Title: Evaluating the Portability of SepsisWatch: A Multi-Site External Validation of a Sepsis Machine Learning Model

Bruno Valan, MENG1; Anusha Prakash, MS2; William Ratliff, MBA1; Michael Gao, MS1; Srikanth Muthya2; Ajit Thomas, MBA2; Jennifer L. Eaton, PhD3; Matt Gardner1; Marshall Nichols, MS1; Mike Revoir1; Dustin Tart, RN4; Cara O’Brien, MD4,5; Manesh Patel, MD5; Suresh Balu, MS, MBA1; Mark Sendak, MPP, MD1

Address Correspondence to:
Mark Sendak
1Duke Institute for Health Innovation
300 West Morgan St. Durham, North Carolina
Phone: (919) 684-4389
Email: mark.sendak@duke.edu

Affiliations
2Cohere Med Inc.
571 Euclid Ave
Berkeley, CA 94708

3Summa Health Research & Innovation,
Summa Health, Akron, OH 44310

4Duke University Hospital
Durham, NC 27701

5Department of Medicine,
Duke University School of Medicine
Durham, NC, United States

Study Type: Diagnostic/Prognostic Study

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Importance: Sepsis accounts for a substantial portion of global deaths and healthcare costs. Early detection using machine learning (ML) models offers a critical opportunity to improve care and reduce the burden of sepsis.

Objective: To externally validate the SepsisWatch ML model, initially developed at Duke University, in a community healthcare and assess its performance and clinical utility in early sepsis detection.

Design: This retrospective external validation study evaluated the performance of the SepsisWatch model in a new environment. Data from patient encounters at Summa Health’s emergency departments between 2020 and 2021 were used. The study analyzed the model's ability to predict sepsis using a combination of static and dynamic patient data.

Setting: The study was conducted at Summa Health, a nonprofit healthcare system in Northeast Ohio, covering two emergency departments (EDs) associated with acute care hospitals, and two standalone EDs.

Participants: Encounters associated with adult patients in any of Summa Health’s four EDs were included. Encounters lasting <1 hour were excluded. Only the first 36 hours of each encounter were used in model evaluation.

Intervention(s)/Exposure(s): The SepsisWatch model was used to predict sepsis based on patient data.

Main Outcome(s) and Measure(s): The primary outcomes measured were the model's area under the precision-recall curve (AUPRC), and area under the receiver operator curve (AUROC).

Results: The study included 205,005 encounters from 101,584 unique patients. 54.7% (n = 112,223) patients were female and the mean age was 50 (IQR, [38,71]). The model demonstrated strong performance across the Summa Health system, with little variation across different sites. The AUROC ranged from 0.906 to 0.960, and the AUPRC ranged from 0.177 to 0.252 across the four sites.

Conclusions and Relevance: The external validation of the SepsisWatch model in a community health system setting confirmed its robust performance and portability across different geographical and demographic contexts. The study underscores the potential of advanced ML models in improving sepsis detection in both academic and community hospital settings, paving the way for prospective studies to measure the clinical and operational impact of such models in healthcare.
Introduction

Sepsis is a systemic inflammatory response to disseminated infection that affects millions of people worldwide.\(^1\) Despite medical advancements, sepsis continues to be a major health concern, accounting for 19.7% of all global deaths.\(^2\) In the US, sepsis is the cause of up to half of all in-hospital fatalities.\(^3\) The economic impact of sepsis is equally staggering. Between 2012 and 2018, the annual cost of inpatient hospitalizations related to sepsis for Medicare beneficiaries increased 26.12% from $17.8 billion to $22.4 billion.\(^4\) Costs for each sepsis-related encounter were estimated to be $18,023 when identified before admission and $51,022 if recognized post-admission.\(^5\) Importantly, early detection and prompt antibiotic treatment can significantly reduce both the cost and the health impacts of sepsis.\(^6-11\) Sepsis care bundles have been developed and diffused to advance the standard of care for sepsis management, but the outcomes of these interventions are mixed.\(^12-14\) Leveraging machine learning (ML) for early sepsis identification has become a pivotal opportunity to further improve sepsis care.

ML models can effectively identify sepsis early and implementation of these models can influence treatment and outcomes.\(^15-18\) Prospective implementation of ML and non-ML driven sepsis alert systems have been shown to decrease in-hospital mortality, organ failure, and length of stay.\(^19,22,24\) However, other studies demonstrate that while these systems may be able to predict impending sepsis they can fall short of changing care delivery and patient outcomes.\(^23\) The role of ML in sepsis treatment merits continued study of models and implementation approaches in different healthcare settings and patient populations.

A key challenge to scaling ML models in healthcare is lack of portability, in part due to significant variability of patient populations and care delivery processes across geographies and points in time. Performance of ML models across diverse healthcare settings and patient
populations is highly variable.\textsuperscript{24-27} Model deployment itself brings forth its own set of challenges. From data curation to quality assurance and continuous monitoring, successful integration requires a multifaceted and deliberate approach.\textsuperscript{28,29} In addition, designing and implementing a workflow solution for clinicians to act on the outputs of an ML model requires additional training and personnel.\textsuperscript{30} Clinical perceptions of ML interventions will also vary depending on intuition and understanding, demonstrating that model accuracy does not necessarily equate to trust by providers and overall solution effectiveness.\textsuperscript{31} Taken together with the privacy concerns related to use and sharing of protected healthcare data, external validations of ML models are limited.

In this paper, we describe the external validation of the SepsisWatch model in Summa Health, a hospital system in Ohio—a setting both geographically and temporally distinct from the context of initial development at Duke University. SepsisWatch is the first deep learning model implemented in routine clinical care in the United States and has been in continuous operations at Duke University since November 2018. This study has two primary objectives. First, we aim to assess the model's performance in this new environment. Second, we aim to estimate the potential workload and benefit to patients associated with the prospective implementation of SepsisWatch in this new environment. Through this study, we hope to contribute meaningful insights into the broader applicability of ML models for sepsis detection, emphasizing the importance of context and adaptation.

**Methods**

**Setting**

Summa Health, a nonprofit integrated healthcare delivery system in Northeast Ohio, encompasses two acute care hospitals: Summa Health Akron City Hospital (ACH) and Summa
Health Barberton (SHB), each equipped with their own emergency department (ED). The system is further complemented by two standalone EDs—ACH Green ED and SHB Wadsworth ED. These facilities form a 1,300-bed system that facilitates over one million patient encounters annually. Summa Health primarily serves Summit, Wayne, Medina, Portage and Stark Counties in Ohio. The service area encompasses urban, suburban, and rural areas. The payer mix of Summa Health patients is 5% uninsured, 25% privately insured, 30% Medicaid, and 40% Medicare.

Data
SepsisWatch was externally evaluated on encounters for adult patients (age ≥ 18) who presented to one of Summa Health’s four emergency departments (EDs) between 1/1/2020 and 12/31/2021. Encounters started at the time of presentation to the ED and ended at time of discharge or death. Encounters were attributed to the site of origination, such that an encounter that began at a free-standing ED and included a transfer to an acute hospital was assigned to the free-standing ED. Each individual visit to the emergency department was considered as a separate encounter, regardless of the number of visits made by the same patient. Encounters with a length of stay less than 1-hour were excluded from the cohort and only the first 36 hours of each encounter were used in the model evaluation.

The model combines static data that remains unchanged throughout an encounter and dynamic data that is updated during an encounter. Static variables included patient demographics, encounter details, and comorbidity data, which looked at ICD-10 codes documented at any encounter in the 12 months prior to ED presentation. Dynamic data used by the model includes...
analyte results, vital signs, and medication administrations. Dynamic variables were considered between the encounter start time and end time.

**Outcome Definition**

We used a previously developed sepsis phenotype as our outcome label. Specifically, we defined sepsis as the co-occurrence of all 3 following criteria: 1) At least two Systemic Inflammatory Response Syndrome (SIRS) criteria, which is valid for 24 hours and includes temperature anomalies (>100.4 F or <96.8 F), a heart rate above 90, a respiration rate exceeding 20, and an abnormal white blood cell count (>12 or <4); (2) a blood culture order; (3) indication of any end-organ damage, characterized by elevated creatinine (>2.0), INR (>1.5), total bilirubin (>2.0), decreased platelet count (<100), lactate levels of 2 or higher, or systolic blood pressure below 90 mmHg or a drop of 40 mmHg in systolic blood pressure within 6 hours. The varying sampling rates for medical measurements were accounted for by adjusting the relevant time window for each criterion. Vital sign documentation values (temperature, heart rate, respiration rate, blood pressure) were valid for a six-hour time window, whereas analyte measurements (white blood cell count, creatinine, bilirubin, platelet count, lactate) and orders (blood culture) were valid for a 24-hour time window. In addition, patients who met the criteria for sepsis within 1 hour of presentation to the ED were excluded.

SepsisWatch was evaluated using a detection window of 12-hours, meaning once a prediction breached the threshold, the prediction would only be classified as a true positive if the patient met sepsis criteria within 12 hours. If the prediction breached the threshold more than 12 hours prior to sepsis, the prediction was classified as a false positive. SepsisWatch was ran hourly on the hour, and produced predictions for all encounters between presenting to the ED
and the minimum of time of sepsis, time of death, time of discharge, and 36 hours after ED presentation. Once a prediction breached the threshold, predictions over the next 8 hours were suppressed or snoozed. This 8-hour snooze window was designed to reduce false positive alerts and downstream potential alert fatigue. The snooze window also avoided inflating performance metrics by repetitively counting true positives.

**Model**

The original model was designed at Duke University Hospital using EHR encounter data from October 1\(^{st}\) 2014 – December 1\(^{st}\) 2015. It is a recurrent neural network (RNN) model and is hereby referred to as SepsisWatch. The data used to train the model was split 80:10:10 for training, internal validation and testing, respectively. The original model performed well in multiple settings. Specifically, on an internal validation cohort SepsisWatch achieved an area under the receiver operator curve (AUROC) of 0.882 and on a temporal validation cohort the model achieved an AUROC of 0.943. The full details of the model development and evaluation can be accessed in the original development, internal validation manuscripts and implementation manuscripts.\(^1\),\(^2\),\(^8\),\(^3\),\(^2\)

**Evaluation**

To answer our first research objective, we evaluated the model performance in the new healthcare setting using precision, recall, area under the precision-recall curve (AUPRC), and area under the receiver operator curve (AUROC). To better understand the performance of SepsisWatch across the different ED and hospital sites, we separately evaluated the model at each location within Summa Health.
To answer our second research objective, we conduct several additional analyses. First, to estimate the potential effect of model integration on clinical care, we quantified the average ‘lead time’, defined as the amount of time between a ‘high risk’ model prediction and a patient meeting sepsis criterion. This measure helps quantify the potential opportunity for earlier intervention. Lead time is only measured for true positive cases identified early by SepsisWatch.

Second, to assess the workflow burden placed on staff, we assessed the number of alerts that would be sent to either a charge nurse or rapid response team (RRT) at a given model threshold. The lead time and number of alerts are calculated separately for each Summa Health site.

**Results**

**Cohort Characteristics**

In total 205,005 encounters from 101,584 unique patients met inclusion criteria for the study. Most patients were female (54.7%, n = 112,223) and the mean age was 50 (IQR, [38,71]). The incidence of sepsis within the first 36 hours of encounters was 3.38% (n = 6,920). The majority of encounters were initiated at the two EDs co-located with acute care hospitals: ACH Emergency Department (ED) (59.08%, n=121,131) and SHB ED (23.53%, n = 48,244). The remaining encounters were initiated at the two standalone EDs: ACH Green ED (11.11%, n = 22,893) and SHB Wadsworth ED (6.21%, n = 12,737). Patient characteristics are broken down by site in Table 1.

**Research Objective 1 - Model Performance**

Overall, the SepsisWatch model demonstrated robust performance on the geographically and temporally distinct Summa Health patient population. Moreover, there was little variation in
model performance across the different emergency departments. Area under the precision-recall
curve (AUPRC) for the four sites was 0.252 for ACH ED, 0.248 for SHB ED, 0.177 for ACH
Green ED and 0.216 for SHB Wadsworth ED. The area under the receiver operator curve
(AUROC) for the four sites was 0.919 for ACH ED, 0.906 for SHB ED, 0.960 for ACH Green
ED and 0.928 for SHB Wadsworth ED. The AUROC and AUPRC for the four sites are
visualized in Figure 1 and Figure 2. The model performed similarly at each distinct location
when stratifying patient population by White and Black races (Table 2).

Research Objective 2 – Evaluating Potential Clinical Benefit and Alert Fatigue

Model performance measures, including precision, recall, average number of alerts per day, and
average lead time prior to meeting sepsis criteria vary based on model threshold. Performance
measures across thresholds are illustrated in Table 3. If a threshold is set at each site to fix
precision (positive predictive value) at 20%, the recall (sensitivity) is 76.2% at ACH ED, 70.9%
at SHB ED, 27.0% at ACH Green ED, and 61.2% at SHB Wadsworth ED. At this same
threshold, the average number of alerts per day is 7 at ACH ED, 3 at SHB ED, 2 at ACH Green
ED, and 1 at SHB Wadsworth ED. Lastly, the average lead time (hours) is 4.06 at ACH ED, 3.79
at SHB ED, 5.07 at ACH Green ED, and 3.58 at SHB Wadsworth ED (eTable 1 in the
Supplement).

Alternatively, if a threshold is set at each site to fix recall (sensitivity) at 60%, the precision
(posi...
average lead time (hours) is 3.94 at ACH ED, 3.69 at SHB ED, 4.89 at ACH Green ED, and 3.42 at SHB Wadsworth ED (eTable 2 in the Supplement).

Discussion

In this study we present the first external validation of a sepsis ML model in a community based health setting. The model was originally developed at a tertiary academic medical center in North Carolina and maintained robust performance across four EDs at a community health system in Ohio. The original model achieved an AUROC of 0.83 and an AUPRC of 0.257.\(^{18}\) This strong performance was maintained in the external validation presented in this study, in which SepsisWatch achieved an AUROC of 0.92 and an AUPRC of 0.24. Moreover, SepsisWatch performed strongly across two EDs that are co-located with acute care hospitals and two standalone EDs, marking the first known validation of a sepsis machine learning algorithm in a standalone ED. A ‘Model Facts’ sheet containing details of this external validation and the original model development can be seen in eFigure 1 in the Supplement.

Most machine learning algorithms are not externally validated and those that are demonstrate varied performance. Wong \textit{et al.} demonstrated that the proprietary Epic Sepsis Model (ESM) had substantially worse calibration and discrimination among adult patients within one US academic health system. A later study by Lyons \textit{et al.} also revealed and varied performance of ESM across nine hospitals and found that hospitals with lower sepsis incidence had worse AUROC.\(^{24}\) On the other hand, Brajer \textit{et al.} demonstrated robust performance of an in-hospital mortality ML model across multiple hospitals within the same health system.\(^{25}\) Moor \textit{et al.} used a novel validation technique pooling predictions of models developed on different data, achieving an AUROC of 0.76 on external validation cohorts verse 0.84 on internal validation.\(^{33}\) While Moor \textit{et al.} were able improve external validation performance after fine tuning the
models using a small set of data from the target testing site, performance remained stronger in
the internal validation cohort. The new version of the ESM similarly includes fine tuning on
local data to improve performance, but improved performance of that model has not yet been
reported in the peer-reviewed literature. In addition, local fine tuning necessitates additional
expertise, personnel, and compute infrastructure.

Beyond generalizing across time and geographic settings, SepsisWatch also exhibited
strong and robust performance across demographic subgroups. Unfortunately, examination of
ML model performance across racial subgroups has not been reported in many prior external
validation studies in sepsis and other clinical domains. There are significant concerns when
models are trained on datasets with minimal diversity and tested on populations that include
more representation of historically marginalized populations. The current study builds on prior
work from dermatology that found that training on more diverse datasets led to improved
performance on diverse populations. SepsisWatch was trained on a cohort of adults in North
Carolina that were 30% Black and performed well on cohorts of adults in Ohio that were 3% –
30% Black. The robust performance of the SepsisWatch model presented in the current study
emphasizes the importance of training models on diverse datasets.

Balancing alert fatigue with clinically meaningful predictions is a well-established
challenge in operationalizing sepsis prediction models. Moreover, given expected trade-
offs between model precision and recall, choosing a prediction threshold that balances both
overall performance with clinical impact becomes exceedingly important. For example, at the
ACH Emergency Department, when prediction threshold was set to 0.1, the model would send
on average 147 alerts per day. At this threshold the model would identify almost every instance
of sepsis and provide an average lead time of 7.1 hours. But the precision of the model at this
threshold is 2.1% and approximately 47 alerts need to be screened for every case that goes on to
develop sepsis. In contrast, a prediction threshold of 0.6 would afford a recall of 84%, precision
of 20% and average lead time of 4.23 hours. The number of alerts at this threshold is reduced
80% to an average of 29 alerts per day. Thus, operational leaders within each setting can titrate
the model threshold to align with the capacity of front-line clinicians to respond to alerts.
Ultimately, the threshold should be determined after a silent trial in which front-line clinicians
can help test SepsisWatch and ensure that patients flagged by the model are appropriate for
review.

Our study had several limitations. First, this was a retrospective study and while model
performance and potential opportunity to improve care could be assessed, SepsisWatch will need
to be prospectively tested to measure impact. Second, this study features a single health system
in a single geography. The generalizability of SepsisWatch beyond North Carolina and Ohio
remains unknown. Third, this study did not directly compare SepsisWatch to other available
sepsis models, such as the Epic Sepsis Model or TREWS. Those technologies were not available
to include in the analysis and additional contracting and potential costs would be required to
compare multiple algorithms. Future comparative effectiveness research will be required to
better understand the tradeoffs associated with use of each model. Fourth, this study evaluated
SepsisWatch during a time window that included the COVID-19 pandemic. The time window
was two years and covered multiple variant waves of the virus. Future research will be needed to
characterize variability of model performance during viral outbreaks. Finally, this study only
evaluated the performance of SepsisWatch with an 8-hour snooze window. Future studies are
necessary to optimize and tailor this time window for different implementation sites.
Conclusions

This study illustrates the first successful external validation of a sepsis deep learning model, SepsisWatch, within a community health system. SepsisWatch demonstrated remarkable consistency in performance despite variations in time, geography, and patient demographics. The strong performance of SepsisWatch highlights the possibility of effectively transporting advanced ML models from academic to community based hospital settings. Additionally, the study addresses the crucial balance between minimizing alert fatigue and maintaining clinical relevance, emphasizing the importance of carefully selecting prediction thresholds tailored to each deployment site and clinical context. The findings suggest that while there are challenges in creating more broadly applicable clinical prediction models, careful evaluation in different health contexts is feasible and can yield promising results. This study will pave the way for future prospective studies to measure the clinical and operational effect of SepsisWatch and other sepsis machine learning models integrated into clinical care.


### Table 1: Description of cohort characteristics across all four emergency departments in the Summa Health system.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>ACH EMERGENCY DEPT</th>
<th>SHB EMERGENCY DEPT</th>
<th>ACH GREEN ED</th>
<th>SHB WADSWORTH ED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Encounters, N (%)</strong></td>
<td>121131 (59.08%)</td>
<td>48244 (23.53%)</td>
<td>22893 (11.11%)</td>
<td>12737 (6.21%)</td>
</tr>
<tr>
<td><strong>Age (years), mean ±SD</strong></td>
<td>50.92 ± 20.12</td>
<td>50.71 ± 20.18</td>
<td>47.36 ± 19.28</td>
<td>48 ± 20.05</td>
</tr>
<tr>
<td><strong>Sex Male, N (%)</strong></td>
<td>56540 (46.67%)</td>
<td>21,284 (44.12%)</td>
<td>9,552 (41.72%)</td>
<td>5,551 (43.58%)</td>
</tr>
<tr>
<td><strong>Sex Female, N (%)</strong></td>
<td>64470 (53.22%)</td>
<td>26,960 (55.88%)</td>
<td>13,341 (58.28%)</td>
<td>7,184 (56.40%)</td>
</tr>
<tr>
<td><strong>Sex (Others/Missing), N (%)</strong></td>
<td>121 (0.10%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>2 (0.02%)</td>
</tr>
<tr>
<td><strong>Race, N (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>36431 (30.08%)</td>
<td>6,545 (13.57%)</td>
<td>2,436 (10.64%)</td>
<td>279 (2.19%)</td>
</tr>
<tr>
<td>Caucasian/White</td>
<td>75833 (62.60%)</td>
<td>40,754 (84.47%)</td>
<td>20,086 (87.74%)</td>
<td>12,242 (96.11%)</td>
</tr>
<tr>
<td>Missing/other</td>
<td>8867 (7.32%)</td>
<td>945 (1.96%)</td>
<td>371 (1.62%)</td>
<td>216 (1.70%)</td>
</tr>
<tr>
<td><strong>Comorbidities, N (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>4515 (3.73%)</td>
<td>1,222 (2.53%)</td>
<td>257 (1.12%)</td>
<td>192 (1.51%)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>5805 (4.80%)</td>
<td>1,581 (3.28%)</td>
<td>534 (2.33)</td>
<td>479 (3.80%)</td>
</tr>
<tr>
<td>Pulmonary circulation disorders</td>
<td>5441 (4.49%)</td>
<td>2,777 (5.76%)</td>
<td>633 (2.77%)</td>
<td>459 (3.60%)</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>3702 (3.06%)</td>
<td>1,012 (2.10%)</td>
<td>283 (1.24%)</td>
<td>205 (1.61%)</td>
</tr>
<tr>
<td>Fluid and electrolyte disorders</td>
<td>9324 (7.70%)</td>
<td>3,705 (7.68%)</td>
<td>995 (4.35%)</td>
<td>642 (5.04%)</td>
</tr>
<tr>
<td>Depression</td>
<td>2521 (2.08%)</td>
<td>612 (1.27%)</td>
<td>141 (0.62%)</td>
<td>161 (1.26%)</td>
</tr>
<tr>
<td>In-hospital mortality, N (%)</td>
<td>1139 (0.94%)</td>
<td>381 (0.79%)</td>
<td>0.16%</td>
<td>0.28%</td>
</tr>
<tr>
<td><strong>Median Length of Stay (25%–75%)</strong></td>
<td>6.38 (3.5, 56.7)</td>
<td>4 (2.46, 28.8)</td>
<td>2.63 (1.71, 4.15)</td>
<td>2.43 (1.6, 4.31)</td>
</tr>
<tr>
<td><strong>Overall rate of ICU admission, N (%)</strong></td>
<td>5487 (4.53%)</td>
<td>2644 (5.48%)</td>
<td>272 (1.19%)</td>
<td>320 (2.51%)</td>
</tr>
<tr>
<td><strong>Septic, N (%)</strong></td>
<td>4639 (3.83%)</td>
<td>1785 (3.70%)</td>
<td>240 (1.05%)</td>
<td>245 (1.92%)</td>
</tr>
</tbody>
</table>
Table 2: Site-specific performance metrics stratified by race including area under the precision recall curve (AU-PR) and area under the receiver operator curve (AUROC)

<table>
<thead>
<tr>
<th>Cohort</th>
<th>AU-PR</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACH ED</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entire Population</td>
<td>0.252</td>
<td>0.919</td>
</tr>
<tr>
<td>White Adults</td>
<td>0.255</td>
<td>0.918</td>
</tr>
<tr>
<td>Black Adults</td>
<td>0.238</td>
<td>0.911</td>
</tr>
<tr>
<td><strong>SHB ED</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entire Population</td>
<td>0.248</td>
<td>0.906</td>
</tr>
<tr>
<td>White Adults</td>
<td>0.246</td>
<td>0.901</td>
</tr>
<tr>
<td>Black Adults</td>
<td>0.286</td>
<td>0.920</td>
</tr>
<tr>
<td><strong>ACH Green ED</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entire Population</td>
<td>0.177</td>
<td>0.960</td>
</tr>
<tr>
<td>White Adults</td>
<td>0.175</td>
<td>0.957</td>
</tr>
<tr>
<td>Black Adults</td>
<td>0.18</td>
<td>0.983</td>
</tr>
<tr>
<td><strong>SHB- Wadsworth ED</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entire Population</td>
<td>0.216</td>
<td>0.928</td>
</tr>
<tr>
<td>White Adults</td>
<td>0.213</td>
<td>0.926</td>
</tr>
<tr>
<td>Black Adults</td>
<td>0.325</td>
<td>0.974</td>
</tr>
</tbody>
</table>
Table 3: Site specific model performance including precision, recall, average number of alerts per day and average lead time based on model threshold.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>ACH Emergency Department</th>
<th>SHB Emergency Department</th>
<th>ACH Green Emergency Department</th>
<th>SHB Wadsworth Emergency Department</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alerts/Day (95% CI)</td>
<td>Average lead time (95% CI)</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>0.1</td>
<td>147 (144, 148)</td>
<td>7.1 (6.9, 7.31)</td>
<td>0.021</td>
<td>0.986</td>
</tr>
<tr>
<td>0.2</td>
<td>42 (41, 43)</td>
<td>4.97 (4.81, 5.15)</td>
<td>0.051</td>
<td>0.949</td>
</tr>
<tr>
<td>0.3</td>
<td>34 (32, 35)</td>
<td>4.60 (4.44, 4.78)</td>
<td>0.161</td>
<td>0.922</td>
</tr>
<tr>
<td>0.4</td>
<td>32 (30, 32)</td>
<td>4.43 (4.27, 4.6)</td>
<td>0.186</td>
<td>0.899</td>
</tr>
<tr>
<td>0.5</td>
<td>30 (29, 31)</td>
<td>4.31 (4.15, 4.48)</td>
<td>0.193</td>
<td>0.872</td>
</tr>
<tr>
<td>0.6</td>
<td>29 (28, 30)</td>
<td>4.23 (4.07, 4.4)</td>
<td>0.195</td>
<td>0.838</td>
</tr>
<tr>
<td>0.7</td>
<td>29 (28, 30)</td>
<td>4.12 (3.96, 4.29)</td>
<td>0.195</td>
<td>0.794</td>
</tr>
<tr>
<td>0.8</td>
<td>25 (24, 25)</td>
<td>4.02 (3.85, 4.19)</td>
<td>0.211</td>
<td>0.730</td>
</tr>
<tr>
<td>0.9</td>
<td>14 (14, 15)</td>
<td>3.86 (3.69, 4.04)</td>
<td>0.258</td>
<td>0.548</td>
</tr>
</tbody>
</table>

Notes: The table includes site-specific model performance with precision, recall, average alerts per day, and average lead time based on model threshold. The results are averaged across different thresholds, indicating the model's effectiveness in predicting ACH events at the Emergency Departments of ACH, SHB, ACH Green, and SHB Wadsworth.
**Figure 1:** Area under the receiver operator curve at each of Summa Health’s four emergency departments, A) ACH Emergency Department (ED), B) SHB ED, C) ACH Green ED, D) SHB Wadsworth ED
Figure 2: Area under precision-recall curve at each of Summa Health’s four emergency departments A) ACH Emergency Department (ED) B) SHB ED C) ACH Green ED D) SHB Wadsworth ED