Journal of Critical Care 55 (2020) 163–170

Contents lists available at ScienceDirect

Journal of Critical Care

journal homepage: www.journals.elsevier.com/journal-of-critical-care

Machine learning for prediction of septic shock at initial triage in emergency department

Joonghee Kim [a,1], HyungLan Chang [b,1], Doyun Kim [a], Dong-Hyun Jang [a], Inwon Park [a], Kyuseok Kim [c,*]

[a] Department of Emergency Medicine, Seoul National University Bundang Hospital, 166 Gumi-ro, Bundang-gu, Gyeonggi-do, Seongnam-si 463-707, Republic of Korea
[b] Department of Emergency Medicine, CHA Bundang Medical Center, CHA University, 59, Yatap-ro, Bundang-gu, Gyeonggi-do, Seongnam-si 463-712, Republic of Korea
[c] College of Medicine, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Sepsis
Septic shock
Machine learning
Clinical decision support tool
Prediction
Diagnosis
Emergency department triage tool

ABSTRACT

Background: We hypothesized utilizing machine learning (ML) algorithms for screening septic shock in ED would provide better accuracy than qSOFA or MEWS.
Methods: The study population was adult (≥20 years) patients visiting ED for suspected infection. Target event was septic shock within 24 h after arrival. Demographics, vital signs, level of consciousness, chief complaints (CC) and initial blood test results were used as predictors. CC were embedded into 16-dimensional vector space using singular value decomposition. Six base learners including support vector machine, gradient-boosting machine, random forest, multivariate adaptive regression splines and least absolute shrinkage and selection operator and ridge regression and their ensembles were tested. We also trained and tested MLP networks with various setting.
Results: A total of 49,560 patients were included and 4817 (9.7%) had septic shock within 24 h. All ML classifiers significantly outperformed qSOFA score, MEWS and their age-sex adjusted versions with their AUROC ranging from 0.883 to 0.929. The ensembles of the base classifiers showed the best performance and addition of CC embedding was associated with statistically significant increases in performance.
Conclusions: ML classifiers significantly outperforms clinical scores in screening septic shock at ED triage.

1. Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection. It is a significant health problem with a global estimated incidence of 148 per 100,000 person-year with an overall mortality of 26% [1].

Emergency departments (EDs) play an important role in early sepsis care. Greater than one half of sepsis cases are first identified in EDs [2]. EDs are taking an increasingly active role in early sepsis care via hemodynamic resuscitation and initiation of antibiotic treatment [3,4].

Septic shock is a sepsis subcategory in which underlying circulatory, cellular, and metabolic abnormalities are associated with a greater risk of mortality [5]. It is operationally defined as a septic condition requiring vasopressors to maintain mean arterial pressure ≥65 mmHg with a serum lactate level > 2 mmol/L after adequate fluid resuscitation. Given its high mortality and hemodynamic instability, patients suspected of having septic shock should be managed in specialized areas with adequate monitoring capacity. However, it is not easy to reliably screen septic shock during triage [6], and the increasing demand for critical care in EDs [4] despite limited resources [7-9] makes it difficult to provide appropriate care to such patients.

An effective screening tool may help to alleviate this problem by helping healthcare professionals to focus their available resources to patients with a high risk of septic shock. However, no clinical tool has been developed to screen for septic shock. The quick sepsis-related organ failure assessment (qSOFA) score, which was initially developed for screening sepsis, or modified early warning score (MEWS), is commonly used by rapid response teams to detect patient deterioration and may be useful for septic shock screening [10]. However, these methods are not optimized for screening septic shock and utilize only small number of variables without considering any interactions among them.

Machine learning (ML) algorithms are being actively studied in healthcare for various applications [11-22]. ML algorithms can detect sepsis with high accuracy using cumulated data in wards, intensive care units (ICUs) and EDs [23-27]. However, whether these algorithms

* Corresponding author.
E-mail address: dreinstein70@gmail.com (K. Kim).
[1] These authors contributed equally to the article.

https://doi.org/10.1016/j.jcrc.2019.09.024
0883-9441/

are useful for screening septic shock at ED triage remains unknown. The primary objective of the study is to assess the performance of ML-based triage tools in screening patients with septic shock in ED.

2. Materials and methods

2.1. Study design

This is a single-center observational study utilizing an electronic health record (EHR) database of patients who visited the ER for suspected infection from 2008 to 2016. The study facility is a tertiary academic hospital located in South Korea with an annual ED visits from >80,000 patients a year. The institutional review boards of the study site approved the study and provided a waiver of informed consent.

2.2. Study population and primary outcome event

The study population was defined as adult (aged ≥20 years) patients in the ED with a suspected infection. A suspected infection was defined as taking cultures and a prescription of systemic antibiotics within 24 h of ED arrival [10]. Trauma visits, transferred cases and cases with >50% of data entry missing were excluded. Recurrent visits were treated as independent cases. The primary outcome was a development of septic shock (or cardiac arrest) within 24 h of ED arrival. Septic shock was defined following the clinical criteria of SEPSIS-3 [5,10] with an assumption that adequate volume resuscitation is provided within three hours. Therefore, a septic shock case is clinically defined as dependence on vasopressor (cumulative vasopressor index [CVI] ≥2) [28] at or after three hours of ED treatment until 24 h accompanied by an increased level of serum lactate (≥2) measured within the same time window (3 to 24 h). If no serum lactate measurement was obtained in the time window (3 to 24 h), we assumed that the case is positive for septic shock if the patient was still vasopressor dependent (CVI ≥2) at or after 12 h (12 to 24 h) of ED treatment.

2.3. Predictor variables and preprocessing

The following predictors were retrieved from the EHR; age; sex; chief complaints (CC); initial vital signs, including systolic blood pressure (SBP, mmHg), diastolic blood pressure (DBP, mmHg), pulse rate (PR, beats per minute), respiratory rate (RR, breaths per minute) and body temperature (BT, measured in Celsius); initial level of consciousness measured on the AVPU scale (AVPU: Alert, Verbal, Pain and Unresponsive); initial O_2 saturation measured by pulse oximetry (SpO_2, %); and the results of initial blood test, including white blood cell count (WBC, 10^9/L), differential counts, red blood cell distribution weight (RDW, %), platelet count (10^9/L), prothrombin time international normalized ratio (PT INR), fibrinogen (mg/dL), blood urea nitrogen (BUN, mg/dL), sodium (mmol/L), potassium (mmol/L), chloride (mmol/L), creatinine (mg/dL), aspartate aminotransferase (AST, U/L), alanine aminotransferase (ALT, U/L), alkaline phosphatase (ALP, U/L), total bilirubin (mg/dL), albumin (g/dL) and C-reactive protein (CRP, mg/dL).

Categorical variables with low cardinality (sex and AVPU) were one-hot encoded. CC, which had high cardinality (over 1500), was embedded into 16-dimensional vector space using singular value decomposition (SVD) [29]. The CC embeddings were projected to a 2-D scatter plot using t-distributed stochastic neighbor embedding (t-SNE) for visualization. Continuous variables were transformed by the Yeo-Johnson method, which is an extension of the Box-Cox transformation, to improve normality [30]. Missing values were imputed by means or modes as appropriate with inclusion of indicator variables. Specifically, we applied mean imputation for continuous predictors and mode imputation for categorical predictors. The mean/mode values were calculated using the training dataset and applied uniformly to both training and test datasets.

2.4. Training of ML classifiers

Performance of a classifier is dependent on its predictor set. In a typical ED, basic information, such as age, sex, CC, vital signs, SpO_2 (if initially measured) and AVPU, is available at triage. Unlike other predictors, CC needs to be embedded to lower dimensional space given its high cardinality. Laboratory test results require several hours to be reported; however, these results are occasionally available at triage if a patient is referred from outpatient departments or other healthcare facilities. Therefore, we assumed four scenarios of different predictor availability: 1) baseline predictors (age, sex, vital signs, SpO_2 and AVPU), 2) baseline predictors and CC embedding, 3) baseline predictors plus initial laboratory test results, and 4) baseline predictors plus both (CC embeddings and laboratory test results). All ML algorithms were optimized and tested in these four different conditions. All the nonlaboratory variables used in this study were available at triage. However, the laboratory variables were not available at triage in most cases.

The study population was randomly partitioned into training and test datasets at a ratio of 6:4 using patient identification numbers. Six base ML algorithms, including support vector machine with radial basis function kernel (SVM), gradient-boosting machine with Bernoulli loss (GBM), random forest (RF), multivariate adaptive regression splines (MARS), least absolute shrinkage and selection operator (lasso) and ridge regression, were assessed. In addition, we also constructed two ensemble classifiers utilizing SVM, GBM, RF, MARS and lasso as base learners (ridge was not utilized given its high correlation with lasso). The first algorithm used simple averaging, and the second algorithm used MARS utilizing cross-validated prediction of the base learners.

Training was performed using the mlr package, which streamlines building predictive models with R's various ML packages [31]. Hyperparameters were tuned using a grid search with 5-fold cross-validation. AUROC was used to select the best hyperparameter combination. The search space and model specifications are described in supplementary table 1.

In addition, we trained multiple multilayer perceptron (MLP) classifiers of various combinations of network sizes, hyperparameters and training schemes to identify the best combination. Using a grid search, each combination was tested in a validation set, which included 40% of cases in the training dataset. A detailed description of the search space is provided in supplementary table 2.

2.5. Statistical analyses

Categorical variables were reported using frequencies and proportions. Continuous variables were reported using median and inter quartile range (IQR). *t*-test, Wilcoxon's rank-sum test, chi-square test or Fisher's exact test were performed as appropriate for comparison between groups.

Performance of ML classifiers was assessed by calculating the area under the receiver operating characteristic curve (AUROC) and the area under precision-recall curve (AUPRC) in the test dataset. 95% confidence intervals (CIs) of both AUROC and AUPRC were computed with 2000 stratified bootstrap replicates of the test dataset. Sensitivity (recall), specificity, positive predictive value (PPV, precision) and negative predictive value (NPV) were assessed after setting specificity at 90%. The choice of 90% specificity was designed to achieve >95% NPV and was made during the data analysis step. 95% CI values of sensitivity, specificity, PPV and NPV were assessed by calculating exact binomial confidence limits. qSOFA score and MEWS and their age and sex-adjusted values (adjusted qSOFA score and MEWS) were used as baseline for comparison. In addition, we assessed variable importance ranking of the predictors by measuring the reduction in prediction accuracy after permutations of target predictors [32].

The clinical definition of septic shock used in this study assumes adequate fluid resuscitation within three hours. However, it is possible this goal could not be achieved in some patients. Therefore, we performed a sensitivity analysis with an alternative assumption that adequate fluid resuscitation can be delayed up to six hours. With this new assumption, some of the cases that were classified as septic shock prematurely will be reclassified as nonseptic shock. We constructed all the ML classifiers again in this condition and reassessed AUROC and AUPRC values of the models to evaluate whether the main hypothesis that ML classifiers are superior to conventional clinical scores still holds.

P-values <0.05 were considered significant. All data handling and statistical analyses were performed using R-packages version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

There were 55,313 ED visits with suspected infection from 2008 to 2016. After exclusion of trauma visits (1582, 2.9%), transferred cases (1917, 3.5%) and those with >50% missing values (2515, 4.5%), a total of 49,299 patients were included as the study population (Fig. 1). Among the population, a total of 4782 patients (9.7%) were identified as having septic shock within 24 h of ED arrival (Table 1). Six-month in-hospital mortality was 31.1% in the septic shock group and 7.6% in patients without septic shock. The median duration of ED stay was 576 min in patients with septic shock and 1036.5 min in those without septic shock.

Table 2 shows the discriminatory performance of non-ML classifiers including qSOFA score and MEWS and their age and sex-adjusted values. The adjusted qSOFA showed the highest AUROC of 0.832 (95% confidence interval [CI], 0.822–0.842), and the unadjusted qSOFA showed the highest AUPRC of 0.395 (95% CI, 0.371–0.419).

We constructed 9 ML classifiers (6 base, 2 ensemble and 1 MLP classifier) for each of the four predictor set which were 1) baseline predictors, 2) baseline predictors plus CC embedding, 3) baseline predictors plus initial laboratory test results, and 4) baseline predictors plus both. The hyperparameters used for the model building are presented in supplementary table 1 and 2 and the 2-D projections of the CC embeddings are visualized in Fig. 2. ML classifiers using baseline predictors outperformed the adjusted qSOFA with their AUROCs ranging from 0.882 to 0.902. The maximum AUROC was found in the ensemble classifiers with AUROC values of 0.902 (95% CI, 0.895–0.909) for simple averaging and 0.902 (95% CI, 0.895–0.908) for MARS-based ensemble. Similarly, they also showed higher AUPRC values of 0.556 (0.531–0.580) and 0.554 (0.529–0.578), respectively.

The inclusion of CC embedding was associated with a small but statistically significant increase of AUROC in GBM and the two ensemble classifiers (all $p < 0.001$) and AUPRC in all the ML classifiers (all $p < 0.001$ except RF, whose $p = 0.035$; supplemental table 4). The addition of initial laboratory findings to the baseline predictor set was associated with significantly increased performance (both AUROC and AUPRC) in all of the ML classifiers (all $p < 0.001$). With inclusion of both CC embedding and initial laboratory findings, the AUROC ranged from 0.904 to 0.924, and the ensemble classifier utilizing MARS showed the highest performance (AUROC, 0.924; 95% CI, 0.918–0.929; AUPRC, 0.623; 95% CI, 0.599–0.645). The inclusion of CC embedding when initial laboratory test results are available was associated with increased AUROC in GBM and the two ensemble models ($p = 0.019$, 0.001 and 0.003, respectively) and increased AUPRC in GBM, RF, MARS, Lasso, ridge and the two ensemble models (all $p < 0.001$). However, decreased AUPRC was noted in the SVM and MLP models (all $p < 0.001$). Supplementary Fig. S1 shows the calibration plots of the whole classifiers.

Table 3 shows the variable importance rankings up to fifth for each of the ML classifiers. SBP was the most consistently important nonlaboratory variable appearing in every classifier (28/28) followed by AVPU: Alert (19/28), SpO_2 (9/28), DBP (8/28) and BT (8/28). For laboratory variables, segmented neutrophils (4/14), monocytes (4/14), WBC (3/14), lymphocytes (3/14) and creatinine (3/14) were the most consistently important variables.

A sensitivity analysis was performed with an alternative assumption that adequate fluid resuscitation can be delayed up to six hours. With

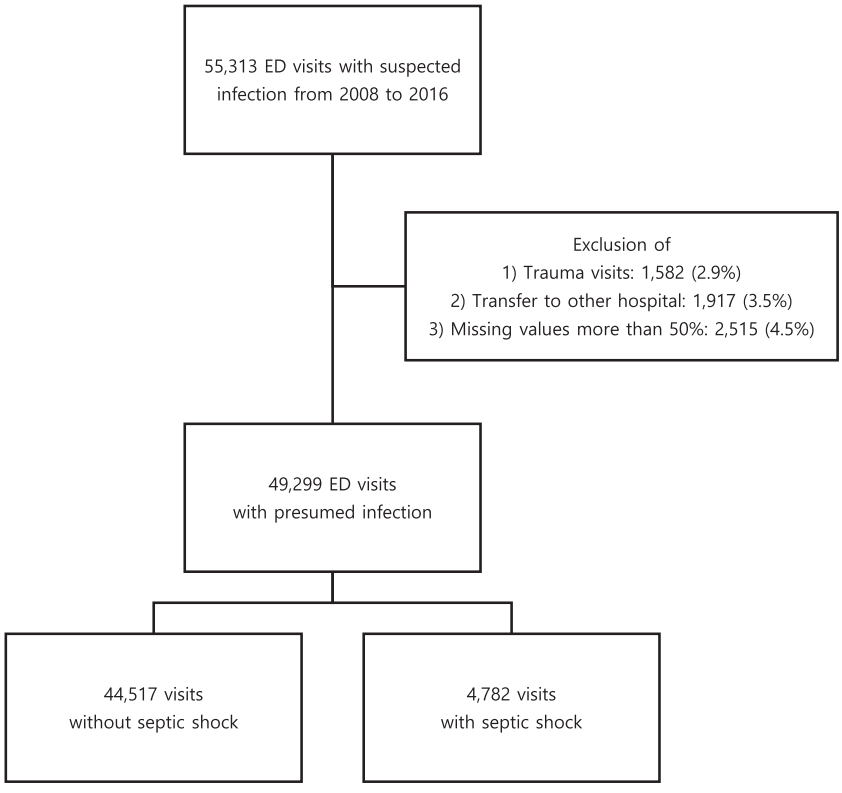

Fig. 1. A flowchart of the study population.

Table 1
Baseline characteristics of the study population.

	Patients without septic shock ($N = 44,517$)	Patients with septic shock ($N = 4782$)	*p*
Age, years (IQR)	68.0 (51.0–79.0)	75.0 (66.0–83.0)	<0.001
Sex, Male (%)	22,230 (49.9%)	2687 (56.2%)	<0.001
Vital signs			
Systolic blood pressure, mmHg (IQR)	129.0 (114.0–145.0)	103.0 (86.0–125.0)	<0.001
Diastolic blood pressure, mmHg (IQR)	73.0 (63.0–82.0)	57.0 (47.0–70.0)	<0.001
Pulse rate, beats per minute (IQR)	96.0 (82.0–110.0)	105.0 (88.0–122.0)	<0.001
Respiratory rate, breaths per minute (IQR)	20.0 (18.0–20.0)	20.0 (18.0–25.0)	<0.001
Body temperature, Celsius (IQR)	37.1 (36.6–38.1)	37.1 (36.4–38.2)	<0.001
SpO2, % (IQR)	97.0 (94.0–98.0)	95.0 (88.0–97.0)	<0.001
Consciousness			<0.001
Alert (%)	41,028 (92.2%)	2910 (60.9%)	
Verbal (%)	1148 (2.6%)	413 (8.6%)	
Pain (%)	2006 (4.5%)	1010 (21.1%)	
Unresponsive (%)	335 (0.8%)	449 (9.4%)	
qSOFA score			<0.001
0 (%)	31,844 (71.5%)	928 (19.4%)	
1 (%)	10,412 (23.4%)	1921 (40.2%)	
2 (%)	2077 (4.7%)	1418 (29.7%)	
3 (%)	184 (0.4%)	515 (10.8%)	
Blood test results			
White blood cell count, 10^9/L (IQR)	10.1 (7.0–13.7)	11.1 (6.5–16.9)	<0.001
Segmented neutrophils, % (IQR)	79.6 (70.0–86.4)	86.0 (76.7–91.2)	<0.001
Monocytes, % (IQR)	6.3 (4.1–8.9)	4.2 (2.2–7.1)	<0.001
Lymphocytes, % (IQR)	11.7 (7.1–19.1)	7.9 (4.3–15.0)	<0.001
Hemoglobin, % (IQR)	12.5 (10.9–13.8)	11.2 (9.6–12.8)	<0.001
RDW, % (IQR)	13.8 (13.1–15.1)	14.9 (13.8–16.7)	<0.001
Prothrombin time, INR (IQR)	1.1 (1.0–1.2)	1.2 (1.1–1.4)	<0.001
Platelets, 10^9/L (IQR)	212.0 (159.0–275.0)	173.0 (109.0–247.0)	<0.001
Fibrinogen, mg/dL (IQR)	525.0 (401.0–665.0)	521.0 (385.0–665.0)	0.001
Blood urea nitrogen, mg/dL (IQR)	15.0 (11.0–22.0)	26.0 (17.0–41.0)	<0.001
Creatinine, mg/dL (IQR)	0.8 (0.7–1.1)	1.3 (0.9–2.1)	<0.001
Sodium, mmol/L (IQR)	136.0 (134.0–139.0)	135.0 (131.0–138.0)	<0.001
Potassium, mmol/L (IQR)	4.1 (3.8–4.4)	4.2 (3.7–4.7)	<0.001
Chloride, mmol/L (IQR)	100.0 (97.0–103.0)	100.0 (95.0–104.0)	<0.001
GOT, U/L (IQR)	25.0 (18.0–41.0)	33.0 (22.0–73.0)	<0.001
GOT, U/L (IQR)	20.0 (12.0–36.0)	21.0 (12.0–47.0)	<0.001
ALP, U/L (IQR)	88.0 (68.0–123.0)	104.0 (75.0–161.0)	<0.001
Total Bilirubin, mg/dL (IQR)	0.8 (0.5–1.2)	0.9 (0.6–1.5)	<0.001
Albumin, g/dL (IQR)	3.8 (3.4–4.2)	3.1 (2.7–3.6)	<0.001
C-reactive protein, mg/dL (IQR)	5.4 (1.4–12.4)	11.4 (4.1–19.5)	<0.001
Source of infection			
Central nervous system (%)	233 (0.5%)	15 (0.3%)	0.066
Upper respiratory tract (%)	1193 (2.7%)	11 (0.2%)	<0.001
Lower respiratory tract (%)	8497 (19.1%)	1236 (25.8%)	<0.001
Gastrointestinal (%)	7659 (17.2%)	710 (14.8%)	<0.001
Genitourinary (%)	7569 (17.0%)	573 (12.0%)	<0.001
Endocarditis (%)	55 (0.1%)	12 (0.3%)	0.039
Musculoskeletal (%)	362 (0.8%)	28 (0.6%)	0.109
Skin and soft tissue (%)	1470 (3.3%)	46 (1.0%)	<0.001
Device-related (%)	142 (0.3%)	8 (0.2%)	0.095
Length of stain in ED, minutes (IQR)	576.0 (289.0–1547.0)	1036.5 (490.0–1791.0)	<0.001
Disposition in ED			<0.001
Admission (%)	23,271 (52.3%)	2789 (58.3%)	
Discharge (%)	19,636 (44.1%)	176 (3.7%)	
Death (%)	41 (0.1%)	302 (6.3%)	
ICU (%)	1111 (2.5%)	1412 (29.5%)	
OR (%)	458 (1.0%)	103 (2.2%)	
Six-month inhospital mortality (%)	3380 (7.6%)	1488 (31.1%)	<0.001

IQR, interquartile range; RDW, red blood cell distribution width; GOT, glutamate oxaloacetate transaminase; GPT, glutamate pyruvate transaminase; ALP, alkaline phosphatase; ED, emergency department; ICU, intensive care unit; OR, operating room.

this assumption, the total number of visits complicated by septic shock decreased to 4636/49,299 from 4782/49,299. The results of the sensitivity analysis did not differ to those of the main analysis with all the ML classifiers showing superior performances compared with clinical scores and the two ensemble models showing the best performances (Supplementary Fig. S2, supplementary table 5).

4. Discussion

In this study, we evaluated the performance of various ML classifiers in four different conditions of data availability. We observed these classifiers have high discriminatory power even when provided with only baseline data and outperform traditional scores, such as qSOFA or

Table 2
Performances of clinical scores.

	AUROC	AUPRC	Specificity	Sensitivity (Recall)	PPV (Precision)	NPV
qSOFA	0.813 (0.803–0.824)	0.395 (0.371–0.419)	95.0 (94.7–95.3)	42.3 (40.1–44.6)	47.2 (44.8–49.6)	94.0 (93.6–94.3)
MEWS	0.790 (0.779–0.800)	0.335 (0.313–0.358)	93.3 (92.9–93.7)	35.2 (33.0–37.4)	35.8 (33.6–38.0)	93.2 (92.8–93.5)
Adj.qSOFA[a]	0.832 (0.822–0.842)	0.391 (0.368–0.414)	90.0 (89.6–90.5)	50.4 (48.1–52.7)	34.8 (33.0–36.6)	94.5 (94.2–94.8)
Adj.MEWS[a]	0.813 (0.803–0.823)	0.355 (0.333–0.379)	90.0 (89.6–90.5)	47.6 (45.3–49.9)	33.5 (31.7–35.3)	94.2 (93.8–94.6)

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision recall curve; PPV, positive predictive value; NPV, negative predictive value; qSOFA, quick sepsis-related organ failure assessment; MEWS, modified early warning score.
[a] Age and sex-adjusted qSOFA score and MEWS.

MEWS. To the best of our knowledge, this is the first report of the feasibility of using ML algorithms for screening septic shock at ED triage.

There are numerous related studies applying ML algorithms to identify or predict various sepsis-related conditions, including infection, infection with systemic inflammatory response syndrome (SIRS), sepsis (severe sepsis before SEPSIS-3 definition, an infectious condition accompanied by organ dysfunction) and septic shock. Thiel et al. built recursive partitioning models using measurements randomly sampled from predefined time windows to identify septic shock (infection with organ dysfunction based on ICD-9 code criteria plus vasopressor use within 24 h of ICU transfer) among hospitalized patients [33]. Tang et al. developed SVM classifiers using the first three principal components of clinical variables to identify patients with sepsis (severe sepsis before SEPSIS-3 definition) among ED patients with positive SIRS criteria [34]. Nachimuthu et al. employed a dynamic Bayesian network to predict sepsis (retrospectively annotated by a clinician) using clinical data collected within 3, 6, 12 and 24 h after ED admission [35]. Paxton et al. used SVM with a linear kernel to predict septic shock in ICU patients (Multiparameter Intelligent Monitoring in Intensive Care [MIMIC]-II data set) [36]. Henry et al. built a Cox-regression model to identify septic shock (prolonged hypotension over 30 min after fluid therapy of ≥20 ml/kg or ≥1200 ml over the past 24 h) among ICU patients (MIMIC-II dataset) [37]. Desautels et al. applied regularized regression (elastic net) using curated features of single/combined clinical variables of the last 2 h to predict sepsis (Sepsis-3 definition) in ICU patients (MIMIC III dataset) [18]. Haug et al. built a Bayesian network to diagnose sepsis (based on ICD-9 criteria) in ED using a set of locally developed clinical variables [38]. Nemati et al. employed Weibull-Cox regression to identify sepsis (SEPSIS-3 definition) in an ICU population (MIMIC-III and Emory University hospitals) [24]. Horng et al. built an infection screening system for ED using SVM with bigram (extracted from free text) and/or conventional features [39]. Zhang et al. developed

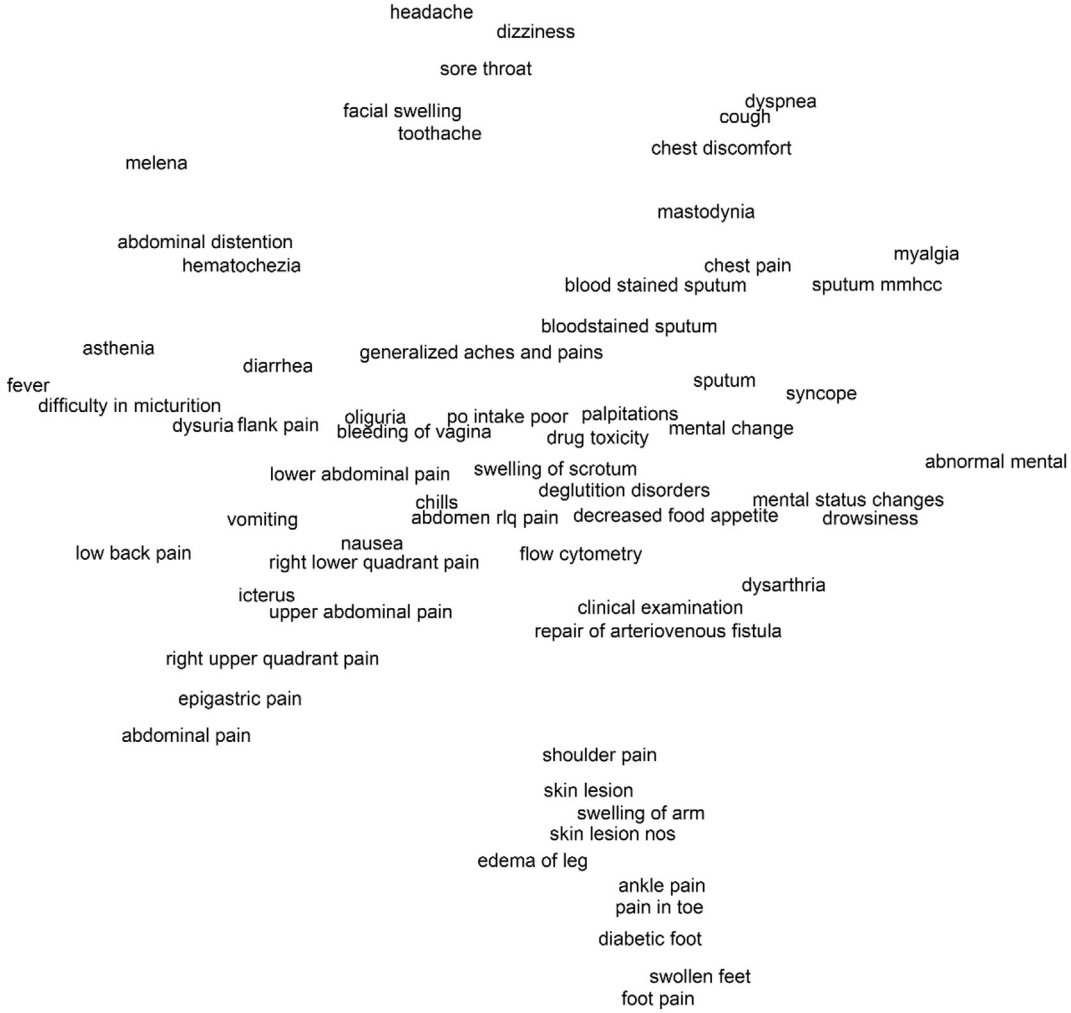

Fig. 2. Non-linear 2-D projection of common chief complaints (up to 100th) visualized using t-distributed stochastic neighbor embedding.

Table 3
Variable importance. (Fig. 3)

Initial predictor set	Variable rank	SVM	GBM	RF	MARS	Lasso	Ridge	MLP
Baseline (Age, sex, vital signs, consciousness) only	1st	SBP	SBP	SBP	SBP	SBP	SBP	SBP
	2nd	AVPU: Alert	DBP	DBP	BT	AVPU: Alert	AVPU: Alert	DBP
	3rd	AVPU: Pain	PR	AVPU: Alert	AVPU: Alert	SpO2.dummy	AVPU: Pain	SpO2
	4th	Age	BT	SpO2	Sex: Male	SpO2	Sex: Male	AVPU: U
	5th	Sex: Female	Age	PR	Sex: Female	PR	Sex: Female	PR
Baseline + CC embedding	1st	SBP	SBP	SBP	SBP	SBP	SBP	V1
	2nd	V12	DBP	V8	V16	V12	SpO2	SBP
	3rd	V8	AVPU: Alert	V1	AVPU: Alert	V15	SpO2. dummy	V10
	4th	AVPU: Alert	PR	DBP	BT	V1	AVPU: Alert	Sex: Female
	5th	V15	BT	SpO2	V9	AVPU: Alert	V16	V12
Baseline + Initial laboratory variables	1st	SBP	SBP	SBP	SBP	SBP	SBP	SBP
	2nd	BT	WBC	Segmented neutrophil	SpO2	Segmented neutrophil	SpO2	Fibrinogen
	3rd	Sodium	Lymphocyte	DBP	Monocyte	Creatinine	AVPU: Alert	Lymphocyte
	4th	Hemoglobin	BT	AVPU: Alert	AVPU: Alert	Monocyte	Monocyte	RR
	5th	Age	AVPU: Alert	Creatinine	BT	SpO2	Platelet	Hemoglobin
Baseline + both	1st	V12	SBP	SBP	SBP	SBP	AVPU: Alert	V12
	2nd	SBP	Lymphocyte	AVPU: Alert	AVPU: Alert	Segmented neutrophil	SBP	SBP
	3rd	V8	DBP	Segmented neutrophil	BT	V15	Monocyte	V11
	4th	Creatinine	WBC	BUN	DBP	V12	V2	V1
	5th	V15	AVPU: Alert	SpO2	WBC	AVPU: Alert	V12	Sex: Female

SVM, support vector machine; GBM, gradient-boosting machine; RF, random forest; MARS, multivariate adaptive regression splines; Lasso, least absolute shrinkage and selection operator; MLP, multilayer perceptron; CC, chief complaint; SBP, systolic blood pressure; DBP, diastolic blood pressure; PR, pulse rate; RR, respiratory rate; BT, body temperature; AVPU, alert, verbal, pain and unresponsive; WBC, white blood cell; BUN, blood urea nitrogen.
Note: V1-V16 indicate the coordinates of chief complaints in their embedding space.

a LSTM model using imperfect data by jointly training on both visit-level labels of septic shock (ICD-9 based) and event-level labels (persistent hypotension or vasopressor use) to early detect septic shock among general patient population [40]. Culliton et al. applied ridge regression to predict sepsis (severe sepsis before SEPSIS-3 definition) utilizing GloVe embedding features extracted from free text EHR and/or structured predictors [41].

Various ML algorithms provide improved predictive performance over conventional regression due to their varying abilities to automate feature selection and to handle nonlinearities and interactions. Recently, these algorithms hold great promise in healthcare due to the significant increase in computing power and the availability of "big data", such as EHR database. However, training a ML classifier requires a different skillset compared with conventional regression modeling [42]. Practitioners are needed to optimize relevant hyperparameters to achieve the best performance. This "tuning process" utilizes interim reports from a performance test within a dedicated subset of training dataset (validation set) or multiple subsets produced by a cross-validation scheme or bootstrapping. The main purpose of the process is to handle the bias-variance trade-off problem [43]. The bias is an error from erroneous assumptions in the learning algorithm. The variance refers to the ability of a model to fit complex data; thus, high variance indicates high flexibility, which is the main strength of many black box ML algorithms. However, this high flexibility can make these algorithms register even meaningless noises, making them prone to overfitting. Larger training datasets can alleviate the problem; however, practitioners are often required to tune the hyperparameters to identify the optimal balance between the bias and the variance.

However, consideration of how a ML product will be implemented in a real world situation has more practical importance. For example, consideration of which predictor will become available when is required because data are collected cumulatively in the real world, especially in clinical situations. In ED triage, basic information, such as demographic information, CC, vital signs, SpO_2 and consciousness, is available at the triage. In contrast, detailed clinical findings and most of the laboratory test results would not be available from the start unless the patient was already evaluated in outpatient departments or other facilities. Even when such data are available, there is a model complexity issue in which too many predictors make a model slow and vulnerable to overfitting.

At the beginning of this study, we had high hopes for the MLP algorithm, considering the recent reports of successful applications of deep learning in various fields [14]. Layers in a "deep" neural network can sequentially transform input vectors nonlinearly, extracting multiple levels of representations that correspond to different levels of abstraction. However, despite our extensive search, the MLP classifiers consistently required only one or two layers for best performance in various conditions of predictor availability. In addition, it consistently underperformed compared with tree-based classifiers (e.g., GBM and RF) and was not improved by inclusion of CC embedding. These findings suggest screening septic shock is not a "deep" problem and thus does not require a large neural network for best performance. We suggest using conventional ML algorithms or their ensembles for screening septic shock at ED triage.

This study has several limitations. First, we used an EHR database instead of a clinical registry. Therefore, the identification of the study population and the outcome events (septic shock) were based on time-stamped EHR records instead of clinician's comprehensive evaluation. However, the objective of this study was to provide a practical description of the performance of ML classifiers. Because most practical ML classifiers would depend on EHR database for its development and operation, we think our approach has some merit. Second, patients' underlying conditions were not incorporated into predictors. We think its inclusion may improve overall classification performance of the classifiers. Third, we tested too many algorithms in large hyperparameter spaces. In addition, grid search is an inefficient method to optimize the hyperparameters and consumes considerable computing resource and time. Fourth, the 60:40 data split ratio was arbitrary, and increasing

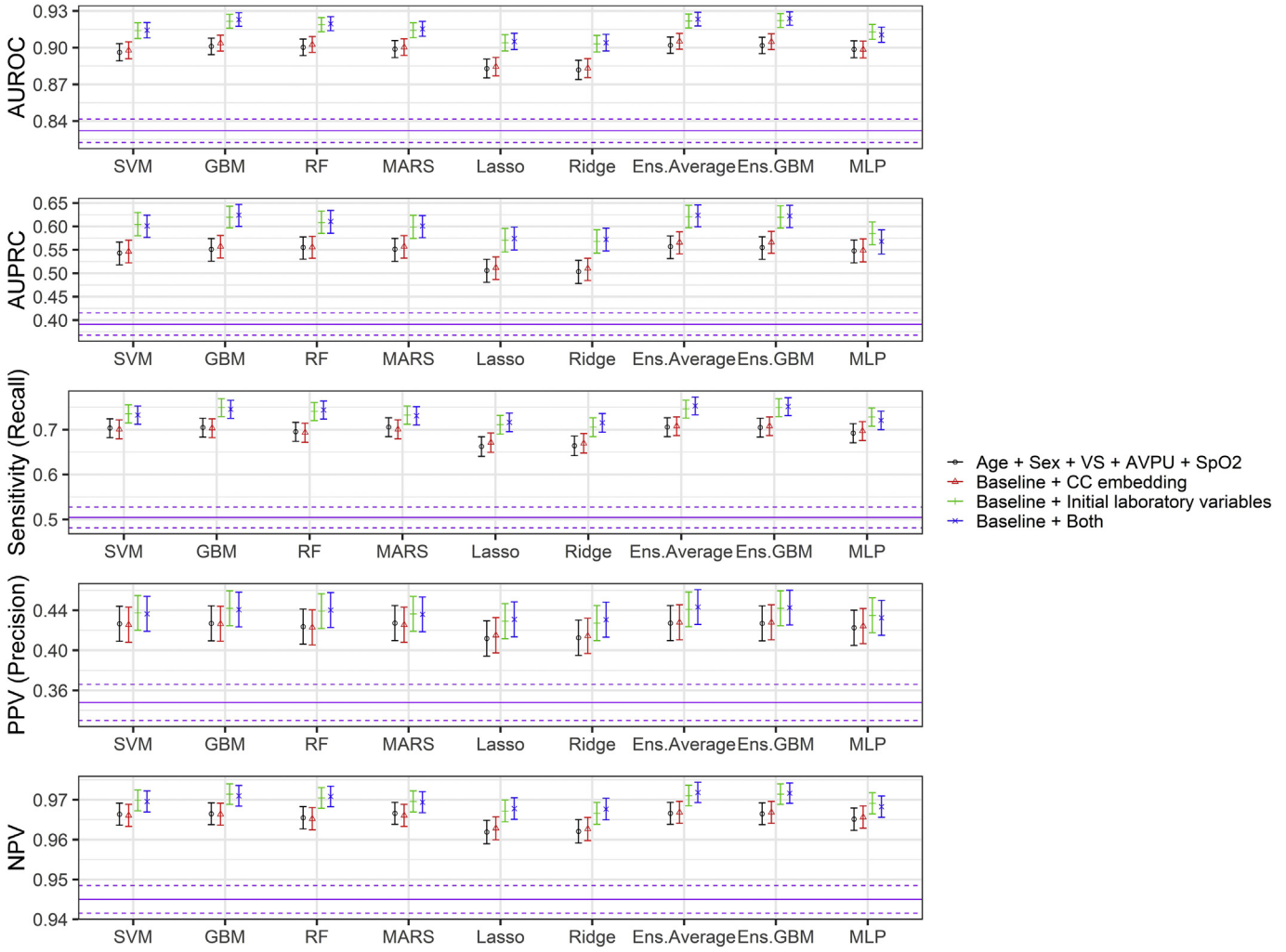

Fig. 3. Performance of ML classifiers compared to age and sex-adjusted qSOFA score (The purple horizontal solid and dashed lines indicate the mean and 95% confidence interval of age and sex-adjusted qSOFA score; AUROC, area under receiver operating characteristic; AUPRC, area under the precision–recall curve; PPV, positive predictive value; NPV, negative predictive value; VS, vital sign; AVPU, alert/verbal/pain/unresponsive); CC, chief complaint; SVM, support vector machine; GBM, gradient boosting machine; RF, random forest; MARS, multivariate adaptive regression splines; LASSO, least absolute shrinkage and selection operator; MLP, multilayer perceptron).

the training portion might lead to better ML model performances. Finally, this is a single center study, which could represent a concern for generalizability.

5. Conclusion

Developing ML-based classifiers for screening septic shock at ED triage using EHR database was feasible. The performance of ML classifiers was high enough for practical use. Ensembles of base classifiers showed the best performance and additional information from CC embedding provided relatively small gain.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcrc.2019.09.024.

Funding source

This work was supported by Research Resettlement Fund for the new faculty of Seoul National University.

Disclosures

The authors declare no conflict of interest.

Author contributions

Research conception & design: J Kim, K Kim. Data analysis and interpretation: J Kim, H Chang. Drafting of the manuscript: J Kim, H Chang. Critical revision and editing: J Kim, H Chang, You Jo, K Kim. Approval of final manuscript: all authors.

Acknowledgements

This work was supported by Research Resettlement Fund for the new faculty of Seoul National University

References

[1] Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, et al. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. Am J Respir Crit Care Med 2016;193:259–72. https://doi.org/10.1164/rccm.201504-0781OC.

[2] Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger PR, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program*. Crit Care Med 2014;42:1749. https://doi.org/10.1097/CCM.0000000000000330.

[3] Filbin MR, Arias SA, Camargo CA, Barche A, Pallin DJ. Sepsis visits and antibiotic utilization in U.S. emergency departments*. Crit Care Med 2014;42:528. https://doi.org/10.1097/CCM.0000000000000037.

[4] Herring AA, Ginde AA, Fahimi J, Alter HJ, Maselli JH, Espinola JA, et al. Increasing critical care admissions from U.S. Emergency Departments, 2001–2009*. Crit Care Med 2013;41:1197. https://doi.org/10.1097/CCM.0b013e31827c086f.
[5] Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, et al. Developing a new definition and assessing new clinical criteria for septic shock: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315:775–87. https://doi.org/10.1001/jama.2016.0289.
[6] Villar J, Clement JP, Stotts J, Linnen D, Rubin DJ, Thompson D, et al. Many emergency department patients with severe sepsis and septic shock do not meet diagnostic criteria within 3 hours of arrival. Ann Emerg Med 2014;64:48–54. https://doi.org/10.1016/j.annemergmed.2014.02.023.
[7] Velt K, Cnossen M, Rood PP, Steyerberg EW, Polinder S, Lingsma HF, et al. Emergency department overcrowding: a survey among European neurotrauma centres. Emerg Med J 2018;35. https://doi.org/10.1136/emermed-2017-206796 emermed-2017-206796.
[8] Richardson DB, Mountain D. Myths versus facts in emergency department overcrowding and hospital access block. Med J Aust 2009;190:369–74.
[9] Bernstein SL, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, et al. The effect of emergency department crowding on clinically oriented outcomes. Acad Emerg Med 2009;16:1–10. https://doi.org/10.1111/j.1553-2712.2008.00295.x.
[10] Singer M, Deutschman CS, Seymour C, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315:801–10. https://doi.org/10.1001/jama.2016.0287.
[11] Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017;97:120–7. https://doi.org/10.1016/j.ijmedinf.2016.09.014.
[12] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE 2017;12:e0174944. https://doi.org/10.1371/journal.pone.0174944.
[13] Rav D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017;21:4–21. https://doi.org/10.1109/jbhi.2016.2636665.
[14] Miotto R, Wang F, Wang S, Jiang X. Deep learning for healthcare: review, opportunities and challenges. Briefings In 2017. https://doi.org/10.1093/bib/bbx044/3800524.
[15] Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 2017;38:500–7. https://doi.org/10.1093/eurheartj/ehw188.
[16] Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci 2017;5:457–69. https://doi.org/10.1177/2167702617691560.
[17] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med 2016;44:368. https://doi.org/10.1097/ccm.0000000000001571.
[18] Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016;4:e28. https://doi.org/10.2196/medinform.5909.
[19] Kessler R, van Loo H, Wardenaar K, Bossarte R, Brenner L, Cai T, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry 2016;21:1366–71. https://doi.org/10.1038/mp.2015.198.
[20] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
[21] Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis I. Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform 2015;84:189–97. https://doi.org/10.1016/j.ijmedinf.2014.10.002.
[22] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Inform Assoc 2014;21:e278–86. https://doi.org/10.1136/amiajnl-2013-002512.
[23] Burdick H, Pino E, Gabel-Comeau D, Gu C, Huang H, Lynn-Palevsky A, et al. Evaluating a sepsis prediction machine learning algorithm in the emergency department and intensive care unit: a before and after comparative study. BioRxiv 2018:224014. https://doi.org/10.1101/224014.
[24] Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate orediction of sepsis in the ICU. Crit Care Med 2017;1. https://doi.org/10.1097/CCM.0000000000002936 Publish Ahead of Print.
[25] Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4:e000234. https://doi.org/10.1136/bmjresp-2017-000234.
[26] Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. Comput Biol Med 2016;74:69–73. https://doi.org/10.1016/j.compbiomed.2016.05.003.
[27] Back J, Jin Y, Jin T, Lee S. Development and validation of an automated sepsis risk assessment system. Res Nurs Health 2016;39:317–27. https://doi.org/10.1002/nur.21734.
[28] Trzeciak S, McCoy JV, Dellinger PR, Arnold RC, Rizzuto M, Abate NL, et al. Early increases in microcirculatory perfusion during protocol-directed resuscitation are associated with reduced multi-organ failure at 24 h in patients with sepsis. Intensive Care Med 2008;34:2210–7. https://doi.org/10.1007/s00134-008-1193-6.
[29] Jang D-H, Kim J, Jo Y, Lee J, Hwang J, Park S, et al. Developing neural network models for early detection of cardiac arrest in emergency department. Am J Emerg Med 2019. https://doi.org/10.1016/j.ajem.2019.04.006.
[30] Kuhn M, Johnson K. Applied predictive modeling. , vol. 26Springer; 2013.
[31] Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. J Mach Learn Res n.d.;17:1–5.
[32] Fisher, Aaron, Rudin, Cynthia, Dominici, Francesca. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon" Perspective.; 2018.
[33] Thiel SW, Rosini JM, Shannon W, Doherty JA, Micek ST, Kollef MH. Early prediction of septic shock in hospitalized patients. J Hosp Med 2010;5:19–25. https://doi.org/10.1002/jhm.530.
[34] Tang CH, Middleton PM, Savkin AV, Chan GS, Bishop S, Lovell NH. Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study. Physiol Meas 2010;31:775–93. https://doi.org/10.1088/0967-3334/31/6/004.
[35] Nachimuthu SK, Haug PJ. Early detection of sepsis in the emergency department using Dynamic Bayesian Networks. AMIA. Annual Symposium Proceedings / AMIA Symposium AMIA Symposium 2012; 2012. p. 653–62.
[36] Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. AMIA Annual Symposium Proceedings AMIA Symposium 2013; 2013. p. 1109–15.
[37] Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015;7. https://doi.org/10.1126/scitranslmed.aab3719 299ra122-299ra122.
[38] Haug P, Ferraro J. Using a semi-automated Modeling environment to construct a Bayesian, Sepsis diagnostic system. Acm 2016:571–8. https://doi.org/10.1145/2975167.2985841.
[39] Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLOS ONE 2017;12:e0174708. https://doi.org/10.1371/journal.pone.0174708.
[40] Zhang Y, Lin C, Chi M, Ivy J, Capan M, Huddleston JM. LSTM for septic shock: adding unreliable labels to reliable predictions. IEEE International Conference on Big Data (Big Data) 2017; 2017. p. 1233–42. https://doi.org/10.1109/BigData.2017.8258049.
[41] Culliton, Phil, Levinson, Michael, Ehresman, Alice, Wherry, Joshua, Steingrub, Jay, Gallant, Steve. Predicting Severe Sepsis Using Text from the Electronic Health Record, 10; 2017.
[42] Goldstein BA, Navar A, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J n.d.:ehw302. doi:https://doi.org/10.1093/eurheartj/ehw302.
[43] Trevor H, Robert T, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer; 2009.