A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice*

Heather M. Giannini, MD[1]; Jennifer C. Ginestra, MD[1]; Corey Chivers, PhD[2]; Michael Draugelis, BS[2]; Asaf Hanish, MPH[2]; William D. Schweickert, MD[2,3]; Barry D. Fuchs, MD, MS[2,3]; Laurie Meadows, RN, CCRN[4]; Michael Lynch, RN, CEN[4]; Patrick J. Donnelly, RN, MS, CCRN[5]; Kimberly Pavan, MSN, CRNP[6]; Neil O. Fishman, MD[2]; C. William Hanson, MD, III[2]; Craig A. Umscheid, MD, MSCE[2,7,8]

***See also p. 1650.**

[1]Department of Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA.

[2]University of Pennsylvania Health System, Philadelphia, PA.

[3]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA.

[4]Department of Nursing, Hospital of the University of Pennsylvania, Philadelphia, PA.

[5]Department of Clinical Informatics, Pennsylvania Hospital, Philadelphia, PA.

[6]Penn Presbyterian Medical Center, Philadelphia, PA.

[7]Division of General Internal Medicine, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

[8]Center for Evidence-based Practice, University of Pennsylvania Health System, Philadelphia, PA.

Drs. Giannini, Ginestra, Chivers, Draugelis, Schweickert, Fuchs, Meadows, Lynch, Donnelly, Pavan, Fishman, Hanson, and Umscheid contributed to conception and design. Drs. Giannini, Ginestra, Chivers, Draugelis, Hanish, Meadows, Lynch and Donnelly contributed to data collection. Drs. Giannini, Ginestra, Chivers, Draugelis, and Umscheid contributed to analysis and interpretation of data. Drs. Giannini, Ginestra, Chivers, and Umscheid contributed to drafting of article. Drs. Giannini, Ginestra, and Umscheid contributed to critical revision of article for important intellectual content.

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 National Center for Research Resources (grant no. UL1RR024134), which is now at the National Center for Advancing Translational Sciences (grant noUL1TR000003). Dr. Umscheid received support for article research from the National Institutes of Health (NIH). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Dr. Schweickert received funding from Arjo, Hill Rom, the Society of Critical Care Medicine (consulting), and American College of Physicians. Dr. Umscheid's institution received funding from National Center for Research Resources (grant no. UL1RR024134), which is now at the National Center for Advancing Translational Sciences (grant no. UL1TR000003); Agency for Healthcare Research and Quality Contracts Evidence-based Practice Center; and the U.S. Food and Drug Administration. He received funding from the Patient-Centered Outcomes Research Institute Advisory Panel and Northwell Health (honoraria for grand rounds). The remaining authors have disclosed that they do not have any potential conflicts of interest.

Address requests for reprints to: Craig A. Umscheid, MD, MS, Office of Clinical Excellence, University of Chicago Medicine, American School Building, 850 E. 58th Street, Suite 123, Office 128, MC 1135, Chicago, IL 60637. E-mail: craigumscheid@medicine.bsd.uchicago.edu

DOI: 10.1097/CCM.0000000000003891

Objectives: Develop and implement a machine learning algorithm to predict severe sepsis and septic shock and evaluate the impact on clinical practice and patient outcomes.

Design: Retrospective cohort for algorithm derivation and validation, pre-post impact evaluation.

Setting: Tertiary teaching hospital system in Philadelphia, PA.

Patients: All non-ICU admissions; algorithm derivation July 2011 to June 2014 (n = 162,212); algorithm validation October to December 2015 (n = 10,448); silent versus alert comparison January 2016 to February 2017 (silent n = 22,280; alert n = 32,184).

Interventions: A random-forest classifier, derived and validated using electronic health record data, was deployed both silently and later with an alert to notify clinical teams of sepsis prediction.

Measurement and Main Result: Patients identified for training the algorithm were required to have *International Classification of Diseases*, 9th Edition codes for severe sepsis or septic shock and a positive blood culture during their hospital encounter with either a lactate greater than 2.2 mmol/L or a systolic blood pressure less than 90 mm Hg. The algorithm demonstrated a sensitivity of 26% and specificity of 98%, with a positive predictive value of 29% and positive likelihood ratio of 13. The alert resulted in a small statistically significant increase in lactate testing and IV fluid administration. There was no significant difference in mortality, discharge disposition, or transfer to ICU, although there was a reduction in time-to-ICU transfer.

Conclusions: Our machine learning algorithm can predict, with low sensitivity but high specificity, the impending occurrence of severe sepsis and septic shock. Algorithm-generated predictive alerts modestly impacted clinical measures. Next steps include describing clinical perception of this tool and optimizing algorithm design and delivery. (*Crit Care Med* 2019; 47:1485–1492)

Key Words: early warning system; electronic medical record; machine learning; predictive medicine; septic shock; severe sepsis

Sepsis continues to be a leading cause of death among hospitalized patients, affecting up to 6% of all admissions and conferring in-hospital mortality greater than 15% (1, 2). Early detection of sepsis has the potential to reduce mortality by facilitating timely implementation of evidence-based interventions (3).

Many studies have used multivariate models based on electronic health record (EHR) data for detection of sepsis or clinical deterioration (4–8). Our team previously developed and implemented one such detection algorithm based on the systemic inflammatory response syndrome (SIRS) (7). In that study, a nonsignificant improvement in mortality and an increase in discharge to home was observed. More recently, our team as well as others have begun to use machine learning (ML) approaches (9) to improve the accuracy of sepsis detection and prediction, both in the emergency department (10–12) and the inpatient setting (13–15).

When applied to retrospective data, ML algorithms designed to predict sepsis have performed well (10, 14, 16, 17, 19). However, implementation has largely been focused on ICU populations, where robust staffing and a high index of suspicion already prompt early recognition of sepsis. For example, ML algorithms have been linked to decreased mortality and length of stay in a small ICU-based randomized trial (16) and decreased sepsis-related mortality at a small private hospital (18). However, to date, the large-scale application of ML algorithms to predict sepsis in the non-ICU inpatient setting has not been reported. Here, we describe the development of a ML algorithm and alert for prediction of severe sepsis and septic shock in hospitalized non-ICU patients and the subsequent clinical impact of this tool when implemented across our multihospital healthcare system.

METHODS

Setting and Data Sources

At the time of the study, the University of Pennsylvania Health System included three urban acute care hospitals with a capacity of over 1,500 beds and 70,000 annual admissions. All hospitals used the EHR Sunrise Clinical Manager version 5.5 (Allscripts, Chicago, IL). Data were retrieved from the Penn Data Store, which includes clinical data from our EHR and administrative data coded by clinical documentation specialists.

Algorithm Derivation and Validation

A cohort of all inpatients discharged between July 2011 and June 2014 (n = 162,212) from our three hospitals was used to train our algorithm. From this population, a total of 950 inpatient encounters met "Sepsis Training Criteria", which required the following: 1) *International Classification of Diseases,* 9th Edition codes 995.92 (severe sepsis) or 785.52 (septic shock); 2) a positive blood culture; and 3) a lactate greater than 2.2 mmol/L or a systolic blood pressure less than 90, all occurring within a 1 hour window. The time of earliest measured elevated lactate or hypotension was considered "sepsis onset" for training purposes. These cases were used to train a random-forest classifier to predict severe sepsis and septic shock. The random forest approach has been described previously and used in similar studies (11). Our model considered a total of 587 features, consisting of demographics, vital signs, and laboratory results. For selected labs and vitals, we also derived time-series features, describing the minimum, maximum, mean, and rate of change over the preceding 24 hours. We used 100 estimators (trees) and gini criteria for splits.

The resulting algorithm was retrospectively validated on hospitalized non-ICU patients from October 1 to December 1 2015 (n = 10,448 discharges). A set point from the algorithm derivation area under the curve (AUC) was selected to produce an average of 10 alerts per day across the three hospitals in our healthcare system. During the validation period, the algorithm identified 347 patients predicted to develop severe sepsis or septic shock. The outcome of "Severe Sepsis" was defined as having: 1) greater than two SIRS criteria, 2) lactate greater than 2.2, and 3) positive blood or urine culture. The outcome of "Septic Shock" was defined as having "Severe Sepsis" plus a systolic blood pressure less than 90 mm Hg. These variables had to collectively occur within a 1-hour time window. We will refer to patients who trigger the algorithm prediction as "screen positive".

AUC with k-fold cross validation (k = 10) was estimated using the derivation population. Test characteristics, including sensitivity, specificity, predictive values, and likelihood ratios, were estimated from the validation population. All model construction and analyses were conducted using the open source Python programming language and Scikit-learn v0.15.2 (http://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/about.html#citing-scikit-learn).

Implementation of "Early Warning System 2.0"

As a successor to our prior SIRS-based sepsis detection tool, Early Warning System (EWS) 1.0 (7), we named this new ML algorithm-based sepsis prediction tool "EWS 2.0". After derivation and validation, EWS 2.0 was deployed in the production environment over a 14-month period. Patients eligible for algorithm screening included non-ICU inpatients who were in the hospital greater than 24 hours (which included any time spent in the emergency department). Patient data were resampled hourly, with new predictions made any time a new observation (data point) was recorded.

During an initial 6-month "silent period" (January 1 to June 15, 2016), process and outcome measures were collected on screen positive patients, but no accompanying alert was sent to the care team. For the subsequent 8-month "alert period" (June 16, 2016 to February 6, 2017), the algorithm was paired with an automated alert sent to the covering care team. Alerts stated that EWS 2.0 had fired for a given patient and included relevant recent laboratory data along with 48 hours of vital sign trends. Nurses received EHR-based alerts. Text messages were sent to providers

TABLE 1. Demographics of Intervention Population

Demographic	Total Intervention Population			Screen Positive Population		
	Silent Period (*n* = 22,280)	Alert Period (*n* = 32,184)	*p*	Silent Period (*n* = 1,540)	Alert Period (*n* = 2,137)	*p*
Age, mean, yr	58.5	58.7	0.22	61.3	62.5	0.02
Female, %	48.7	49.0	0.60	47.7	47.4	0.88
Body mass index, mean	29.0	28.8	0.07	27.9	28.0	0.93
Race/ethnicity, *n* (%)			< 0.01			0.66
White	11,802 (53.0)	16,987 (52.8)		825 (53.6)	1,186 (55.5)	
Black	8,664 (38.9)	12,234 (38.0)		551 (35.8)	739 (34.6)	
Other	492 (2.2)	679 (2.1)		51 (3.3)	62 (2.9)	
Unknown	1,322 (5.9)	2,284 (7.1)		113 (7.3)	150 (7.0)	
Hospital, *n* (%)			0.03			0.67
Hospital of the University of Pennsylvania	13,990 (62.8)	19,918 (61.9)		1,189 (77.2)	1,636 (76.6)	
Penn Presbyterian Medical Center	8,290 (37.2)	12,266 (38.1)		351 (22.8)	501 (23.4)	
Admission type, *n* (%)			< 0.01			0.02
Elective	8,560 (38.4)	12,086 (37.6)		433 (28.1)	515 (24.1)	
Emergency	9,583 (43.0)	13,750 (42.7)		831 (54.0)	1,197 (56.0)	
Transfer	3,963 (17.8)	6,336 (19.7)		275 (17.9)	424 (19.8)	
Hospital length of stay, median (IQR), d	4 (2–7)	4 (2–7)	< 0.01	9 (5–18)	9 (5–18)	0.39
Diagnosis-related group weight, median (IQR)	1.45 (0.97–2.23)	1.48 (0.97–2.20)	0.27	1.88 (1.35–4.23)	1.79 (1.38–3.49)	0.05

IQR = interquartile range.

and a rapid response coordinator (a critical-care nursing professional who monitors and responds to hospital emergencies 24 hr daily). The team was asked to perform a bedside assessment of the patient, but no specific interventions were required.

Clinical characteristics of screen positive patients were compared with those of a random population of screen negative non-ICU inpatients during the alert period. Hourly data following algorithm trigger for screen positive patients were compared with hourly data from screen negative patients following a randomly selected time. Process and outcome measures were collected for screen positive patients until discharge from the hospital during both implementation periods. Because one of our three hospitals transitioned to a new EHR during the intervention study period, it was excluded from implementation analysis; data from the two other hospitals in our system (including our flagship teaching hospital) were used for silent and alert period analyses.

To assess the alert's impact on care, we estimated proportions with CIs, means with SDs, and medians with interquartile ranges (IQRs) for descriptive characteristics, process measures, and clinical outcomes in the silent and alerted periods. Unadjusted analyses using the chi-square test for dichotomous variables and the Wilcoxon rank-sum test for continuous variables compared demographics and process and outcome measures in all study populations, including training, validation, silent and alert periods. *p* values less than 0.05 were considered significant.

To better assess the impact of the alert when the care team suspected sepsis, we also performed analyses stratified by "Suspected" versus "Unsuspected Sepsis". "Suspected Sepsis" was defined by active orders for at least two of the following within 12 hours prior to the alert: broad-spectrum antibiotics, blood cultures, and/or lactate testing.

Institutional Review

This study received expedited approval, HIPAA waiver, and informed consent waiver from the University of Pennsylvania Institutional Review Board (protocol number 826028).

RESULTS

Algorithm Derivation and Validation

Demographics of the derivation, validation, and implementation period populations were clinically similar (**Supplemental Table 1**, Supplemental Digital Content 1, http://links.lww.com/CCM/E782) (**Table 1**). During algorithm derivation, the estimated AUC

for the study outcomes of Severe Sepsis or Septic Shock was 0.88 (SD ± 0.03) following k-fold validation (k = 10). Test characteristics estimated with the validation cohort demonstrated sensitivity of 26% and specificity of 98%. Positive and negative predictive values were 29% and 97%, respectively. Positive and negative likelihood ratios were 13 and 0.75, respectively. The clinical variables with the greatest contribution to the algorithm predictions are shown in **Table 2** (for a full list of included variables, see **Supplemental Table 2**, Supplemental Digital Content 2, http://links.lww.com/CCM/E783). SIRS criteria and markers of end-organ dysfunction contributed most to the prediction, consistent with recent sepsis consensus guidelines and definitions (20).

Algorithm and Alert Implementation

Clinical Characteristics of Screen Positive Patients. Demographics of the total study population in the silent and alert periods were clinically similar, as were the characteristics of screen positive patients from each group (Table 1). EWS 2.0 triggered for 7.4% of admissions (n = 1,540) during the silent period and 7.1% of admissions (n = 2,137) during the alert period. During the silent period, the tool triggered a median of 6 hours and 34 minutes (IQR, 0 hr:50 min to 53 hr:19 min) prior to the onset of severe sepsis or septic shock. This was similar to the alert period (median of 5 hr and 25 min) (IQR, 0 hr:45 min to 45 hr:0 min). Almost 60% of screen positive patients met two of four SIRS criteria at the time of alert, increasing to 84% by 48 hours post alert (**Fig. 1**). Only 11% of patients met study criteria for the outcomes of severe sepsis or septic shock at time of alert. By 48 hours after the alert, 30% of screen positive patients met study criteria for the outcomes of severe sepsis or septic shock. Screen positive patients demonstrated marked abnormalities in vital signs and laboratory data compared with those who did not trigger EWS 2.0 (**Supplemental Fig. 1**, Supplemental Digital Content 3, http://links.lww.com/CCM/E784; legend, Supplemental Digital Content 10, http://links.lww.com/CCM/E791).

Process Measures. The alert prompted a modest but statistically significant increase in lactate testing, administration of IV fluid boluses, and complete blood count or basic metabolic panel testing within 3 hours following the alert (**Table 3**). Increases in lactate testing and IV fluid bolus administration were sustained at 6 hours post alert, but only lactate testing remained significantly increased at the 48-hour mark (**Supplemental Table 3**, Supplemental Digital Content 4, http://links.lww.com/CCM/E785). Transfusion of packed RBCs was also significantly increased in the first 6 hours post alert (Supplemental Table 3, Supplemental Digital Content 4, http://links.lww.com/CCM/E785). Frequency of blood cultures or initiation of antibiotics did not significantly differ between the silent and alert periods (Supplemental Table 3, Supplemental Digital Content 4, http://links.lww.com/CCM/E785). Time to administration of broad-spectrum sepsis antibiotics also did not differ significantly (silent period: median 11 hr:12 min; IQR, 2 hr:27 min to 36 hr:34 min; alert period: median 9 hr:47 min, IQR, 2 hr:37 min to 35 hr:39 min; p = 0.59).

In the alert period, 27% of screen positive patients met criteria for Suspected Sepsis. Postalert increases in lactate testing were significant in both the Suspected and Unsuspected Sepsis groups (**Supplemental Table 4**, Supplemental Digital Content 5, http://links.lww.com/CCM/E786; and **Supplemental Table 5**, Supplemental Digital Content 6, http://links.lww.com/CCM/E787). Increases in IV fluid bolus administration, telemetry, and laboratory testing were primarily observed in Unsuspected Sepsis (Supplemental Table 5, Supplemental Digital Content 6, http://links.lww.com/CCM/E787). Antibiotic initiation did not significantly differ for either group.

Outcome Measures. Compared with screen positive patients during the silent period, screen positive patients during the alert period had a statistically significant decrease in time to ICU transfer, but no significant change in the frequency of ICU transfer or median length of stay in the ICU (**Table 4**). There were also no statistically significant differences in the development of severe sepsis or septic shock, all-cause mortality, or discharge disposition.

The observed decrease in median time-to-ICU-transfer among patients in the alert group was primarily driven by the Unsuspected Sepsis cohort (24 hr [IQR, 3–117] hr vs 8 hr [IQR, 2–73] hr; p < 0.01) (**Supplemental Table 6**, Supplemental Digital

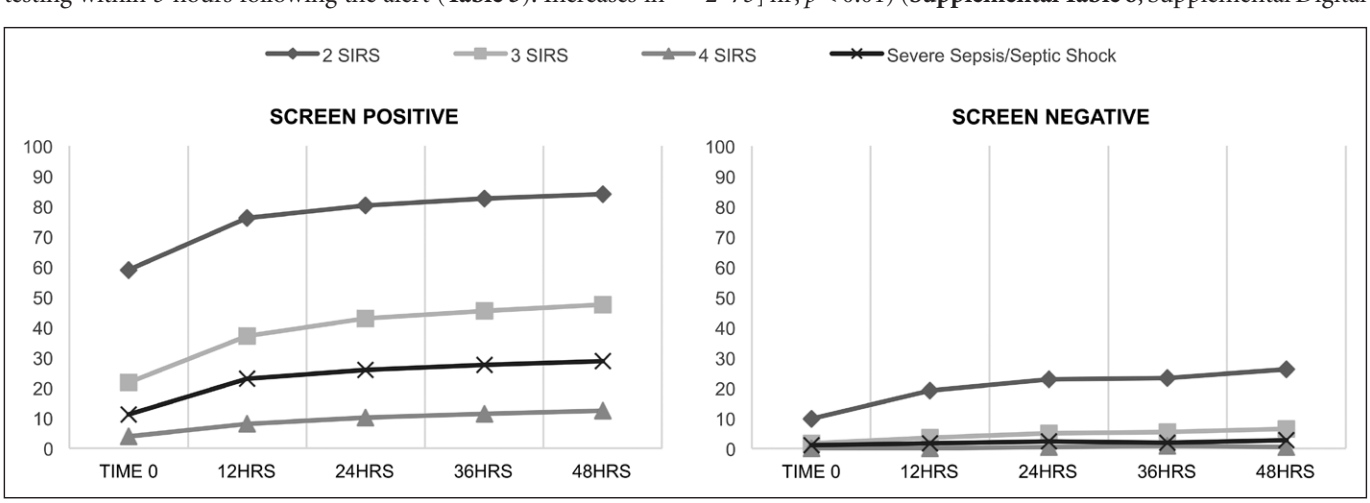

Figure 1. Proportion of screen positive patients meeting systemic inflammatory response syndrome (SIRS) criteria in the hours following algorithm detection, compared with controls. SIRS criteria include the following: (1) temperature greater than 38°C (100.4°F) or less than 36°C (96.8°F); (2) heart rate greater than 90; (3) respiratory rate greater than 20 or $PaCO_2$ less than 32 mm Hg; and (4) WBC greater than 12,000/mm^3, less than 4,000/mm^3, or greater than 10% bands. Criteria for severe sepsis: greater than two SIRS and positive blood or urine culture and lactate greater than 2.2; septic shock: severe sepsis and systolic blood pressure less than 90 mm Hg.

TABLE 2. Top Twenty Variables Contributing to Algorithm Prediction and Corresponding Weight

Variables	Time Variation	Weight
BP noninvasive diastolic (mm Hg)	Most recent	0.01624320
BP noninvasive systolic (mm Hg)	24 hr Minimum	0.01606099
Pulmonary service	Not applicable	0.01554681
Heart rate (beats/min)	24 hr rate of change	0.01455791
Blood urea nitrogen	Most recent	0.01372632
BP noninvasive systolic (mm Hg)	24hr variation from the mean	0.01370169
Temperature	Most recent	0.01358034
Temperature	24 hr maximum	0.01325225
% Monocytes	Most recent	0.01315597
Temperature	24 hr variation	0.01266883
Blood urea nitrogen	24 hr mean	0.01264225
Heart rate (beats/min)	Most recent	0.01182879
Blood urea nitrogen	24 hr minimum	0.01165007
Blood urea nitrogen	24 hr maximum	0.01141574
Age	Most recent	0.01108977
BP noninvasive diastolic (mm Hg)	24 hr minimum	0.01092703
Co_2	Most recent	0.01057971
Creatinine	Most recent	0.01047186
Absolute lymphocyte count	Most recent	0.01046288
Temperature (degrees Fahrenheit)	24 hr variation from the mean	0.00959570

BP = blood pressure.

Content 7, http://links.lww.com/CCM/E788; and **Supplemental Table 7**, Supplemental Digital Content 8, http://links.lww.com/CCM/E789). There was no significant change observed for the Suspected Sepsis cohort (Supplemental Table 6, Supplemental Digital Content 7, http://links.lww.com/CCM/E788). Neither cohort had postalert changes in frequency of ICU transfer, median length of stay in ICU, or mortality; however, we did observe increased frequency of discharge to inpatient hospice among patients with Suspected Sepsis at the time of the alert (3.0% vs 5.9%; $p = 0.04$) (Supplemental Table 6, Supplemental Digital Content 7, http://links.lww.com/CCM/E788).

DISCUSSION

We developed a ML algorithm to predict severe sepsis and septic shock and implemented the tool on non-ICU services across our multihospital healthcare system. Here, we confirmed the feasibility of widespread implementation of a ML predictive alert but observed a limited impact on clinical practice and outcomes.

TABLE 3. Clinical Process Measures in Screen Positive Patients Within 3 Hours of Alert

Process Measure, %	Silent ($n = 1,540$)	Alert ($n = 2,137$)	p
Complete blood count or basic metabolic panel	46.9	51.0	0.01
IV fluid bolus	21.7	25.5	< 0.01
Any antibiotic	17.3	16.9	0.76
Sepsis antibiotic(s)[a]	15.6	15.2	0.76
Blood cultures	14.0	15.7	0.18
Telemetry or electrocardiogram	12.8	14.5	0.15
Chest radiograph	9.4	10.0	0.62
Lactate	8.0	11.7	< 0.01
CT imaging[b]	5.3	4.6	0.38
RBC transfusion	3.8	4.2	0.67
Diuretic	3.2	3.8	0.43
Atrioventricular nodal blockade	2.9	3.4	0.39
Arterial blood gas	2.8	3.5	0.29
Vasopressors	2.2	2.8	0.30
Naloxone	0.1	0.2	0.40

See Supplemental Table 3 (Supplemental Digital Content 4, http://links.lww.com/CCM/E785) for process measures at 3, 6, and 48-hr time intervals.

[a]List available in **Supplemental Table 8** (Supplemental Digital Content 9, http://links.lww.com/CCM/E790).

[b]Includes CT chest, head, or abdomen.

Algorithm Design

To train our algorithm, we sought to identify patients with unequivocal sepsis physiology. Our selected Sepsis Training Criteria included hypotension and lactic acidosis as markers of impaired perfusion and shock (20) and a positive blood culture as a specific marker of infection. Although recent sepsis definitions do not include bacteremia, and in fact up to 50% of sepsis cases have no confirmed source of infection, we used narrower criteria to improve the specificity and predictive value of our resulting algorithm. SIRS criteria and clinical data related to end-organ dysfunction were heavily weighted in the algorithm, thus supporting our approach to algorithm development. However, this study's results may be limited by our use of more specific sepsis definitions that have not been externally validated.

The resulting algorithm accurately identified hospitalized patients at risk for developing severe sepsis or septic shock,

TABLE 4. Outcomes in Screen Positive Patients

Outcome Measures	Silent ($n = 1,540$)	Alert ($n = 2,137$)	p
Hospital length of stay, median (IQR), d	9 (5–18)	9 (5–18)	0.39
ICU transfer < 6 hr after alert, %	9.2	12.0	0.14
ICU transfer < 24 hr after alert, %	14.4	16.8	0.19
ICU transfer < 48 hr after alert, %	16.4	18.9	0.20
ICU transfer any time after alert, %	25.6	26.1	0.80
Time to ICU transfer after alert, median (IQR), hr	16 (2–108)	8 (2–62)	< 0.01
ICU length of stay, median (IQR), hr	71 (38–163)	85 (43–179)	0.11
Mortality ≤ 30 d after trigger, %	9.8	9.4	0.81
In-hospital mortality, %	10.6	10.3	0.88
Discharged to home, %	59.9	58.4	0.42
Discharged to nursing facility, %	15.3	15.2	0.93
Discharged to inpatient hospice, %	3.4	4.6	0.51
Severe Sepsis or septic shock[a], %	20.5	18.6	0.32

IQR = interquartile range.

[a]Severe sepsis: > 2 SIRS and positive blood or urine culture and lactate > 2.2; septic shock: severe sepsis and systolic blood pressure < 90 mm Hg.

despite the inherent limitations of EHR data, which can be plagued by missingness, inaccuracies, and changes in practice patterns over time. Importantly, the sensitivity of the tool was limited to minimize alert fatigue given that hospital providers are estimated to receive greater than 50 EHR alerts on average per day (21), leading to providers declining, ignoring, or deferring a majority of the alerts they encounter (22). Our lower sensitivity resulted in higher specificity and an excellent positive likelihood ratio.

Alert Impact

Ultimately, EWS 2.0 did not significantly improve our main outcome measures. We hypothesize that the alert's impact on clinical processes and patient outcomes was limited by multiple factors, including a lack of prespecified interventions, limited alert format, long alert lead-times, and perhaps most importantly, minimal predictive value beyond predictions already made by the clinical teams.

Clinician Response

Despite good predictive values, in many cases, the postalert bedside evaluation resulted in minimal changes to clinical care or outcomes. Ambiguity may have arisen about how to manage patients in the setting of a positive screen but apparent clinical stability. Prior to apparent disease, the utility of further laboratory testing, fluid administration, and empiric antibiotics is unclear. This may contribute to the low level of practice change following the alert. In addition, our study is limited in that we did not evaluate whether observed practice changes were appropriate for the clinical context of each patient.

We previously reported that with our prior alert system, EWS 1.0, clinical teams already strongly suspected sepsis in greater than 50% of cases (23). A survey administered to alerted providers and nurses during the alert period of this current study revealed a similar sentiment (24). However, in the subgroup analysis of patients who were not suspected of having sepsis at the time of alert, there was a statistically significant decrease in time to ICU transfer. On the other hand, there was an increase in referral to inpatient hospice for patients already suspected of sepsis. Thus, it appears that the algorithm may have provided information that supported either escalating care for those not suspected of having sepsis or adjusting overall goals of care for those initially suspected of having sepsis.

Intervention Format

The format of our intervention, as a one-time alert, may have affected the alert's impact on clinical care and outcomes. There may be multiple critical opportunities for clinical teams to integrate clinical information with a sepsis risk assessment. Yet, we did not require reevaluation of alerted patients at later time-points, even though in many cases our one-time alert triggered hours prior to the onset of sepsis physiology. For these cases, the lead-time of the alert and evaluation prior to clinically overt disease may have been too long. The question remains as to whether alerts are the most effective method of communicating real-time predictive information or whether a continuous score may more dynamically support clinical decision-making. Furthermore, although some clinical data were reported with the alerts, the variables and logic leading to alert trigger were not clearly delineated, creating what has been referred to as a ML "black box model" (25). This lack of transparency may have reduced overall trust in the algorithm and may

have affected the clinician perception of the reliability of the prediction.

Future Algorithm Optimization

Considerations must also be made for optimizing algorithm design. Recent studies have shown that ML predictions in sepsis and critical care may be strengthened by incorporating free text from provider documentation using natural language processing (10, 26, 27). Importantly, we derived our algorithm using criteria that although guided by sepsis consensus guidelines, were defined by the study team, and not externally validated. Although we prioritized specificity, a more sensitive algorithm may pick up subtle clinical trends for patients who are less likely to be captured by clinician's usual risk assessments (although at the risk of alert fatigue). Additionally, the most actionable moment in the course of a patient's sepsis trajectory may be the time just prior to, or during, the onset of clinical change. In this case, our alert frequently fired at a time when the patient appeared clinically well, sometimes many hours ahead of later decompensation. Finally, severe sepsis and septic shock may not be the most relevant outcomes to target when predicting unsuspected active clinical deterioration requiring a response from frontline providers. Algorithms trained for general decline, which may predict ICU transfer (17) or even mortality (28), might be more impactful with respect to changing process and outcome measures and preventing these critical events.

CONCLUSIONS

This study demonstrates the feasibility of implementing a ML algorithm for real-time analysis of EHR data to accurately predict the development of severe sepsis or septic shock. We have also shown the potential implications of alerting clinicians to this prediction throughout a multihospital healthcare system. In this study, the alert did not significantly alter clinical practice or outcomes. Training the algorithm on more traditional definitions of clinical deterioration, enhancing ML algorithms through incorporation of natural language processing, and effectively communicating risk while avoiding alerts in patients already suspected of clinical deterioration, represent potential opportunities to improve the impact of sepsis prediction on clinical care outcomes.

ACKNOWLEDGMENTS

We thank Mark E. Mikkelsen, MD, and Joanne Resnic, MBA, BSN, RN, for their contributions to the design, testing, and implementation of the clinical decision support intervention examined in this article.

REFERENCES

1. 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
2. Torio CM, Andrews RM: National Inpatient Hospital Costs: The most Expensive Conditions by Payer, 2011. HCUP Statistical Brief #160. August 2013. Agency for Healthcare Research and Quality, Rockville, MD. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb160.pdf. Accessed July 9, 2019
3. 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
4. Escobar GJ, LaGuardia JC, Turk BJ, et al: Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *J Hosp Med* 2012; 7:388–395
5. Bellomo R, Ackerman M, Bailey M, et al; Vital Signs to Identify, Target, and Assess Level of Care Study (VITAL Care Study) Investigators: A controlled trial of electronic automated advisory vital signs monitoring in general hospital wards. *Crit Care Med* 2012; 40:2349–2361
6. Churpek MM, Yuen TC, Winslow C, et al: Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014; 190:649–655
7. Umscheid CA, Betesh J, VanZandbergen C, et al: Development, implementation, and impact of an automated early warning and response system for sepsis. *J Hosp Med* 2015; 10:26–31
8. Khurana HS, Groves RH Jr, Simons MP, et al: Real-time automated sampling of electronic medical records predicts hospital mortality. *Am J Med* 2016; 129:688–698.e2
9. Deo RC: Machine learning in medicine. *Circulation* 2015; 132:1920–1930
10. 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
11. Taylor RA, Pare JR, Venkatesh AK, et al: Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23:269–278
12. Berger T, Birnbaum A, Bijur P, et al: A computerized alert screening for severe sepsis in emergency department patients increases lactate testing but does not improve inpatient mortality. *Appl Clin Inform* 2010; 1:394–407
13. Churpek MM, Yuen TC, Winslow C, et al: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44:368–374
14. 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
15. Hackmann G, Chen M, Chipara O, et al: Toward a two-tier clinical warning system for hospitalized patients. *AMIA Annu Symp Proc* 2011; 2011:511–519
16. Shimabukuro DW, Barton CW, Feldman MD, et al: 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
17. Wellner B, Grand J, Canzone E, et al: Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements. *JMIR Med Inform* 2017; 5:e45
18. McCoy A, Das R: Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual* 2017; 6:e000158
19. Mao Q, Jay M, Hoffman JL, et al: Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018; 8:e017833
20. 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
21. Murphy DR, Reis B, Sittig DF, et al: Notifications received by primary care practitioners in electronic health records: A taxonomy and time analysis. *Am J Med* 2012; 125:209.e1–209.e7
22. Ancker JS, Edwards A, Nosal S, et al; with the HITEC Investigators: Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med Inform Decis Mak* 2017; 17:36

23. Guidi JL, Clark K, Upton MT, et al: Clinician perception of the effectiveness of an automated early warning and response system for sepsis in an academic medical center. *Ann Am Thorac Soc* 2015; 12:1514–1519
24. 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 May 24. [Epub ahead of print]
25. Cabitza F, Rasoini R, Gensini GF: Unintended consequences of machine learning in medicine. *JAMA* 2017; 318:517–518
26. Weissman GE, Hubbard RA, Ungar LH, et al: Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Crit Care Med* 2018; 46:1125–1132
27. Marafino BJ, Park M, Davies JM, et al: Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw Open* 2018; 1:e185097
28. Ryan DP, Daley BJ, Wong K, et al: Prediction of ICU in-hospital mortality using a deep Boltzmann machine and dropout neural net. IEEE, 2013, pp 1–4. Available at: http://ieeexplore.ieee.org/document/6618491/