibration, whereas the RF model underestimated the mortality risk of high-risk sepsis patients. Meanwhile, the SAPS II model showed the poorest calibration.

Discussion

In this study, sepsis patient data were first extracted from the MIMIC III database. Then, four data-based tools, including LASSO, GBM, RF and LR models, were developed to perform in-hospital mortality prediction. Finally, the prediction performances of the fours tools were compared with the existent scoring tool SAPS II.

In general, the criteria for modelling method selection include model performance and interpretability. As for the model performance, ensemble-learning models usually have better prediction performance than single-learning models [46]. Regarding model interpretability, traditional regression models have considerably better interpretability than black-box machine learning models. Therefore, we considered machine learning and regression models and ensemble- and single-learning models in model selection. Finally, we selected the LASSO, GBM, RF and LR models. As an ensemble model, the RF model has good prediction performance and good interpretability. The GBM model is also an ensemble model with good performance, but its interpretability is poorer than that of the RF model. The LASSO model is a type of regression model with good interpretability but fair performance, whereas the LR model is a traditional regression model with good interpretability. We attempted to find a trade-off between the model performance and interpretability in model selection.

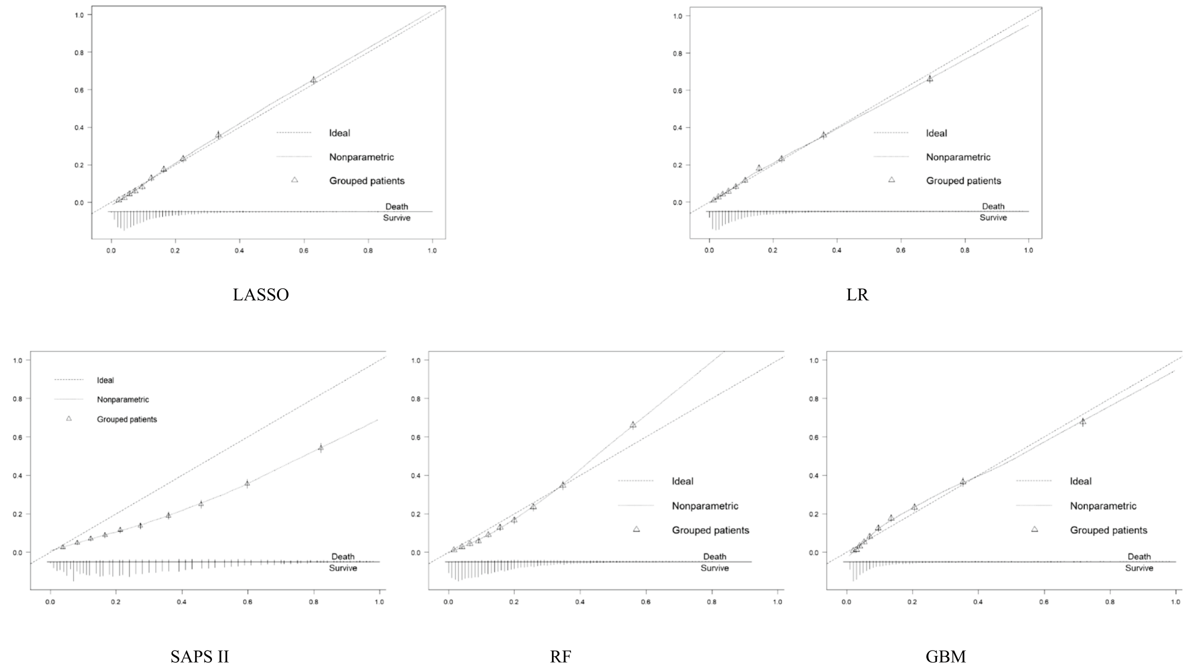
Regarding identification of sepsis patients using Sepsis-3, a SOFA score can be calculated for each patient in MIMIC III database and assessed whether the number is larger than 2. However, as infections have varied causes and infected sites, different diagnoses may be provided for infected patients. Thus, various diseases are related to infections. In this study, we employed the ICD codes provided by Angus et al. [38] to identify infected patients from MIMIC III. These ICD codes for infections have already been tested by a study [47]. Therefore, ICD codes can be reasonably used to identify sepsis patients from MIMIC III.

In this study, the observed mortality rate of the whole dataset was around 17.7%. However, we performed no processing on the imbalance because the prediction performance of all models developed from the dataset were acceptable. To eliminate the effect caused by data imbalance on the trained models, we employed the AUROC, which considers sensitivity and specificity and cannot be wavered by data imbalance, to evaluate prediction performance. Therefore, data imbalance would not affect the identification of optimal prediction models.

As for the predictor variable inclusion, as this study aimed to develop in-hospital mortality prediction models based on all available physiology, chronic health status,

Table 5 Performance comparison of five models

Model	Overall performance		Discrimination					
	Brier score	95% CI	AUROC	95% CI	Sensitivity	95% CI	Specificity	95% CI
LASSO	0.108	0.107–0.109	0.829	0.827–0.831	0.744	0.721–0.767	0.754	0.731–0.777
GBM	0.104	0.102–0.105	0.845	0.837–0.853	0.771	0.750–0.792	0.755	0.722–0.789
RF	0.109	0.108–0.109	0.829	0.823–0.834	0.765	0.756–0.774	0.740	0.719–0.761
LR	0.107	0.105–0.108	0.833	0.830–0.838	0.760	0.740–0.780	0.748	0.724–0.772
SAPS II	0.146	0.142–0.150	0.77	0.760–0.780	0.697	0.668–0.725	0.714	0.695–0.734



LASSO

LR

SAPS II

RF

GBM

Fig. 1 Calibration plots of the LASSO, LR, SAPS II, RF and GBM models

demographics and several other hospitalisation-related variables during the first 24 h after ICU admission, we excluded medications or procedures, which would alter the course of health trajectory of sepsis patients. All predictor variables used in this study aimed for mortality prediction instead of sepsis diagnosis. Thus, these predictors may include both pre-sepsis and post-sepsis variables but are restricted to the first 24 h after ICU admission. Similar to the established scoring tools SAPS II and APACHE III, the models developed in this study are stationary and can be used after the first 24 h.

Amongst the five prediction models, the GBM model had the best discrimination and overall performance, and it also had a good calibration as illustrated by the calibration plot. Unlike the LASSO and LR models, the RF and GBM models are ensemble learning models with decision tree as base models; both belong to a type of non-parametric machine learning technique having no requirements of distribution or parameter of the training dataset [48]. Ensemble learning methods based on decision tress are superior to learning techniques with parametric requirements, such as LR, because non-parametric methods have advantages in handling high-volume data without specific distribution patterns. Although the basic idea behind the GBM and RF models is to aggregate many individual weaker decision trees into an ensemble and stronger learner, a GBM model generally produces better performance than a RF model [42, 49, 50]. A RF model trains each tree independently and uses a random sample of the data for training, whereas a GBM model builds an ensemble of decision trees in a sequential way, in which each new decision tree is fitted through correcting for the residuals of all its preceding trees. This condition means that each new tree in a GBM model corrects the errors made by all previously trained trees in the model [42]. However, the GBM model has poorer interpretability than RF. The LR model performed worse than the GBM model by all measures but showed better performance than the RF and LASSO models in terms of Brier score and AUROC. The SAPS II model exhibited the poorest prediction performance amongst all these models. This condition implies that the SAPS II model requires customisation when applied to a different patient population.

In effect, machine learning-based prediction models have advantages in handling high-dimension data, which indicates that more clinical variables can be considered as model inputs than those used in existing severity scoring systems, with the benefit of discovering meaningful clinical variables that have prediction effects on in-hospital mortality.

Compared with existing studies, our research has several strengths. Firstly, we extracted sepsis patient data from the MIMIC-III dataset using the latest Sepsis-3. Secondly, we used three different machine learning methods LASSO, RF and GBM for prediction model development and compared their prediction performances with those of traditional LR and SAPS II models. The originality of this study lies in clinical application other than methodology. For the first time in the literature, the GBM model was compared with RF, LASSO, LR and SAPS II models in developing in-hospital mortality prediction models for sepsis patient in ICU.

This study also has limitations. Firstly, the MIMIC-III dataset is a dataset containing only patients from a single medical centre. The application of the developed GBM-based model to other datasets or population needs further clinical evaluation. Secondly, the data used for prediction model training and test was a subset of MIMIC-III dataset. The datasets regarding high-resolution waveforms were excluded in the analysis in this study. Thirdly, the developed models were based on the baseline data during the first 24 h in ICU and cannot be used to provide dynamic prediction.

Conclusions

This study contributes to clinical areas with an optimal GBM-based in-hospital mortality prediction model for sepsis patients in ICU. The prediction model has the potential to aid ICU physicians to determine which patients have a high mortality risk and who should be prioritised in treatment, thus enabling them to make optimal clinical interventions and improve prognoses of sepsis patients. As all the variables used in this study for model construction are routinely collected by the information systems in ICU, utilising machine learning-based mortality prediction models to aid ICU physicians in clinical practice is feasible. We plan to computerise a GBM-based tool for predicting the risk of in-hospital death amongst patients with sepsis in the future and integrate it into the existing ICU information system for real-time patient mortality monitoring and clinical decision support. Our future studies would prospectively evaluate the effectiveness of this mortality prediction model and system and check whether it improves the outcome of sepsis patients in clinical practice.

Abbreviations

ICU: Intensive Care Unit; MIMIC: Medical Information Mart for Intensive Care; LASSO: Least Absolute Shrinkage and Selection Operator; GBM: Gradient Boosting Machine; RF: Random Forest; LR: Logistic Regression; AUROC: Area Under the Receiver Operating Characteristic curve; APACHE: Acute Physiology And Chronic Health Evaluation; SAPS: Simplified Acute Physiology Score; SOFA: Sequential Organ Failure Assessment; SSS: Sepsis Severity Score; BIDMC: Beth Israel Deaconess Medical Center; ICD-9-CM: International Classification of Diseases, ninth revision, Clinical Modification

Acknowledgements

The authors would like to thank the reviewers of this manuscript for their valuable comments, which have helped the authors improve the paper significantly.

Declarations of interest

None.

Competing interest

The authors declare that they have no competing interests.

Authors' contributions

Conception and design: GK and KL; Collection and assembly of data: KL, YH and GK; Data analysis and interpretation: KL; Manuscript writing: GK and KL. All authors provided valuable inputs and comments on manuscript revision. The final manuscript was written by GK, and all authors have read and approved the final manuscript.

Funding

This study was supported by the National Natural Science Foundation of China under Grant Nos. 81771938 and 91846101, by Peking University in China under Grant Nos. BMU2018MX020 and PKU2017LCX05, and by the University of Michigan Health System-Peking University Health Science Center Joint Institute for Translational and Clinical Research under Grant No. BMU2020JI011 in the whole research process.

Availability of data and materials

The dataset used in this study, MIMIC III, is available at https://mimic.physionet.org/.

Ethics approval and consent to participate

Research use of MIMIC III has been approved by the Institutional Review Board of the BIDMC and Massachusetts Institute of Technology [33].

Consent for publication

Not applicable.

Competing interests

All authors have declared that they do not have any potential conflicts of interest.

Author details

[1]National Institute of Health Data Science, Peking University, Beijing, China. [2]Center for Data Science in Health and Medicine, Peking University, Beijing, China. [3]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China. [4]Medical Informatics Center, Peking University, Beijing, China.

Received: 13 January 2020 Accepted: 20 September 2020
Published online: 02 October 2020

References

1. Singer M, Deutschman CS, Seymour CW, Shankarhari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J, Coopersmith CM. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016; 315(8):801–10.
2. Zimmerman JJ. Pediatric sepsis from start to finish. Pediatr Crit Care Med. 2015;16(5):479–80.
3. Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, Angus DC, Reinhart K. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. Am J Respir Crit Care Med. 2016;193(3):259–72.
4. Ou L, Chen J, Hillman K, Flabouris A, Parr M, Assareh H, Bellomo R. The impact of post-operative sepsis on mortality after hospital discharge among elective surgical patients: a population-based cohort study. Crit Care. 2017;21(1):34.
5. Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med. 2003;348(16):1546–54.
6. Sheetrit E, Nissim N, Klimov D, Fuchs L, Elovici Y, Shahar Y. Temporal pattern discovery for accurate sepsis diagnosis in ICU patients. arXiv:170901720; 2017.
7. Dremsizov T, Kellum J, Angus D. Incidence and definition of sepsis and associated organ dysfunction. Int J Artif Organs. 2004;27(5):352–9.
8. Andaluz-Ojeda D, Iglesias V, Bobillo F, Almansa R, Rico L, Gandía F, Loma AM, Nieto C, Diego R, Ramos E. Early natural killer cell counts in blood predict mortality in severe sepsis. Crit Care. 2011;15(5):R243.
9. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, Schein RMH, Sibbald WJ. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest. 1992;101(6):1644–55.
10. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent J-L, Ramsay G, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS international Sepsis definitions conference. Intensive Care Med. 2003;29(4):530–8.
11. Vincent J-L, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C, Suter P, Thijs L. The SOFA (Sepsis-related organ failure

assessment) score to describe organ dysfunction/failure. Intensive Care Med. 1996;22:707–10.

12. Le Gall J-R, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/north American multicenter study. JAMA. 1993;270(24):2957–63.
13. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818–29.
14. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A. The APACHE III prognostic system: risk prediction of hospital mortality for critically Ill hospitalized adults. Chest. 1991;100(6):1619–36.
15. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34(5):1297–310.
16. Zygun DA, Laupland KB, Fick GH, Sandham JD, Doig CJ. Limited ability of SOFA and MOD scores to discriminate outcome: a prospective evaluation in 1,436 patients. Can J Anesth. 2005;52(3):302–8.
17. Khwannimit B, Bhurayanontachai R, Vattanavanit V. Validation of the sepsis severity score compared with updated severity scores in predicting hospital mortality in sepsis patients. Shock. 2017;47(6):720–5.
18. Osborn TM, Phillips G, Lemeshow S, Townsend S, Schorr CA, Levy MM, Dellinger RP. Sepsis severity score: an internationally derived scoring system from the surviving sepsis campaign database. Crit Care Med. 2014;42(9):1969–76.
19. Masson S, Caironi P, Spanuth E, Thomae R, Panigada M, Sangiorgi G, Fumagalli R, Mauri T, Isgrò S, Fanizza C. Presepsin (soluble CD14 subtype) and procalcitonin levels for mortality prediction in sepsis: data from the albumin Italian outcome Sepsis trial. Crit Care. 2014;18(1):R6.
20. Adrie C, Francais A, Alvarez-Gonzalez A, Mounier R, Azoulay E, Zahar J-R, Clec'h C, Goldgran-Toledano D, Hammer L, Descorps-Declere A. Model for predicting short-term mortality of severe sepsis. Crit Care. 2009;13(3):R72.
21. Ripoll VJR, Vellido A, Romero E, Ruiz-Rodríguez JC. Sepsis mortality prediction with the quotient basis kernel. Artif Intell Med. 2014;61(1):45–52.
22. Fang W-F, Douglas IS, Chen Y-M, Lin C-Y, Kao H-C, Fang Y-T, Huang C-H, Chang Y-T, Huang K-T, Wang Y-H. Development and validation of immune dysfunction score to predict 28-day mortality of sepsis patients. PLoS One. 2017;12(10):e0187088.
23. Xie Y, Yabes J, Parker R, Clermont G. 1485: using clinical features and biomarkers to predict 60-day mortality of sepsis patients. Crit Care Med. 2018;46(1):726.
24. Richards G, Rayward-Smith VJ, Sönksen P, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med. 2001;22(3):215–31.
25. Mahdi MA, Al_Janabi S. A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers, vol. 2020. Cham: Springer International Publishing; 2020. p. 320–39.
26. Al-Janabi S, Mahdi MA. Evaluation prediction techniques to achievement an optimal biomedical analysis. Int J Grid Utility Comput. 2019;10(5):512–27.
27. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
28. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2000;29(5):1189–232.
29. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
30. Gurm HS, Kooiman J, LaLonde T, Grines C, Share D, Seth M. A random forest based risk model for reliable and accurate prediction of receipt of transfusion in patients undergoing percutaneous coronary intervention. PLoS One. 2014;9(5):e96385.
31. Van Poucke S, Zhang Z, Schmitz M, Vukicevic M, Laenen MV, Celi LA, De Deyne C. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PLoS One. 2016;11(1):e0145791.
32. Zhang Z, Hong Y. Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: the use of electronic healthcare records with LASSO regression. Oncotarget. 2017;8(30):49637–45.
33. 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. 2016;3:160035.
34. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with Sepsis: a local big data-driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269–78.
35. Pregernig A, Müller M, Held U, Beck-Schimmer B. Prediction of mortality in adult patients with sepsis using six biomarkers: a systematic review and meta-analysis. Ann Intensive Care. 2019;9(1):125.
36. Baxt WG. Complexity, chaos and human physiology: the justification for non-linear neural computational analysis. Cancer Lett. 1994;77(2–3):85–93.
37. Scott H, Colborn K. Machine learning for predicting sepsis in-hospital mortality: an important start. Acad Emerg Med. 2016;23(11):1307.
38. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med. 2001;29(7):1303–10.
39. Ding X, Gellad ZF, Mather C, Barth P, Poon EG, Newman M, Goldstein BA. Designing risk prediction models for ambulatory no-shows across different specialties and clinics. J Am Med Inform Assoc. 2018;25(8):924–30.
40. Venables WN, Ripley BD. Tree-based methods. In: Modern applied statistics with S. New York: Springer-Verlag; 2002. p. 251–69.
41. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform. 2019; 125:55–61.
42. Lee H-C, Yoon H-K, Nam K, Cho Y, Kim T, Kim W, Bahk J-H. Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. J Clin Med. 2018;7(10):322.
43. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer-Verlag; 2001.
44. Roulston MS. Performance targets and the brier score. Meteorol Appl. 2007; 14(2):185–94.
45. Jiménez-Valverde A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob Ecol Biogeogr. 2012;21(4):498–507.
46. Dietterich TG. Ensemble methods in machine learning, vol. 2000. Berlin: Springer Berlin Heidelberg; 2000. p. 1–15.
47. Fang X, Wang Z, Yang J, Cai H, Yao Z, Li K, Fang Q. Clinical evaluation of Sepsis-1 and Sepsis-3 in the ICU. Chest. 2018;153(5):1169–76.
48. Newgard CD, Lewis RJ, Jolly BT. Use of out-of-hospital variables to predict severity of injury in pediatric patients involved in motor vehicle crashes. Ann Emerg Med. 2002;39(5):481–91.
49. Johnson AE, Mark RG. Real-time mortality prediction in the Intensive Care Unit. In: AMIA Annual Symposium Proceedings: 2017. Washington, D.C.: American Medical Informatics Association; 2017. p. 994.
50. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: ICML'06: Proceedings of The 23rd International Conference on Machine Learning, vol. 2006. New York: Association for Computing Machinery; 2006. p. 161–8.

Publisher's Note

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