

Why Health AI Implementations Fail

Karandeep Singh, MD, MMSc

Incoming faculty, Division of Biomedical Informatics

Chief Health AI Officer

Associate CMIO for Inpatient Care

Disclosures

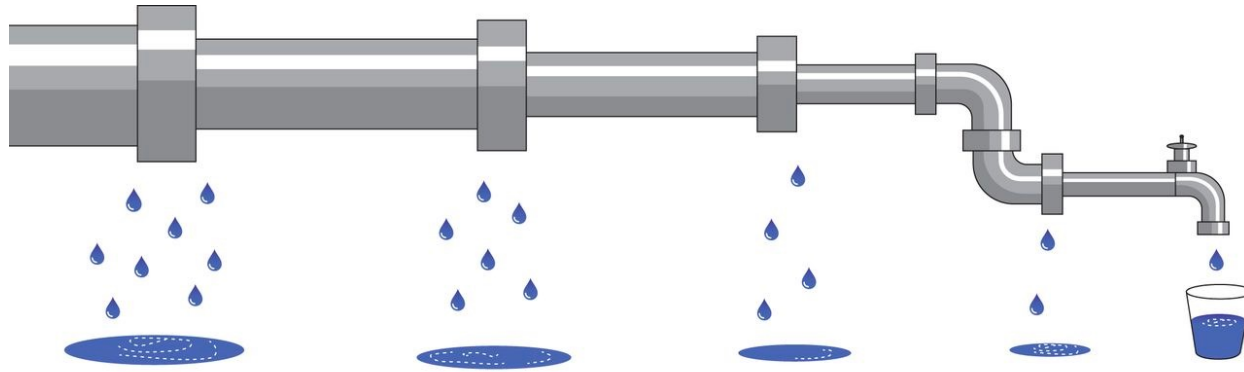
- Current grant funding from NIDDK
- Prior grant funding from Blue Cross Blue Shield of Michigan, and Teva Pharmaceuticals
- Previously served as a consultant and on a scientific advisory board for Flatiron Health

Objectives

- Describe the components of an ideal AI implementation
- Review common issues with **models** that contribute to failure
- Review common issues with **workflows** that contribute to failure
- Discuss strategies to make AI better and safer

The AI Paradox in Health:

Researched AI tools aren't implemented
 Implemented AI tools aren't researched



Not fit for purpose	No validation	No implementation	Not adopted
Developed on wrong patient population	Lack of data or incentive to pursue validation studies	No impact on decision making or patient (health) outcomes	Prediction (perceived as) not useful
Expensive or non-available predictors	Incompletely reported prediction model	No software developed to implement and use the model	Predictions not trusted
Time intensive to use model	Poorly developed or overfitted model	Requirements for adherence to (medical device) regulations	Model not transparent enough, or no tools available to enhance its use in practice
Outcome measured unreliably	Proprietary model code	Cost(-effectiveness) of using proprietary model	Model (perceived as) outdated

Models
coming from
the research
pipeline



Models
coming from
industry

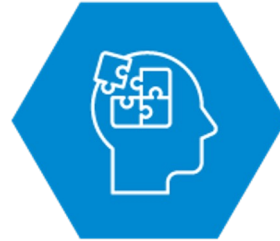
What is an AI implementation?



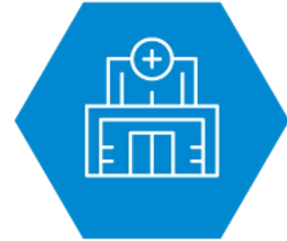
AI
model



...produces a
recommendation...



...that is reviewed
by a person...



...and is
implemented into a
clinical workflow.

What is an AI implementation?



AI
model



...produces a
recommendation...



...and is
implemented into a
clinical workflow...



...to reorganize or
streamline a
person's work.

An ideal AI implementation

- Produces recommendations that are informative and can be acted upon
- Allocates clinician and staff time more effectively
- Allocates more resources towards people in our communities who may benefit from them

Why do AI implementations fail?

- Issues related to **AI models**
 - ...and their recommendations
- Issues related to **workflow**
 - ...and how care is reorganized and reallocated

Why do AI models fail?

- Lack of transparency
- Lack of reproducibility
- Lack of transportability
- Net benefit of model worse than no model
- Predicting non-modifiable risk

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

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

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

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

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

EXPOSURE The ESM score, calculated every 15 minutes.

MAIN OUTCOMES AND MEASURES Sepsis, as defined by a composite of (1) the Centers for

- [← Editorial page 1040](#)
- [+ Multimedia](#)
- [+ Supplemental content](#)
- [+ CME Quiz at \[jamacmelookup.com\]\(https://jamacmelookup.com\) and CME Questions page 1148](#)

What goes into this sepsis model?

eMethods

The Epic Sepsis Model

The Epic Sepsis Model (ESM) is a penalized logistic regression model developed from a pooled sample of 405,000 patient encounters across three health care organizations between 2013 and 2015. Data was collected from the electronic health record in 30 minute observation intervals, up to 24 hours prior to the time of clinical intervention, defined as initiation of antibiotics, documentation of sepsis or suspicion of sepsis, usage of a sepsis-related order set, or an order for a lactate lab. Data elements included vital signs, medication orders, lab values, comorbidities, and demographic information. For model development, sepsis was defined as any encounter associated with an International Classification of Diseases (ICD-9) code indicating diagnosis of sepsis. Time of sepsis onset was defined as 6 hours prior to clinical intervention, with any time point falling prior to 6 hours before the time of clinical intervention labeled as negative for sepsis. Site-specific models were separately trained at each of the three institutions and the model coefficients were averaged to create a final 80-variable model. Model performance for the final model was separately assessed at each site, and the area under the receiver operating characteristic curve (AUC) ranged between 0.76 to 0.83.

Lack of transparency

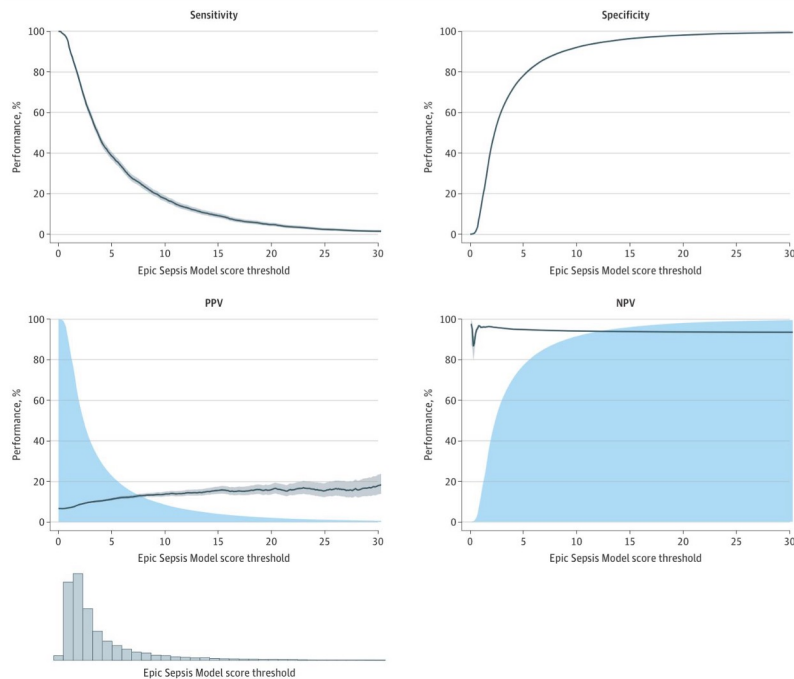
Table 4. Adherence Rates to Entire Reporting Guidelines Across Model Briefs

Epic Systems Corporation model briefs, %											
Model reporting guideline	Deterioration index	Early detection of sepsis	Risk					Risk			
			Unplanned readmission	Patient no-show	Pediatric risk of hospital admission or ED visit	Risk of hospital admission or ED visit	Inpatient risk of falls	Projected block utilization	Remaining length of stay	Admission for heart failure	Hospital admission or ED visit for asthma
Model cards	66	47	63	51	40	69	51	45	50	47	41
Model facts labels	77	71	80	89	71	80	71	71	82	60	63
Guidelines	64	66	66	66	57	74	62	49	70	64	64
MI-CLAIM	55	58	63	58	47	68	53	34	51	53	45
MINIMAR	71	71	79	61	68	86	71	46	67	75	61
TRIPOD	63	63	61	48	42	61	47	36	57	48	44
CONSORT-AI	63	43	63	60	33	67	53	47	47	49	42
SPIRIT-AI	61	55	54	54	38	61	44	49	51	41	39

Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Pfeffer MA, Shah NH. Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: A systematic review. JAMA Netw Open. 2022 Aug 1;5(8):e2227779. PMID: PMC9391954

So how well did the model perform?

Figure 1. Threshold Performance Plots for the Epic Sepsis Model at the Hospitalization Level



The distribution of predictions is displayed at the bottom. NPV indicates negative predictive value; PPV, positive predictive value. In the PPV plot, the blue-shaded region refers to the percentage of patients classified as positive.

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

Predicting the onset of sepsis?

AUC 0.63

Predicting sepsis up to 3 hours *after* its onset?

AUC 0.80

Is this a reproducibility issue?

Our AUC on
our data

0.63

The vendor's AUC on
our data

0.88

Is this a transportability issue?

Research Letter

ONLINE FIRST

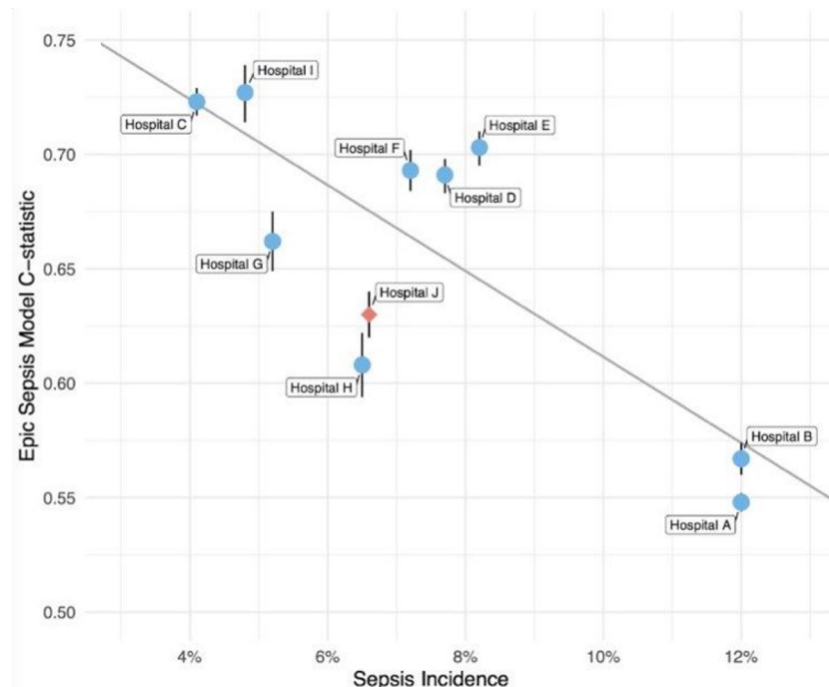
April 3, 2023

Factors Associated With Variability in the Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the US

Patrick G. Lyons, MD, MSc^{1,2}; Mackenzie R. Hofford, MD³; Sean C. Yu, PhD³; Andrew P. Michelson, MD¹; Philip R. O. Payne, PhD³; Catherine L. Hough, MD, MSc⁴; Karandeep Singh, MD, MMSc⁵

» Author Affiliations | Article Information

JAMA Intern Med. Published online April 3, 2023. doi:10.1001/jamainternmed.2022.7182



Is this a transparency issue?


A STAT INVESTIGATION

STAT+

Epic's sepsis algorithm is going off the rails in the real world. The use of these variables may explain why

By Casey Ross  Sept. 27, 2021

Reprints



MIKE REDDY FOR STAT

The illustration depicts a hospital room. In the center is a patient lying in a bed. To the left, a large monitor displays the 'Epic' logo and a patient's profile. To the right of the bed is a red warning sign with a white exclamation mark. Below the bed is a red heart icon. In the bottom left corner, there is a graph showing 'RDC' (Real-World Data) and 'AUC' (Area Under the Curve) with a dashed line representing a baseline and a solid line showing a significant drop. The overall scene is rendered in a stylized, hand-drawn manner with a color palette of purples, blues, and reds.

But STAT has learned it is using a curious piece of data to make its prediction: whether a doctor has already ordered antibiotics.

The use of that information, which has not been publicly disclosed by the company, is contributing to a discrepancy between the accuracy of the algorithm in Epic's internal testing and its performance in the outside world. Those problems came into view after multiple health systems attempted to validate the tool, but found it performed significantly worse than Epic advertised in their own hospitals.

Another example: Predicting acute kidney injury

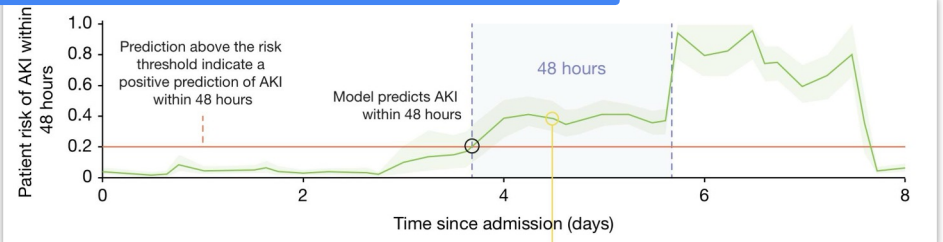
LETTER

<https://doi.org/10.1038/s41586-019-1390-1>

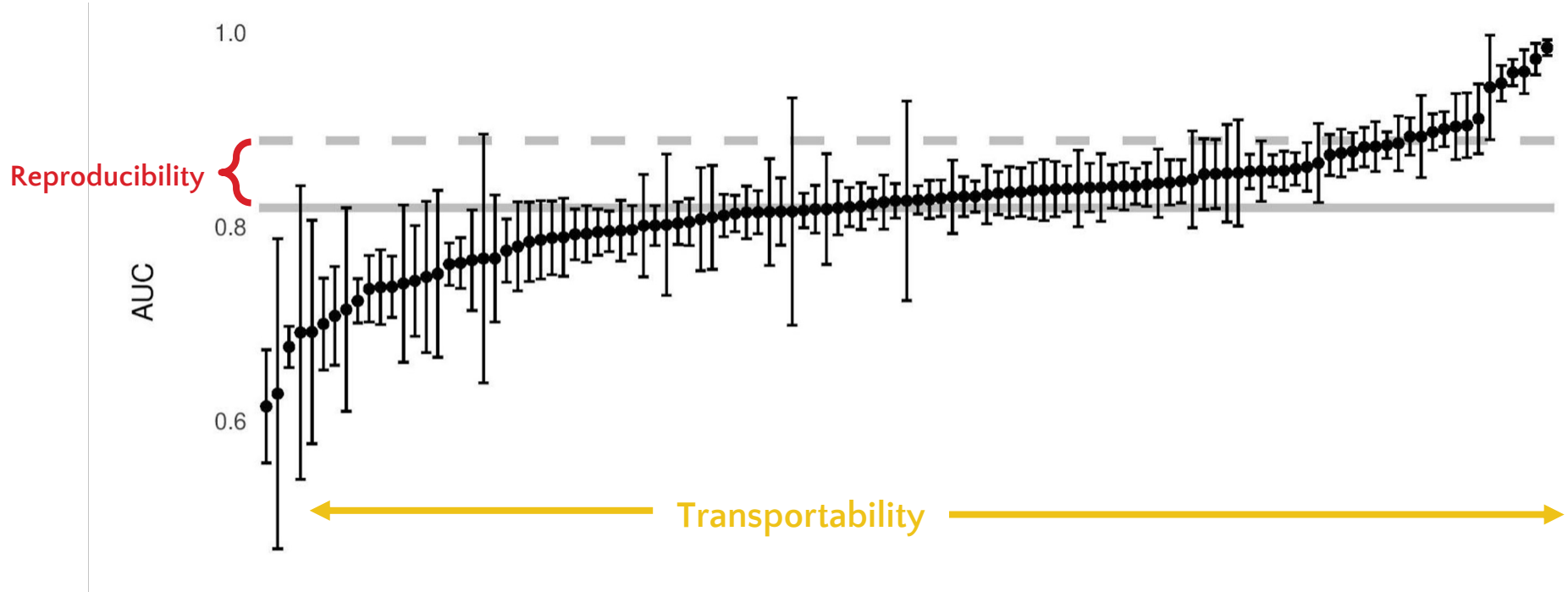
A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev^{1*}, Xavier Glorot¹, Jack W. Rae^{1,2}, Michał Zieliński¹, Harry Askham¹, Andre Saraiva¹, Anne Mottram¹, Clemens Meyer³, Suman Ravuri¹, Ivan Protsyuk¹, Alistair Connell¹, Cian O. Hughes³, Alan Karthikesalingam¹, Julien Cornebise^{1,12}, Hugh Montgomery³, Geraint Rees⁴, Chris Laing⁵, Clifton R. Baker⁶, Kelly Peterson^{7,8}, Ruth Reeves⁹, Demis Hassabis¹, Dominic King⁴, Mustafa Suleyman¹, Trevor Back^{1,13}, Christopher Nielson^{10,11,13}, Joseph R. Ledsam^{1,13*} & Shakir Mohamed^{1,13}

Despite the state-of-the-art retrospective performance of our model compared to existing literature, future work should now prospectively evaluate and independently validate the proposed model to establish its clinical utility and effect on patient outcomes, as well as explore the role of the model in researching strategies for delivering preventative care for AKI.




Reproducibility and Transportability



Cao J, Zhang X, Shahinian V, Yin H, Steffick D, Saran R, Crowley S, Mathis M, Nadkarni GN, Heung M, Singh K. Generalizability of an acute kidney injury prediction model across health systems. *Nat Mach Intell*. Springer Science and Business Media LLC; 2022 Dec 1;4(12):1121–1129.

Net benefit: Can a model be worse than no model?

$$NB = \frac{TP}{N} - \frac{FP}{N} \times \frac{Pt}{1 - Pt}$$


Benefit conferred by intervening on true positives

Harm conferred by intervening on false positives

Scaling of harm based on cost/benefit ratio

Figure 1. Formula for calculating net benefit (*NB*) based on the number of true positives (*TP*), false positives (*FP*), sample size (*N*), and the threshold probability (*Pt*). Source: Karandeep Singh.

Predicting non-modifiable risk

If trying to reduce readmissions, should we use AI to predict...

- ...the risk of readmission?
- ...the chance that readmission can be avoided with an intervention?

Why do AI interventions fail?

- Lack of efficacy of the intervention
- Wrong end-users
- Increased workload
- Culturally not ready for change
- Resource constraints

Lack of efficacy: Good model ≠ effective intervention

[AMIA Annu Symp Proc](#). 2018; 2018: 295–304.
Published online 2018 Dec 5.

PMCID: PMC6371247
PMID: [30815068](#)

Towards a Learning Health System to Reduce Emergency Department Visits at a Population Level

[Elliott Brannon](#), MPH, ¹ [Tianshi Wang](#), ² [Jeremy Lapedis](#), DrPH, MHPS, ³ [Paul Valenstein](#), MD, ⁴ [Michael Klinkman](#), MD, MS, ⁵ [Ellen Bunting](#), MA, ⁶ [Alice Stanulis](#), ⁶ and [Karandeep Singh](#), MD, MMSc ^{1, 2, 7}

▶ Author information ▶ Copyright and License information [Disclaimer](#)

Abstract

Go to: 

High utilizers of the Emergency Department (ED) often have complex needs that require coordination of care between multiple organizations. We describe a Learning Health Systems (LHS) approach to reducing ED visits, in which an intervention is delivered to a cohort of high utilizers identified using population-level data and predictive modeling. We focus on the development and validation of a random forest model that utilizes electronic health record data from three health systems across two counties in Michigan to predict the number of ED visits each resident will incur in the next six months. Using 5-fold cross-validation, the model achieves a root-mean-squared-error of 0.51 visits and a mean absolute error of 0.24 visits. Using time-based validation, the model achieves a root-mean-squared error of 0.74 visits and a mean absolute error of 0.29 visits. Patients projected to have high ED utilization are being enrolled in a community-wide care coordination intervention using twelve sites across two counties. We believe that the repeated cycles of modeling and intervention demonstrate an LHS in action.

Original Research | Published: 10 March 2021

Predictive Model-Driven Hotspotting to Decrease Emergency Department Visits: a Randomized Controlled Trial

[Brady Post](#) PhD, [Jeremy Lapedis](#) DrPH, [Karandeep Singh](#) MD, [Paul Valenstein](#) MD, [Ayşe G. Büyüktür](#) PhD, [Karin Teske](#) MPH & [Andrew M. Ryan](#) PhD 

[Journal of General Internal Medicine](#) (2021) | [Cite this article](#)

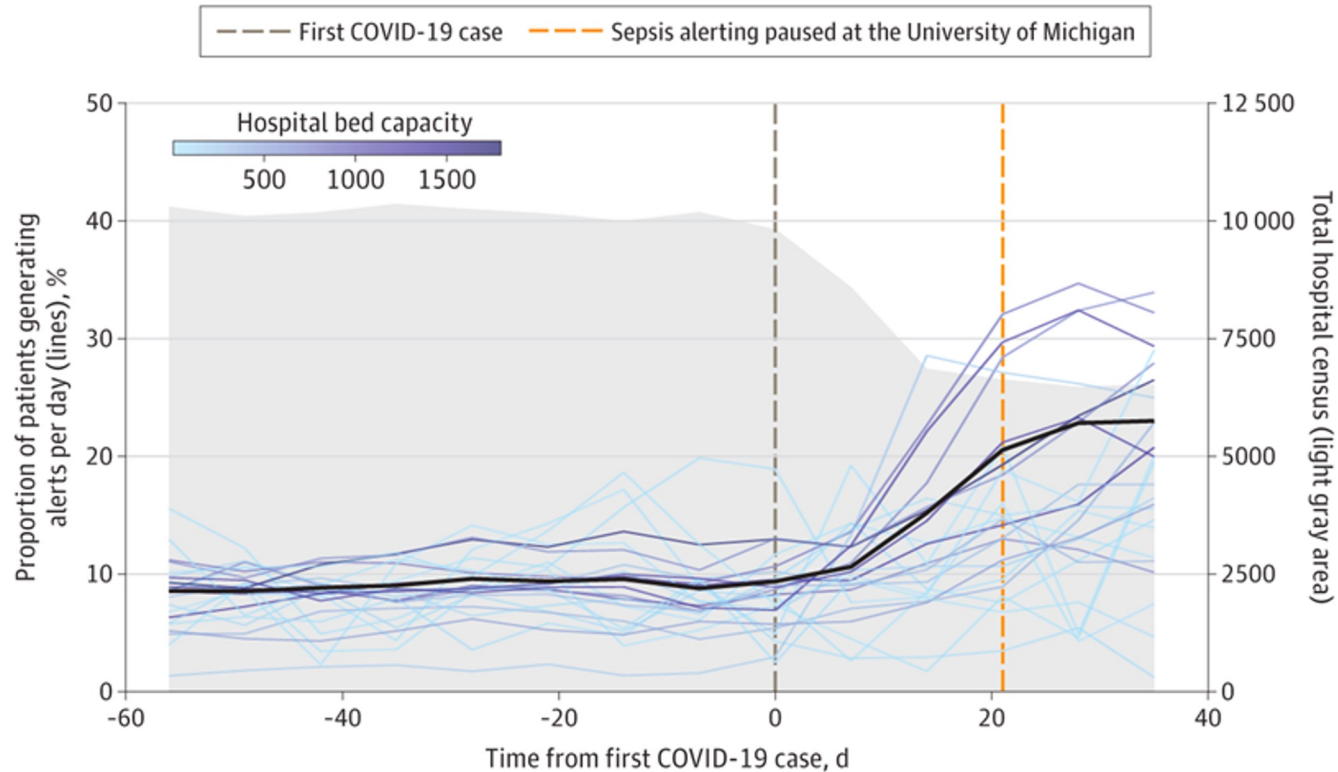
Conclusions

The community case management intervention targeting ED visits was not associated with reduced utilization. Future case management interventions may benefit from additional patient engagement strategies and longer evaluation time periods.

Trial Registration

[Clinicaltrials.gov](#) Identifier: NCT03293160.

Workload



Wong A, Cao J, Lyons PG, Dutta S, Major VJ, Ötles E, Singh K. Quantification of Sepsis Model Alerts in 24 US Hospitals Before and During the COVID-19 Pandemic. JAMA Netw Open. 2021 Nov 1;4(11):e2135286. doi: 10.1001/jamanetworkopen.2021.35286. PMID: 34797372; PMCID: PMC8605481.

Readiness for change

Current state

- Nurses screen **all** patients for falls in the hospital
- We would like to use AI to prevent falls and reduce workload

Future state

- If we could accurately predict falls using an AI model...
- ...are we willing to **not** screen low-risk patients?

Resource constraints

Patient #	Actually required ICU	Based on model		
		Predicted risk (sorted from high to low)	Predicted to require ICU (based on $Pt = 0.2$)	Predicted to require ICU (based on $Pt = 0.2$ and capacity = 3)
1	Yes	0.8	Yes	Yes
2	Yes	0.7	Yes	Yes
3	No	0.6	Yes	Yes
4	Yes	0.5	Yes	No
5	Yes	0.3	Yes	No
6	No	0.2	Yes	No
7	No	0.1	No	No
8	No	0.1	No	No
9	Yes	0.05	No	No
10	No	0.01	No	No

How do we make AI better and safer?



Clinical Intelligence Committee

★ Following Share

Home

Documents

Request Form

Model Requests

Model Cards

Team Calendar

Discussion

Tasks

Team Members

Recycle bin

Edit

[Return to classic SharePoint](#)

+ New

Edit in grid view

Share

Export

Automate

...

Model Summary View



Model Name
Early Detection of Sepsis (Epic Adult S...

Model Description and Purpose
Predicts early onset sepsis.

Target Audience
Physician, nurse and provider point of conta...

Clinical Owner
UMHS Sepsis Committee

Clinical Point of Contact
Jessie King

Flowsheet ID
304011076

Model Name
M-CURES

Model Description and Purpose
Predicts the risk of deterioration in hospitaliz ed patients.

Target Audience
Rapid Response Teams (RRT)

Clinical Owner
Formally Rapid Response Team

Clinical Point of Contact
Formally Rapid Response Team

Flowsheet ID
3040321254

Model Name
PICTURE-General - Adult

Model Description and Purpose
PICTURE is a suite of machine learning algorit hms that utilizes electronic health record (EH

Target Audience
Rapid Response Team

Clinical Owner
Rapid Response Team

Clinical Point of Contact
Rapid Response Team

Flowsheet ID
3040321255

Model Name
Epic Deterioration Index

Model Description and Purpose
Early warning system that predicts patient de terioration.

Target Audience
Rapid Response Team. Physician, nurse and...

Clinical Owner
Rapid Response Team

Model Name
PICTURE- General - Pediatric

Model Description and Purpose
PICTURE is a suite of machine learning algorit hms that utilizes electronic health record (EH

Target Audience
Pediatric nurses, clinicians and providers

Clinical Owner
-

Model Name
Epic Inpatient Fall Risk Model

Model Description and Purpose
Predicts the risk of fall. Based on various scor ing systems to "bucket" risk levels (MORSE a

Target Audience
All clinicians

Clinical Owner
Falls Committee

EXCLUSIVE

Epic overhauls popular sepsis algorithm criticized for faulty alarms



By [Casey Ross](#) Oct. 3, 2022



ADOBE

STAT+

Reprints

Epic overhauls sepsis model amid scrutiny

[The Wisconsin health record giant](#) is revamping its flagship sepsis prediction model following

Casey's [investigations](#) revealing high rates of false alarms and failure to reliably predict sepsis, according to documents obtained by STAT. In a major policy shift, **Epic** is now recommending that hospital customers train the model on their own data before deploying it, and has adjusted its definition of “sepsis onset” to align with a more commonly accepted standard.

Federal regulation

HEALTH TECH

STAT+

In new guidance, FDA says AI tools to warn of sepsis should be regulated as devices



By [Casey Ross](#) Sept. 27, 2022

[Reprints](#)



ALEX HOGAN/STAT

Contains Nonbinding Recommendations

Clinical Decision Support Software Guidance for Industry and Food and Drug Administration Staff

Document issued on September 28, 2022.

The draft of this document was issued on September 27, 2019.

For questions about this document regarding CDRH-regulated devices, contact the Division of Digital Health via email at DigitalHealth@fda.hhs.gov. For questions about this document regarding CBER-regulated devices, contact the Office of Communication, Outreach, and Development (OCOD) at 1-800-835-4709 or 240-402-8010, or by email at ocod@fda.hhs.gov. For questions about this document regarding CDER-regulated products, contact Center for Drug Evaluation and Research, Food and Drug Administration, 10903 New Hampshire Ave., Bldg. 51, Rm. 6158, Silver Spring, MD 20993-0002, 301-796-8936. For questions about this document regarding combination products, contact the Office of Combination Products at combination@fda.gov.



DEPART

This document is scheduled to be published in the Federal Register on 04/18/2023 and available online at [federalregister.gov/d/2023-07229](https://www.federalregister.gov/d/2023-07229), and on [govinfo.gov](https://www.govinfo.gov)

Office of the Secretary

45 CFR Parts 170, 171

RIN: 0955-AA03

Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing

AGENCY: Office of the National Coordinator for Health Information Technology (ONC),
Department of Health and Human Services (HHS).








ACTION: Proposed rule.

Algorithmic Transparency:

Being Clear with the Public

This letter was written by an interdisciplinary group at the University of Michigan and includes co-signatories from other institutions. It was developed and informed by research conducted by those involved in building predictive decision support interventions (DSIs), those studying public attitudes about data and artificial intelligence in clinical decision support systems, and those in current practice at academic medical centers. The views expressed in this letter represent those of the signatories and not of their institutions.

BMJ Open Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence

Gary S Collins ,^{1,2} Paula Dhiman ,^{1,2} Constanza L Andaur Navarro ,³
Jie Ma ,¹ Lotty Hooft,^{3,4} Johannes B Reitsma,³ Patricia Logullo ,^{1,2}
Andrew L Beam ,^{5,6} Lily Peng,⁷ Ben Van Calster ,^{8,9,10}
Maarten van Smeden ,³ Richard D Riley ,¹¹ Karel GM Moons^{3,4}

Partnerships and Education



Thank you

Karandeep Singh, MD, MMSc
karandeep@ucsd.edu