Research on Time Window Prediction and Scoring Model for Trauma-Related Sepsis

Ke Luo[1], Jing Li[1(✉)], and Yuzhuo Zhao[2]

[1] School of Economics and Management, Beijing Jiaotong University, Beijing, China
{17711020,jingli}@bjtu.edu.cn
[2] Department of Emergency, Chinese PLA General Hospital, Beijing, China

Abstract. Based on the MIMIC-III database of the Massachusetts Institute of Technology, this paper studies and analyzes the symptoms of trauma-related sepsis. Use SOFA score as the Inclusion and Exclusion Criteria, extract the relevant patient medical index data with the guidance of a professional clinician. Sequential forward search is applied to search the optimal index combination based on the eXtreme Gradient Boosting (XGBoost) algorithm. Twenty independent replicates perform to obtain 7 key risk indicators (Urea Nitrogen, Prothrombin Time, PO2, Sodium, Red Blood Cells, Carbon Dioxide, International Normalized Ratio). The time window prediction model builds by four machine learning algorithms (decision tree, random forest, decision tree-based adaptive reinforcement (Adaboost) algorithm, XGBoost). The results show that the time window prediction model of trauma-related sepsis has good generalization ability. The prediction effect of the random forest and XGBoost algorithm is better than the other two. Finally, using the multi-factor Logistic regression method build the risk scoring tool for sepsis-induced by trauma-related infection base on the key risk indicators and the opinions of professional clinicians. The results show that the data-driven risk scoring tool can effectively predict the outcome of patients with trauma-related sepsis, which has high clinical significance.

Keywords: Trauma-related Sepsis · Big data · Key risk indicators · Time window forecast · Risk score · Machine learning

1 Introduction

With the development of medicine and the application of Smart healthcare, many diseases are under control. But the number of trauma patients is increasing, and the number of deaths is also increasing. Trauma has become the first cause of death for patients between 1 and 44 years old [1]. Trauma-related infection is one of the most common complications of trauma patients. Due to the difference in traumatic environment and degree of trauma, most trauma patients will have different degrees of infection, among which sepsis is

This work was partly supported by the National Key Research and Development Plan for Science and Technology Winter Olympics of the Ministry of Science and Technology of China (2019YFF0302301).

X. Shi et al. (Eds.): *LISS 2021*, LNOR, pp. 111–123, 2022.
https://doi.org/10.1007/978-981-16-8656-6_10

the result of out-of-control infection of trauma patients in the late stage. Difficulty in predicting, diagnosis, and treatment are the reasons for the extremely high mortality of sepsis. It can highly improve the treatment rate of sepsis if sepsis prediction in the early stage. Many researchers applied Smart healthcare in the study of sepsis, combined with the computer, big data, and clinical medicine, analyze the risk factors for sepsis, trying to find out the relationship between the specific indicators and sepsis, focus on the point in a certain time of sepsis judgment, lack of the forecast on time interval judgment [2]. This paper will focus on the possibility of sepsis in trauma patients over time. Combined with the MIMIC-III database, this paper will conduct data mining and statistical analysis, and machine-learning algorithms to construct time window prediction and risk scoring tools for trauma-related sepsis to reduce the risk of sepsis and improve the working efficiency of medical staff.

2 Methods

2.1 Study Population and Data Sources

All data used in this study obtained from the Medical Information Mart for Intensive Care III database (MIMIC-III). The data are patients aged 18 years or older who had been in the ICU for 4 h or more due to trauma. Blood culture should be performed within 24 h if the patient uses antibiotics first. If the patient is performed within blood culture first, antibiotics should use within 72 h, and the priority project time recorded as $T_{suspicion}$. When the SOFA score of the patients at 12 h after $T_{suspicion}$ minus the SOFA score at 24 h before $T_{suspicion}$ is higher than 2, the patients are identified as having sepsis, that is, the experimental group, otherwise for the control group.

2.2 Data Processing

The SOFA score is used as the inclusion and exclusion criteria of the research experiment. Data is preprocessed by data transpose, outlier processing, missing value analysis, and data filling. Missing no module of Python is used as the main tool for missing value analysis, and indicators with a missing ratio of more than 80% are removed. And then time series data are filled based on two dimensions, namely linear interpolation and distance filling.

2.3 Feature Selection and Machine Learning

Feature Selection is the preliminary step of machine learning and data mining, and it is a process of data preprocessing. It eliminates redundant or irrelevant features to identify the most important features, thus reducing the complexity of the problem [3]. In this study, the greedy algorithm is used to design a feature selection algorithm, and the XGBoost algorithm is used to select features of 35 indicators by sequential forward search strategy search. XGBoost has good anti-over-fitting characteristics and high computational efficiency [4]. The tree model of XGBoost is characterized by providing a basis for quantitative feature selection and forming encapsulated feature selection. The

time-series data of trauma-related sepsis are input into the key indicator screening model for iteration, and the results of each iteration are recorded to select the index with the highest performance.

Compared with the black-box model (uninterpreted algorithm taking neural network as an example), the Decision Tree is based on if-then-else rules and is easier to understand, realize, explain and visualize [5]. The neural network (the black-box model representation) has certain defects: difficult to optimize, result in the local-optimal solution rather than the global-optimal solution, and low generalization leads to overfitting problems, etc. To sum up, this study uses the decision tree algorithm to build the time window prediction model. This study also uses the random forest and Boosting method which is derived from the decision tree to carry out multiple groups of experiments, to improve the accuracy of time window prediction of trauma-related sepsis [6].

In this study, grid parameter iteration is used for parameter adjustment. Given a set of parameters, the enumeration search method is used to iterate over all possibilities to select the best result (Fig. 1).

```
def train_model(k, name, feature, n_cluster, timestep, delay, paths, path1):
    list_y_true, list_y_pred=[], []
    results=[]
    if  name=='CART':
        clf=DecisionTreeClassifier(criterion='gini')
        param_grid={'max_depth':range(1,5),'max_features':range(5,10)}
    elif name=='RF':
        clf=RandomForestClassifier(criterion='gini')
        param_grid={'n_estimators':range(5,10,15),'max_depth':range(1,10)}
    elif name=='Adb':
        clf=AdaBoostClassifier(algorithm='SAMME.R',base_estimator=None,random_state=None)
        param_grid={'n_estimators':range(50,100,150),'learning_rate':[0.5,1,1.5,2]}
    elif name=='XGB':
        clf=XGBClassifier(subsample=0.8,colsample_bytree=1,objective='binary:logistic',
                          min_child_weight=1,scale_pos_weight=1)
        param_grid={'n_estimators':range(100,150,200),'max_dapth':[1,6],'learning_rate':[0.5,1,1.5],'gamma':[0.01,0.1,1]}
```

Fig. 1. Time prediction model parameter adjustment and grid search setting.

The logistic regression model is a multivariate statistical method to study the relationship between explanatory variables and observed results. In this study, the scoring tool is based on the key index set, using the multi-factor logistic regression method and the clinical grading consensus of the indicators to quantitatively calculated the severity of the patient's illness, namely the score. And then the score is corresponding to the outcome probability. The specific steps of model construction are as follows:

1) Calculate the risk index regression coefficient β of each index;
2) Combined with the inherent medical knowledge to determine the scoring threshold of each index, and determine the reference value w_{ij} of each group;
3) The basic risk reference value w_{iREF} of each risk indicator is determined. In the subsequent scoring model construction, w_{iREF} is recorded as 0 points, and when it is higher than w_{iREF}, it is recorded as positive points; otherwise, the higher the score, the higher the risk is.
4) Calculate the distance D between the reference value w_{ij} of each risk indicator and basic risk reference value w_{iREF}:

$$D = \left(w_{ij} - w_{iREF}\right) * \beta_i \tag{1}$$

5) Set constant B, the change value of the index corresponding to the change of 1 point in the risk scoring tool;
6) Calculate the score $Points_{ij}$, of each group of risk indicators, and round the calculated value as the corresponding score value of this group:

$$Points_{ij} = \frac{D}{B} = \left(w_{ij} - w_{iREF}\right) * \beta_i / B \tag{2}$$

7) Calculate the total score and risk prediction probability:

$$\hat{p} = \frac{1}{1 + \exp(\beta_i X_i)} \tag{3}$$

3 Results

This study uses PostgreSQL to extract data from the MIMIC-III database. And finally obtained data of 177 patients in the experimental group and 369 patients in the control group, with 35 various examination and laboratory indicators, amount to 201189 records.

The Hosmer Lemeshow goodness of fit index (H-L) [7] is used to verify the time series data after filling, and the significance of the result is 0.553 which greater than 0.05. There is no significant difference between the predicted value and the observed value, which proves that the model has a good fit.

3.1 Key Risk Indicators

The key indexes of trauma-related sepsis are extracted by feature selection. After 20 separate repeated experiments, the key indexes with retention times more than or equal to 16 times are Urea Nitrogen, PTT, PO2, Sodium, Red Blood Cells, CO2, and INR. The key indicators for retention between 12 and 16 times are Lactate and White Blood Cells. Hematocrit, Chloride, Hemoglobin, Temperature, Base Excess are the key indicators of retention between 10 and 12 times. Key indicators with retention times between 8 and 10 are Heart Rate, PCO2, Glucose, Platelet, Creatinine, and Calcium. The more times a single indicator is retained, the greater the influence of this indicator on the outcome and the stronger the ability to identify patients.

In this study, the feature_importances function of XGBoost is used to obtain the characteristic importance of each index. The results are shown in the following Table 1.

Compared with the calculation results, find that Urea Nitrogen and PTT, with the most retention times of key indicators, rank the first and the sixth in the ranking of characteristic importance, respectively, which proved that there is a causal relationship between key indicators and the accuracy of the predicted results of trauma-related sepsis. The weight of characteristic importance does not represent the correlation degree between the indicators and the predicted results, but it proves the correlation between the indicators and the predicted results, which lays a foundation for the subsequent prediction of trauma-related sepsis (Table 2).

Table 1. Indicator weights and rankings

Rank	Label	Feature_importances	Rank	Label	Feature_importances
1	Urea Nitrogen	0.06	14	Heart Rate	0.039
2	Hemoglobin	0.059	15	Calcium	0.038
3	Lactate	0.052	16	Chloride	0.037
4	CO2	0.049	17	Potassium	0.036
5	INR	0.048	18	Magnesium	0.035
6	PTT	0.047	19	Red Blood Cells	0.033
7	PO2	0.046	20	Base Excess	0.031
8	White Blood Cells	0.046	21	Hematocrit	0.03
9	Glucose	0.045	22	ph	0.03
10	Creatinine	0.045	23	Respiratory Rate	0.024
11	Platelet	0.042	24	Systolic Pressure	0.02
12	Sodium	0.041	25	Diastolic Pressure	0.016
13	PCO2	0.04	26	Temperature	0.01

Table 2. Forecast the distribution of key indicators

Category	Vital signs	Coagulation function	Arterial blood gas	Blood routine	Blood biochemical
Number	2	2	4	5	7
Total Weight	0.049	0.095	0.169	0.21	0.308

By summing up the weights of all the key indicators, the calculation results show that the weight summation of Blood Biochemical and Blood Routine is the highest, and the distribution is 0.308 and 0.21. To a certain extent, it proved the important value of Blood Biochemical and Blood Routine in the prediction of trauma-related sepsis, followed by Arterial Blood Gas and Coagulation Function, and Vital Signs had the least influence in the prediction of trauma-related sepsis (Fig 2).

3.2 Time Window Forecast

Under the time window model of the full index data set, the accuracy rate, recall rate, and precision rate of the four model algorithms are between 64% and 83%, which meet the requirements of clinical medicine. The best prediction effect is the Random Forest. From the perspective of time, although the prediction effect fluctuates slightly, the overall accuracy decreases with the increase of time, which is in line with the actual prediction logic. Moreover, the overall model performance increase with time, but the changing trend is not obvious, which proved the stability of the model and is more conducive to the earlier prediction and early warning of trauma-related sepsis (Table 3, Fig. 3).

Table 3. Comparison of prediction time parameters of all indexes in different Time

Predicted time	Method	F1.5	Acc	Pre	Rec
1 h	Decision Tree	0.6645	0.6452	0.6364	0.6785
	Random Forest	0.8119	0.7941	0.7762	0.8299
	Adaboost	0.6818	0.6681	0.6618	0.6921
	XGBoost	0.7770	0.7492	0.7266	0.8028
2 h	Decision Tree	0.6631	0.6514	0.6464	0.6718
	Random Forest	0.8123	0.7881	0.7637	0.8367
	Adaboost	0.6586	0.6562	0.6559	0.6610
	XGBoost	0.7753	0.7452	0.7209	0.8028
3 h	Decision Tree	0.6559	0.6427	0.6366	0.6655
	Random Forest	0.8099	0.7895	0.7704	0.8305
	Adaboost	0.6808	0.6633	0.6541	0.6944
	XGBoost	0.7718	0.7492	0.7331	0.7921
4 h	Decision Tree	0.6598	0.6511	0.6477	0.6667
	Random Forest	0.8073	0.7845	0.7634	0.8299
	Adaboost	0.6752	0.6684	0.6653	0.6802
	XGBoost	0.7690	0.7469	0.7296	0.7887

According to the model AUC value, all the models are higher than 0.64 under different time window parameter Settings, which can meet the dynamic requirements (Table 4).

Under the time window model of "Key Indicator Set 1–4", the performance results of each machine learning method are still above 63%, and the accuracy, recall, and precision rate of Random Forest are all the best. Although the effect of different indicators fluctuated slightly, in general, the prediction effect decreased with the decrease of the number of key indicators (Fig. 4).

According to the model AUC value, all models are higher than 0.63 under different time window parameter Settings, among which Random Forest performed the best, followed by XGBoost, Adaboost, and Decision Tree. With the decrease of the amount

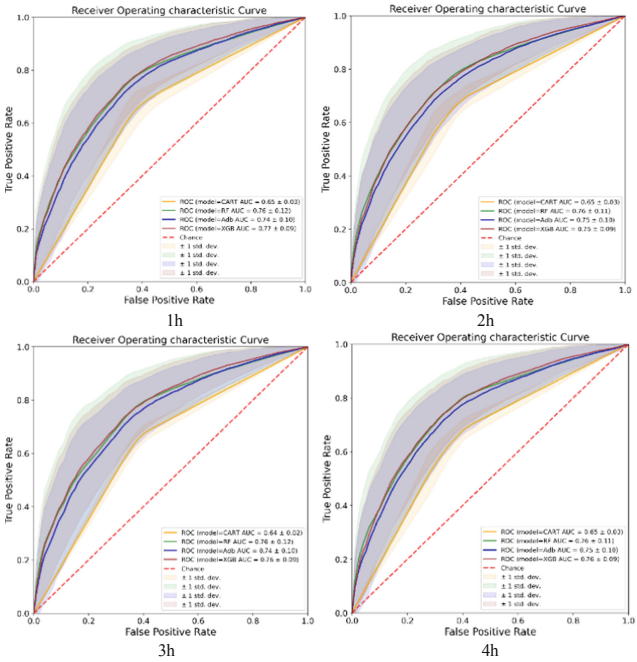

Fig. 2. ROC curve and AUC of internal validation of each model in the full index data set.

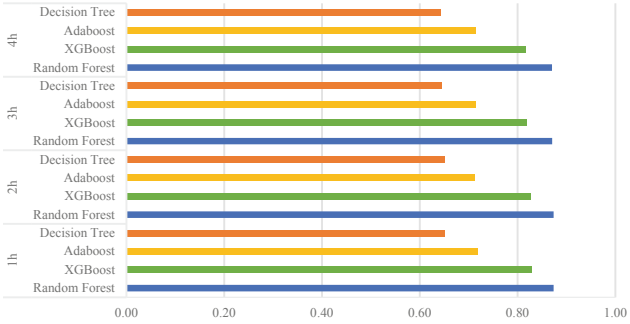

Fig. 3. AUC comparison of internal validation among models in the full indicator set.

of key index data, the prediction effect has a certain tendency to decrease, but it is not obvious, which does not affect the application requirements of the dynamic real-time time window prediction model for trauma-related sepsis, which is in line with clinical practice, and proves the generalization ability of the time window prediction model for trauma-related sepsis (Fig. 5).

Table 4. Comparison of prediction time parameters of key index datasets in different sets

Method	Set	N	F1.5	Acc	Pre	Rec
Decision Tree	1	20	0.6468	0.6339	0.6284	0.6559
Random Forest	1	20	0.8132	0.7910	0.7695	0.8356
Adaboost	1	20	0.6774	0.6655	0.6585	0.6870
XGBoost	1	20	0.7834	0.7545	0.7317	0.8102
Decision Tree	2	14	0.6645	0.6506	0.6449	0.6740
Random Forest	2	14	0.8106	0.7881	0.7672	0.8328
Adaboost	2	14	0.6748	0.6562	0.6473	0.6881
XGBoost	2	14	0.7729	0.7486	0.7279	0.7955
Decision Tree	3	9	0.6817	0.6610	0.6509	0.6966
Random Forest	3	9	0.8083	0.7898	0.7716	0.8266
Adaboost	3	9	0.6887	0.6740	0.6669	0.7006
XGBoost	3	9	0.7873	0.7565	0.7317	0.8164
Decision Tree	4	7	0.6834	0.6548	0.6424	0.7040
Random Forest	4	7	0.7966	0.7712	0.7485	0.8209
Adaboost	4	7	0.6655	0.6455	0.6363	0.6797
XGBoost	4	7	0.7658	0.7398	0.7200	0.7893

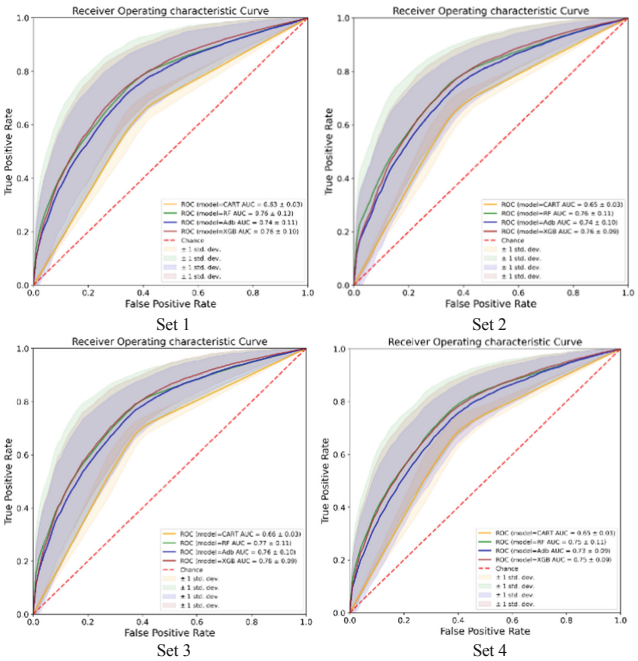

Fig. 4. ROC curve and AUC of internal validation of each model in different sets.

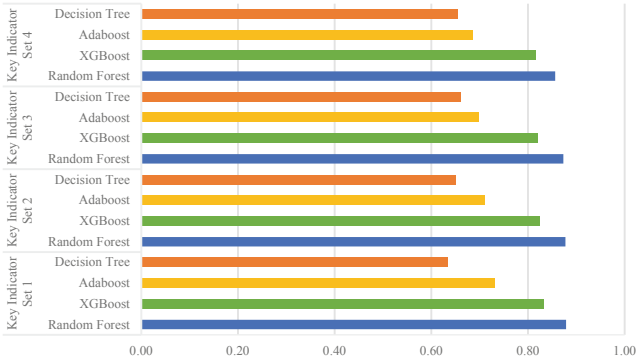

Fig. 5. AUC comparison of internal validation among models in the key indicator set.

3.3 Risk Score

In this study, the two indexes with the highest weighted sum of feature importance, blood biochemical and blood routine, are selected to construct a risk scoring tool. Creatinine is taken as a constant reference index, and the results are as follows (Table 5).

Table 5. Trauma-related sepsis risk scoring tool

Key indicators	Group	w_{ij}	w_{iREF}	β	D	B	S
Platelet	$0 \leq x < 20$	10		−0.0107	2.5650	1.0185	3
	$20 \leq x < 50$	35			2.2978	1.0185	2
	$50 \leq x < 100$	75			1.8703	1.0185	2
	100 ≤ x < 400	***250***	***250***		**0**	1.0185	0
	$x \geq 400$	*658*			−4.3605	1.0185	−4
Creatinine	$0 \leq x < 88$	44		0.0078	−0.6894	1.0185	−1
	88 ≤ x < 176	***132***	***132***		**0**	1.0185	0
	$176 \leq x < 308$	242			0.8618	1.0185	1
	$308 \leq x < 440$	374			1.8960	1.0185	2
	$x \geq 440$	410			2.1780	1.0185	2
Urea Nitrogen	$0 \leq x < 1.8$	1.79		0.0942	0.6701	1.0185	1
	1.8 ≤ x < 7.1	***8.9***	***8.9***		**0**	1.0185	0
	$7.1 \leq x < 21$	14			0.4806	1.0185	0
	$x \geq 21$	25.5			1.5645	1.0185	2

(*continued*)

Table 5. (*continued*)

Key indicators	Group	w_{ij}	w_{iREF}	β	D	B	S
Sodium	$0 \leq x < 135$	107.5		−0.0377	1.2239	1.0185	1
	$135 \leq x \leq 145$	***140***	***140***		**0**	1.0185	0
	$x > 145$	155			0.5649	1.0185	1
RBC	$0 \leq x < 2$	1		0.3224	1.2088	1.0185	1
	$2 \leq x < 4$	3			0.5641	1.0185	1
	4 ≤ x < 5.5	***4.75***	***4.75***		**0**	1.0185	0
	$x \geq 5.5$	4.48			0.0870	1.0185	0
Chloride	$0 \leq x < 75$	60		−0.055	2.2099	1.0185	2
	$75 \leq x < 95$	85			0.8287	1.0185	1
	95 ≤ x < 105	***100***	***100***		**0**	1.0185	0
	$105 \leq x < 125$	115			−0.8287	1.0185	−1
	$x \geq 125$	122			−1.2155	1.0185	−1
CO2	$0 \leq x < 23$	20.5		−0.1138	0.7397	1.0185	1
	$23 \leq x \leq 31$	***27***	***27***		**0**	1.0185	0
	$x > 31$	34			0.7966	1.0185	1
WBC	$0 \leq x < 2$	1		−0.1624	0.9741	1.0185	1
	$2 \leq x < 4$	3			0.6494	1.0185	1
	4 ≤ x < 10	***7***	***7***		**0**	1.0185	0
	$10 \leq x < 20$	15			−1.2988	1.0185	−1
	$x \geq 20$	26			−3.0847	1.0185	−3
Hematocrit	$0 \leq x < 20$	20.5		0.0943	2.1699	1.0185	2
	20 ≤ x < 37	*28.5*			1.4152	1.0185	1
	37 ≤ x < 50	***43.5***	***43.5***		**0**	1.0185	0
	$x \geq 50$	46			0.2359	1.0185	0
Calcium	$0 \leq x < 2.06$	1.03		−0.4104	0.5335	1.0185	1
	$2.06 \leq x \leq 2.6$	***2.33***	***2.33***		**0**	1.0185	0
	$x > 2.6$	2.5			0.0698	1.0185	0
Hemoglobin	$0 \leq x < 80$	60		0.0309	2.3207	1.0185	2
	$80 \leq x < 110$	95			1.2377	1.0185	1
	110 ≤ x < 160	***135***	***135***		**0**	1.0185	0
	$x \geq 160$	151			−0.4951	1.0185	0
Glucose	$0 \leq x < 3.9$	1.95		0.2356	0.7186	1.0185	1
	3.9 ≤ x < 6.1	***5***	***5***		**0**	1.0185	0
	$6.1 \leq x < 11.1$	8.6			0.8482	1.0185	1
	$x \geq 11.1$	12.7			1.8141	1.0185	2

Taking a wounded patient as an example, the sum of all index scores is 8, so the probability of the patient suffering from sepsis at that moment is 91.58% (Table 6).

Table 6. Comparison table of the total score and risk prediction probability of trauma-related sepsis

Score	Probability	Score	Probability
−9	0.00003%	6	58.65514%
−8	0.00009%	7	79.70920%
−7	0.00025%	8	91.58080%
−6	0.00070%	9	96.78666%
−5	0.00193%	10	98.81521%
−4	0.00535%	11	99.56886%
−3	0.01482%	12	99.84387%
−2	0.04103%	13	99.94356%
−1	0.11353%	14	99.97961%
0	0.31374%	15	99.99264%
1	0.86396%	16	99.99734%
2	2.35630%	17	99.99904%
3	6.26353%	18	99.99965%
4	15.61375%	19	99.99987%
5	33.87739%	20	99.99995%

Input the time series data of trauma patients into the risk scoring model can get the change of the risk probability of the trauma patients suffering from sepsis, which reflects the increase in the severity of trauma-related infection over time. Taking a certain trauma patient as an example, the data from 4 h before to 24 h after the treatment of antibiotics showed that the patient's condition improved after the treatment of antibiotics, but only lasted for a while. The subsequent changes in the patient's condition are not detected in time, leading to a rapid increase in the possibility of the patient suffering from sepsis (Table 7, Fig. 6).

Table 7. Changes of risk scores and risk probability in a period of A trauma patienT

Time	Score	Probability
10:00	8	91.58080%
12:00	8	91.58080%

(*continued*)

Table 7. (*continued*)

Time	Score	Probability
14:00	7	79.70920%
16:00	4	15.61375%
18:00	3	6.26353%
20:00	3	6.26353%
22:00	3	6.26353%
0:00	3	6.26353%
2:00	1	0.86396%
4:00	4	15.61375%
6:00	8	91.58080%
8:00	11	99.56886%
10:00	11	99.56886%
12:00	13	99.94356%
14:00	14	99.97961%

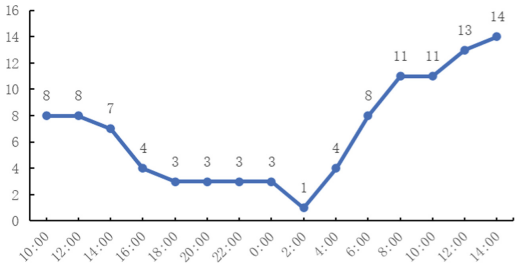

Fig. 6. Changes of risk scores in a period of one trauma patient.

The AUC value of the trauma-related sepsis risk scoring tool is 0.79, which proved that the model has a certain generalization ability and can meet the clinical needs.

4 Conclusion

Sepsis is one of the main causes of death in patients with trauma-related infection in clinical practice. The onset of sepsis is serious and develops rapidly. Once diagnosed, a large amount of time for diagnosis and treatment has been lost, and it is difficult to control the development of sepsis. At the same time, due to the particularity of sepsis, the diagnosis process is complicated and needs a long time, which leads to the delay in the judgment of trauma-related sepsis by medical staff. Therefore, this study realizes the trauma-related sepsis predict warning, as well as the risk score. The model has good generalization ability, and can basically meet the clinical practical application, could

help doctors perceive ahead of trauma-related sepsis in patients development trend and extent, thus early medical intervention on the patients, Controlling the development of sepsis will greatly improve the treatment rate of sepsis.

The next step is to validate the model in clinical trials. The diagnosis of sepsis needs to base on the changes of various indicators of patients over some time, so the data obtained in this study is limited, and the conditions of sepsis patients are diverse. As the limited data cannot accurately display all the clinical manifestations of sepsis, the model obtained in this study cannot achieve higher performance. More effective and real clinical data can optimize and improve the model.

Acknowledgment. Thanks to Dr. Zhao of the Chinese PLA General Hospital for his help. Medical big data analysis cannot separate from the support of clinical medicine. Dr. Zhao guided and verified my research content with professional clinical experience, which made me realize the clinical significance of the research content fundamentally. I would like to express my gratitude to Ms. Li Jing of the Intelligent Diagnostic Team of the Winter Olympics for her help and guidance in the whole research stage of my thesis.

References

1. Brunicardi, F., et al.: Schwartz's Principles of Surgery, 10th edn., p. 161. McGraw-Hill Education, New York (2015)
2. Desautels, T., et al.: Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med. Inf. **4**(3), e28 (2016)
3. Lieber, D., Stolpe, M., Konrad, B., Deuse, J., Morik, K.: Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. Procedia CIRP **7**, 193–198 (2013)
4. Jalal Mussa, D., Jameel, N.M.: Relevant SMS spam feature selection using wrapper approach and XGBoost algorithm. Kurdistan J. Appl. Res. **4**(2), 110–120 (2019)
5. Tofan, C.: Optimization techniques of decision making - decision tree. Adv. Soc. Sci. Res. J. **1**(5), 142–148 (2014)
6. Sagi, O., Rokach, L.: Approximating XGBoost with an interpretable decision tree. Information Sciences (2021)
7. Hosmer, D., Lemesbow, S.: Goodness of fit tests for the multiple logistic regression model. Commun. Stat. Theory Methods **9**(10), 1043–1069 (1980)