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 2,552 (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 2,552 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 1,709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6,971 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.
Earl detection and appropriate treatment of sepsis have been associated with a significant mortality benefit in hospitalized patients. Several methods have been developed to improve timely identification of sepsis, but their lack of adoption has led to an implementation gap in early warning systems for sepsis. This gap has largely been filled by systems reporting using electronic alerts, with nearly all using an alert system for sepsis.

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.

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 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). 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 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 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 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%).

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

---

**Table 1. Characteristics of Patients**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Hospitalizations, No. (%)a</th>
<th>P valueb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall (N = 38 455)</td>
<td>No sepsis (n = 35 903)</td>
</tr>
<tr>
<td>Age, median (IQR), y</td>
<td>56 (35-69)</td>
<td>55 (34-68)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>21 904 (57)</td>
<td>20 781 (58)</td>
</tr>
<tr>
<td>Male</td>
<td>16 551 (43)</td>
<td>15 122 (42)</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>4638 (12)</td>
<td>4303 (12)</td>
</tr>
<tr>
<td>White</td>
<td>30 812 (80)</td>
<td>28 742 (80)</td>
</tr>
<tr>
<td>Otherc</td>
<td>2526 (7)</td>
<td>2405 (7)</td>
</tr>
<tr>
<td>Unknown</td>
<td>479 (1)</td>
<td>453 (1)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>7062 (18)</td>
<td>6189 (17)</td>
</tr>
<tr>
<td>Chronic obstructive</td>
<td>7513 (20)</td>
<td>6763 (19)</td>
</tr>
<tr>
<td>pulmonary disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>7313 (19)</td>
<td>6332 (18)</td>
</tr>
<tr>
<td>Depression</td>
<td>8140 (21)</td>
<td>7478 (21)</td>
</tr>
<tr>
<td>Type 1 and 2 diabetes</td>
<td>8267 (22)</td>
<td>7325 (20)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>18 066 (47)</td>
<td>16 369 (46)</td>
</tr>
<tr>
<td>Obesity</td>
<td>8422 (22)</td>
<td>7799 (22)</td>
</tr>
<tr>
<td>Liver disease</td>
<td>3213 (8)</td>
<td>2611 (7)</td>
</tr>
<tr>
<td>Solid tumor, nonmetastatic</td>
<td>5973 (16)</td>
<td>5473 (15)</td>
</tr>
<tr>
<td>Metastatic cancer</td>
<td>3255 (9)</td>
<td>2950 (8)</td>
</tr>
</tbody>
</table>

Abbreviation: IQR, interquartile range.

a For some patients, there were repeated hospitalizations.

b Statistical tests performed: Wilcoxon rank sum test and χ² test of independence.

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

**Table 2. ESM Performance**

<table>
<thead>
<tr>
<th>Model performance</th>
<th>Hospitalization</th>
<th>Time horizons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24 h</td>
<td>12 h</td>
</tr>
<tr>
<td>Outcome incidence, %</td>
<td>6.6</td>
<td>0.43</td>
</tr>
<tr>
<td>Area under the receiver operating characteristic curve (95% CI)</td>
<td>0.63 (0.62-0.64)</td>
<td>0.72 (0.72-0.72)</td>
</tr>
<tr>
<td>Positive predictive value (ESM score ≥6), %</td>
<td>12</td>
<td>2.4</td>
</tr>
<tr>
<td>No. needed to evaluate (ESM score ≥6)a</td>
<td>15</td>
<td>42</td>
</tr>
</tbody>
</table>

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.
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 after 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 38455) 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 15 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).
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). 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 (i.e., 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.22 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.


