

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

3
4
5 Bruno Valan, MENG¹; Anusha Prakash, MS²; William Ratliff, MBA¹; Michael Gao, MS¹;
6 Srikanth Muthya²; Ajit Thomas, MBA²; Jennifer L. Eaton, PhD³; Matt Gardner¹; Marshall
7 Nichols, MS¹; Mike Revoir¹; Dustin Tart, RN⁴; Cara O'Brien, MD^{4,5}; Manesh Patel, MD⁵;
8 Suresh Balu, MS, MBA¹; Mark Sendak, MPP, MD¹

9
10 **Address Correspondence to:**

11 Mark Sendak

12 ¹Duke Institute for Health Innovation
13 300 West Morgan St. Durham, North Carolina

14 Phone: (919) 684-4389

15 Email: mark.sendak@duke.edu

16
17 **Affiliations**

18
19 ²Cohere Med Inc.

20 571 Euclid Ave

21 Berkeley, CA 94708

22
23 ³Summa Health Research & Innovation,
24 Summa Health, Akron, OH 44310

25
26 ⁴Duke University Hospital

27 Durham, NC 27701

28
29 ⁵Department of Medicine,

30 Duke University School of Medicine

31 Durham, NC, United States

32
33
34
35
36
37
38
39
40 **Study Type: Diagnostic/Prognostic Study**

46 Abstract

47
48 **Importance:** Sepsis accounts for a substantial portion of global deaths and healthcare costs.
49 Early detection using machine learning (ML) models offers a critical opportunity to improve care
50 and reduce the burden of sepsis.

51
52 **Objective:** To externally validate the SepsisWatch ML model, initially developed at Duke
53 University, in a community healthcare and assess its performance and clinical utility in early
54 sepsis detection.

55
56 **Design:** This retrospective external validation study evaluated the performance of the
57 SepsisWatch model in a new environment. Data from patient encounters at Summa Health's
58 emergency departments between 2020 and 2021 were used. The study analyzed the model's
59 ability to predict sepsis using a combination of static and dynamic patient data.

60
61 **Setting:** The study was conducted at Summa Health, a nonprofit healthcare system in Northeast
62 Ohio, covering two emergency departments (EDs) associated with acute care hospitals, and two
63 standalone EDs.

64
65 **Participants:** Encounters associated with adult patients in any of Summa Health's four EDs
66 were included. Encounters lasting <1 hour were excluded. Only the first 36 hours of each
67 encounter were used in model evaluation.

68
69 **Intervention(s)/Exposure(s):** The SepsisWatch model was used to predict sepsis based on
70 patient data.

71
72 **Main Outcome(s) and Measure(s):** The primary outcomes measured were the model's area
73 under the precision-recall curve (AUPRC), and area under the receiver operator curve (AUROC).

74
75 **Results:** The study included 205,005 encounters from 101,584 unique patients. 54.7% (n =
76 112,223) patients were female and the mean age was 50 (IQR, [38,71]). The model demonstrated
77 strong performance across the Summa Health system, with little variation across different sites.
78 The AUROC ranged from 0.906 to 0.960, and the AUPRC ranged from 0.177 to 0.252 across the
79 four sites.

80
81 **Conclusions and Relevance:** The external validation of the SepsisWatch model in a community
82 health system setting confirmed its robust performance and portability across different
83 geographical and demographic contexts. The study underscores the potential of advanced ML
84 models in improving sepsis detection in both academic and community hospital settings, paving
85 the way for prospective studies to measure the clinical and operational impact of such models in
86 healthcare.

87

88

89

90

91 **Introduction**

92
93
94 Sepsis is a systemic inflammatory response to disseminated infection that affects millions of
95 people worldwide.¹ Despite medical advancements, sepsis continues to be a major health
96 concern, accounting for 19.7% of all global deaths.² In the US, sepsis is the cause of up to half of
97 all in-hospital fatalities.³ The economic impact of sepsis is equally staggering. Between 2012 and
98 2018, the annual cost of inpatient hospitalizations related to sepsis for Medicare beneficiaries
99 increased 26.12% from \$17.8 billion to \$22.4 billion.⁴ Costs for each sepsis-related encounter
100 were estimated to be \$18,023 when identified before admission and \$51,022 if recognized post-
101 admission.⁵ Importantly, early detection and prompt antibiotic treatment can significantly reduce
102 both the cost and the health impacts of sepsis.⁶⁻¹¹ Sepsis care bundles have been developed and
103 diffused to advance the standard of care for sepsis management, but the outcomes of these
104 interventions are mixed.¹²⁻¹⁴ Leveraging machine learning (ML) for early sepsis identification
105 has become a pivotal opportunity to further improve sepsis care.

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

113 A key challenge to scaling ML models in healthcare is lack of portability, in part due to
114 significant variability of patient populations and care delivery processes across geographies and
115 points in time. Performance of ML models across diverse healthcare settings and patient

116 populations is highly variable.²⁴⁻²⁷ Model deployment itself brings forth its own set of
117 challenges. From data curation to quality assurance and continuous monitoring, successful
118 integration requires a multifaceted and deliberate approach.^{28,29} In addition, designing and
119 implementing a workflow solution for clinicians to act on the outputs of an ML model requires
120 additional training and personnel.³⁰ Clinical perceptions of ML interventions will also vary
121 depending on intuition and understanding, demonstrating that model accuracy does not
122 necessarily equate to trust by providers and overall solution effectiveness.³¹ Taken together with
123 the privacy concerns related to use and sharing of protected healthcare data, external validations
124 of ML models are limited.

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

135 136 **Methods**

137 138 *Setting*

139 Summa Health, a nonprofit integrated healthcare delivery system in Northeast Ohio,
140 encompasses two acute care hospitals: Summa Health Akron City Hospital (ACH) and Summa

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

149

150 *Data*

151 SepsisWatch was externally evaluated on encounters for adult patients (age ≥ 18) who
152 presented to one of Summa Health's four emergency departments (EDs) between 1/1/2020 and
153 12/31/2021. Encounters started at the time of presentation to the ED and ended at time of
154 discharge or death. Encounters were attributed to the site of origination, such that an encounter
155 that began at a free-standing ED and included a transfer to an acute hospital was assigned to the
156 free-standing ED. Each individual visit to the emergency department was considered as a
157 separate encounter, regardless of the number of visits made by the same patient. Encounters with
158 a length of stay less than 1-hour were excluded from the cohort and only the first 36 hours of
159 each encounter were used in the model evaluation.

160

161 The model combines static data that remains unchanged throughout an encounter and dynamic
162 data that is updated during an encounter. Static variables included patient demographics,
163 encounter details, and comorbidity data, which looked at ICD-10 codes documented at any
164 encounter in the 12 months prior to ED presentation. Dynamic data used by the model includes

165 analyte results, vital signs, and medication administrations. Dynamic variables were considered
166 between the encounter start time and end time.

167

168 *Outcome Definition*

169 We used a previously developed sepsis phenotype as our outcome label. Specifically, we defined
170 sepsis as the co-occurrence of all 3 following criteria: 1) At least two Systemic Inflammatory
171 Response Syndrome (SIRS) criteria, which is valid for 24 hours and includes temperature
172 anomalies (>100.4 F or <96.8 F), a heart rate above 90, a respiration rate exceeding 20, and an
173 abnormal white blood cell count (>12 or <4); (2) a blood culture order; (3) indication of any end-
174 organ damage, characterized by elevated creatinine (>2.0), INR (>1.5), total bilirubin (>2.0),
175 decreased platelet count (<100), lactate levels of 2 or higher, or systolic blood pressure below 90
176 mmHg or a drop of 40 mmHg in systolic blood pressure within 6 hours. The varying sampling
177 rates for medical measurements were accounted for by adjusting the relevant time window for
178 each criterion. Vital sign documentation values (temperature, heart rate, respiration rate, blood
179 pressure) were valid for a six-hour time window, whereas analyte measurements (white blood
180 cell count, creatinine, bilirubin, platelet count, lactate) and orders (blood culture) were valid for a
181 24-hour time window. In addition, patients who met the criteria for sepsis within 1 hour of
182 presentation to the ED were excluded.

183 SepsisWatch was evaluated using a detection window of 12-hours, meaning once a
184 prediction breached the threshold, the prediction would only be classified as a true positive if the
185 patient met sepsis criteria within 12 hours. If the prediction breached the threshold more than 12
186 hours prior to sepsis, the prediction was classified as a false positive. SepsisWatch was ran
187 hourly on the hour, and produced predictions for all encounters between presenting to the ED

188 and the minimum of time of sepsis, time of death, time of discharge, and 36 hours after ED
189 presentation. Once a prediction breached the threshold, predictions over the next 8 hours were
190 suppressed or snoozed. This 8-hour snooze window was designed to reduce false positive alerts
191 and downstream potential alert fatigue. The snooze window also avoided inflating performance
192 metrics by repetitively counting true positives.

193

194 *Model*

195 The original model was designed at Duke University Hospital using EHR encounter data from
196 October 1st 2014 – December 1st 2015. It is a recurrent neural network (RNN) model and is
197 hereby referred to as SepsisWatch. The data used to train the model was split 80:10:10 for
198 training, internal validation and testing, respectively. The original model performed well in
199 multiple settings. Specifically, on an internal validation cohort SepsisWatch achieved an area
200 under the receiver operator curve (AUROC) of 0.882 and on a temporal validation cohort the
201 model achieved an AUROC of 0.943. The full details of the model development and evaluation
202 can be accessed in the original development, internal validation manuscripts and implementation
203 manuscripts.^{18,28,32}

204

205 *Evaluation*

206 To answer our first research objective, we evaluated the model performance in the new
207 healthcare setting using precision, recall, area under the precision-recall curve (AUPRC), and
208 area under the receiver operator curve (AUROC). To better understand the performance of
209 SepsisWatch across the different ED and hospital sites, we separately evaluated the model at
210 each location within Summa Health.

211
212 To answer our second research objective, we conduct several additional analyses. First, to
213 estimate the potential effect of model integration on clinical care, we quantified the average ‘lead
214 time’, defined as the amount of time between a ‘high risk’ model prediction and a patient
215 meeting sepsis criterion. This measure helps quantify the potential opportunity for earlier
216 intervention. Lead time is only measured for true positive cases identified early by SepsisWatch.
217 Second, to assess the workflow burden placed on staff, we assessed the number of alerts that
218 would be sent to either a charge nurse or rapid response team (RRT) at a given model threshold.
219 The lead time and number of alerts are calculated separately for each Summa Health site.

220

221 **Results**

222

223 *Cohort Characteristics*

224

225 In total 205,005 encounters from 101,584 unique patients met inclusion criteria for the study.

226 Most patients were female (54.7%, n = 112,223) and the mean age was 50 (IQR, [38,71]). The

227 incidence of sepsis within the first 36 hours of encounters was 3.38% (n = 6,920). The majority

228 of encounters were initiated at the two EDs co-located with acute care hospitals: ACH

229 Emergency Department (ED) (59.08%, n=121,131) and SHB ED (23.53%, n = 48,244). The

230 remaining encounters were initiated at the two standalone EDs: ACH Green ED (11.11%, n =

231 22,893) and SHB Wadsworth ED (6.21%, n = 12,737). Patient characteristics are broken down

232 by site in **Table 1**.

233

234 *Research Objective 1 - Model Performance*

235

236 Overall, the SepsisWatch model demonstrated robust performance on the geographically and

237 temporally distinct Summa Health patient population. Moreover, there was little variation in

238 model performance across the different emergency departments. Area under the precision-recall
239 curve (AUPRC) for the four sites was 0.252 for ACH ED, 0.248 for SHB ED, 0.177 for ACH
240 Green ED and 0.216 for SHB Wadsworth ED. The area under the receiver operator curve
241 (AUROC) for the four sites was 0.919 for ACH ED, 0.906 for SHB ED, 0.960 for ACH Green
242 ED and 0.928 for SHB Wadsworth ED. The AUROC and AUPRC for the four sites are
243 visualized in **Figure 1** and **Figure 2**. The model performed similarly at each distinct location
244 when stratifying patient population by White and Black races (**Table 2**).

245

246 *Research Objective 2 – Evaluating Potential Clinical Benefit and Alert Fatigue*

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

256

257 Alternatively, if a threshold is set at each site to fix recall (sensitivity) at 60%, the precision
258 (positive predictive value) is 23.9% at ACH ED, 23.1% at SHB ED, 16.3% at ACH Green ED,
259 and 20.1% at SHB Wadsworth ED. At this same threshold, the average number of alerts per day
260 is 7 at ACH ED, 3 at SHB ED, 1 at ACH Green ED, and 1 at SHB Wadsworth ED. Lastly, the

261 average lead time (hours) is 3.94 at ACH ED, 3.69 at SHB ED, 4.89 at ACH Green ED, and 3.42
262 at SHB Wadsworth ED (**eTable 2 in the Supplement**).

263
264 **Discussion**

265
266 In this study we present the first external validation of a sepsis ML model in a community
267 based health setting. The model was originally developed at a tertiary academic medical center in
268 North Carolina and maintained robust performance across four EDs at a community health
269 system in Ohio. The original model achieved an AUROC of 0.83 and an AUPRC of 0.257.¹⁸ This
270 strong performance was maintained in the external validation presented in this study, in which
271 SepsisWatch achieved an AUROC of 0.92 and an AUPRC of 0.24. Moreover, SepsisWatch
272 performed strongly across two EDs that are co-located with acute care hospitals and two
273 standalone EDs, marking the first known validation of a sepsis machine learning algorithm in a
274 standalone ED. A ‘Model Facts’ sheet containing details of this external validation and the
275 original model development can be seen in **eFigure 1 in the Supplement**.

276 Most machine learning algorithms are not externally validated and those that are
277 demonstrate varied performance. Wong *et al.* demonstrated that the proprietary Epic Sepsis
278 Model (ESM) had substantially worse calibration and discrimination among adult patients within
279 one US academic health system. A later study by Lyons *et al.* also revealed and varied
280 performance of ESM across nine hospitals and found that hospitals with lower sepsis incidence
281 had worse AUROC.²⁴ On the other hand, Brajer *et al.* demonstrated robust performance of an in-
282 hospital mortality ML model across multiple hospitals within the same health system.²⁵ Moor *et*
283 *al.* used a novel validation technique pooling predictions of models developed on different data,
284 achieving an AUROC of 0.76 on external validation cohorts verse 0.84 on internal validation.³³
285 While Moor *et al.* were able improve external validation performance after fine tuning the

286 models using a small set of data from the target testing site, performance remained stronger in
287 the internal validation cohort.³³ The new version of the ESM similarly includes fine tuning on
288 local data to improve performance, but improved performance of that model has not yet been
289 reported in the peer-reviewed literature. In addition, local fine tuning necessitates additional
290 expertise, personnel, and compute infrastructure.

291 Beyond generalizing across time and geographic settings, SepsisWatch also exhibited
292 strong and robust performance across demographic subgroups. Unfortunately, examination of
293 ML model performance across racial subgroups has not been reported in many prior external
294 validation studies in sepsis and other clinical domains.^{27,33} There are significant concerns when
295 models are trained on datasets with minimal diversity and tested on populations that include
296 more representation of historically marginalized populations.³⁴ The current study builds on prior
297 work from dermatology that found that training on more diverse datasets led to improved
298 performance on diverse populations.³⁵ SepsisWatch was trained on a cohort of adults in North
299 Carolina that were 30% Black and performed well on cohorts of adults in Ohio that were 3% –
300 30% Black. The robust performance of the SepsisWatch model presented in the current study
301 emphasizes the importance of training models on diverse datasets.

302 Balancing alert fatigue with clinically meaningful predictions is a well-established
303 challenge in operationalizing sepsis prediction models.^{27,36,37} Moreover, given expected trade-
304 offs between model precision and recall, choosing a prediction threshold that balances both
305 overall performance with clinical impact becomes exceedingly important. For example, at the
306 ACH Emergency Department, when prediction threshold was set to 0.1, the model would send
307 on average 147 alerts per day. At this threshold the model would identify almost every instance
308 of sepsis and provide an average lead time of 7.1 hours. But the precision of the model at this

309 threshold is 2.1% and approximately 47 alerts need to be screened for every case that goes on to
310 develop sepsis. In contrast, a prediction threshold of 0.6 would afford a recall of 84%, precision
311 of 20% and average lead time of 4.23 hours. The number of alerts at this threshold is reduced
312 80% to an average of 29 alerts per day. Thus, operational leaders within each setting can titrate
313 the model threshold to align with the capacity of front-line clinicians to respond to alerts.
314 Ultimately, the threshold should be determined after a silent trial in which front-line clinicians
315 can help test SepsisWatch and ensure that patients flagged by the model are appropriate for
316 review.

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

331

332 **Conclusions**

333

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

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363 **References**

- 364
- 365 1. Levy MM, Fink MP, Marshall JC, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International
366 Sepsis Definitions Conference. *Intensive Care Med.* 2003;29(4):530-538. doi:10.1007/s00134-
367 003-1662-x
- 368 2. Liu V, Escobar GJ, Greene JD, et al. Hospital deaths in patients with sepsis from 2
369 independent cohorts. *JAMA.* 2014;312(1):90-92. doi:10.1001/jama.2014.5804
- 370 3. Rudd KE, Johnson SC, Agesa KM, et al. Global, regional, and national sepsis incidence and
371 mortality, 1990-2017: analysis for the Global Burden of Disease Study. *Lancet Lond Engl.*
372 2020;395(10219):200-211. doi:10.1016/S0140-6736(19)32989-7
- 373 4. Buchman TG, Simpson SQ, Sciarretta KL, et al. Sepsis Among Medicare Beneficiaries: 1. The
374 Burdens of Sepsis, 2012-2018. *Crit Care Med.* 2020;48(3):276-288.
375 doi:10.1097/CCM.0000000000004224
- 376 5. Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and Costs of Sepsis in
377 the United States-An Analysis Based on Timing of Diagnosis and Severity Level. *Crit Care*
378 *Med.* 2018;46(12):1889-1897. doi:10.1097/CCM.0000000000003342
- 379 6. Rhodes A, Evans LE, Alhazzani W, et al. Surviving Sepsis Campaign: International
380 Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Med.*
381 2017;43(3):304-377. doi:10.1007/s00134-017-4683-6
- 382 7. Cohen J, Vincent JL, Adhikari NKJ, et al. Sepsis: a roadmap for future research. *Lancet Infect*
383 *Dis.* 2015;15(5):581-614. doi:10.1016/S1473-3099(15)70112-X
- 384 8. Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective
385 antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care*
386 *Med.* 2006;34(6):1589-1596. doi:10.1097/01.CCM.0000217961.75225.E9
- 387 9. Ferrer R, Martin-Loeches I, Phillips G, et al. Empiric antibiotic treatment reduces mortality in
388 severe sepsis and septic shock from the first hour: results from a guideline-based performance
389 improvement program. *Crit Care Med.* 2014;42(8):1749-1755.
390 doi:10.1097/CCM.0000000000000330
- 391 10. Liu VX, Fielding-Singh V, Greene JD, et al. The Timing of Early Antibiotics and Hospital
392 Mortality in Sepsis. *Am J Respir Crit Care Med.* 2017;196(7):856-863.
393 doi:10.1164/rccm.201609-1848OC
- 394 11. Peltan ID, Brown SM, Bledsoe JR, et al. ED Door-to-Antibiotic Time and Long-term
395 Mortality in Sepsis. *Chest.* 2019;155(5):938-946. doi:10.1016/j.chest.2019.02.008
- 396 12. Rhee C, Yu T, Wang R, et al. Association Between Implementation of the Severe Sepsis
397 and Septic Shock Early Management Bundle Performance Measure and Outcomes in Patients
398 With Suspected Sepsis in US Hospitals. *JAMA Netw Open.* 2021;4(12):e2138596.
399 doi:10.1001/jamanetworkopen.2021.38596

- 400 13. Townsend SR, Phillips GS, Duseja R, et al. Effects of Compliance With the Early
401 Management Bundle (SEP-1) on Mortality Changes Among Medicare Beneficiaries With
402 Sepsis: A Propensity Score Matched Cohort Study. *Chest*. 2022;161(2):392-406.
403 doi:10.1016/j.chest.2021.07.2167
- 404 14. Barbash IJ, Davis BS, Yabes JG, Seymour CW, Angus DC, Kahn JM. Treatment Patterns
405 and Clinical Outcomes After the Introduction of the Medicare Sepsis Performance Measure
406 (SEP-1). *Ann Intern Med*. 2021;174(7):927-935. doi:10.7326/M20-5043
- 407 15. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score
408 (TREWScore) for septic shock. *Sci Transl Med*. 2015;7(299):299ra122.
409 doi:10.1126/scitranslmed.aab3719
- 410 16. Desautels T, Calvert J, Hoffman J, et al. Prediction of Sepsis in the Intensive Care Unit
411 With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med*
412 *Inform*. 2016;4(3):e28. doi:10.2196/medinform.5909
- 413 17. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an
414 automated trigger for sepsis clinical decision support at emergency department triage using
415 machine learning. *PLOS ONE*. 2017;12(4):e0174708. doi:10.1371/journal.pone.0174708
- 416 18. Bedoya AD, Futoma J, Clement ME, et al. Machine learning for early detection of sepsis:
417 an internal and temporal validation study. *JAMIA Open*. 2020;3(2):252-260.
418 doi:10.1093/jamiaopen/ooaa006
- 419 19. Tarabichi Y, Cheng A, Bar-Shain D, et al. Improving Timeliness of Antibiotic
420 Administration Using a Provider and Pharmacist Facing Sepsis Early Warning System in the
421 Emergency Department Setting: A Randomized Controlled Quality Improvement Initiative.
422 *Crit Care Med*. 2022;50(3):418-427. doi:10.1097/CCM.0000000000005267
- 423 20. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes
424 after implementation of the TREWS machine learning-based early warning system for sepsis.
425 *Nat Med*. 2022;28(7):1455-1460. doi:10.1038/s41591-022-01894-0
- 426 21. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine
427 learning-based severe sepsis prediction algorithm on patient survival and hospital length of
428 stay: a randomised clinical trial. *BMJ Open Respir Res*. 2017;4(1):e000234.
429 doi:10.1136/bmjresp-2017-000234
- 430 22. Nelson JL, Smith BL, Jared JD, Younger JG. Prospective trial of real-time electronic
431 surveillance to expedite early care of severe sepsis. *Ann Emerg Med*. 2011;57(5):500-504.
432 doi:10.1016/j.annemergmed.2010.12.008
- 433 23. Giannini HM, Ginestra JC, Chivers C, et al. A Machine Learning Algorithm to Predict
434 Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical
435 Practice. *Crit Care Med*. 2019;47(11):1485-1492. doi:10.1097/CCM.0000000000003891

- 436 24. Lyons PG, Hofford MR, Yu SC, et al. Factors Associated With Variability in the
437 Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the
438 US. *JAMA Intern Med.* 2023;183(6):611-612. doi:10.1001/jamainternmed.2022.7182
- 439 25. Brajer N, Cozzi B, Gao M, et al. Prospective and External Evaluation of a Machine
440 Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. *JAMA Netw*
441 *Open.* 2020;3(2):e1920733. doi:10.1001/jamanetworkopen.2019.20733
- 442 26. Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal
443 Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration.
444 *Crit Care Med.* 2019;47(1):49-55. doi:10.1097/CCM.0000000000003439
- 445 27. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented
446 Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med.*
447 2021;181(8):1065-1070. doi:10.1001/jamainternmed.2021.2626
- 448 28. Sendak MP, Ratliff W, Sarro D, et al. Real-World Integration of a Sepsis Deep Learning
449 Technology Into Routine Clinical Care: Implementation Study. *JMIR Med Inform.*
450 2020;8(7):e15182. doi:10.2196/15182
- 451 29. Sendak M, Sirdeshmukh G, Ochoa T, et al. Development and Validation of ML-DQA -- a
452 Machine Learning Data Quality Assurance Framework for Healthcare. Published online
453 August 4, 2022. doi:10.48550/arXiv.2208.02670
- 454 30. Topiwala R, Patel K, Twigg J, Rhule J, Meisenberg B. Retrospective Observational Study
455 of the Clinical Performance Characteristics of a Machine Learning Approach to Early Sepsis
456 Identification. *Crit Care Explor.* 2019;1(9):e0046. doi:10.1097/CCE.0000000000000046
- 457 31. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician Perception of a Machine
458 Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock.
459 *Crit Care Med.* 2019;47(11):1477-1484. doi:10.1097/CCM.0000000000003803
- 460 32. Futoma J, Hariharan S, Heller K, et al. An Improved Multi-Output Gaussian Process
461 RNN with Real-Time Validation for Early Sepsis Detection. In: *Proceedings of the 2nd*
462 *Machine Learning for Healthcare Conference.* PMLR; 2017:243-254. Accessed September
463 26, 2023. <https://proceedings.mlr.press/v68/futoma17a.html>
- 464 33. Moor M, Bennett N, Plečko D, et al. Predicting sepsis using deep learning across
465 international sites: a retrospective development and validation study. *eClinicalMedicine.*
466 2023;62:102124. doi:10.1016/j.eclinm.2023.102124
- 467 34. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine
468 Learning in Healthcare. *Annu Rev Biomed Data Sci.* 2021;4:123-144. doi:10.1146/annurev-
469 biodatasci-092820-114757
- 470 35. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance
471 on a diverse, curated clinical image set. *Sci Adv.* 2022;8(32):eabq6147.
472 doi:10.1126/sciadv.abq6147

473 36. Gregory ME, Russo E, Singh H. Electronic Health Record Alert-Related Workload as a
474 Predictor of Burnout in Primary Care Providers. *Appl Clin Inform.* 2017;08(3):686-697.
475 doi:10.4338/ACI-2017-01-RA-0003

476 37. Borowski M, Görge M, Fried R, Such O, Wrede C, Imhoff M. Medical device alarms.
477 2011;56(2):73-83. doi:10.1515/bmt.2011.005

478

479
480

Tables

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

481
482
483

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

484

485 **Table 2:** Site-specific performance metrics stratified by race including area under the precision

486 recall curve (AU-PR) and area under the receiver operator curve (AUROC)

Cohort	AU-PR	AUROC
ACH ED		
Entire Population	0.252	0.919
White Adults	0.255	0.918
Black Adults	0.238	0.911
SHB ED		
Entire Population	0.248	0.906
White Adults	0.246	0.901
Black Adults	0.286	0.920
ACH Green ED		
Entire Population	0.177	0.960
White Adults	0.175	0.957
Black Adults	0.18	0.983
SHB- Wadsworth ED		
Entire Population	0.216	0.928
White Adults	0.213	0.926
Black Adults	0.325	0.974

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516
517
518

Threshold	ACH Emergency Department				SHB Emergency Department				ACH Green Emergency Department				SHB Wadsworth Emergency Department			
	Alerts/Day (95% CI)	Average lead time (95% CI)	Precision	Recall	Alerts per day (95% CI)	Lead time (hours) (95% CI)	Precision	Recall	Alerts/Day (95% CI)	Average lead time (95% CI)	Precision	Recall	Alerts/Day (95% CI)	Average lead time (95% CI)	Precision	Recall
0.1	147 (144, 148)	7.1 (6.9, 7.31)	0.021	0.986	60 (58, 61)	6.45 (6.16, 6.76)	0.024	0.975	28 (27, 28)	7.77 (6.82, 8.73)	0.009	0.988	15 (15,16)	6.05 (5.24, 6.87)	0.015	0.992
0.2	42 (41, 43)	4.97 (4.81, 5.15)	0.051	0.949	17 (17, 18)	4.69 (4.43, 4.95)	0.047	0.946	5 (4, 5)	6.11 (5.27, 6.97)	0.014	0.980	3 (3, 4)	4.79 (4.03, 5.56)	0.025	0.951
0.3	34 (32, 35)	4.60 (4.44, 4.78)	0.161	0.922	14 (13, 14)	4.31 (4.06, 4.57)	0.149	0.901	4 (3, 4)	5.59 (4.79, 6.41)	0.092	0.959	3 (3, 3)	4.27 (3.54, 5.01)	0.134	0.927
0.4	32 (30, 32)	4.43 (4.27, 4.6)	0.186	0.899	13 (12, 13)	4.13 (3.88, 4.39)	0.176	0.873	4 (3, 4)	5.47 (4.67, 6.29)	0.118	0.951	3 (3, 3)	4.01 (3.32, 4.72)	0.171	0.910
0.5	30 (29, 31)	4.31 (4.15, 4.48)	0.193	0.872	13 (12, 13)	4.04 (3.79, 4.3)	0.183	0.849	3 (3, 4)	5.37 (4.57, 6.18)	0.125	0.934	3 (3, 3)	3.84 (3.16, 4.52)	0.177	0.890
0.6	29 (28, 30)	4.23 (4.07, 4.4)	0.195	0.838	12 (11, 12)	3.94 (3.69, 4.2)	0.186	0.821	3 (3, 4)	5.25 (4.44, 6.07)	0.128	0.910	3 (3, 3)	3.72 (3.06, 4.4)	0.179	0.861
0.7	29 (28, 30)	4.12 (3.96, 4.29)	0.195	0.794	12 (11, 12)	3.85 (3.6, 4.11)	0.184	0.772	3 (3, 4)	5.13 (4.33, 5.94)	0.127	0.857	3 (3, 3)	3.65 (2.98, 4.32)	0.172	0.796
0.8	25 (24, 25)	4.02 (3.85, 4.19)	0.211	0.730	10 (9, 11)	3.73 (3.48, 3.99)	0.199	0.710	3 (2, 3)	4.96 (4.16, 5.76)	0.137	0.775	2 (2, 2)	3.51 (2.84, 4.19)	0.182	0.743
0.9	14 (14, 15)	3.86 (3.69, 4.04)	0.258	0.548	6 (5, 6)	3.62 (3.35, 3.89)	0.255	0.545	2 (2, 2)	4.74 (3.92, 5.57)	0.168	0.578	2 (2, 2)	3.15 (2.47, 3.84)	0.199	0.518
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

519

520 **Table 3:** Site specific model performance including precision, recall, average number of alerts per day and average lead time based on
521 model threshold.

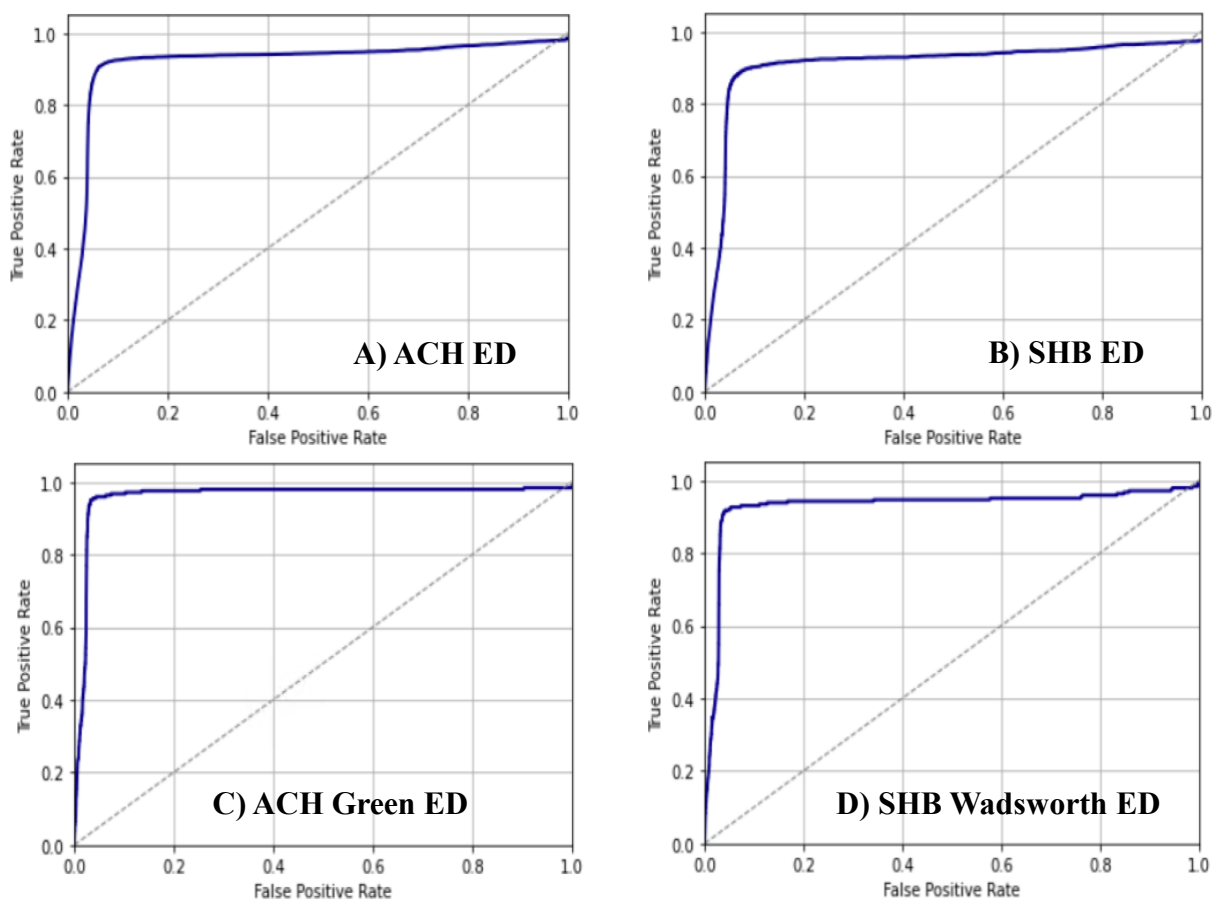


Figure 1: Area under the receiver operator curve at each of Summa Health’s four emergency departments, A) ACH Emergency Department (ED), B) SHB ED, C) ACH Green ED, D) SHB Wadsworth ED

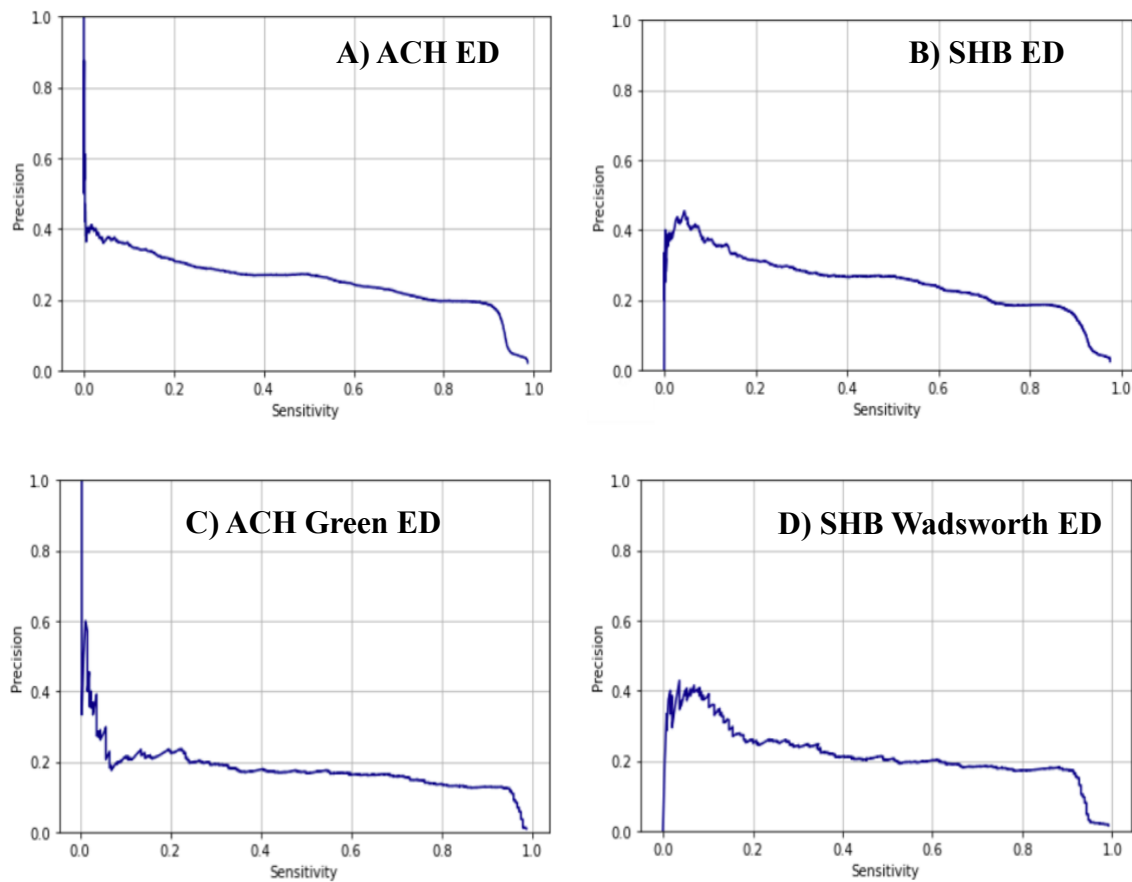


Figure 2: Area under precision-recall curve at each of Summa Health's four emergency departments A) ACH Emergency Department (ED) B) SHB ED C) ACH Green ED D) SHB Wadsworth ED