









ORIGINAL ARTICLE

Evaluation of Sepsis Prediction Models before Onset of Treatment

Fahad Kamran , Ph.D.,¹ Donna Tjandra , M.S.,¹ Andrew Heiler , M.B.A.,² Jessica Virzi , M.S.N.,³ Karandeep Singh , M.D.,^{3,4} Jessie E. King , M.D., Ph.D.,⁵ Thomas S. Valley , M.D., M.Sc.,^{6,7} and Jenna Wiens , Ph.D.^{1,3}

Received: July 10, 2023; Revised: November 9, 2023; Accepted: November 15, 2023; Published: February 7, 2024

Abstract

BACKGROUND Timely interventions, such as antibiotics and intravenous fluids, have been associated with reduced mortality in patients with sepsis. Artificial intelligence (AI) models that accurately predict risk of sepsis onset could speed the delivery of these interventions. Although sepsis models generally aim to predict its onset, clinicians might recognize and treat sepsis before the sepsis definition is met. Predictions occurring after sepsis is clinically recognized (i.e., after treatment begins) may be of limited utility. Researchers have not previously investigated the accuracy of sepsis risk predictions that are made before treatment begins. Thus, we evaluate the discriminative performance of AI sepsis predictions made throughout a hospitalization relative to the time of treatment.

METHODS We used a large retrospective inpatient cohort from the University of Michigan's academic medical center (2018–2020) to evaluate the Epic sepsis model (ESM). The ability of the model to predict sepsis, both before sepsis criteria are met and before indications of treatment plans for sepsis, was evaluated in terms of the area under the receiver operating characteristic curve (AUROC). Indicators of a treatment plan were identified through electronic data capture and included the receipt of antibiotics, fluids, blood culture, and/or lactate measurement. The definition of sepsis was a composite of the Centers for Disease Control and Prevention's surveillance criteria and the severe sepsis and septic shock management bundle definition.

RESULTS The study included 77,582 hospitalizations. Sepsis occurred in 3766 hospitalizations (4.9%). ESM achieved an AUROC of 0.62 (95% confidence interval [CI], 0.61 to 0.63) when including predictions before sepsis criteria were met and in some cases, after clinical recognition. When excluding predictions after clinical recognition, the AUROC dropped to 0.47 (95% CI, 0.46 to 0.48).

CONCLUSIONS We evaluate a sepsis risk prediction model to measure its ability to predict sepsis before clinical recognition. Our work has important implications for future

The author affiliations are listed at the end of the article.

Dr. Wiens can be contacted at wiensj@umich.edu or at the Bob and Betty Beyster Building, 2260 Hayward Street, Ann Arbor, MI 48109.

work in model development and evaluation, with the goal of maximizing the clinical utility of these models. (Funded by Cisco Research and others.)

Introduction

Sepsis contributes to approximately one in three in-hospital deaths in the United States.¹⁻⁶ Timely identification and treatment of patients with sepsis have led to significant improvements in mortality rates among hospitalized patients.⁷⁻¹³ To enhance clinical decision-making, researchers have recently focused on developing predictive models that use electronic health record (EHR) data to identify patients at risk of sepsis.¹⁴⁻¹⁸ An example is the Epic sepsis model (ESM), one of the most widely implemented systems in U.S. hospitals.^{19,20} Similar to most sepsis prediction models, the ESM aims to flag patients at high risk for developing sepsis before sepsis onset.²¹⁻²⁴

Models such as the ESM make predictions throughout an individual's hospitalization, incorporating into its predictions relevant changes in a patient's health status on the basis of the content of the EHR. The algorithm can then be used to identify individuals at greatest risk for acquiring sepsis and to help target clinical resources. Discriminative metrics, such as the area under the receiver operating characteristic curve (AUROC), are commonly used to assess the model's performance.^{15,25,26} Typically, with the goal of evaluating the model's ability to predict sepsis before it occurs, researchers calculate AUROCs at the hospitalization level on the basis of the model's predictions before patients meet sepsis definitions.^{15,27} However, clinicians may recognize or begin to treat sepsis well before it is definitively diagnosed. This may be particularly true in the United States, where recent quality measures, such as the severe sepsis and septic shock management bundle (SEP-1) performance measure, incentivize clinicians and hospitals to administer treatments for sepsis early rather than late.²⁸ Consequently, predictions occurring after treatment might not be as clinically useful as those occurring before clinical recognition. This phenomenon, referred to as label leakage, can be exacerbated in models that include treatments (e.g., antibiotics) as predictors, and the accuracy of such models may largely be derived from predictions made after clinical recognition of sepsis.^{29,30} Thus, predictions using treatment indicators may lead to greater apparent performance but do not provide

clinicians with new information, and instead, they may contribute to alert fatigue.³¹ In such cases, many patients correctly identified by the model as high risk have already been identified as such by health care practitioners.³² In contrast, a model that identifies patients at high risk for developing sepsis before a clinician recognizes its signs or symptoms could enable the more timely delivery of care.

Given the rising adoption and use of sepsis prediction models, we need a more comprehensive evaluation protocol that captures the clinical utility of a model. Although it is difficult to retrospectively identify which patients a clinician might have otherwise missed, we can evaluate the accuracy of artificial intelligence (AI) predictions made in advance of indicators for sepsis treatment because such indicators can serve as proxies for clinical recognition. In this study, we introduce a new model evaluation framework that incorporates the timing of various indicators of sepsis treatment and that enables us to evaluate to what extent the ESM may rely on data pertaining to the clinical recognition of sepsis in making its predictions. An improved understanding of the utility of AI prediction methods can help prioritize the integration of useful models in clinical care.

Methods

STUDY COHORT

Our retrospective cohort included adult inpatients admitted to the University of Michigan's academic medical center, Michigan Medicine. ESM scores were calculated for all adult hospitalizations between October 2018 and December 2020. We included all hospitalizations in this period for evaluation, excluding hospitalizations from psychiatric and rehabilitation units. This study was approved by the institutional review board at Michigan Medicine (HUM: 00176141), and the need for consent was waived because there was minimal risk to participants.

SEPSIS CRITERIA AND INDICATORS OF SEPSIS TREATMENT

We used multiple criteria for a composite definition of sepsis that was on the basis of either the clinical surveillance definition from the Centers for Disease Control and Prevention (CDC)^{33,34} or the U.S. Centers for Medicare and Medicaid Services (CMS) definition, which involves meeting two criteria for systemic inflammatory response syndrome (SIRS) and one criterion for organ dysfunction

within 6 hours of one another (i.e., SEP-1).^{35,36} Regarding the CMS definition of sepsis, the time of meeting the sepsis criteria was defined as the latter of either meeting the SIRS criteria or meeting the organ dysfunction criteria. Regarding the CDC definition, the time of meeting the sepsis criteria was defined as the first time at which this definition was met. The time of onset was indeterminable in a small portion of patients with sepsis, and these patients were removed from the study cohort.

Through electronic data capture, we compared the timing of ordering and administration of treatments suggesting possible sepsis with the time at which sepsis criteria were met.^{9,11} Initial validation of the data capture was completed manually by a team of clinical analysts and engineers to ensure nearly perfect accuracy in identifying relevant treatment indicators. We included treatment and diagnostic orders as indicating the initiation of a treatment plan: intravenous fluids, antibiotics, lactate measurement (captured from both order sets and individual orders), and blood culture. We reported the percentage of individuals with sepsis who received each type of treatment as well as the timing of orders with respect to the time of meeting sepsis criteria. We excluded patients admitted with sepsis or who met sepsis criteria or received treatment before the first measurement of vital signs.

THE ESM

The ESM is a proprietary sepsis risk model developed by Epic Systems Corporation, Verona, Wisconsin.^{15,37} The

model was developed with a pooled sample of data collected by Epic Systems across three anonymous health care organizations, not including Michigan Medicine. The ESM uses data recorded within the EHR to make predictions every 20 minutes during a hospitalization. The ESM is a penalized logistic regression model that outputs a continuous score between 0 and 100, where 0 represents the lowest possible risk and 100 represents the highest. At Michigan Medicine, ESM scores of six or higher are used to generate alerts, a threshold chosen by the hospital operations committee. This threshold is within the recommended score range of five to eight suggested by Epic Systems. We included individuals with ESM scores before the first of meeting the sepsis criteria, ordering of any treatment indicating potential clinical recognition of sepsis, or death or discharge.

EVALUATION FRAMEWORK

We evaluated the ESM's predictions at key time points during hospitalization to understand its performance in relation to clinical recognition of sepsis (Fig. 1). To mimic a situation in which a clinician has not yet recognized sepsis and has not initiated treatment, we used predictions occurring before sepsis criteria were met and before the first indicator of treatment. We compared the predictive performance resulting from this evaluation with the performance achieved in predictions occurring before sepsis criteria were met, regardless of any treatment indicator. In addition, as an upper bound on performance, we evaluated the model's performance when it made sepsis

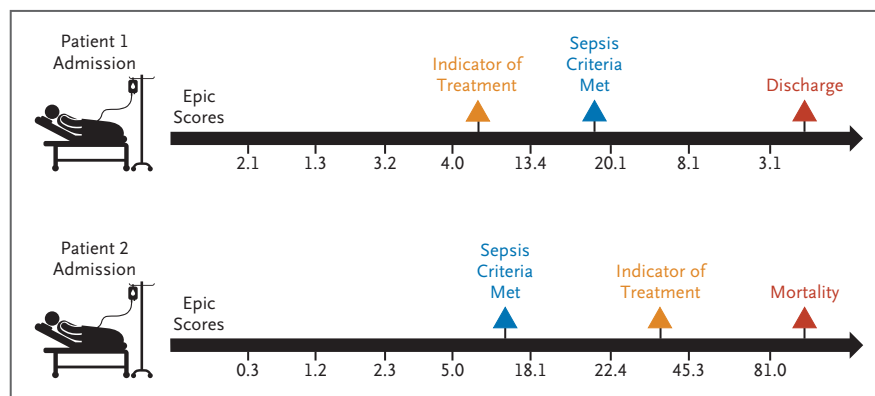


Figure 1. Overview of Different Evaluation Schemes.

In patient 1, indicators of treatment for sepsis occur before sepsis criteria are met. In patient 2, sepsis criteria are met before any treatment indication. In the case of patient 1, if the artificial intelligence model is relying on treatment indicators, then the accuracy of the Epic sepsis model (ESM) should decrease when data collected after the initiation of treatment are excluded. However, for patient 2, the ESM's accuracy should not change because no treatments were ordered before sepsis criteria were met. For both patients, the highest accuracy should occur when the ESM uses all data up to the time of discharge.

predictions up until discharge, including predictions made after sepsis criteria may have been met. We calculated a hospitalization-level AUROC, where an AUROC of 0.5 means that a model's performance is no better than random. To calculate a hospitalization-level AUROC, we took the maximum ESM score as the hospitalization-level score for predictions before each evaluation time point for each hospitalization. This evaluation mimicked how the ESM would generate alerts in a real-life clinical setting, where the ESM would fire an alert for a particular hospitalization the first time the sepsis threshold was exceeded before the time point of interest.^{15,27} We estimated the 95% confidence interval (CI) of the AUROC with 1000 bootstrap samples. We also measured the positive predictive value (PPV) and sensitivity of the ESM identifying high-risk individuals at a threshold of six or higher. This is the same score threshold used by Michigan Medicine to alert clinicians as to which patients are at the highest risk of developing sepsis, thus allowing clinicians to tailor care and provide extra monitoring for these individuals. Institutional priorities and resources can help determine meaningful PPV and sensitivity values for generating alerts, with the goal of balancing alert fatigue, missing cases of sepsis, and resource overuse in individuals without sepsis. For example, past work has indicated that a PPV of 60% might provide a meaningful trade-off between alert fatigue and improved clinical decision-making.³⁸

In a sensitivity analysis, we separately evaluated ESM performance with respect to the timing of each treatment or diagnostic indicator (antibiotics, lactate measurement, fluids, blood culture) to better understand which indicators drove changes in performance. In addition, we evaluated ESM performance with respect to the timing of diagnostic orders (lactate and blood culture collection) and treatment orders (antibiotic and fluid administration) separately.

SECONDARY ANALYSES: ADJUSTING FOR THE AMOUNT OF DATA AVAILABLE

Evaluating sepsis prediction models only on the basis of the timing of indicators of possible sepsis treatment could exclude significant portions of a patient's hospitalization, thereby reducing the amount of data available for making predictions. In many cases, clinicians might order treatments well in advance of sepsis criteria being met. To measure the amount of data available at different points of evaluation, we quantified the number of laboratory and medication orders for admissions in which the sepsis

criteria were met at each time point of our evaluation. Because the type of data available to the ESM might differ across different care settings, we also reported a breakdown of where ESM scores were generated (see the Supplementary Appendix for details). To control for the amount of data available to the algorithm, we stratified the number of orders available at treatment initiation into quintiles, and we evaluated the ESM within each quintile. We compared ESM performance before the first indicator for treatment with before the sepsis criteria were met. Across all quintiles, we included all individuals without sepsis as negative examples to calculate the AUROC.

In addition, to further isolate the effect of using reduced and earlier data per hospitalization and to better understand the effect of the model's use of data after the time of treatment, we compared the ESM with a simple Baseline Model that did not rely on clinicians' behaviors (i.e., orders) to make a prediction. The Baseline Model was a penalized logistic regression model that made predictions every 15 minutes using only data pertaining to vital signs, nursing scores, demographics, and information from past visits. More information on the features and training procedure for the Baseline Model can be found in the Supplementary Appendix. We assessed and compared the performance of the two models with respect to the time that sepsis criteria were met and with respect to the time that treatment indicators were ordered.

Results

We identified 77,582 hospitalizations (41,191 females [55%]; median age, 58 years [interquartile range, 37 to 70]) that met the inclusion/exclusion criteria for the study cohort ([Table 1](#)). Sepsis occurred in 3766 people (4.9%). Patients whose illness met sepsis criteria during their stay tended to be sicker as measured by comorbidities, with significantly longer hospital stays. A total of 3538 hospitalizations with sepsis had some indicator of sepsis treatment (93.9%). Over 70% of hospitalizations with sepsis involved orders for antibiotics (76.4%), blood cultures (72.4%), or lactate measurements (77.6%) as part of a treatment plan for sepsis, whereas orders for some level of fluids were made in only 29.0% of hospitalizations. For lactate measurements, 38.0% of orders came from an order set as opposed to an individual order. Over 45% of hospitalizations meeting sepsis criteria involved orders for antibiotics, blood cultures, or lactate measurements before sepsis criteria were met, with median lead times of 55, 46, and

Table 1. Characteristics of the Cohort of Adult Inpatients Used Throughout Evaluation.*				
Characteristic	Hospitalizations (n)			
	Overall (n=77,582)	No Sepsis (n=73,816)	Sepsis (n=3,766)	SMD (95% CI)
Median age — yr (IQR)	58 (37–70)	57 (36–69)	62 (50–72)	0.30 (0.27 to 0.34)
Female sex — n (%)	42,774 (55.13)	41,191 (55.80)	1,583 (42.03)	–0.28 (–0.31 to –0.24)
Race — n (%)				
American Indian or Alaska Native	363 (0.47)	340 (0.46)	23 (0.61)	0.02 (–0.01 to 0.05)
Asian	2,261 (2.91)	2,188 (2.96)	73 (1.94)	–0.06 (–0.09 to –0.03)
Black	9,451 (12.18)	8,866 (12.01)	585 (15.53)	0.11 (0.07 to 0.14)
Native Hawaiian or other Pacific Islander	69 (0.09)	63 (0.09)	6 (0.16)	0.02 (–0.01 to 0.06)
Not recorded	276 (0.36)	251 (0.34)	25 (0.66)	0.05 (0.02 to 0.09)
Other	2,281 (2.94)	2,160 (2.93)	121 (3.21)	0.02 (–0.02 to 0.05)
Refused	243 (0.31)	235 (0.32)	8 (0.21)	–0.02 (–0.05 to 0.01)
White	62,118 (80.07)	59,221 (80.23)	2,897 (76.93)	–0.08 (–0.12 to –0.05)
Unknown	520 (0.67)	492 (0.67)	28 (0.74)	0.01 (–0.02 to 0.04)
Ethnicity — n (%)				
Hispanic	2,276 (2.93)	2,147 (2.91)	129 (3.43)	0.03 (–0.002 to 0.06)
Not Hispanic	73,753 (95.06)	70,188 (95.09)	3,565 (94.66)	–0.02 (–0.05 to 0.01)
Not recorded	348 (0.45)	321 (0.43)	27 (0.72)	0.04 (0.01 to 0.07)
Refused	282 (0.36)	270 (0.37)	12 (0.32)	–0.01 (–0.04 to 0.02)
Unknown	923 (1.19)	890 (1.21)	33 (0.88)	–0.03 (–0.06 to 0.002)
Comorbidities — n (%)				
Hypertension	52,179 (67.26)	49,048 (66.45)	3,131 (83.14)	0.36 (0.32 to 0.39)
Obesity	31,663 (40.81)	29,995 (40.63)	1,668 (44.29)	0.07 (0.04 to 0.11)
Diabetes	28,324 (36.51)	26,508 (35.91)	1,816 (48.22)	0.26 (0.22 to 0.29)
Cancer	24,220 (31.21)	22,845 (30.95)	1,375 (36.51)	0.12 (0.09 to 0.15)
Chronic kidney disease	22,146 (28.55)	20,430 (27.68)	1,716 (45.57)	0.40 (0.36 to 0.43)
Congestive heart failure	20,905 (26.95)	19,290 (26.13)	1,615 (42.88)	0.38 (0.35 to 0.41)
Chronic obstructive pulmonary disease	14,064 (18.13)	13,066 (17.70)	998 (26.50)	0.23 (0.20 to 0.26)
Chronic liver disease	5,453 (7.03)	4,841 (6.56)	612 (16.25)	0.38 (0.35 to 0.41)
Dementia	4,511 (5.81)	4,162 (5.64)	349 (9.27)	0.15 (0.12 to 0.19)
Human immunodeficiency virus	392 (0.51)	366 (0.50)	26 (0.69)	0.03 (–0.01 to 0.06)
Time to meeting the sepsis criteria — median hours (IQR)	N/A	N/A	5 (1–37)	N/A
Discharge time — median hours (IQR)				
From admission time	91 (53–161)	86 (52–149)	268 (147–515)	1.65 (1.61 to 1.68)
From sepsis time	N/A	N/A	225 (125–445)	N/A
Antibiotics				
No. (%)	N/A	N/A	2,878 (76.4)	N/A
Order time from time of meeting sepsis criteria — median minutes (IQR)		N/A	–55 (–688 to 54)	N/A
Lactate measurement				
No. (%)	N/A	N/A	2,924 (77.6)	N/A
Order time from time of meeting sepsis criteria — median minutes (IQR)	N/A	N/A	–43 (–115 to –8)	N/A
Blood cultures				
No. (%)	N/A	—	2,727 (72.4)	—
Order time from time of meeting sepsis criteria — median minutes (IQR)	N/A		–46 (–288 to 17)	

(continued)

Characteristic	Hospitalizations (n)			
	Overall (n=77,582)	No Sepsis (n=73,816)	Sepsis (n=3,766)	SMD (95% CI)
Fluids				
No. (%)	N/A	N/A	1,072 (28.5)	N/A
Order time from time of meeting sepsis criteria — median minutes (IQR)	N/A	N/A	4 (-56 to 53)	N/A

* CI denotes confidence interval; IQR, interquartile range; N/A, not available; and SMD, standardized mean difference (Cohen's d).

43 minutes, respectively (Fig. 2). Indicators of sepsis treatment preceded the time of meeting sepsis criteria in 3193 hospitalizations (84.8%). Lactate was the first treatment indicator ordered in 47.1% of hospitalizations, followed by antibiotics (23.7%) and blood cultures (20.0%).

PRIMARY ANALYSIS: EVALUATION OF ESM PREDICTIONS WITH RESPECT TO VARYING DEGREES OF CLINICAL RECOGNITION

Using all predictions up until discharge during a hospitalization, the ESM achieved an AUROC of 0.87 (95% CI, 0.86 to 0.87), a PPV of 16% (95% CI, 16 to 17%), and a sensitivity of 79% (95% CI, 78 to 80%). Using only predictions from before sepsis criteria were met, the ESM model had an AUROC of 0.62 (95% CI, 0.61 to 0.63), a PPV of 8% (95% CI, 8 to 9%), and a sensitivity of 38% (95% CI, 36 to 39%). Further restricting the analysis to predictions made before treatment indicators, performance decreased, with an AUROC of 0.47 (95% CI, 0.46 to 0.48), a PPV of 5% (95% CI, 4 to 5%), and a sensitivity of 20% (95% CI, 19 to 22%). Performance dropped the most when predictions

were restricted to those made before blood cultures were ordered (AUROC, 0.53 [95% CI, 0.52 to 0.54]; PPV, 5.9%) and dropped the least when predictions were restricted to the time before fluids were ordered (AUROC, 0.61 [95% CI, 0.60 to 0.62]; PPV, 8.0%) (Fig. 3). Regarding predictive performance before the first diagnostic order (i.e., for lactate and blood culture collection), the ESM achieved an AUROC of 0.49 (95% CI, 0.48 to 0.50). Meanwhile, regarding the performance of the ESM before the first treatment orders (i.e., for antibiotics and fluid administration), the AUROC was 0.55 (95% CI, 0.54 to 0.56).

SECONDARY ANALYSES: ADJUSTING FOR THE AMOUNT OF DATA

For most cases, treatment indicators occurred before sepsis criteria were met (84.8%). There were significantly fewer clinical orders at treatment indicator time than when sepsis criteria were met (median count, 22 [interquartile range, 9 to 92] vs. 79 [interquartile range, 28 to 187]). However, the percentage of scores filed in the emergency department did not meaningfully differ across the

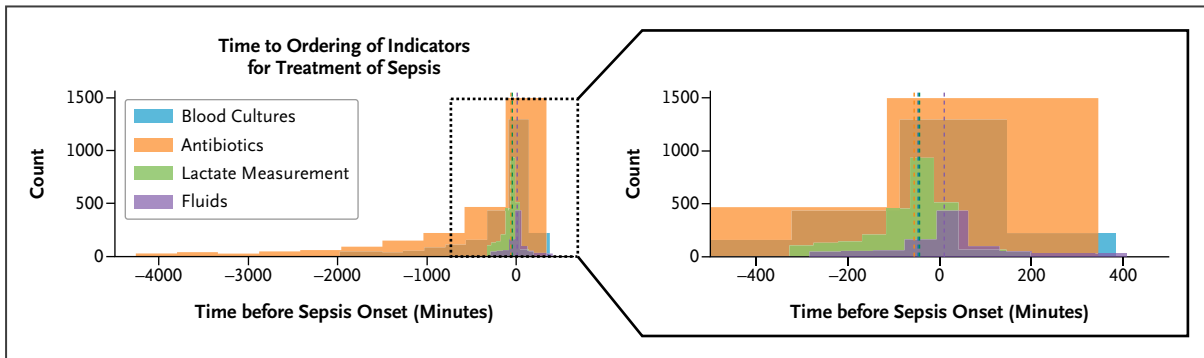


Figure 2. Temporal Distribution of Indicators of Treatment with Respect to the Time of Meeting Sepsis Criteria.

The dashed vertical bars represent the median time for each treatment. Antibiotics, blood culture collections, and lactate measurements are ordered substantially before the time of sepsis. Nearly half of the population has orders for lactate measurement, antibiotics, or blood cultures before the onset of sepsis.

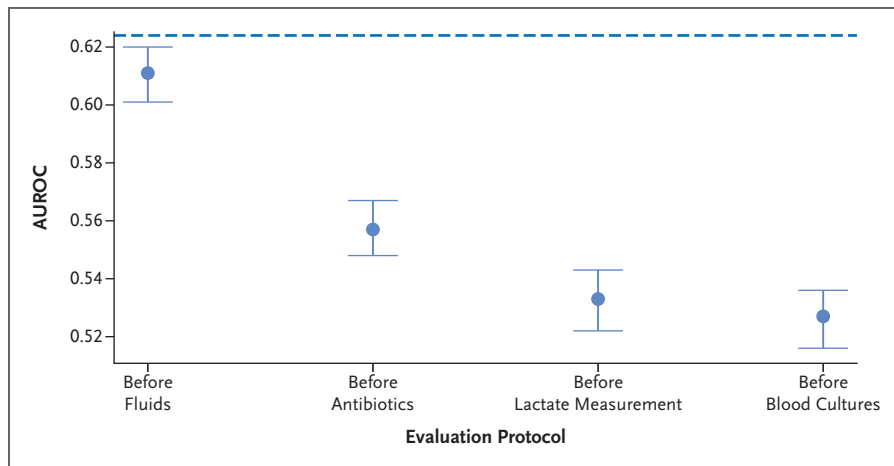


Figure 3. Evaluating the Accuracy of the ESM with Respect to Different Treatments.

We visualize the model's performance with 95% confidence intervals for each evaluation. The blue dashed line denotes the ESM's performance before the time of meeting the sepsis criteria. Its performance drops the most when predictions are made only before the time of blood culture orders, achieving nearly random performance. Meanwhile, the model's performance drops only slightly when using predictions before orders for fluids. AUROC denotes area under the receiver operating characteristic curve; and ESM, Epic sepsis model.

two evaluation time points (23.5 and 23.2%) (Table S1 in the Supplementary Appendix). In analyses adjusting for the amount of available data, the ESM model consistently performed worse when evaluating cases before treatment indicators than before sepsis criteria were met across all levels of available data (Fig. S1). However, the gap decreased as more data (i.e., more orders) became available to the ESM.

AUROC of the Baseline Model was similar to the ESM when evaluating hospitalization data before sepsis criteria were met (ESM, 0.62 [95% CI, 0.61 to 0.63]; Baseline, 0.60 [95% CI, 0.59 to 0.61]) (Table 2). Regarding model performance when sepsis predictions were on the basis of earlier data from before treatment indicators, both models performed worse, although the ESM's drop in performance was greater (ESM, 0.47 [95% CI, 0.46 to 0.48]; Baseline, 0.51 [95% CI, 0.50 to 0.52]).

Discussion

Recent work has emphasized the need to accurately frame the evaluation of machine-learning models to understand their potential for clinical impact.³⁹ In line with this idea, we evaluated a commonly used sepsis risk prediction model, the ESM, with respect to when a clinician places an order for treatment indicating possible sepsis. We found that for most individuals who developed sepsis,

clinicians sent an order for an indicator of sepsis treatment before the criteria for sepsis were met. In our analyses excluding predictions after treatment indicators, the ESM's performance was no better than random and performed significantly worse than its predictions using data up to meeting the sepsis criteria. This suggests that the ESM does not help to identify cases before clinical recognition has occurred.

We focused on the ESM because of its prevalence in health care systems across the United States. The clinical utility and performance of the ESM have been topics of recent interest across institutions.^{25,40-42} In our study, the ESM demonstrated strong performance when using predictions across an entire hospitalization because these predictions included data after sepsis definitions were met. Despite the ESM's high AUROC, PPV, and sensitivity, none of these predictions are clinically meaningful. The ESM also demonstrated performance significantly better than random for identifying cases before sepsis criteria were met. These results were consistent with previous evaluations, where the ESM performed similarly at the same institution using predictions before sepsis onset, and other institutions found strong performance when using all predictions during a stay.¹⁵ However, in contrast to past work, we focused on measuring the performance of the ESM with respect to when a clinician placed an order for a treatment indicator. In line with the type of evaluation proposed by Beaulieu-Jones et al.,⁴³ this helped to shed

Table 2. Performance of the Epic Sepsis Model and Baseline Model under Different Evaluation Schemes.*

Evaluation Scheme	ESM Performance, AUROC (95% CI)	Baseline Model Performance, AUROC (95% CI)
Using predictions before sepsis criteria are met	0.62 (0.61 to 0.63)	0.60 (0.59 to 0.61)
Using predictions before treatment indicators	0.47 (0.46 to 0.48)	0.51 (0.50 to 0.52)

* AUROC denotes area under the receiver operating characteristic curve; CI, confidence interval; and ESM, Epic sepsis model.

light on whether the ESM was simply relying on human intuition (i.e., looking over the shoulders of clinicians) or actually augmenting clinical knowledge in predicting the likelihood of sepsis. Through our analysis, we found that most of the discriminative ability of the ESM during standard evaluation was attributable to predictions made after clinicians recognized sepsis and initiated treatment. When evaluating predictions made before clinical recognition, the model achieved a PPV of 5%, far below the recommended threshold of 60%.³⁸ This means that 95% of alerts sent to clinicians by the ESM were for individuals who did not acquire sepsis during their stay, which could contribute to a large burden of alert fatigue. Moreover, false alarms could result in potentially harmful overuse of treatments, such as antibiotics. False alarms might also divert resources away from patients who acquired sepsis during their stay but were not flagged as high risk, thus contributing to delays in treatment. In our evaluation, the ESM flagged only one in five cases of sepsis as high risk. Hence, if clinicians were to follow the ESM, the remaining 80% of sepsis cases might not receive timely monitoring and treatment. Once allowed to make predictions up until the time of the sepsis criteria being met, the model flagged nearly twice as many sepsis cases correctly. However, at that point of clinical recognition, the model would not be providing clinicians with any new knowledge. The poor performance of the ESM within this context helps to explain other recent work evaluating the ESM, which found that the ESM scores often did not cross the alert threshold for sepsis until after antibiotics were given or after lactate was measured.¹⁵ Through our new evaluation framework, we determined that the ESM was unable to identify patients with sepsis before treatments were initiated.

In a large majority of the sepsis population, clinicians ordered treatment before sepsis criteria were met. We found that the ESM performance dropped most when we evaluated on the basis of predictions made before orders for blood cultures, followed by predictions before orders for lactate measurements and antibiotics. The drop in the ESM's performance when relying on only predictions made before blood culture orders might in part be explained by how clinicians use blood cultures in practice because

clinicians might order these when they have a strong suspicion that a person will develop sepsis.³⁴ The poor performance when predictions used only data from before antibiotic orders helped explain findings in recent work by Burgin et al.,⁴¹ who reported no improvement in the time to antibiotics for patients with sepsis when using the ESM. These findings suggest that the ESM might not be useful in guiding the timing of antibiotic administration before clinicians have made that decision. We hypothesize that this poor performance might be attributed to the inclusion of antibiotics as a predictor in the ESM. Moreover, ESM performance deteriorated more when it was evaluated with respect to diagnostic orders compared with treatment orders. This suggested that the ESM's predictive accuracy was improved by relying on human clinical intuition: that is, when diagnostic orders were placed. In addition, diagnostic orders tended to occur before treatment indicators (Fig. 2), making the task more difficult from a timing perspective.

The degradation in performance of the ESM before treatment indicators was likely because of a combination of the model's reliance on treatment indicators and the limited availability of data earlier in a patient's hospitalization. There were fewer orders earlier than later in the visit, which in turn, resulted in less data available as input to the algorithm and thus, worse performance. Nonetheless, even before treatment, the ESM had a nonnegligible amount of data to use in making its predictions. To further disentangle the effects of utilizing less data earlier in a patient's hospitalization and removing data that included treatment on performance, we compared the ESM with a simple Baseline Model for sepsis risk prediction. The Baseline Model was designed to focus on representations of patient physiology rather than clinician behavior. The greater drop in performance of the ESM compared with the drop in performance of the Baseline Model suggested its higher sensitivity to the timing of treatment, potentially because of the ESM's inclusion of features that incorporated clinical intuition.

Our study is not without limitations. First, we identified sepsis on the basis of a composite definition and identified

the ordering time of indicators for sepsis treatment according to this definition. However, sepsis definitions are still debated.^{36,44,45} Next, we used our proposed evaluation scheme to assess only two models: the ESM and our Baseline Model. Importantly, our goal was not to find the best sepsis risk model but rather, to gain insights into how our proposed evaluation scheme would affect our understanding of a model's performance. Furthermore, our study was performed at a single medical center. However, our cohort was large and represented many subpopulations of interest, providing meaningful representation. Researchers could consider using our evaluation scheme to assess the performance of multiple models in a variety of institutions and settings and to understand a model's performance before it used data indicating clinical intuition. Moreover, we focused on understanding how our new evaluation scheme assessed the discriminative performance of the ESM; in our evaluation, we did not account for the uncertainty of individual predictions because the ESM only provided point estimates of risk. Researchers have begun to consider how to quantify prediction uncertainty at an individual level for classification tasks.⁴⁶ Incorporating these models for sepsis risk prediction could help us understand the effect of different evaluation milestones on individual-level uncertainty. Finally, we evaluated the ESM using scores that could generate Epic alerts. However, past work has shown that the discriminative performance of the Epic model is similar for patients with Covid-19 and patients without Covid-19, despite the large difference in alerts.⁴² Moreover, the ESM's performance in our analysis was similar to that in a previous study that removed alert-eligible hospitalizations.¹⁵ Hence, we expect that our findings would not be affected by the potential of Epic to generate alerts from the data.

This study has significant implications for the development and evaluation of clinically useful models for sepsis prediction. Although we have focused on sepsis, our work also has implications for the development and evaluation of predictive models for other clinically relevant diseases. When designing AI models, researchers need to consider that features could inadvertently encode clinical suspicion, such as medications, laboratory tests, and treatments, thus suggesting strong performance in retrospective studies but failing to identify new cases that a clinician has not yet recognized. Model developers should carefully consider the inputs and observation period used during model training. Incorporating data from windows after treatment onset may encourage the model to rely on

features that could leak clinical suspicion to maximize model performance. Our results emphasize the importance of rigorous evaluation to better understand how these models can affect clinical care and improve patient outcomes.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Supported by Cisco Research, the National Science Foundation (award no. IIS 2124127), and Precision Health at the University of Michigan.

We thank the anonymous reviewers for their valuable feedback. In addition, we thank early team members from Michigan Medicine who helped with initial data pulls related to the Epic sepsis model and Erkin Otles for his guidance in navigating these initial data.

Author Affiliations

¹ Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor

² Division of Quality Analytics, Michigan Medicine, Ann Arbor

³ Precision Health, University of Michigan, Ann Arbor

⁴ Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor

⁵ Division of Hospital Medicine, Departments of Internal Medicine and of Quality and Safety, University of Michigan, Ann Arbor

⁶ Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor

⁷ Center for Clinical Management and Research, Ann Arbor VA Healthcare System, Department of Veterans Affairs, Ann Arbor, MI

References

- Centers for Disease Control and Prevention. What is sepsis? August 24, 2023 (<https://www.cdc.gov/sepsis/what-is-sepsis.html>).
- Rhee C, Jones TM, Hamad Y, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Netw Open* 2019;2:e187571. DOI: [10.1001/jamanetworkopen.2018.7571](https://doi.org/10.1001/jamanetworkopen.2018.7571).
- Watson RS, Carcillo JA. Scope and epidemiology of pediatric sepsis. *Pediatr Crit Care Med* 2005;6(Suppl):S3-S5. DOI: [10.1097/01.PCC.0000161289.22464.C3](https://doi.org/10.1097/01.PCC.0000161289.22464.C3).
- Angus DC, van der Poll T. Severe sepsis and septic shock. *N Engl J Med* 2013;369:840-851. DOI: [10.1056/NEJMr1208623](https://doi.org/10.1056/NEJMr1208623).
- Schlapbach LJ, Kisson N, Alhawsawi A, et al. World Sepsis Day: a global agenda to target a leading cause of morbidity and mortality. *Am J Physiol Lung Cell Mol Physiol* 2020;319:L518-L522. DOI: [10.1152/ajplung.00369.2020](https://doi.org/10.1152/ajplung.00369.2020).
- Liu V, Escobar GJ, Greene JD, et al. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA* 2014;312:90-92. DOI: [10.1001/jama.2014.5804](https://doi.org/10.1001/jama.2014.5804).

7. 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. DOI: [10.1001/jama.2016.0287](https://doi.org/10.1001/jama.2016.0287).
8. Ferrer R, Artigas A, Suarez D, et al. Effectiveness of treatments for severe sepsis: a prospective, multicenter, observational study. *Am J Respir Crit Care Med* 2009;180:861-866. DOI: [10.1164/rccm.200812-1912OC](https://doi.org/10.1164/rccm.200812-1912OC).
9. Levy MM, Evans LE, Rhodes A. The Surviving Sepsis Campaign Bundle: 2018 update. *Intensive Care Med* 2018;44:925-928. DOI: [10.1007/s00134-018-5085-0](https://doi.org/10.1007/s00134-018-5085-0).
10. Rhodes A, Evans LE, Alhazzani W, et al. Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive Care Med* 2017;43:304-377. DOI: [10.1007/s00134-017-4683-6](https://doi.org/10.1007/s00134-017-4683-6).
11. 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:R764-R770. DOI: [10.1186/cc3909](https://doi.org/10.1186/cc3909).
12. Yealy DM, Kellum JA, Huang DT, et al. A randomized trial of protocol-based care for early septic shock. *N Engl J Med* 2014;370:1683-1693. DOI: [10.1056/NEJMoa1401602](https://doi.org/10.1056/NEJMoa1401602).
13. Rivers E, Nguyen B, Havstad S, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med* 2001;345:1368-1377. DOI: [10.1056/NEJMoa010307](https://doi.org/10.1056/NEJMoa010307).
14. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015;7:299ra122. DOI: [10.1126/scitranslmed.aab3719](https://doi.org/10.1126/scitranslmed.aab3719).
15. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065-1070. DOI: [10.1001/jamainternmed.2021.2626](https://doi.org/10.1001/jamainternmed.2021.2626).
16. 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;73:334-344. DOI: [10.1016/j.annemergmed.2018.11.036](https://doi.org/10.1016/j.annemergmed.2018.11.036).
17. Williams JM, Greenslade JH, McKenzie JV, Chu K, Brown AFT, Lipman J. 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. DOI: [10.1016/j.chest.2016.10.057](https://doi.org/10.1016/j.chest.2016.10.057).
18. Kijpaisalratana N, Sanglertsinlapachai D, Techaratsami S, Musikatavom K, Saoraya J. Machine learning algorithms for early sepsis detection in the emergency department: a retrospective study. *Int J Med Inform* 2022;160:104689. DOI: [10.1016/j.ijmedinf.2022.104689](https://doi.org/10.1016/j.ijmedinf.2022.104689).
19. Rolnick JA, Weissman GE. Early warning systems: the neglected importance of timing. *J Hosp Med* 2019;14:445-447. DOI: [10.12788/jhm.3229](https://doi.org/10.12788/jhm.3229).
20. Makam AN, Nguyen OK, Auerbach AD. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review. *J Hosp Med* 2015;10:396-402. DOI: [10.1002/jhm.2347](https://doi.org/10.1002/jhm.2347).
21. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med* 2022;28:1455-1460. DOI: [10.1038/s41591-022-01894-0](https://doi.org/10.1038/s41591-022-01894-0).
22. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform* 2020;8:e15182. DOI: [10.2196/15182](https://doi.org/10.2196/15182).
23. Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know.” *NPJ Digit Med* 2021;4:134. DOI: [10.1038/s41746-021-00504-6](https://doi.org/10.1038/s41746-021-00504-6).
24. Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE — an interpretable and recurrent neural survival model for early prediction of sepsis. *Artif Intell Med* 2021;113:102036. DOI: [10.1016/j.artmed.2021.102036](https://doi.org/10.1016/j.artmed.2021.102036).
25. Lyons PG, Ramsey B, Simkins M, Maddox TM. How useful is the Epic sepsis prediction model for predicting sepsis? Poster presented at American Thoracic Society 2021 International Conference, May 14–19, 2021.
26. Schinkel M, van der Poll T, Wiersinga WJ. Artificial intelligence for early sepsis detection: a word of caution. *Am J Respir Crit Care Med* 2023;207:853-854. DOI: [10.1164/rccm.202212-2284VP](https://doi.org/10.1164/rccm.202212-2284VP).
27. 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;39:425-433. DOI: [10.1017/ice.2018.16](https://doi.org/10.1017/ice.2018.16).
28. Gesten F, Evans L. SEP-1 — taking the measure of a measure. *JAMA Netw Open* 2021;4:e2138823. DOI: [10.1001/jamanetworkopen.2021.38823](https://doi.org/10.1001/jamanetworkopen.2021.38823).
29. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337-1340. DOI: [10.1038/s41591-019-0548-6](https://doi.org/10.1038/s41591-019-0548-6).
30. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc* 2020;2020:191-200.
31. Cash JJ. Alert fatigue. *Am J Health Syst Pharm* 2009;66:2098-2101. DOI: [10.2146/ajhp090181](https://doi.org/10.2146/ajhp090181).
32. Ross C. Epic’s sepsis algorithm is going off the rails in the real world. The use of these variables may explain why. *Stat*, September 27, 2021 (<https://www.statnews.com/2021/09/27/epic-sepsis-algorithm-antibiotics-model>).
33. Rhee C, Zhang Z, Kadri SS, et al. Sepsis surveillance using adult sepsis events simplified eSOFA criteria versus Sepsis-3 sequential organ failure assessment criteria. *Crit Care Med* 2019;47:307-314. DOI: [10.1097/CCM.0000000000003521](https://doi.org/10.1097/CCM.0000000000003521).
34. 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;28:305-309. DOI: [10.1136/bmjqs-2018-008331](https://doi.org/10.1136/bmjqs-2018-008331).

35. Venkatesh AK, Slesinger T, Whittle J, et al. Preliminary performance on the new CMS Sepsis-1 national quality measure: early insights from the Emergency Quality Network (E-QUAL). *Ann Emerg Med* 2018;71:10-15.e1. DOI: [10.1016/j.annemergmed.2017.06.032](https://doi.org/10.1016/j.annemergmed.2017.06.032).
36. Kalantari A, Mallemat H, Weingart SD. Sepsis definitions: the search for gold and what CMS got wrong. *West J Emerg Med* 2017; 18:951-956. DOI: [10.5811/westjem.2017.4.32795](https://doi.org/10.5811/westjem.2017.4.32795).
37. Tarabichi Y, Cheng A, Bar-Shain D, et al. Improving timeliness of antibiotic administration using a provider and pharmacist facing sepsis early warning system in the emergency department setting: a randomized controlled quality improvement initiative. *Crit Care Med* 2022;50:418-427. DOI: [10.1097/CCM.0000000000005267](https://doi.org/10.1097/CCM.0000000000005267).
38. 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;28:762-768. DOI: [10.1136/bmjqs-2018-008765](https://doi.org/10.1136/bmjqs-2018-008765).
39. Lauritsen SM, Thiesson B, Jørgensen MJ, et al. The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. *NPJ Digit Med* 2021;4:158. DOI: [10.1038/s41746-021-00529-x](https://doi.org/10.1038/s41746-021-00529-x).
40. Bennett TD, Russell S, King J, et al. Accuracy of the Epic sepsis prediction model in a regional health system. February 19, 2019 (<http://arxiv.org/abs/1902.07276>). Preprint.
41. Burgin D, O'Neal HR, Hamer D, Thomas CB, Jagneaux T. Assessment of Epic sepsis predictive analytic impact on antibiotic use in the ED. *Chest* 2022;162(Suppl):A775. DOI: [10.1016/j.chest.2022.08.611](https://doi.org/10.1016/j.chest.2022.08.611).
42. Lyons PG, Hofford MR, Yu SC, et al. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. *JAMA Intern Med* 2023; 183:611-612. DOI: [10.1001/jamainternmed.2022.7182](https://doi.org/10.1001/jamainternmed.2022.7182).
43. Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit Med* 2021;4:62. DOI: [10.1038/s41746-021-00426-3](https://doi.org/10.1038/s41746-021-00426-3).
44. Saria S, Henry KE. Too many definitions of sepsis: can machine learning leverage the electronic health record to increase accuracy and bring consensus? *Crit Care Med* 2020;48:137-141. DOI: [10.1097/CCM.0000000000004144](https://doi.org/10.1097/CCM.0000000000004144).
45. Gül F, Arslantaş MK, Cinel İ, Kumar A. Changing definitions of sepsis. *Turk J Anaesthesiol Reanim* 2017;45:129-138. DOI: [10.5152/TJAR.2017.93753](https://doi.org/10.5152/TJAR.2017.93753).
46. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Abstract presented at the Thirty-First Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, December 4-9, 2017.