An Explainable Artificial Intelligence Predictor for Early Detection of Sepsis

Meicheng Yang, BS[1]; Chengyu Liu, PhD[1]; Xingyao Wang, MS[1]; Yuwen Li, PhD[1]; Hongxiang Gao, BS[1]; Xing Liu, MD, PhD[2]; Jianqing Li, PhD[1,3]

Objectives: Early detection of sepsis is critical in clinical practice since each hour of delayed treatment has been associated with an increase in mortality due to irreversible organ damage. This study aimed to develop an explainable artificial intelligence model for early predicting sepsis by analyzing the electronic health record data from ICU provided by the PhysioNet/Computing in Cardiology Challenge 2019.

Design: Retrospective observational study.

Setting: We developed our model on the shared ICUs publicly data and verified on the full hidden populations for challenge scoring.

Patients: Public database included 40,336 patients' electronic health records sourced from Beth Israel Deaconess Medical Center (hospital system A) and Emory University Hospital (hospital system B). A total of 24,819 patients from hospital systems A, B, and C (an unidentified hospital system) were sequestered as full hidden test sets.

Interventions: None.

Measurements and Main Results: A total of 168 features were extracted on hourly basis. Explainable artificial intelligence sepsis predictor model was trained to predict sepsis in real time. Impact of each feature on hourly sepsis prediction was explored in-depth to show the interpretability. The algorithm demonstrated the final clinical utility score of 0.364 in this challenge when tested on the full hidden test sets, and the scores on three separate test sets were 0.430, 0.422, and −0.048, respectively.

Conclusions: Explainable artificial intelligence sepsis predictor model achieves superior performance for predicting sepsis risk in a real-time way and provides interpretable information for understanding sepsis risk in ICU. (*Crit Care Med* 2020; 48:e1091–e1096)

[1]The State Key Laboratory of Bioelectronics, School of Instrument Science and Engineering, Southeast University, Nanjing, China.

[2]Department of Anesthesiology, The third Xiangya Hospital, Central South University, Changsha, China.

[3]School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China.

DOI: 10.1097/CCM.0000000000004550

Key Words: artificial intelligence; intensive care unit; PhysioNet challenge; prediction; sepsis

Sepsis is a serious complication of infection in emergency department and represents a major cause of maternal and neonatal morbidity and mortality (1). Reliable and early detection of sepsis is clinically important, facilitating the active antibiotic therapy and fluid resuscitation (2–4). Artificial intelligence (AI) algorithms have been adopted in learning electronic health record (EHR) data for early detection of sepsis or septic shock (5–8). However, interpretability for the developed models faces a difficulty in clinic, resulting in the poor practicality for clinical decision support.

The PhysioNet/Computing in Cardiology (CinC) Challenge 2019 addressed this issue and promoted the development of open-source AI algorithms for real-time and early detection of sepsis (9). In this study, we trained an explainable AI sepsis predictor (EASP) to predict sepsis risk hour-by-hour and focused on its interpretability for the clinical EHR data sourced from ICU patients. The EASP model was verified on the full hidden test data from separate hospital systems.

METHODS

Data and Features

Data were retrieved from the PhysioNet/CinC Challenge 2019 (9), which provided a total of 40,336 patients' EHR data (2,932 septic and 37,404 nonseptic) from three separate hospital systems for public, as well as a hidden test set from 24,819 patients. We randomly split 85% of the public data (34,285 patients, 2,492 septic and 31,793 nonseptic) for algorithm development and 15% (6,051 patients, 440 septic and 5,611 nonseptic) for validation (**Table 1**).

All raw variables were used as input features (total 37) of model except three ones (direct bilirubin, troponin I, and fibrinogen), due to their massive missing values. One hundred thirty-one features were derived from the raw variables and were classified as three subtypes: 62 informative missingness features reflecting measurement frequency or time interval of

TABLE 1. Characteristics of Population for Algorithm Development and Validation

Characteristics	Development Population			Validation Population		
	Septic (*n* = 2,492)	Nonseptic (*n* = 31,793)	*p*[a]	Septic (*n* = 440)	Nonseptic (*n* = 5,611)	*p*[a]
Age, median (IQR), yr	64 (52–73)	63 (51–74)	0.14	63 (50–75)	63 (51–74)	0.61
Male, %	59.0	55.7	< 0.01	60.9	55.4	< 0.01
ICU length of stay (hours since ICU admit), median (IQR), hr	38 (15–82)	39 (25–47)	< 0.01	39 (17–82)	39 (25–47)	< 0.05
Hours between hospital admit and ICU admit, median (IQR), hr	−2.6 (−63.7 to 0.0)	−6.2 (−42.2 to −0.1)	< 0.01	−4.9 (−86.2 to 0.0)	−6.4 (−43.0 to −0.1)	0.85
Administrative identifier for ICU unit, *n* (%)			< 0.01			< 0.01
Unit 1 (medical ICU)	833 (33.4)	9,608 (30.2)		133 (30.2)	1,693 (30.2)	
Unit 2 (surgical ICU)	550 (22.1)	9,990 (31.4)		100 (22.7)	1,812 (32.3)	
Unidentified	1,109 (44.5)	12,195 (38.4)		207 (47.1)	2,106 (37.5)	

IQR = interquartile range.

[a]Statistical analyses using a χ^2 for binary variables and the Wilcoxon rank-sum test for all continuous variables. *p* values < 0.05 were considered significant.

TABLE 2. Top 20 Features Contributing to Prediction and Corresponding Impact

Features	Description	Type	Impact
ICULOS	ICU length of stay (hr)	R	0.2718
HospAdmTime	Time between hospital and ICU admission (hr)	R	0.2219
Temp	Temperature (°C)	R	0.2182
FIO_2	Fraction of inspired oxygen (%)	R	0.2065
FIO_2_interval	Measurement interval of FIO_2 (hr)	D1	0.1657
Lactate	Lactic acid (mg/dL)	R	0.1196
WBC	Leukocyte count (count/L)	R	0.1091
Creatinine	Creatinine (mg/dL)	R	0.1052
Unit1	Medical ICU (0/1)	R	0.0967
BUN	Blood urea nitrogen (mg/dL)	R	0.0817
HR_frequency	Measurement frequency of heart rate	D1	0.0779
Alkalinephos	Alkaline phosphatase (IU/L)	R	0.0772
MAP_window_mean	Mean of mean arterial pressure in a 6-hr window (mm Hg)	D2	0.0738
HR_window_max	Maximum of heart rate in a 6-hr window (beats/min)	D2	0.0716
SBP_window_diffstd	SD after first difference of systolic blood pressure in a 6-hr window (mm Hg)	D2	0.0651
PTT	Partial thromboplastin time (s)	R	0.0577
Resp_window_mean	Mean of respiration rate in a 6-hr window (breaths/min)	D2	0.0564
$ETCO_2$	End-tidal carbon dioxide (mm Hg)	R	0.0555
HR	Heart rate (beats/min)	R	0.0535
Temp_diff	Difference between the current record and last temperature (°C)	D2	0.0527

D = derived features and D1, D2 represent informative missingness, time series features, respectively, R = raw features.

raw variables (10), 61 time series features including the differences between the current record and the previous value, statistics (maximum, minimum, mean, median, SD and differential SD) in a 6-hour sliding window for the selected measurements (heart rate [HR], pulse oximetry, systolic blood pressure [SBP], and respiration rate [Resp]), and eight empiric features scoring for HR, SBP, mean arterial pressure, Resp, temperature, creatinine, platelets, and total bilirubin according to the scoring systems of NEWS (11), Sequential Organ Failure Assessment (SOFA) (12) and quick SOFA (12). Thus, a total of 168 features were obtained. Missing values used a forward-filling strategy.

Algorithm Design

A high performance gradient-boosting-trees model, that is, XGBoost (13), was applied to train the model since its supporting for flexible and complex nonlinear learning. *K*-fold cross validation ($K = 5$) was implemented during training, and five XGBoost models were produced. Ensemble approach by averaging prediction risks from the five models was used for robust determination. Model variables were tuned using a Bayesian optimizer (14). Impacts of features on risk output were quantified by Shapley values (15) to obtain instant interpretability for the developed EASP model. Shapley value was computed

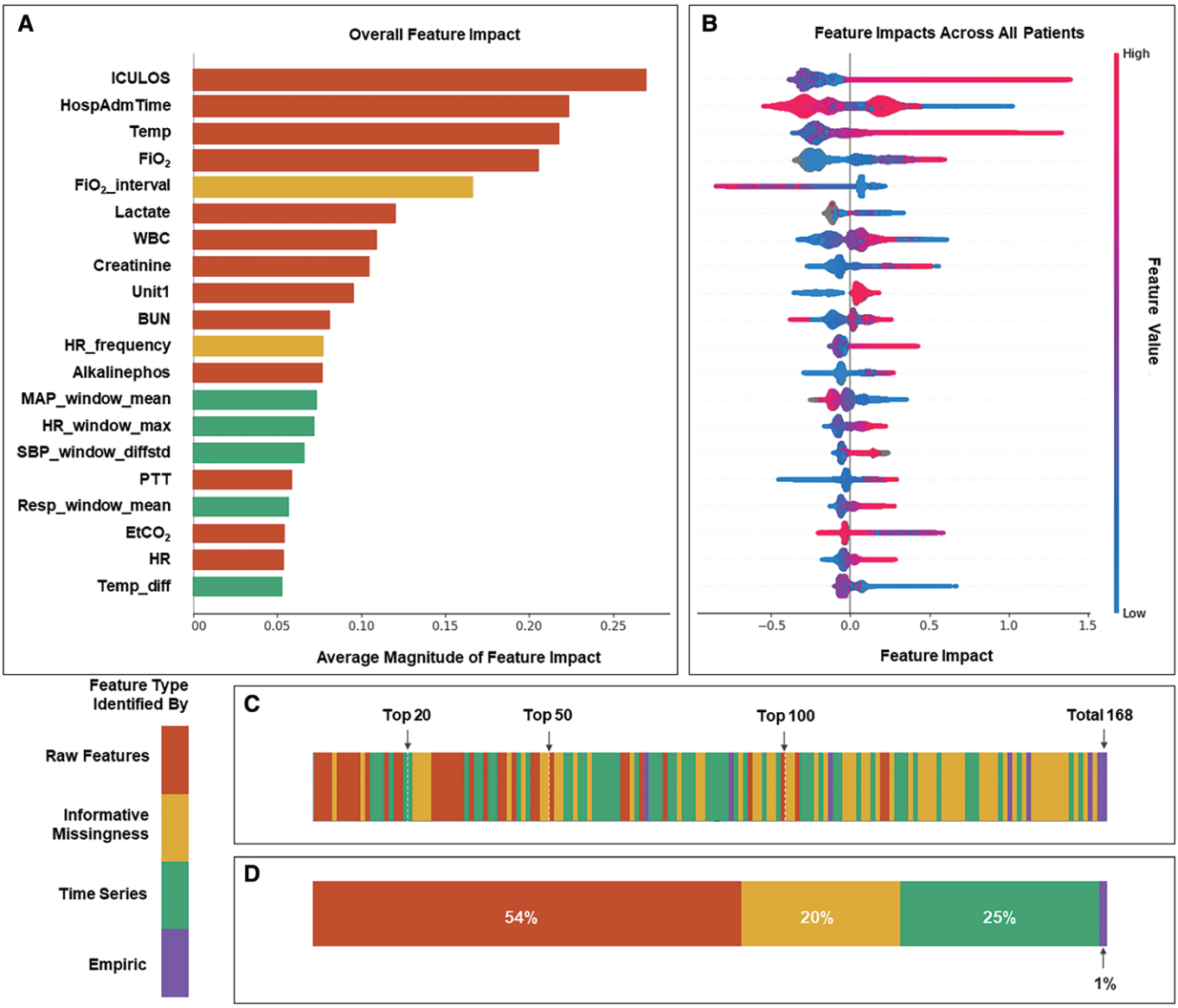

Figure 1. Summary of impact for the employed features. **A**, Overall impacts of the top 20 features. **B**, Beeswarm plots show feature impacts across all patients for the top 20 features where each *dot* indicates the impact of feature for 1-hr sample. *Gray dots* refer to unfilled missing values. When multiple *dots* fall on the same *x* position, they are stacked to show density. Features with positive impact values push the sepsis risk higher, whereas negative push the risk lower. Long *tails* indicate features are extremely important for specific patients. **C**, Overall impacts of all 168 features from highest to lowest. **D**, Overall impacts of the four different feature types. Alkalinephos = alkaline phosphatase, BUN = blood urea nitrogen, Et_{CO_2} = end-tidal carbon dioxide, HospAdmTime = hours between hospital admit and ICU admit, HR_frequency = measurement frequency of heart rate, HR_window_max = maximum of heart rate in a 6-hr window, LOS = length-of-stay, MAP_window_mean = mean of mean arterial pressure in a 6-hr window, PTT = partial thromboplastin time, Resp_window_mean = mean of respiration rate in a 6-hr window, SBP_window_diffstd = SD after first difference of systolic blood pressure in a 6-hr window, Temp = temperature, Temp_diff = difference between the current record and last temperature.

as the change in the expected risk output when a specific feature was versus missing, and it reflected the impact of the current feature on an hourly sepsis risk prediction. Method for fast estimating Shapley values can refer to (16). Averaging the magnitude of the impacts across all patients shows the overall importance of a specific feature.

Open-source code implementation of this algorithm in Python is available online at https://github.com/Meicheng-SEU/EASP.

Algorithm Evaluation

Hourly data that triggered a fixed risk threshold in the EASP model would be identified as a positive prediction for sepsis. A clinical utility score defined by the challenge organizers was used as evaluation index (9). Traditional index of area under the curve (AUC) was also calculated for evaluating all the hourly predictions across all patients. In addition, we referred to the patients as positive cases if they were detected developing sepsis during their ICU stay and negative cases if not. Therefore, the septic/nonseptic predictions for all patients were identified, and the indices of sensitivity and specificity for patients' detection were also reported.

RESULTS

Characteristics from validation populations are clinically similar to those from development populations (Table 1). EASP model yielded the AUC of 0.85 for totally 233,835 hourly predictions in validation set. When optimizing sepsis risk threshold as 0.525 (optimized between 0.4 and 0.6), EASP model achieved a utility score of 0.430. For patients' statistics, we correctly detecting 395 septic patients (sensitivity = 0.90) while falsely identifying 2,034 nonseptic patients as septic (specificity = 0.64). The top 20 most important features for sepsis prediction were summarized in **Table 2**, with visible explanations across all patients in **Figure 1**. Raw features contributed the highest overall impact (total 0.54), followed by time series features (0.25), informative missingness features (0.20), and empiric features (0.01).

Finally, EASP model yielded the highest utility score of 0.364 in the PhysioNet/CinC Challenge 2019 when tested on the full hidden test sets, with utility scores of 0.430, 0.422, and –0.048 for the three separate test sets (**Table 3**).

DISCUSSION

An explainable EASP model for early detection of sepsis was proposed. Reliable sepsis prediction using AI models with quantitative and explainable risk factors is important for medical decision support (16) and is in urgent need by clinician (22). In previous studies, Rosnati et al (23) proposed a deep learning model and added attention mechanism to present feature importance, but it is not suitable for real-time sepsis prediction. Nemati et al (24) used a modified regularized Weibull-Cox analysis and calculated hourly importance of each feature. However, they do not capture the model's overall behavior. The developed EASP can inform important variables contributing to the model prediction, and the utility score (i.e., model performance) on validation set would decrease from 0.430 to 0.399 when the top 20 variables were masked. In addition, it can help doctors to gain insight into how risk prediction score varies according to the contribution from all features in real time. **Figure 2** demonstrates its interpretability for the clinical data. The main risk indicators as shown in Figure 2 including temperature, FIO_2, lactate, etc, providing a real-time interpretation for the sepsis risk. Furthermore, extending analysis of the input features' impact on all hourly predictions can capture their potential interaction effects (16) (examples see **Supplemental Fig. 1**, Supplemental Digital Content 1, http://links.lww.com/CCM/F726).

Limitations should be mentioned. One limitation is the alert fatigue. In validation populations, about 75% of sepsis predictions in true positive patients were penalized when calculating the utility score. In addition, nearly 36% nonseptic patients were falsely predicted as septic. An effective false alarm rejection mechanism should be developed in future work. Another limitation lies in that EASP performed well on two hidden test sets but not on the third one. Thus, enhancing its generalizability also need.

To conclude, this study demonstrates the proposed EASP model could achieve superior performance in the challenge when predicting sepsis risk in a real-time way. EASP can also

TABLE 3. Top Utility Score Teams in the PhysioNet/Computing in Cardiology Challenge 2019

References	Final Utility Score	Score A	Score B	Score C
Yang et al (17)[a]	0.364	0.430	0.422	−0.048
Morrill et al (18)	0.360	0.433	0.434	−0.123
Du et al (19)	0.345	0.409	0.396	−0.042
Guan (20)[b]	0.340	0.422	0.410	−0.166
Zabihi et al (21)	0.339	0.422	0.395	−0.146

[a]Our team as an unofficial entry.

[b]Unofficial entry.

Scores A, B, and C are utility scores on each of the test sets from hospital systems A, B, and C.

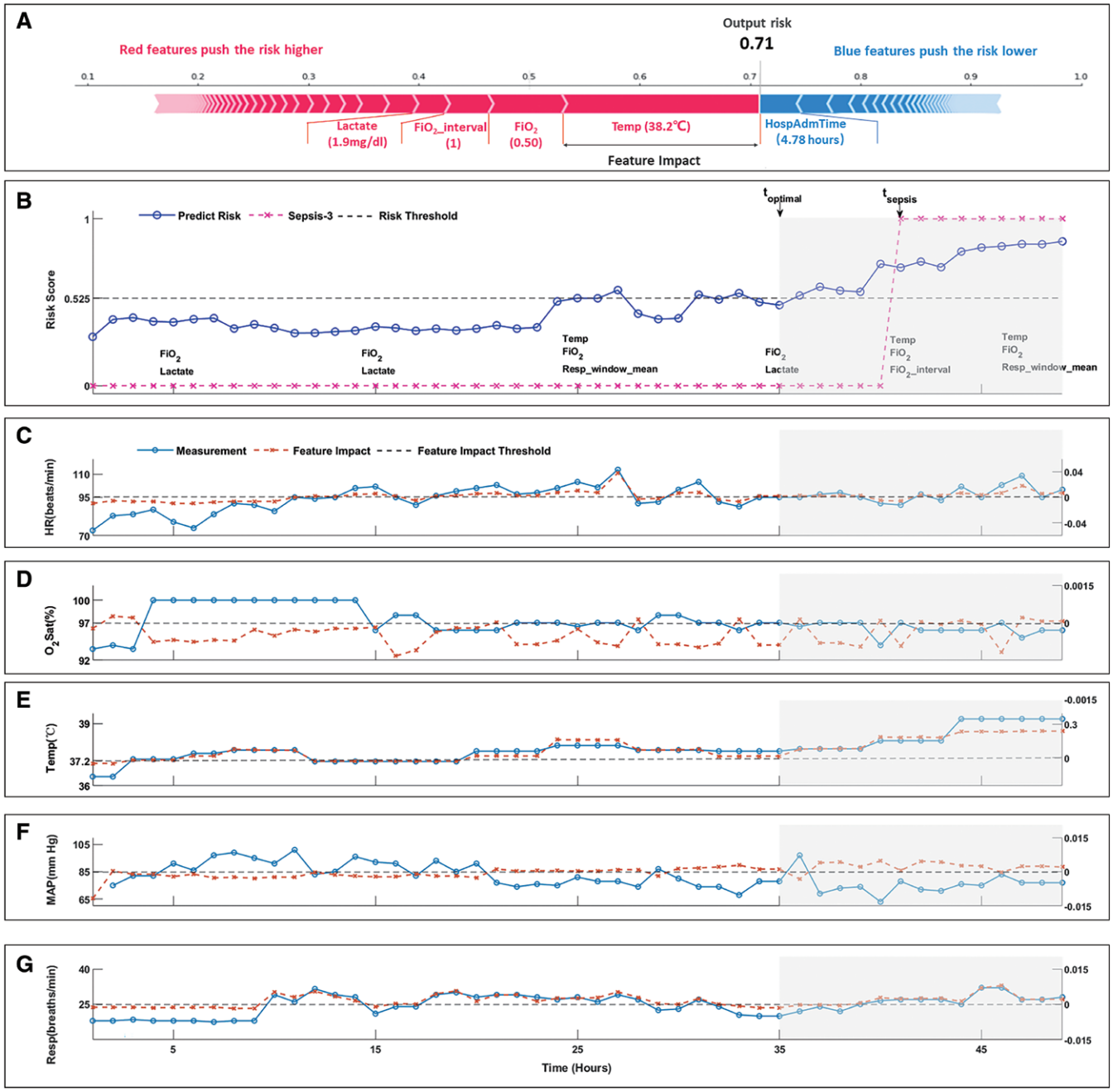

Figure 2. An example of real-time prediction and interpretability of explainable artificial intelligence sepsis predictor (EASP) model. **A**, Explanation of how sepsis risk score is output for a positive prediction. The abnormality of temperature (Temp), FIO_2, measurement interval of FIO_2 (FIO_2_interval), and lactate highlight the patient's sepsis risk. While time between hospital and ICU admission (HospAdmTime) is relative normal. **B**, Hourly calculated EASP sepsis risk score and Sepsis-3 clinical definitions, as well as several features with the greatest contribution to risk scores at selected timestamps. Resp_window_mean indicates mean value of respiration rate in a 6-hr window. The timestamp of 6 hr prior to the onset time of sepsis (t_{sepsis}) is defined as $t_{optimal}$. From **C** to **G**, typical vital signs including heart rate (HR), oxygen saturation (O_2Sat), Temp, mean arterial blood pressure (MAP), and respiratory rate (Resp) after forward-filling, and their feature impacts on prediction are shown over time. The measurements with feature impact above the threshold push the sepsis risk higher, whereas below push risk lower. The *shaded area* represents the targeted alert period.

provide interpretable information for improving clinical understanding of sepsis risk in ICU.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Supported, in part, by the Distinguished Young Scholars of Jiangsu Province (BK20190014), the National Natural Science Foundation of China (81871444), the Key Research & Development Plan of Ministry of science and technology (2017YFB1303200), and the Primary Research & Development Plan of Jiangsu Province (BE2017735).

Dr. Chengyu Liu received support from the Distinguished Young Scholars of Jiangsu Province (BK20190014), the National Natural Science Foundation of China (81871444). Dr. Li received support from the Key Research & Development Plan of Ministry of Science and Technology (2017YFB1303200) and the Primary Research & Development Plan of Jiangsu Province (BE2017735). The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: chengyu@seu.edu.cn; ljq@seu.edu.cn

The content of this article is solely the responsibility of the authors.

REFERENCES

1. World Health Organization: Sepsis. 2018. Available at: https://www.who.int/news-room/fact-sheets/detail/sepsis. Accessed December 10, 2019
2. Liu VX, Fielding-Singh V, Greene JD, et al: The timing of early antibiotics and hospital mortality in sepsis. *Am J Respir Crit Care Med* 2017; 196:856–863
3. Rivers E, Nguyen B, Havstad S, et al; Early Goal-Directed Therapy Collaborative Group: Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med* 2001; 345:1368–1377
4. Seymour CW, Gesten F, Prescott HC, et al: Time to treatment and mortality during mandated emergency care for sepsis. *N Engl J Med* 2017; 376:2235–2244
5. Calvert JS, Price DA, Chettipally UK, et al: A computational approach to early sepsis detection. *Comput Biol Med* 2016; 74:69–73
6. Giannini HM, Ginestra JC, Chivers C, et al: A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice. *Crit Care Med* 2019; 47:1485–1492
7. Henry KE, Hager DN, Pronovost PJ, et al: A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015; 7:299ra122
8. Kam HJ, Kim HY: Learning representations for the early detection of sepsis with deep neural networks. *Comput Biol Med* 2017; 89:248–255
9. Reyna MA, Josef CS, Jeter R, et al: Early prediction of sepsis from clinical data: The PhysioNet/computing in cardiology challenge 2019. *Crit Care Med* 2020; 48:210–217
10. Sterne JA, White IR, Carlin JB, et al: Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 2009; 338:b2393
11. Smith GB, Prytherch DR, Meredith P, et al: The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; 84:465–470
12. Singer M, Deutschman CS, Seymour CW, et al: The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 2016; 315:801–810
13. 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, San Francisco, CA, August 13-17, 2016, pp 785–794
14. Shahriari B, Swersky K, Wang Z, et al: Taking the human out of the loop: A review of Bayesian optimization. *Proc IEEE* 2016; 104:148–175
15. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; 30:4768–4777
16. Lundberg SM, Erion G, Chen H, et al: From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; 2:56–67
17. Yang M, Wang X, Gao H, et al: Early prediction of sepsis using multi-feature fusion based XGBoost learning and Bayesian optimization. *In:* Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, September 8-11, 2019
18. Morrill J, Kormilitzin A, Nevado-Holgado A, et al: The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. *In:* Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, September 8-11, 2019
19. Du JA, Sadr N, Chazal Pd. Automated prediction of sepsis onset using gradient boosted decision trees. *In:* Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, September 8-11, 2019
20. Reyna M, Josef C, Jeter R, et al: Early prediction of sepsis from clinical data–the PhysioNet Computing in Cardiology Challenge 2019 (version 1.0.0). 2019. Available at: https://doi.org/10.13026/v64v-d857. Accessed September 17, 2019
21. Zabihi M, Kiranyaz S, Gabbouj M: Sepsis prediction in intensive care unit using ensemble of XGboost models. *In:* Proceedings of the 2019 Computing in Cardiology (CinC), Singapore, September 8-11, 2019
22. Ginestra JC, Giannini HM, Schweickert WD, et al: Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit Care Med* 2019; 47:1477–1484
23. Rosnati M, Fortuin V: MGP-AttTCN: An Interpretable Machine Learning Model for the Prediction of Sepsis. 2019. Available at: https://arxiv.org/abs/1909.12637. Accessed September 27, 2019
24. Nemati S, Holder A, Razmi F, et al: An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 2018; 46:547–553