JAMA Internal Medicine | **Original Investigation**

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD; Erkin Otles, MEng; John P. Donnelly, PhD; Andrew Krumm, PhD; Jeffrey McCullough, PhD; Olivia DeTroyer-Cooley, BSE; Justin Pestrue, MEcon; Marie Phillips, BA; Judy Konye, MSN, RN; Carleen Penoza, MHSA, RN; Muhammad Ghous, MBBS; Karandeep Singh, MD, MMSc

IMPORTANCE The Epic Sepsis Model (ESM), a proprietary sepsis prediction model, is implemented at hundreds of US hospitals. The ESM's ability to identify patients with sepsis has not been adequately evaluated despite widespread use.

OBJECTIVE To externally validate the ESM in the prediction of sepsis and evaluate its potential clinical value compared with usual care.

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study was conducted among 27 697 patients aged 18 years or older admitted to Michigan Medicine, the academic health system of the University of Michigan, Ann Arbor, with 38 455 hospitalizations between December 6, 2018, and October 20, 2019.

EXPOSURE The ESM score, calculated every 15 minutes.

MAIN OUTCOMES AND MEASURES Sepsis, as defined by a composite of (1) the Centers for Disease Control and Prevention surveillance criteria and (2) *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* diagnostic codes accompanied by 2 systemic inflammatory response syndrome criteria and 1 organ dysfunction criterion within 6 hours of one another. Model discrimination was assessed using the area under the receiver operating characteristic curve at the hospitalization level and with prediction horizons of 4, 8, 12, and 24 hours. Model calibration was evaluated with calibration plots. The potential clinical benefit associated with the ESM was assessed by evaluating the added benefit of the ESM score compared with contemporary clinical practice (based on timely administration of antibiotics). Alert fatigue was evaluated by comparing the clinical value of different alerting strategies.

RESULTS We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

CONCLUSIONS AND RELEVANCE This external validation cohort study suggests that the ESM has poor discrimination and calibration in predicting the onset of sepsis. The widespread adoption of the ESM despite its poor performance raises fundamental concerns about sepsis management on a national level.

Multimedia

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Karandeep Singh, MD, MMSc, Department of Learning Health Sciences, University of Michigan Medical School, 1161H NIB, 300 N Ingalls St, Ann Arbor, MI 48109 (kdpsingh@ umich.edu).

Early detection and appropriate treatment of sepsis have been associated with a significant mortality benefit in hospitalized patients.[1-3] Many models have been developed to improve timely identification of sepsis,[4-9] but their lack of adoption has led to an implementation gap in early warning systems for sepsis.[10,11] This gap has largely been filled by commercial electronic health record (EHR) vendors, who have integrated early warning systems into the EHR where they can be readily accessed by clinicians and linked to clinical interventions.[12,13] More than half of surveyed US health systems report using electronic alerts, with nearly all using an alert system for sepsis.[14]

One of the most widely implemented early warning systems for sepsis in US hospitals is the Epic Sepsis Model (ESM), which is a penalized logistic regression model included as part of Epic's EHR and currently in use at hundreds of hospitals throughout the country. This model was developed and validated by Epic Systems Corporation based on data from 405 000 patient encounters across 3 health systems from 2013 to 2015. However, owing to the proprietary nature of the ESM, only limited information is publicly available about the model's performance, and no independent validations have been published to date, to our knowledge. This limited information is of concern because proprietary models are difficult to assess owing to their opaque nature and have been shown to decline in performance over time.[15,16]

The widespread adoption of the ESM despite the lack of independent validation raises a fundamental concern about sepsis management on a national level. An improved understanding of how well the ESM performs has the potential to inform care for the several hundred thousand patients hospitalized for sepsis in the US each year. We present an independently conducted external validation of the ESM using data from a large academic medical center.

Methods

Study Cohort

Our retrospective study included all patients aged 18 years or older admitted to Michigan Medicine (ie, the health system of the University of Michigan, Ann Arbor) between December 6, 2018, and October 20, 2019. Epic Sepsis Model scores were calculated for all adult hospitalizations. The ESM was used to generate alerts on 2 hospital units starting on March 11, 2019, and expanded to a third unit on August 12, 2019; alert-eligible hospitalizations were excluded from our analysis to prevent bias in our evaluation. The study was approved by the institutional review board of the University of Michigan Medical School, and the need for consent was waived because the research involved no more than minimal risk to participants, the research could not be carried out practically without the waiver, and the waiver would not adversely affect the rights and welfare of the participants.

The Epic Sepsis Model

The ESM is a proprietary sepsis prediction model developed by Epic Systems Corporation using data routinely recorded

Key Points

Question How accurately does the Epic Sepsis Model, a proprietary sepsis prediction model implemented at hundreds of US hospitals, predict the onset of sepsis?

Findings In this cohort study of 27 697 patients undergoing 38 455 hospitalizations, sepsis occurred in 7% of the hosptalizations. The Epic Sepsis Model predicted the onset of sepsis with an area under the curve of 0.63, which is substantially worse than the performance reported by its developer.

Meaning This study suggests that the Epic Sepsis Model poorly predicts sepsis; its widespread adoption despite poor performance raises fundamental concerns about sepsis management on a national level.

within the EHR. Epic Systems Corporation is one of the largest health care software vendors in the world and reportedly includes medical records for nearly 180 million individuals in the US (or 56% of the US population).[17] The eMethods in the Supplement includes more details.

Definition of Sepsis and Timing of Onset

Sepsis was defined based on meeting 1 of 2 criteria: (1) the Centers for Disease Control and Prevention clinical surveillance definition[18-20] or (2) an *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* diagnosis of sepsis accompanied by meeting 2 criteria for systemic inflammatory response syndrome and 1 Centers for Medicare & Medicaid Services criterion for organ dysfunction within 6 hours of one another (eMethods in the Supplement).

External Validation of the ESM Scores

We used scores from the ESM prospectively calculated every 15 minutes, beginning on arrival at the emergency department and throughout the hospitalization, to predict the onset of sepsis. For patients experiencing sepsis, we excluded any scores calculated after the outcome had occurred. We evaluated model discrimination using the area under the receiver operating characteristic curve (AUC), which represents the probability of correctly ranking 2 randomly chosen individuals (one who experienced the event and one who did not). We calculated a hospitalization-level AUC based on the entire trajectory of predictions[21-23] and calculated model performance across the spectrum of ESM thresholds. We also calculated time horizon-based AUCs (eMethods in the Supplement).

Using the entire trajectory of predictions, we calculated a median lead time by comparing when patients were first deemed high risk during their hospitalization (based on our implemented ESM score threshold of ≥6 described below) with when they experienced sepsis. Model calibration was assessed using a calibration plot by comparing predicted risk with the observed risk.

Selection of High-risk Threshold

We evaluated the ESM's performance at a score threshold of 6 or higher because this threshold was selected by our hospital operations committee to generate pages to clinicians

Table 1. Characteristics of Patients

Characteristic	Hospitalizations, No. (%)[a]			
	Overall (N = 38 455)	No sepsis (n = 35 903)	Sepsis (n = 2552)	P value[b]
Age, median (IQR), y	56 (35-69)	55 (34-68)	63 (51-73)	<.001
Sex				
Female	21 904 (57)	20 781 (58)	1123 (44)	<.001
Male	16 551 (43)	15 122 (42)	1429 (56)	
Race/ethnicity				
Black	4638 (12)	4303 (12)	335 (13)	
White	30 812 (80)	28 742 (80)	2070 (81)	<.001
Other[c]	2526 (7)	2405 (7)	121 (5)	
Unknown	479 (1)	453 (1)	26 (1)	
Comorbidities				
Congestive heart failure	7062 (18)	6189 (17)	873 (34)	<.001
Chronic obstructive pulmonary disease	7513 (20)	6763 (19)	750 (29)	<.001
Chronic kidney disease	7313 (19)	6332 (18)	981 (38)	<.001
Depression	8140 (21)	7478 (21)	662 (26)	<.001
Type 1 and 2 diabetes	8267 (22)	7325 (20)	942 (37)	<.001
Hypertension	18 066 (47)	16 369 (46)	1697 (67)	<.001
Obesity	8422 (22)	7799 (22)	623 (24)	.002
Liver disease	3213 (8)	2611 (7)	602 (24)	<.001
Solid tumor, nonmetastatic	5973 (16)	5473 (15)	500 (20)	<.001
Metastatic cancer	3255 (9)	2950 (8)	305 (12)	<.001

Abbreviation: IQR, interquartile range.

[a] For some patients, there were repeated hospitalizations.

[b] Statistical tests performed: Wilcoxon rank sum test and χ^2 test of independence.

[c] Includes American Indian or Alaska Native, Asian, and Native Hawaiian and other Pacific Islander.

Table 2. ESM Performance

Model performance	Hospitalization	Time horizons			
		24 h	12 h	8 h	4 h
Outcome incidence, %	6.6	0.43	0.29	0.22	0.14
Area under the receiver operating characteristic curve (95% CI)	0.63 (0.62-0.64)	0.72 (0.72-0.72)	0.73 (0.73-0.74)	0.74 (0.74-0.75)	0.76 (0.75-0.76)
Positive predictive value (ESM score ≥6), %	12	2.4	1.7	1.4	0.92
No. needed to evaluate (ESM score ≥6)[a]	8	42	59	73	109

Abbreviation: ESM, Epic Sepsis Model.

[a] The number needed to evaluate makes different assumptions at the hospitalization and time horizon levels. At the hospitalization level, the number needed to evaluate assumes that each patient would be evaluated only the first time the ESM score is 6 or higher. For each time horizon, the number needed to evaluate assumes that each patient would be evaluated every time the ESM score is 6 or higher.

and is currently in clinical use at Michigan Medicine (although patients eligible for alerts during the study period were excluded from our evaluation). This threshold is within the recommended score range (5-8) suggested by its developer.

Evaluation of Potential Clinical Benefit and Alert Fatigue

To evaluate potential benefit associated with the ESM, we compared the timing of patients exceeding the ESM score threshold of 6 or higher with their receipt of antibiotics to evaluate the potential added value of the ESM vs current clinical practice. We evaluated the potential impact of alert fatigue by comparing the number of patients who would need to be evaluated using different alerting strategies.

Sensitivity Analysis

To enhance the comparability of our results with other evaluations, we recalculated the hospitalization-level AUC after including ESM scores up to 3 hours after sepsis onset (eMethods in the Supplement). We used R, version 3.6.0 (R Group for Statistical Computing) for all analyses, as well as the pROC and runway packages.[24-26] Statistical tests were 2-sided.

Results

We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) who met inclusion criteria for our study cohort (**Table 1**). Sepsis occurred in 2552 of the hospitalizations (7%).

Figure 1. Threshold Performance Plots for the Epic Sepsis Model at the Hospitalization Level
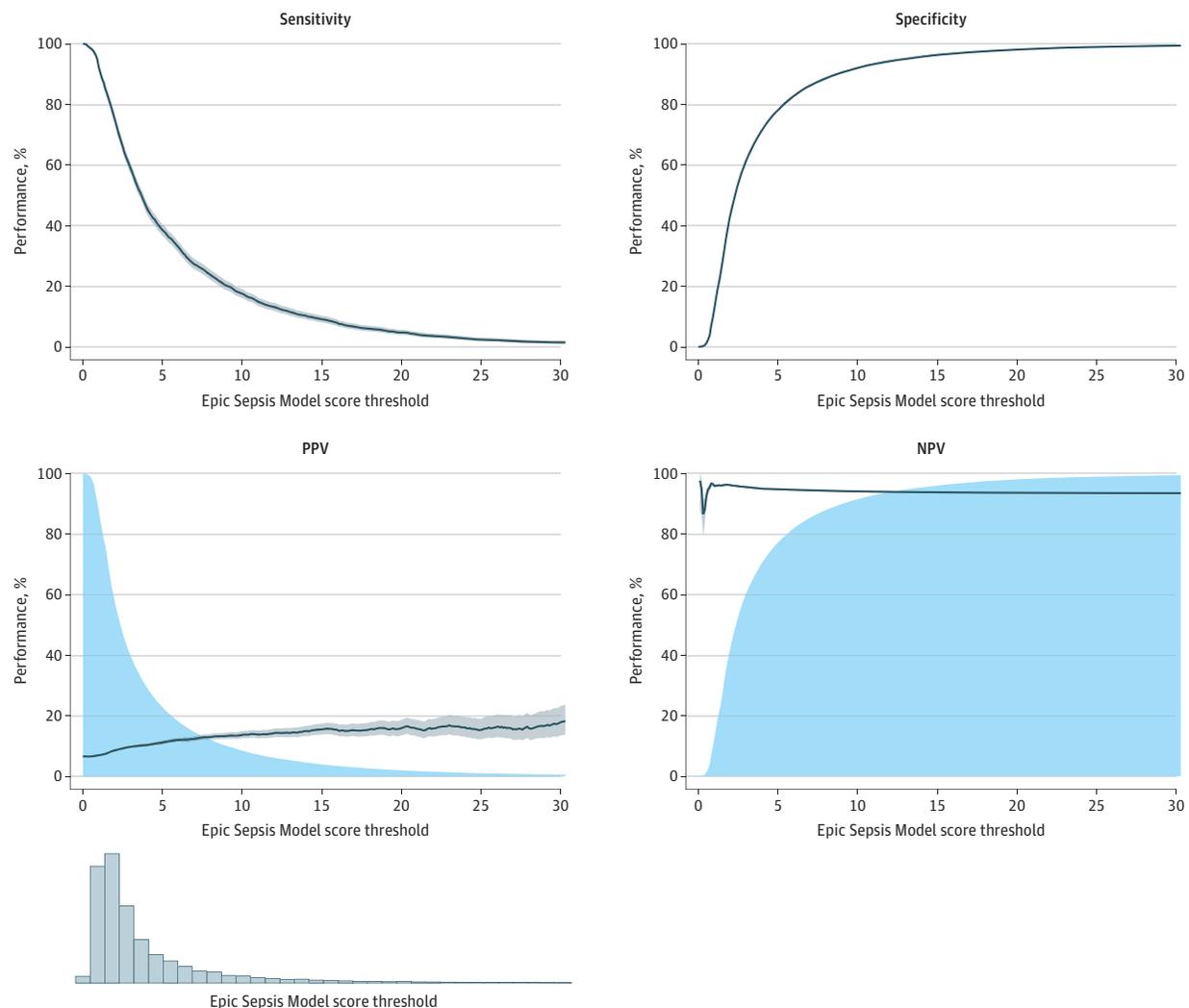
The distribution of predictions is displayed at the bottom. NPV indicates negative predictive value; PPV, positive predictive value. In the PPV plot, the blue-shaded region refers to the percentage of patients classified as positive.

In the NPV plot, the blue-shaded region refers to the percentage of patients classified as negative.

ESM Performance

The ESM had a hospitalization-level AUC of 0.63 (95% CI, 0.62-0.64) (Table 2). The AUC was between 0.72 (95% CI, 0.72-0.72) and 0.76 (95% CI, 0.75-0.76) when calculated at varying time horizons. At our selected score threshold of 6, the ESM had a hospitalization-level sensitivity of 33%, specificity of 83%, positive predictive value of 12%, and negative predictive value of 95% (Figure 1). The median lead time between when a patient first exceeded an ESM score of 6 and the onset of sepsis was 2.5 hours (interquartile range, 0.5-15.6 hours) (Figure 2). The calibration was poor at all time horizons possibly considered by the developer (eFigures 1, 2, and 3 in the Supplement).

Evaluation of Potential Clinical Benefit and Alert Fatigue

Of the 2552 hospitalizations with sepsis, 183 (7%) were identified by an ESM score of 6 or higher, but the patient did not receive timely antibiotics (ie, prior to or within 3 hours af-

ter sepsis). The ESM did not identify 1709 patients with sepsis (67%), of whom 1030 (60%) still received timely antibiotics.

An ESM score of 6 or higher occurred in 18% of hospitalizations (6971 of 38 455) even when not considering repeated alerts. If the ESM were to generate an alert only once per patient when the score threshold first exceeded 6—a strategy to minimize alerts—then clinicians would still need to evaluate 8 patients to identify a single patient with eventual sepsis (Table 2). If clinicians were willing to reevaluate patients each time the ESM score exceeded 6 to find patients developing sepsis in the next 4 hours, they would need to evaluate 109 patients to find a single patient with sepsis.

Sensitivity Analysis

When ESM scores up to 3 hours after the onset of sepsis were included, the hospitalization-level AUC improved to 0.80 (95% CI, 0.79-0.81).

Figure 2. Distribution of Alert Times Based on an Epic Sepsis Model Score Threshold of 6 or Higher
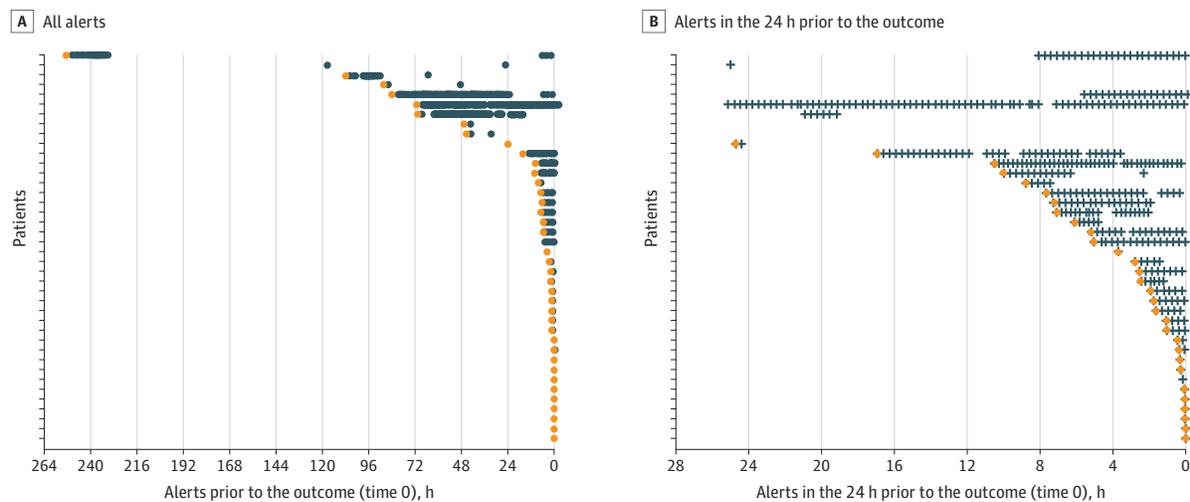
A, All alerts. B, Alerts in the 24 hours prior to the outcome. The first alert is highlighted in orange. Each point represents a hypothetical alert; no actual alerts were generated. Forty randomly selected patients who experienced sepsis and met the alerting threshold of 6 are shown here.

Discussion

In this external validation study, we found the ESM to have poor discrimination and calibration in predicting the onset of sepsis at the hospitalization level. When used for alerting at a score threshold of 6 or higher (within Epic's recommended range), it identifies only 7% of patients with sepsis who were missed by a clinician (based on timely administration of antibiotics), highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 67% of patients with sepsis despite generating alerts on 18% of all hospitalized patients, thus creating a large burden of alert fatigue.

Our observed hospitalization-level model performance (AUC, 0.63) was substantially worse than that reported by Epic Systems (AUC, 0.76-0.83) in internal documentation (shared with permission) and in a prior conference proceeding coauthored with Epic Systems (AUC, 0.73).[27] Although our time horizon–based AUCs were higher (0.72-0.76), they are misleading because they treat each prediction as independent. Even a small number of bad predictions (ie, high scores that result in alerts in patients without sepsis) can cause alert fatigue, but these bad predictions only minimally affect time horizon–based AUCs (eMethods in the Supplement). The large difference in reported AUCs is likely due to our consideration of sepsis timing. A prior study that did not exclude predictions made after development of sepsis found that the ESM produced an alert at a median of 7 hours (interquartile range, 4-22 hours) after the

first lactate level was measured, suggesting that ESM-driven alerts reflect the presence of sepsis already apparent to clinicians.[27] Our sensitivity analysis including predictions made up to 3 hours after the sepsis event found an improved AUC of 0.80, highlighting the importance of considering sepsis timing in the evaluation.

Limitations

Our study has some limitations. Our external validation was performed at a single academic medical center, although the cohort was large and relatively diverse.[28] We used a composite definition to account for the 2 most common reasons why health care organizations track sepsis, namely, surveillance and quality assessment, although sepsis definitions are still debated.

Conclusions

Our study has important national implications. The increase and growth in deployment of proprietary models has led to an underbelly of confidential, non–peer-reviewed model performance documents that may not accurately reflect real-world model performance. Owing to the ease of integration within the EHR and loose federal regulations, hundreds of US hospitals have begun using these algorithms. Medical professional organizations constructing national guidelines should be cognizant of the broad use of these algorithms and make formal recommendations about their use.

Author Affiliations: Department of Internal Medicine, University of Michigan Medical School, Ann Arbor (Wong, Singh); Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor (Otles); Department of Industrial and Operations Engineering, University of Michigan College of Engineering, Ann Arbor (Otles); Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor (Donnelly, Krumm, Ghous, Singh); School of Public Health, University of Michigan, Ann Arbor (McCullough);

Department of Quality, Michigan Medicine, Ann Arbor (DeTroyer-Cooley, Pestrue); Health Information Technology and Services, Michigan Medicine, Ann Arbor (Phillips); Nursing Informatics, Michigan Medicine, Ann Arbor (Konye, Penoza).

REFERENCES

1. Rivers E, Nguyen B, Havstad S, et al; Early Goal-Directed Therapy Collaborative Group. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med*. 2001;345 (19):1368-1377. doi:10.1056/NEJMoa010307

2. Yealy DM, Kellum JA, Huang DT, et al; ProCESS Investigators. A randomized trial of protocol-based care for early septic shock. *N Engl J Med*. 2014;370 (18):1683-1693. doi:10.1056/NEJMoa1401602

3. Gao F, Melody T, Daniels DF, Giles S, Fox S. The impact of compliance with 6-hour and 24-hour sepsis bundles on hospital mortality in patients with severe sepsis: a prospective observational study. *Crit Care*. 2005;9(6):R764-R770. doi:10.1186/cc3909

4. Sawyer AM, Deal EN, Labelle AJ, et al. Implementation of a real-time computerized sepsis alert in nonintensive care unit patients. *Crit Care Med*.

2011;39(3):469-473. doi:10.1097/CCM. 0b013e318205df85

5. Semler MW, Weavind L, Hooper MH, et al. An electronic tool for the evaluation and treatment of sepsis in the ICU: a randomized controlled trial. *Crit Care Med*. 2015;43(8):1595-1602. doi:10.1097/ CCM.0000000000001020

6. Giannini HM, Ginestra JC, Chivers C, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med*. 2019;47(11):1485-1492. doi:10.1097/ CCM.0000000000003891

7. Downing NL, Rolnick J, Poole SF, et al. Electronic health record–based clinical decision support alert for severe sepsis: a randomised evaluation. *BMJ Qual Saf*. 2019;28(9):762-768. doi:10.1136/bmjqs-2018-008765

8. Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. *Ann Emerg Med*. 2019; 73(4):334-344. doi:10.1016/j.annemergmed.2018. 11.036

9. Afshar M, Arain E, Ye C, et al. Patient outcomes and cost-effectiveness of a sepsis care quality improvement program in a health system. *Crit Care Med*. 2019;47(10):1371-1379. doi:10.1097/CCM. 0000000000003919

10. 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(10):1514-1519. doi:10.1513/AnnalsATS.201503-129OC

11. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit Care Med*. 2019;47(11): 1477-1484. doi:10.1097/CCM.0000000000003803

12. Rolnick JA, Weissman GE. Early warning systems: the neglected importance of timing. *J Hosp Med*. 2019;14(7):445-447. doi:10.12788/ jhm.3229

13. Makam AN, Nguyen OK, Auerbach AD. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review. *J Hosp Med*. 2015;10(6):396-402. doi:10.1002/jhm.2347

14. Benthin C, Pannu S, Khan A, Gong M; NHLBI Prevention and Early Treatment of Acute Lung Injury (PETAL) Network. The nature and variability of automated practice alerts derived from electronic health records in a U.S. nationwide critical care research network. *Ann Am Thorac Soc*. 2016;13(10):1784-1788. doi:10.1513/AnnalsATS. 201603-172BC

15. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc*. 2019;26(12):1651-1654. doi:10.1093/ jamia/ocz130

16. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

17. Caldwell P. We've spent billions to fix our medical records, and they're still a mess: here's why. *Mother Jones*. Published October 21, 2015. Accessed April 24, 2020. https://www. motherjones.com/politics/2015/10/epic-systems-judith-faulkner-hitech-ehr-interoperability/

18. Rhee C, Dantes RB, Epstein L, Klompas M. Using objective clinical data to track progress on preventing and treating sepsis: CDC's new "Adult Sepsis Event" surveillance strategy. *BMJ Qual Saf*. 2019;28(4):305-309. doi:10.1136/bmjqs-2018-008331

19. Rhee C, Dantes R, Epstein L, et al; CDC Prevention Epicenter Program. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA*. 2017;318(13): 1241-1249. doi:10.1001/jama.2017.13836

20. Centers for Disease Control and Prevention. Hospital toolkit for adult sepsis surveillance. Published March 2018. Accessed February 11, 2021. https://www.cdc.gov/sepsis/pdfs/Sepsis-Surveillance-Toolkit-Mar-2018_508.pdf

21. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015; 7(299):299ra122. doi:10.1126/scitranslmed.aab3719

22. Oh J, Makar M, Fusco C, et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect Control Hosp Epidemiol*. 2018; 39(4):425-433. doi:10.1017/ice.2018.16

23. Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized COVID-19 patients. *Ann Am Thorac Soc*. 2020. Published online December 24, 2020. doi:10.1513/AnnalsATS.202006-698OC

24. R Core Team. R: a language and environment for statistical computing. Published online 2020. Accessed May 4, 2021. http://www.r-project.org/

25. pROC: Display and analyze ROC curves [R package pROC version 1.16.2]. Accessed April 23, 2020. https://CRAN.R-project.org/package=pROC

26. Singh K. The runway package for R. Accessed October 21, 2020. https://github.com/ML4LHS/ runway

27. Bennett T, Russell S, King J, et al Accuracy of the Epic Sepsis Prediction Model in a regional health system. *arXiv*. Preprint posted online February 19, 2019. https://arxiv.org/abs/1902.07276

28. Healthcare Cost and Utilization Project. HCUP weighted summary statistics report: NIS 2018 core file means of continuous data elements. Accessed March 8, 2021. https://www.hcup-us. ahrq.gov/db/nation/nis/tools/stats/MaskedStats_ NIS_2018_Core_Weighted.PDF