Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis

Ryan J. Delahanty, PhD; JoAnn Alvarez, MS; Lisa M. Flynn, MD; Robert L. Sherwin, MD; Spencer S. Jones, PhD*

*Corresponding Author. E-mail: ssj1364@gmail.com, Twitter: @ssj1364.

Study objective: The Third International Consensus Definitions (Sepsis-3) Task Force recommended the use of the quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA) score to screen patients for sepsis outside of the ICU. However, subsequent studies raise concerns about the sensitivity of qSOFA as a screening tool. We aim to use machine learning to develop a new sepsis screening tool, the Risk of Sepsis (RoS) score, and compare it with a slate of benchmark sepsis-screening tools, including the Systemic Inflammatory Response Syndrome, Sequential Organ Failure Assessment (SOFA), qSOFA, Modified Early Warning Score, and National Early Warning Score.

Methods: We used retrospective electronic health record data from adult patients who presented to 49 urban community hospital emergency departments during a 22-month period (N=2,759,529). We used the Rhee clinical surveillance criteria as our standard definition of sepsis and as the primary target for developing our model. The data were randomly split into training and test cohorts to derive and then evaluate the model. A feature selection process was carried out in 3 stages: first, we reviewed existing models for sepsis screening; second, we consulted with local subject matter experts; and third, we used a supervised machine learning called gradient boosting. Key metrics of performance included alert rate, area under the receiver operating characteristic curve, sensitivity, specificity, and precision. Performance was assessed at 1, 3, 6, 12, and 24 hours after an index time.

Results: The RoS score was the most discriminant screening tool at all time thresholds (area under the receiver operating characteristic curve 0.93 to 0.97). Compared with the next most discriminant benchmark (Sequential Organ Failure Assessment), RoS was significantly more sensitive (67.7% versus 49.2% at 1 hour and 84.6% versus 80.4% at 24 hours) and precise (27.6% versus 12.2% at 1 hour and 28.8% versus 11.4% at 24 hours). The sensitivity of qSOFA was relatively low (3.7% at 1 hour and 23.5% at 24 hours).

Conclusion: In this retrospective study, RoS was more timely and discriminant than benchmark screening tools, including those recommend by the Sepsis-3 Task Force. Further study is needed to validate the RoS score at independent sites. [Ann Emerg Med. 2019;73:334-344.]

Please see page 335 for the Editor's Capsule Summary of this article.

Readers: click on the link to go directly to a survey in which you can provide **feedback** to *Annals* on this particular article.
A **podcast** for this article is available at www.annemergmed.com.

0196-0644/$-see front matter

https://doi.org/10.1016/j.annemergmed.2018.11.036

INTRODUCTION

Sepsis is a syndrome without a criterion standard diagnostic test,[1] and the challenges associated with defining it have made it difficult to quantify the associated morbidity and mortality.[2] The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) incorporated nearly two decades of advances in pathobiology, epidemiology, and management into a new definition of sepsis.[1,3,4] The Sepsis-3 definition has the potential to benefit researchers and public health officials interested in consistently measuring sepsis incidence and trends.[5] In addition to the potential epidemiologic benefits, researchers interested in developing models for the early identification of sepsis welcome the new definitions. The development of screening tools for sepsis that use statistical and, more recently, machine learning methods is an active and important area of investigation,[6-10] and the development of these tools depends on the availability of "labeled data," which in machine learning parlance refers to a data set that includes a reliable "target" (ie, the outcome of interest). Although not a criterion standard diagnostic, the Sepsis-3 definition does provide a consistent consensus-based target on which machine learning models can be developed and validated.

Editor's Capsule Summary

What is already known on this topic

Despite the availability of screening tools, sepsis remains a difficult disease to detect early.

What question this study addressed

This study addressed whether the use of machine learning on a large population (>2 million) could develop a risk score for sepsis identification that outperformed existing screening tools.

What this study adds to our knowledge

The Risk of Sepsis score, which uses commonly available clinical data such as lactate level, neutrophil levels, and shock index, performed superiorly to the Sequential Organ Failure Assessment, the quick Sequential [Sepsis-related] Organ Failure Assessment, and other recommended methods across all points up to 24 hours.

How this is relevant to clinical practice

Although prospective validation is needed, embedding Risk of Sepsis scores in electronic health records could result in improved sepsis identification.

Machine learning models are a departure from traditional screening tools for sepsis that are based on strong conceptual models. Traditional screening tools have the advantage of being relatively easy to describe and can often be calculated without assistance at the bedside. However, there is evidence suggesting that machine learning algorithms outperform traditional alternatives in contexts in which data inputs are abundant and where there is high potential for complex variable interactions.[11-13] Because of these attractive features, machine learning models are supplanting rule-based models in many industries.[12,14] Given the complexity and acknowledged gaps in our understanding of sepsis, screening for sepsis seems like an ideal use case for machine learning. In fact, some studies have already shown that machine learning models offer improved sepsis prognostication compared with the Systemic Inflammatory Response Syndrome (SIRS), quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA), and the Modified Early Warning Score (MEWS) among ICU patients.[6-8,10]

There is evidence that early initiation of treatment for sepsis is associated with significant reductions in morbidity and mortality.[1,15,16] A screening tool that provides more accurate and timely assessment of patients' risk for sepsis could facilitate earlier treatment and improved patient outcomes. In this article, we describe the development and evaluation of a new screening tool for sepsis, the Risk of Sepsis (RoS) score. We aimed to develop a tool that would incorporate the latest definition of sepsis, be applicable to all adult patients presenting to an emergency department (ED), and use machine learning methods to identify sepsis with a high degree of sensitivity and specificity in a timely fashion. We compared the performance of the RoS score with a number of benchmarks, including SIRS, the Sequential Organ Failure Assessment (SOFA), qSOFA, the National Early Warning Score, and MEWS.

MATERIALS AND METHODS

Study Design, Setting, and Selection of Participants

A retrospective cohort study was performed among all patients aged 18 years and older who presented to the ED at 49 urban community hospitals operated by Tenet Healthcare between January 1, 2016, and October 31, 2017. The hospitals are located in 39 cities across 9 states. In total, 2,856,060 patient encounters were eligible for inclusion. A patient encounter is defined as a continuous interaction between a patient and a facility (ie, a scenario in which a patient presents to the ED and is admitted as an inpatient is treated as a single encounter). Approximately 3% of encounters were excluded because no laboratory results or vital signs were documented, and ultimately 2,759,529 encounters (97%) were included in the study. The data were randomly split into training and test cohorts. Two thirds of the encounters were assigned to the training cohort; this group served as the basis for the derivation of the machine learning model, whereas the remaining third of the encounters were assigned to the testing cohort (Figure 1). This study was deemed to meet the conditions for exemption by the MetroWest Medical Center institutional review board.

Data Collection and Processing

Clinical data were captured as part of the usual processes of care in 1 of 7 instantiations of the Cerner Millennium electronic health record. Administrative data were also captured through the usual course of hospital operations. These data were archived in Tenet's enterprise data warehouse and extracted with structured query language.

Outcome Measures

We used the clinical surveillance criteria proposed by Rhee et al[5] as our standard definition of sepsis and as the primary target for developing our model (Figure E1, available online at http://www.annemergmed.com). Data for the entire encounter were used to determine whether the patient met the clinical surveillance criteria for sepsis.

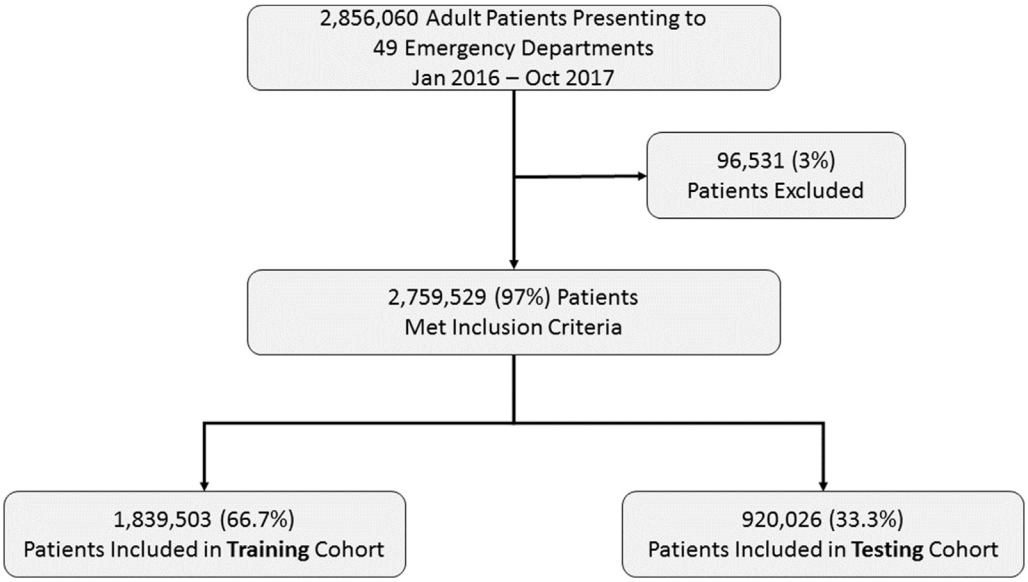

Figure 1. Study population and the split between the training and testing cohorts. *ML*, Machine learning.

Dr. Rhee generously shared SAS code (version 9.3; SAS Institute, Inc., Cary, NC) and instructions, which we applied to our data to identify sepsis-positive patients. To facilitate comparison with previous literature, we also considered inhospital mortality as a secondary outcome, and also evaluated the performance of the RoS score with inhospital morality as the target.

Given that early identification and treatment of sepsis is associated with improved outcomes,[1,15,16] we also believed it was important to evaluate the timeliness of the RoS score relative to benchmarks. Timeliness was assessed by evaluating model performance at thresholds of 1, 3, 6, 12, and 24 hours after an index time. The index time for each patient was represented by the time at which the first vital sign or laboratory result was documented in the electronic health record.

Feature selection is the process of identifying the subset of inputs that will be included in the final model from a larger number of available inputs. Our feature selection process was carried out in 3 stages: first, we reviewed existing models for sepsis screening; second, we consulted with local subject matter experts; and third, we used supervised machine learning, a process that is further described below. The first 2 stages of the process yielded a feature set consisting of 56 potential model inputs. The feature set included a subset of laboratory results, vital signs, demographics, administered medications, nursing documentation, and key words extracted from the ED chief complaint (Table E1, available online at http://www.annemergmed.com).

Feature engineering is the application of domain knowledge to potential model inputs, with the goal of creating a feature set that is optimized to predict the target. For the features that were likely to have multiple observations, we evaluated the inclusion of the first available, last available, mean, minimum, maximum, and overall change in value. We also created some "engineered" features that are a combination of multiple features (eg, shock index multiplied by age, which is the pulse rate divided by the systolic blood pressure multiplied by age)[17] and evaluated the inclusion of the first, last, and overall change in value for these engineered features. We replaced unobserved data points with an extreme value (–9,999). In our experience, extreme values indicating the absence of a feature produce better performance than other approaches for handling unobserved data.

Gradient boosting is a machine learning approach for regression and classification problems.[18] Gradient-boosted tree models are composed of thousands of relatively simple decision trees. These models are trained iteratively by combining individual decision trees to optimize a specified evaluation metric. At each iteration, an additional decision tree is added to the "ensemble" of previously trained decision trees. Each new decision tree takes into account errors made in the previous iterations. In this way, the model "learns" its own shortcomings and introduces a new decision tree to address them.

In total, 217 features were entered into the supervised machine learning feature selection process. Relative influence of the features was determined by evaluating "gain," which is the increase in accuracy brought about by including a given feature as a branch point in the decision trees. Initially, all features were entered into the model, and several rounds of feature selection ensued. At each round, the features with the lowest gain were omitted from the

model. This process was carried out until we observed substantial declines in the area under the receiver operator characteristic curve (AUROC). We used a learning rate of 0.3. We limited decision tree complexity to reach a maximum depth of 3. We also used 5-fold cross validation, a process by which we randomly separated the training set into 5 equal-sized partitions and then trained the model on 4 of the partitions and validated it on the remaining partition. This process was carried out iteratively so that each partition served as the validation set one time. Cross validation reduces the risk of overfitting the model. The model was developed with only the training cohort. The output of the RoS score model is the probability (0% to 100%) of a given patient's meeting the sepsis criteria at any point during their encounter. Model development and evaluation was carried out with R software (xgboost version 0.71.2; R version 3.5.1).[19,20]

Key evaluation metrics included AUROC, alert rate, sensitivity, specificity, and precision (positive predictive value). These metrics were calculated for the RoS score and the benchmarks. A screening threshold, or the value at which the patient would merit further treatment or evaluation for sepsis, was set for each model. For the benchmarks, we used the thresholds recommended in the literature.[4,21-23] We set the screening threshold for the RoS score at greater than or equal to 8.6%. This threshold was established to achieve an alert rate of approximately 5%, which is comparable to that of a commercially available sepsis alerting system currently implemented in our hospitals. To compare the timeliness of the models, evaluation metrics were assessed at 1, 3, 6, 12, and 24 hours after the index time. Using this evaluation framework, each algorithm was allowed to use only input data that were available up to the specified time threshold. However, we did use data for the entire encounter to determine whether patients met the criteria for sepsis.

AUROC and 95% confidence intervals for each algorithm were calculated for each time threshold. Tests of significance for differences between AUROC were performed with bootstrap replicates.[24] Agreement between the RoS score and benchmarks was assessed by Cronbach's α. We plotted curves of the observed and predicted rates of sepsis across risk deciles to evaluate model calibration. All evaluation metrics are based on performance in the testing cohort.

RESULTS

Characteristics of Study Subjects

Average annual ED volumes at the study hospitals ranged from 7,314 to 85,286 (median volume 33,960; interquartile range 25,238 to 43,601). Table 1 illustrates the demographic and geographic diversity of our large cohort of patients. We did not observe significant differences in the demographics, patient type, comorbid diagnoses, or outcomes between the testing and training cohorts. Differing from previous studies, the majority (79.5%) of our denominator population were not admitted as inpatients. Among all encounters, 54,661 (2.0%) met Rhee's criteria for sepsis. Patients who met the Rhee's criteria for sepsis were more likely to have been transferred to an ICU (53.0% versus 3.4%; $P<.001$), to have one of the selected comorbid diagnoses (20.4% versus 3.0%; $P<.001$), and to have died inhospital (11.9% versus 0.4%; $P<.001$). Rhee's criteria for sepsis were met in 39.7% of all inhospital deaths.

The final RoS score model retained 13 features (Table 2), including the first and last values for shock index multiplied by age, 3 vital signs, and 8 laboratory results. Lactic acid was the most influential input to the model, with both the maximum lactic acid result and the change in lactic acid level retained in the model. Table 2 also illustrates that the RoS score is robust to unobserved data and can extract important information based on the presence or absence of a given feature. For example, lactic acid level is influential in the model despite being unobserved for more than 90% of patients.

Figure 2 presents AUROCs and 95% confidence intervals for each screening tool at 24 hours. We observed in the testing cohort that the RoS score (AUROC=0.97) offered significantly better discrimination than all of the benchmarks. SOFA was the second most discriminant (AUROC=0.90), and SIRS was the least discriminant (AUROC=0.77). The RoS score and SOFA scores demonstrated a modest level of agreement (α=.72), whereas agreement was relatively poor between the RoS score and the other benchmarks. Table E2 (available online at http://www.annemergmed.com) presents a comparison of the inputs used in the various models. This table shows substantial overlap in model inputs.

The RoS score demonstrated significantly better discrimination in the testing cohort than the benchmarks across all time thresholds (Figure 3). In fact, it demonstrated better discrimination at 1 hour (AUROC=0.93) than the closest-performing benchmark, SOFA (AUROC=0.90), did at 24 hours. The RoS score became more discriminant over time (AUROC=0.93 at 1 hour versus 0.97 at 24 hours). The change in discrimination is further illustrated in Table E3 (available online at http://www.annemergmed.com), which shows the substantial separation between the RoS scores of patients who did and did not meet Rhee's sepsis criteria. The RoS scores at 1 hour of patients who did

Table 1. Characteristics of patients presenting to any of 49 EDs from January 1, 2016, through October 2017.

	All Encounters (N=2,759,529)		Training Cohort (N=1,839,503)		Testing Cohort (N=920,026)	
	Nonsepsis Encounters (N=2,704,868)	Sepsis Encounters (N=54,661)	Nonsepsis Encounters (N=1,803,045)	Sepsis Encounters (N=36,458)	Nonsepsis Encounters (N=901,823)	Sepsis Encounters (N=18,203)
Demographics						
Age, mean (SD), y	47.6 (20.2)	66.6 (17.1)	47.6 (20.2)	66.66 (17.2)	47.7 (20.2)	66.5 (17.1)
Men, No. (%)	1,138,975 (42.1)	27,859 (51.0)	758,161 (42.0)	18,595 (51.0)	380,814 (42.2)	9,264 (50.9)
Race, No. (%)						
White	1,728,420 (63.9)	38,896 (71.2)	1,151,978 (63.9)	25,950 (71.2)	576,442 (63.9)	12,946 (71.1)
Black	633,766 (23.4)	9,195 (16.8)	422,648 (23.4)	6,153 (16.9)	211,118 (23.4)	3,042 (16.7)
Asian	50,604 (1.9)	1,697 (3.1)	33,688 (1.9)	1,099 (3.0)	16,916 (1.9)	598 (3.3)
Other	292,078 (10.8)	4,873 (8.9)	194,731 (10.8)	3,256 (8.9)	97,347 (10.8)	1,617 (8.9)
Region, No. (%)						
Southeast	1,006,470 (37.2)	21,336 (39.0)	671,141 (37.2)	14,209 (39.0)	335,329 (37.2)	7,127 (39.2)
Midwest	100,483 (3.7)	1,903 (3.5)	66,801 (3.7)	1,280 (3.5)	33,682 (3.7)	623 (3.4)
Southwest	592,255 (21.9)	13,303 (24.3)	395,014 (21.9)	8,881 (24.4)	197,241 (21.9)	4,422 (24.3)
West	1,005,660 (37.2)	18,119 (33.1)	670,089 (37.2)	12,088 (33.2)	335,571 (37.2)	6,031 (33.1)
Patient type, No. (%)						
Outpatient*	2,193,276 (81.1)	158 (0.3)	1,462,066 (81.1)	106 (0.3)	731,210 (81.1)	52 (0.3)
Inpatient	511,592 (18.9)	54,503 (99.7)	340,979 (18.9)	36,352 (99.7)	170,613 (18.9)	18,151 (99.7)
Surgical	76,829 (2.8)	10,169 (18.6)	51,161 (2.8)	6,789 (18.6)	25,668 (2.8)	3,380 (18.6)
ICU	92,689 (3.4)	28,984 (53.0)	61,582 (3.4)	19,382 (53.2)	31,107 (3.4)	9,602 (52.7)
Comorbid diagnoses, No. (%)						
CHF	8,003 (0.3)	1,329 (2.4)	5,312 (0.3)	904 (2.5)	2,691 (0.3)	425 (2.3)
COPD	20,463 (0.8)	1,689 (3.1)	13,652 (0.8)	1,152 (3.2)	6,811 (0.8)	537 (3.0)
Dementia	6,804 (0.3)	1,011 (1.8)	4,569 (0.3)	676 (1.9)	2,235 (0.2)	335 (1.8)
Diabetes	7,316 (0.3)	978 (1.8)	4,905 (0.3)	658 (1.8)	2,411 (0.3)	320 (1.8)
HIV	1,511 (0.1)	75 (0.1)	1,024 (0.1)	54 (0.1)	487 (0.1)	21 (0.1)
Liver disease	3,403 (0.1)	795 (1.5)	2,271 (0.1)	514 (1.4)	1,132 (0.1)	281 (1.5)
Cancer	9,398 (0.3)	1,491 (2.7)	6,244 (0.3)	983 (2.7)	3,154 (0.3)	508 (2.8)
Renal disease	23,071 (0.9)	3,840 (7.0)	15,333 (0.9)	2,522 (6.9)	7,738 (0.9)	1,318 (7.2)
Outcomes						
Inhospital mortality, No. (%)	9,873 (0.4)	6,511 (11.9)	6,589 (0.4)	4,354 (11.9)	3,284 (0.4)	2,157 (11.8)
Outpatient* LOS, median (IQR), h	3.1 (1.9–5.1)	3.8 (2.5–6.0)	3.1 (1.9–5.1)	3.6 (2.1–6.1)	3.10 (1.9–5.1)	4.13 (2.63–6.1)
Inpatient LOS, median (IQR), h	70.8 (43.2–119.4)	161.8 (100.1–265.3)	70.8 (43.1–119.3)	161.4 (100.1–264.9)	70.9 (43.3–119.5)	162.6 (100.0–266.8)

*ED or observation patients.

Table 2. Relative influence and percentage of unobserved of features retained in the RoS score model.

Feature	Median (Interquartile Range)	Relative Influence, %*	Unobserved, %†
Lactic acid (max), mmol/L	1.5 (1.1 to 2.4)	52.4	90.7
Shock index‡×age (last)	27.4 (19.5 to 37.1)	6.5	0.6
WBC count (max), 10^9/L	8.5 (6.6 to 11.1)	5.4	46.4
Lactic acid (change), mmol/L	−0.8 (−1.6 to −0.2)	4.4	90.7
Neutrophils (max), %	68 (58.6 to 77.7)	4.2	49.6
Glucose (max), mg/dL	110 (96 to 142)	4.2	48.0
Blood urea nitrogen (max), mg/dL	14 (10.2 to 20)	3.9	49.7
Shock index‡×age (first)	27.6 (19.7 to 37.5)	3.9	0.6
Respiratory rate (max), breaths/min	18 (18 to 20)	3.8	0.7
Albumin (last), g/dL	4 (3.6 to 4.3)	3.4	58.5
Systolic blood pressure (min), mm Hg	128 (114 to 142)	3.1	0.4
Serum creatinine (max), mg/dL	0.8 (0.7 to 1.1)	2.5	49.7
Temperature (max), °F	98.3 (98 to 98.6)	2.5	2.8

*Relative influence (gain): Takes into account each feature's contribution to each tree in the model. A higher value implies increased importance relative to other model features. All values sum to 100%.
†Unobserved at 24 hours after first laboratory results or vital signs recorded.
‡Shock index=(pulse rate/systolic blood pressure).

not meet Rhee's sepsis criteria were very low (median RoS score=0.18%) compared with that of individuals who met Rhee's sepsis criteria (median RoS score=21.5%). This gap was even wider at 24 hours (median RoS score for no sepsis 0.12% versus 43.6% for sepsis).

Model calibration refers to how well predicted probabilities align with observed probabilities. For example, if the average predicted probability for a subset of the population is 10%, we would expect that approximately 10% of that same subset would meet sepsis criteria if the model is well calibrated. Visual analysis showed relatively close agreement between predicted and observed probabilities (Figure E2, available online at http://www.annemergmed.com); however, the curve shows that the RoS score model has a tendency to overestimate the probability of sepsis.

Previous studies have used inhospital mortality as a target for their models[4,25-29]; therefore, we also assessed the discrimination of the RoS score and the benchmarks for inhospital mortality. The RoS score demonstrated significantly higher discrimination at 24 hours (AUROC=0.92) than all of the benchmarks (Figure E3, available online at http://www.annemergmed.com). As was the case for the identification of sepsis, the SOFA and National Early Warning Score scores were the next most discriminant (AUROC=0.89).

We were most interested in evaluating how the RoS score would perform in the context of clinical decision support. Table 3 shows the alert rate, sensitivity, specificity, precision, and AUROC for the RoS score and the benchmarks at the specified screening and time thresholds. At 1 hour, the RoS score achieved 67.7% sensitivity and precision of 27.6%. SOFA and SIRS were the closest benchmarks in terms of sensitivity (49.2% and 40.4%, respectively), but both had much higher alert rates (SOFA=8.0% and SIRS=7.1%) and much lower precision (SOFA=12.2% and SIRS=11.2%). The alert rate and precision for the RoS score remained fairly stable over time (alert rate 4.9% to 5.8%; precision 27.6% to 31.9%), whereas sensitivity increased substantially over time (67.7% at 1 hour versus 84.6% at 24 hours). qSOFA was very specific (specificity 98.5% to 99.8%) but was one of the least sensitive screening tools (sensitivity 3.7% to 23.5%). Overall, the RoS score performed significantly better across multiple dimensions of prognostic ability than the benchmarks.

LIMITATIONS

We used a set of well-documented clinical criteria for sepsis as the target for this model.[5] We believe that this approach for identifying sepsis-positive patients was the best of available options and enabled us to assemble a large cohort of labeled data. However, we acknowledge that the lack of criterion standard diagnostic for sepsis continues to be a challenge and is a limitation of this analysis.

Machine learning models may be perceived as "black boxes," whereas the benchmark models are easy to describe and calculate. However, there is increasing sentiment that machine learning models are well suited for clinical

	RoS	SOFA	SIRS	NEWS	qSOFA	MEWS
RoS	0.97 (0.97 - 0.98)	0.72 (0.71 - 0.72)	0.51 (0.50 - 0.51)	0.54 (0.53 - 0.54)	0.57 (0.57 - 0.58)	0.55 (0.54 - 0.55)
SOFA		0.90 (0.90 - 0.91)	0.38 (0.38- 0.39)	0.63 (0.62 - 0.63)	0.71 (0.70- 0.71)	0.60 (0.59 - 0.60)
SIRS			0.77 (0.76 - 0.78)	0.58 (0.57 - 0.58)	0.44 (0.44- 0.45)	0.63 (0.63 - 0.64)
NEWS				0.84 (0.83 - 0.84)	0.79 (0.79 - 0.80)	0.83 (0.83 - 0.84)
qSOFA					0.80 (0.79 - 0.80)	0.78 (0.77 - 0.78)
MEWS						0.78 (0.78 - 0.79)

Figure 2. AUROC and 95% confidence intervals of RoS score versus benchmark screening tools for sepsis model performance based on data available up to 24 hours after the index time in the testing cohort. Diagonal cells (blue) contain AUROC with confidence intervals. Cells above the diagonal contain Cronbach α statistics (with 95% confidence intervals), a measure of agreement.

applications.[11] An important limitation of this study is that our key evaluation metrics were technical performance measures, and we did not address how this type of screening tool could be integrated into clinical practice. We hypothesize that the RoS score could replace any of the benchmarks in clinical decision rules, but we acknowledge that the benchmarks, in particular SIRS, have the advantage of being familiar to clinicians and have been used broadly for more than 2 decades to identify and initiate treatment for sepsis. Further study is needed to determine the best ways to operationalize the RoS score for clinical practice.

We have highlighted the value of including all patients presenting to the ED in our denominator population and imply that the RoS could be used to screen all ED patients for sepsis. We believe that this approach could lead to more timely identification of patients at risk for sepsis but also acknowledge that, like other sepsis screening tools, the RoS score has a high false-discovery rate and that its use among a broad population with relatively low pretest probability could lead to overly aggressive and unnecessary treatment. Further study is needed to determine optimal risk thresholds and appropriate clinical actions at different thresholds of risk.

Lactic acid results are an important input to the RoS score model. We are aware that evaluating a patient's lactic acid level is often a first step after sepsis is suspected, and that by including lactic acid level as an input our model may also be subject to the same criticisms we leveled against other models (ie, that by restricting use cases to populations in which there is already cause for suspicion of infection, the utility of the model is limited). To investigate this concern, we performed a sensitivity analysis in which we replaced all lactic acid results in the testing cohort with an extreme value (–9,999), effectively "blinding" the model to

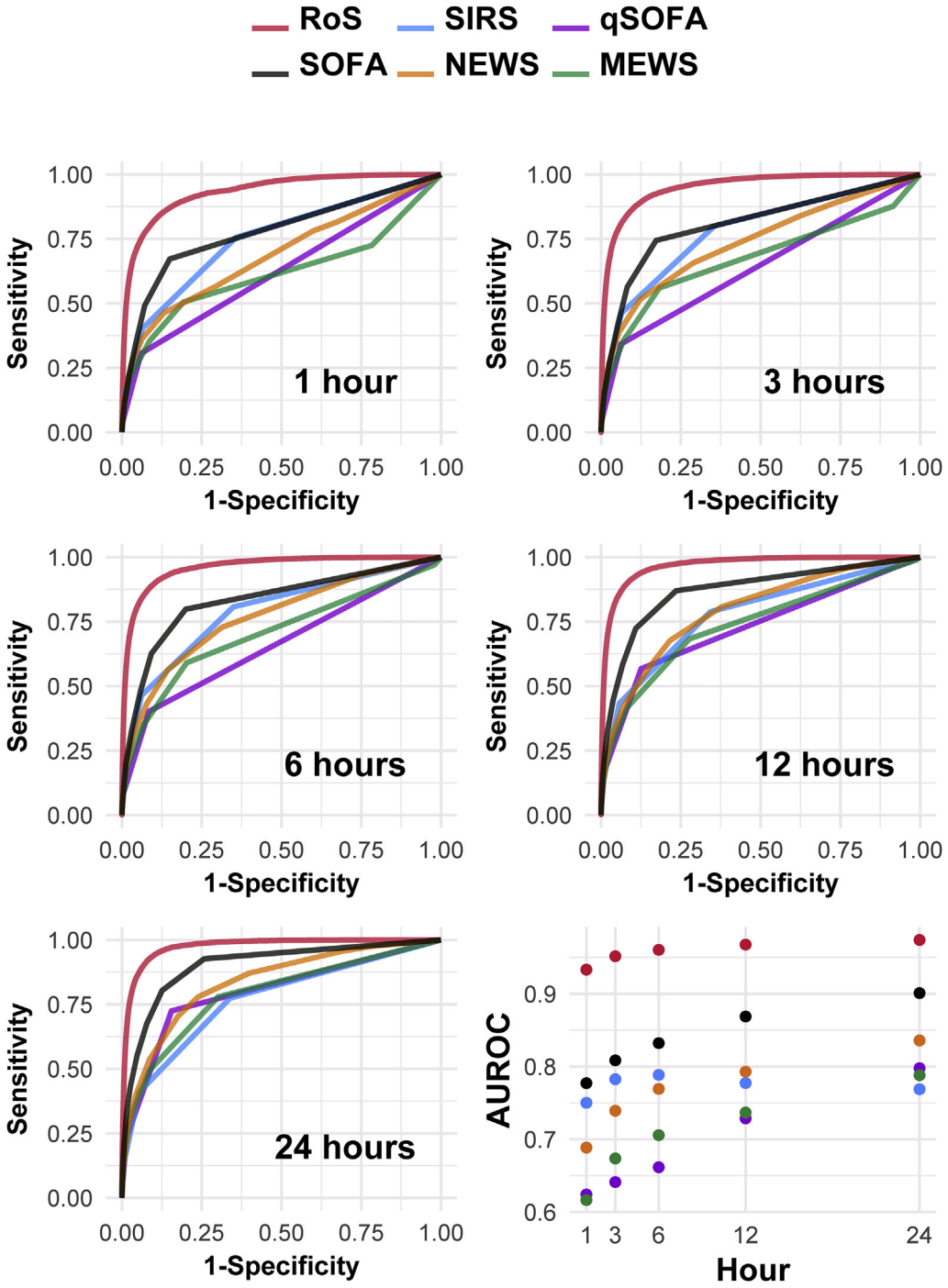

Figure 3. Receiver operating characteristic curves for RoS score compared with benchmark screening tools at specified time thresholds in the testing cohort.

lactic acid measurements, and then reevaluated the performance statistics. Table E4 (available online at http://www.annemergmed.com) presents the results of this analysis and shows that the RoS score model is robust to unobserved lactic acid results and continues to be very discriminant (AUROC=0.91 at 1 hour and 0.96 at 24 hours), and still outperforms the benchmarks.

We used robust and well-validated methods to ensure that the RoS score model would perform well when exposed to new data. However, despite our large cohort of patients, the demonstrable heterogeneity across our hospitals, and our robust training and testing framework, further external and independent evaluations are needed. Existing benchmarks have been developed during the

Table 3. Performance statistics of RoS score compared with benchmarks (1 to 24 hours).

Model	Screening Threshold	Time Threshold, Hours	Alert Rate, %	Sensitivity, %	Specificity, %	Precision, %*	AUROC
RoS	≥8.6%	1	4.9	67.7	96.4	27.6	0.93
qSOFA	≥2	1	0.2	3.7	99.8	31.3	0.62
SOFA	≥2	1	8.0	49.2	92.9	12.2	0.78
SIRS	≥2	1	7.1	40.4	93.6	11.2	0.75
MEWS	≥5	1	1.1	11.5	99.1	21.2	0.62
NEWS	≥7	1	1.7	18.2	98.6	20.7	0.69
RoS	≥8.6%	3	4.8	72.1	96.6	30.0	0.95
qSOFA	≥2	3	0.3	4.4	99.8	32.6	0.64
SOFA	≥2	3	9.2	56.4	91.8	12.2	0.81
SIRS	≥2	3	7.2	46.2	93.6	12.7	0.78
MEWS	≥5	3	0.7	9.7	99.4	25.5	0.67
NEWS	≥7	3	1.5	18.2	98.8	24.1	0.74
RoS	≥8.6%	6	4.7	74.9	96.8	31.9	0.96
qSOFA	≥2	6	0.6	8.2	99.6	27.8	0.66
SOFA	≥2	6	10.3	62.7	90.8	12.1	0.83
SIRS	≥2	6	6.9	46.4	93.9	13.3	0.79
MEWS	≥5	6	0.7	9.1	99.5	27.0	0.71
NEWS	≥7	6	1.9	20.6	98.5	21.6	0.77
RoS	≥8.6%	12	5.1	79.3	96.4	30.8	0.97
qSOFA	≥2	12	1.4	17.1	99.0	24.8	0.73
SOFA	≥2	12	12.1	72.3	89.1	11.8	0.87
SIRS	≥2	12	6.4	43.3	94.3	13.3	0.78
MEWS	≥5	12	0.9	12.6	99.3	26.5	0.74
NEWS	≥7	12	3.8	30.2	96.8	15.8	0.79
RoS	≥8.6%	24	5.8	84.6	95.8	28.8	0.97
qSOFA	≥2	24	1.9	23.5	98.5	24.3	0.80
SOFA	≥2	24	13.9	80.4	87.4	11.4	0.90
SIRS	≥2	24	6.1	41.0	94.6	13.3	0.77
MEWS	≥5	24	1.1	16.9	99.2	30.0	0.79
NEWS	≥7	24	4.8	39.0	95.9	16.0	0.84

NEWS, National Early Warning Score.
*Precision=positive predictive value (number of sepsis-positive encounters/number of screen-positive encounters).

course of decades and have been externally validated in a number of settings. External evaluations are necessary to test whether a level of performance similar to that which was observed in this study could be achieved at independent sites.*

*Interested parties can contact the authors to receive the software necessary to implement and evaluate the RoS score model.

DISCUSSION

In this study, we sought to build a screening tool that would incorporate the latest definitions and surveillance tools for sepsis, be applicable to all adult patients who present to an ED, use machine learning methods, and be more discriminant, sensitive, precise, and timely than available benchmarks. Our objectives were informed by the limitations we identified through our review of the literature.

Table E5 (available online at http://www.annemergmed.com) includes summaries of 11 recent studies that describe and evaluate tools for sepsis screening or prediction. These studies illustrate the challenge of not having a criterion standard for sepsis. More than half of these studies did not use sepsis as the target against which the models were trained and validated; rather, the authors used inhospital mortality as the primary modeling target.[4,25-29] Three of the remaining studies used a combination of SIRS criteria and *International Classification of Diseases, Ninth Revision* codes as their target,[6-8] and the remaining 2 studies used

the Sepsis-3 criteria as their target.[10,30] The lack of a common target across these studies makes it difficult to judge the relative performance of the various algorithms. Our study addresses this limitation by conducting a head-to-head evaluation of the RoS score compared with the benchmarks, using a consistent and current definition of sepsis as the target.

Another important limitation of the existing literature is that many of the proposed sepsis screening models have been validated only in populations with "suspected infection,"[4,25-30] or in some cases have been limited to patients already in the ICU.[6-8,10,28] This is not reflective of the most compelling use case; a screening tool that can identify patients who should be further evaluated or treated for sepsis from among all patients has more utility than a tool that is meant to be used after suspicion of infection has been established or after patient transfer to the ICU. Our study addresses this limitation by evaluating performance among all patients who presented to the ED.

The performance statistics, particularly AUROC, reported in our analysis tended to be better than those reported in previous studies. The difference is the product of the 2 factors described above (ie, differences in the definition of the numerator and differences in the denominator). The implications of these differences are that the AUROCs previously presented will tend to be lower because their analyses were conducted with more homogenous populations (ie, all patients with suspected infection or all patients in the ICU), thus making it more difficult to discriminate between sepsis and nonsepsis cases or in some cases live discharged versus dead patients. In an effort to place our study in the context of existing literature, we analyzed the performance of the RoS score and benchmarks, using a denominator and numerator similar to those that were used by Seymour et al[4] (Appendix E1, available online at http://www.annemergmed.com). This analysis showed that RoS offered improved performance relative to benchmarks and that performance statistics for qSOFA and SOFA were within the range of those previously reported in the literature (Tables E6-E11, available online at http://www.annemergmed.com).

The definitional changes proposed by the Sepsis-3 Task Force have generally been well received; however, the recommendation that SOFA and qSOFA scores supplant SIRS-centric screening algorithms has been more controversial.[31] Some subsequent analyses have indicated that the increased specificity of qSOFA comes at the cost of inferior sensitivity compared with SIRS and that qSOFA is less discriminating than MEWS and the National Early Warning Score.[18,19,21] Our analysis reaffirms these concerns (Figure E4, available online at http://www.annemergmed.com). We observed that qSOFA had the lowest sensitivity of all of the screening tools up to the 6-hour threshold, and then offered slightly better sensitivity than only MEWS at the 12- and 24-hour thresholds.

Our analysis illustrates the performance advantages of using machine learning methods. We used a relatively small set of known risk factors as inputs to our model, many of which are included in the other benchmark models; however, our model significantly outperformed these benchmarks. Compared with the next best benchmark (SOFA), the RoS showed an increase in sensitivity of 18.5 percentage points (67.7% versus 49.2%) at 1 hour and 4.2 percentage points (84.6% versus 80.4%) at 24 hours; likewise, the RoS score offered an increase in precision over SOFA of 15.4 percentage points (27.6% versus 12.2%) at 1 hour and 17.4 percentage points (28.8% versus 11.4%) at 24 hours. We attribute these performance gains to our large training cohort and the flexibility of machine learning methods, which enabled us to more effectively model complex interactions that are not captured in the benchmark models. Our results suggest that the RoS score could be effective as a screening tool and that a RoS-driven screening could be significantly more sensitive than screening driven by any of the benchmarks, as well as reduce the number of false alerts (Figure E5, available online at http://www,annemergmed.com).

Using more than 2.5 million patient encounters from a clinically and geographically diverse group of hospitals, we have shown that the RoS score model performs significantly better than a slate of currently available screening tools for sepsis, including those recommended by the Sepsis-3 Task Force. Our results suggest that the RoS score model could be used in place of existing screening tools such as SOFA, qSOFA, and SIRS, and provide more timely and accurate alerts in regard to which patients should be further evaluated and treated for sepsis.

The authors acknowledge Syed Mateen, Jason Wise, and Dhruti Shah for their assistance with data acquisition, as well as Dr. Huiling Zhang, MD, for institutional support.

Supervising editor: Alan E. Jones, MD. Specific detailed information about possible conflict of interest for individual editors is available at https://www.annemergmed.com/editors.

Author affiliations: From the Tenet Healthcare Corporation, Nashville, TN (Delahanty, Alvarez, Flynn, Jones); and the Department of Emergency Medicine, Wayne State University, Detroit, MI (Sherwin).

Author contributions: RJD and SSJ conceived and designed the study. SSJ supervised the model development and analysis. JA implemented the code to identify sepsis-positive cases. RJD

developed and evaluated the machine learning model and created the tables and figures. SSJ drafted the article, and all authors contributed substantially to its revision. SSJ takes responsibility for the paper as a whole.

All authors attest to meeting the four ICMJE.org authorship criteria: (1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND (2) Drafting the work or revising it critically for important intellectual content; AND (3) Final approval of the version to be published; AND (4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist. Dr. Sherwin has received funding from the Agency for Healthcare Research and Quality (PA-14-001, Exploratory and Developmental Grant to Improve Health Care Quality through Health Information Technology [IT]–R21) for the project titled "Enhancing an EMR-Based Real-Time Sepsis Alert System Performance Through Machine Learning."

Publication dates: Received for publication July 3, 2018. Revisions received October 10, 2018, and November 14, 2018. Accepted for publication November 27, 2018.

REFERENCES

1. 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.
2. Rudd KE, Delaney A, Finfer S. Counting sepsis, an imprecise but improving science. *JAMA*. 2017;318:1228-1229.
3. Shankar-Hari M, Phillips GS, Levy ML, 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;315:775-787.
4. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315:762-774.
5. Rhee C, Dantes R, Epstein L, et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA*. 2017;318:1241-1249.
6. 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.
7. Calvert J, Desautels T, Chettipally U, et al. High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg*. 2016;8:50-55.
8. Calvert JS, Price DA, Chettipally UK, et al. A computational approach to early sepsis detection. *Comput Biol Med*. 2016;74:69-73.
9. Horng S, Sontag DA, Halpern Y, et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One*. 2017;12:e0174708.
10. Desautels T, Calvert J, Hoffman J, 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.
11. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319:1317-1318.
12. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16:199-231.
13. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Crit Care Med*. 2018;46:e481-e488.
14. Lewis-Kraus G. The great AI awakening. *New York Times*. Available at: https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html. Published December 14, 2016. Accessed April 25, 2018.
15. Dellinger RP, Levy MM, Rhodes A, et al. Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Crit Care Med*. 2013;41:580-637.
16. Levy MM, Rhodes A, Phillips GS, et al. Surviving Sepsis Campaign: association between performance metrics and outcomes in a 7.5-year study. *Crit Care Med*. 2015;43:3-12.
17. Bruijns SR, Guly HR, Bouamra O, et al. The value of traditional vital signs, shock index, and age-based markers in predicting trauma mortality. *J Trauma Acute Care Surg*. 2013;74:1432-1437.
18. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189-1232.
19. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. Available at: https://CRAN.R-project.org/package=xgboost. Accessed April 25, 2018.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available at: https://www.R-project.org/. Accessed January 2, 2019.
21. Bone RC, Balk RA, Cerra FB, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest*. 1992;101:1644-1655.
22. Subbe CP, Kruger M, Rutherford P, et al. Validation of a modified Early Warning Score in medical admissions. *QJM*. 2001;94:521-526.
23. 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.
24. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1-26.
25. Amland RC, Sutariya BB. Quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA) and St. John Sepsis Surveillance Agent to detect patients at risk of sepsis: an observational cohort study. *Am J Med Qual*. 2018;33:50-57.
26. Churpek MM, Snyder A, Han X, et al. Quick Sepsis-related Organ Failure Assessment, Systemic Inflammatory Response Syndrome, and Early Warning Scores for detecting clinical deterioration in infected patients outside the intensive care unit. *Am J Respir Crit Care Med*. 2017;195:906-911.
27. Lembke K, Parashar S, Simpson S. Sensitivity and specificity of SIRS, qSOFA and severe sepsis for mortality of patients presenting to the emergency department with suspected infection. *Chest*. 2017;152:A401.
28. Raith EP, Udy AA, Bailey M, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *JAMA*. 2017;317:290-300.
29. Freund Y, Lemachatti N, Krastinova E, et al. Prognostic accuracy of Sepsis-3 criteria for in-hospital mortality among patients with suspected infection presenting to the emergency department. *JAMA*. 2017;317:301-308.
30. Williams JM, Greenslade JH, McKenzie JV, et al. Systemic Inflammatory Response Syndrome, Quick Sequential Organ Function Assessment, and organ dysfunction: insights from a prospective database of ED patients with infection. *Chest*. 2017;151:586-596.
31. Simpson SQ. SIRS in the time of Sepsis-3. *Chest*. 2018;153:34-38.