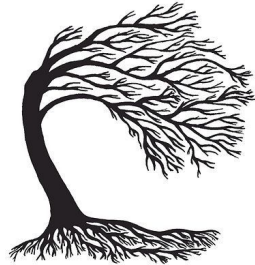


# Reliability and Fairness Audits of Clinical AI Models using STARR-OMOP



**Jonathan Lu**, Amelia Sattler, Samantha Wang, Ali Raza Khaki, Alison Callahan, Scott Fleming, Rebecca Fong, Benjamin Ehlert, Ron C. Li, Lisa Shieh, Kavitha Ramchandran, Michael F. Gensheimer, Sarah Chobot, Stephen Pfohl, Siyun Li, Kenny Shum, Nitin Parikh, Priya Desai, Briththa Seevaratnam, Melanie Hanson, Margaret Smith, Yizhe Xu, Arjun Gokhale, Winifred Teuteberg, Nigam H. Shah

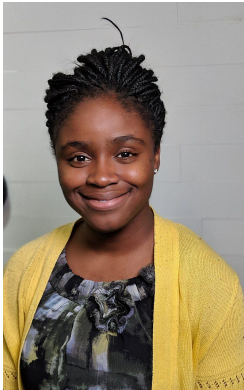
Contact: [jhlu@stanford.edu](mailto:jhlu@stanford.edu)

***“We propose **model cards** as a step towards the responsible democratization of machine learning and related artificial intelligence technology, **increasing transparency into how well artificial intelligence technology works.**”***

- Margaret Mitchell, ..., Timnit Gebru, et al. 2019

[Model Cards for Model Reporting](#)

They have been leading voices for fairness in AI, and were unjustly fired by Google in 2019 for raising concerns about harms of AI, including environmental/financial harms and harms toward Black people and women.



***“Audits are evaluations with an expectation for accountability.”***

- Inioluwa Deborah Raji, 2022

[It's Time to Develop the Tools We Need to Hold Algorithms Accountable](#)

# Introduction

- Deployed AI models in healthcare systems have been found to be unreliable and unfair

FASTCOMPANY

05-28-21

## How a largely untested AI algorithm crept into hundreds of hospitals

During the pandemic, the electronic health record giant Epic quickly rolled out an algorithm to help doctors decide which patients needed the most immediate care. Doctors believe it will change how they practice.

[Khetpal 2021](#)

## Dissecting racial bias in an algorithm used to manage the health of populations

 Ziad Obermeyer<sup>1,2,\*</sup>, Brian Powers<sup>3</sup>, Christine Vogeli<sup>4</sup>,  Sendhil Mullainathan<sup>5,\*†</sup>

[Obermeyer 2019](#)

# Introduction

- 15 Model Reporting Guidelines published since 2012 (!)

<b>Model Facts</b>	<b>Model name:</b> Deep Sepsis	<b>Locale:</b> Duke University Hospital				
<b>Approval Date:</b> 09/22/2019	<b>Last Update:</b> 01/13/2020	<b>Version:</b> 1.0				
<b>Summary</b>						
This model uses EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.						
<b>Mechanism</b>						
<ul style="list-style-type: none"> <li>▪ <b>Outcome</b> .....sepsis within the next 4 hours, see outcome definition in "Other Information"</li> <li>▪ <b>Output</b> .....0% - 100% probability of sepsis occurring in the next 4 hours</li> <li>▪ <b>Target population</b> .....all adult patients &gt;18 y.o. presenting to DUH ED</li> <li>▪ <b>Time of prediction</b> .....every hour of a patient's encounter</li> <li>▪ <b>Input data source</b> .....electronic health record (EHR)</li> <li>▪ <b>Input data type</b> .....demographics, analytes, vitals, medication administrations</li> <li>▪ <b>Training data location and time-period</b> .....DUH, diagnostic cohort, 10/2014 – 12/2015</li> <li>▪ <b>Model type</b>..... Recurrent Neural Network</li> </ul>						
<b>Validation and performance</b>						
	Prevalence	AUC	PPV @ Sensitivity of 60%	Sensitivity @ PPV of 20%	Cohort Type	Cohort URL / DOI
Local Retrospective	18.9%	0.88	0.14	0.50	Diagnostic	<a href="https://arxiv.org/abs/1708.05894">arxiv.org/abs/1708.05894</a>
Local Temporal	6.4%	0.94	0.20	0.66	Diagnostic	<a href="https://jmir.org/preprint/15182">jmir.org/preprint/15182</a>
Local Prospective	TBD	TBD	TBD	TBD	TBD	TBD
External	TBD	TBD	TBD	TBD	TBD	TBD
Target Population	6.4%	0.94	0.20	0.66	Diagnostic	<a href="https://jmir.org/preprint/15182">jmir.org/preprint/15182</a>

[Sendak 2020](#)

# Introduction

- 15 Model Reporting Guidelines published since 2012 (!)
  - Only 1 completed for a model in use for a health system
- We assessed if commonly used Epic models adhere to the guidelines

<b>Model Facts</b>	<b>Model name:</b> Deep Sepsis	<b>Locale:</b> Duke University Hospital				
<b>Approval Date:</b> 09/22/2019	<b>Last Update:</b> 01/13/2020	<b>Version:</b> 1.0				
<b>Summary</b> This model uses EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.						
<b>Mechanism</b> <ul style="list-style-type: none"><li>▪ <b>Outcome</b> .....sepsis within the next 4 hours, see outcome definition in "Other Information"</li><li>▪ <b>Output</b> .....0% - 100% probability of sepsis occurring in the next 4 hours</li><li>▪ <b>Target population</b> .....all adult patients &gt;18 y.o. presenting to DUH ED</li><li>▪ <b>Time of prediction</b> .....every hour of a patient's encounter</li><li>▪ <b>Input data source</b> .....electronic health record (EHR)</li><li>▪ <b>Input data type</b> .....demographics, analytes, vitals, medication administrations</li><li>▪ <b>Training data location and time-period</b> .....DUH, diagnostic cohort, 10/2014 – 12/2015</li><li>▪ <b>Model type</b> ..... Recurrent Neural Network</li></ul>						
<b>Validation and performance</b>						
	<b>Prevalence</b>	<b>AUC</b>	<b>PPV @ Sensitivity of 60%</b>	<b>Sensitivity @ PPV of 20%</b>	<b>Cohort Type</b>	<b>Cohort URL / DOI</b>
<b>Local Retrospective</b>	18.9%	0.88	0.14	0.50	Diagnostic	<a href="https://arxiv.org/abs/1708.05894">arxiv.org/abs/1708.05894</a>
<b>Local Temporal</b>	6.4%	0.94	0.20	0.66	Diagnostic	<a href="https://jmir.org/preprint/15182">jmir.org/preprint/15182</a>
<b>Local Prospective</b>	TBD	TBD	TBD	TBD	TBD	TBD
<b>External</b>	TBD	TBD	TBD	TBD	TBD	TBD
<b>Target Population</b>	6.4%	0.94	0.20	0.66	Diagnostic	<a href="https://jmir.org/preprint/15182">jmir.org/preprint/15182</a>



[Sendak 2020](#)

# Introduction

- Low reporting of items related to:
  - Reliability
    - External Validation (33%)
    - Confidence Intervals (0%)
    - Calibration Plots (0%)
  - Fairness
    - Summary Statistics: Sex (33%), Ethnicity/Race (33%)
    - Subgroup Analyses (33%)
- **How hard is it to report these for a model in use?**



# Models

How hard is it to do a reliability and fairness audit for Advance Care Planning models?

<b>Epic End-of-life Index</b>	<b>Stanford Hospital Medicine ACP Model</b>
Input: 46 features (age, sex, insurance status, comorbidities, medications)	Input: 13189 features (age, sex, lab orders, procedure orders)
12-month mortality	12-month mortality
Logistic regression	Gradient Boosted Tree
All patients within health system	Hospitalized patients

# Study Design

## 1. Solicit Clinician Labels

Would you be surprised if the last patient you saw in clinic passed away in the next 2 years?



*Expect a treat of appreciation! :)*

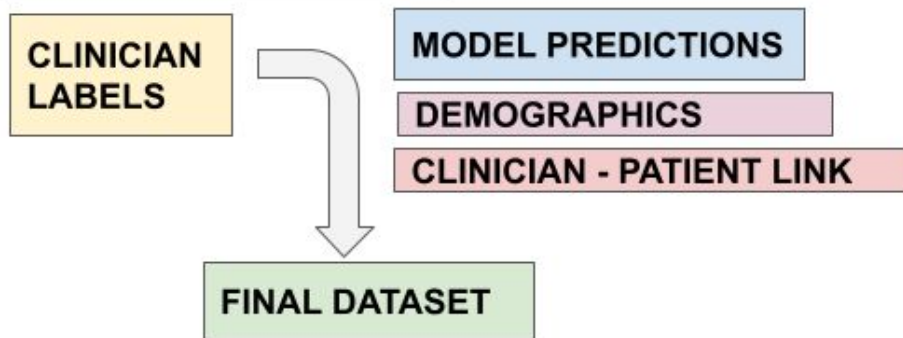
Please help us by answering this question for **5** of your patients

*Request coming soon via Epic staff message*

Thank you!

Dr. Amelia Sattler and Winnie Teuteberg and the Serious Illness Conversation Program (SICP)

## 2. Link to Model Predictions, Demographics, Clinician-Patient Link



## 3. Perform Reliability and Fairness Audit

## 4. Survey Decisionmakers, Assess Time and Requirements



# ***Fairness Audit: Epic EOL High Threshold in Inpatient Oncology***

## **A. SUMMARY STATISTICS**

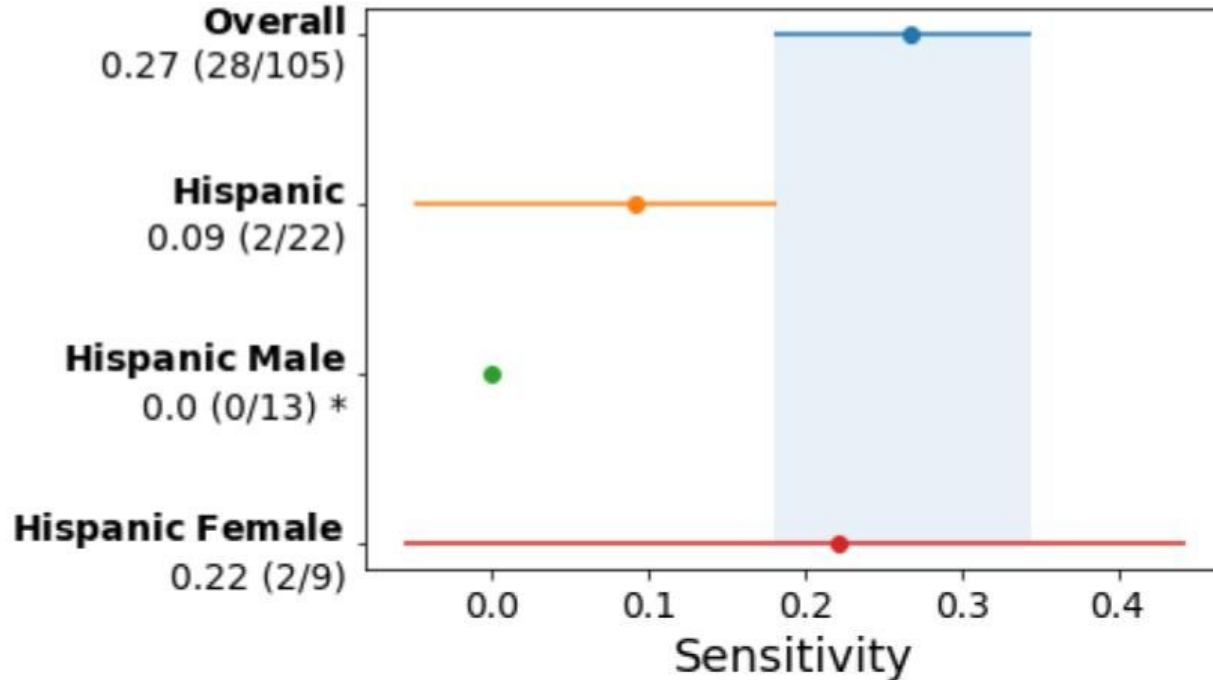
No significant differences in prevalence of positive label for Hispanic patients.

	<b># Patients</b>	<b>Positive label prevalence (fraction)</b>	<b>95% CI for prevalence</b>
<b>Overall</b>	150	0.7 (105/150)	0.62-0.77
<b>Hispanic</b>	30	0.73 (22/30)	0.54-0.88
<b>Hispanic Male</b>	17	0.76 (13/17)	0.5-0.93
<b>Hispanic Female</b>	13	0.69 (9/13)	0.39-0.91

# Fairness Audit: Epic EOL High Threshold in Inpatient Oncology

## B. SUBGROUP PERFORMANCE

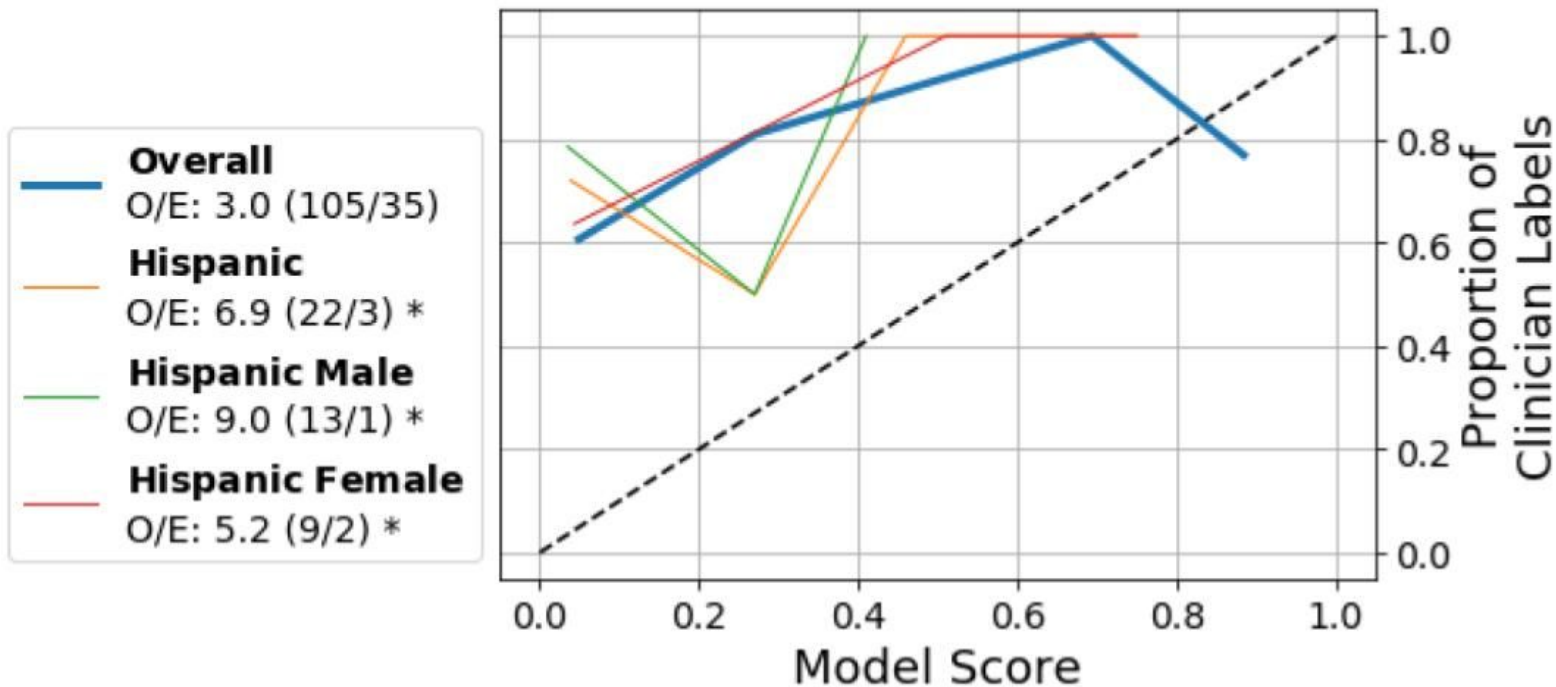
Sensitivity is significantly lower for Hispanic male patients.



# Fairness Audit: Epic EOL High Threshold in Inpatient Oncology

## C. SUBGROUP CALIBRATION

Significant underprediction of events for Hispanic patients.



# Caveat: all results are probably wrong



## Validating self-identified race/ethnicity at an academic family medicine clinic



Randy Nhan, BS; Samantha Lane, Lupe Barragan, Jeremy Valencia, A Sattler, MD; K Taylor, MD

Stanford Family Medicine, Stanford University School of Medicine

### Introduction

Healthy disparities based on race/ethnicity are rampant throughout the United States that affect specific groups in attaining proper care.

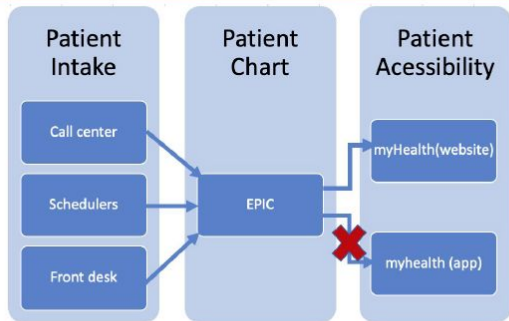
In order to address these disparities, a critical step is having accurate race/ethnicity data to understand and act on those disparities. Prior studies suggest missing data is a significant problem to accurately report and understand health disparities.

We sought to understand the accuracy of race/ethnicity data collection in an academic family medicine clinic as a step toward addressing race/ethnicity disparities.

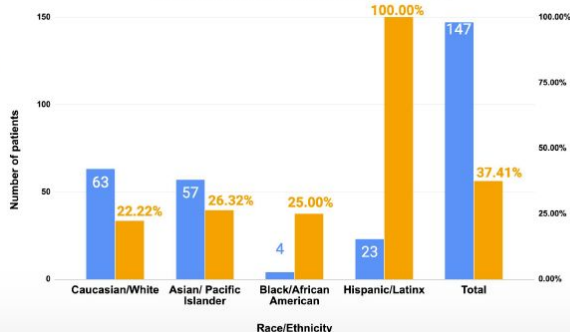
### Methods

To test for the validity of the data, our team worked with a PCC (primary care coordinator) and front desk staff to survey individual patients over a total of 3 weeks who had either a video or in-person visit. They were all asked to self-identify their race/ethnicity. We tallied the number and type of mismatch between self-report versus EMR recorded data.

### Results



Percent mismatched of Race/Ethnicity



### Conclusion & Discussion

Patients were misclassified almost 37% of the time in the EMR. The most common misclassification was “other” and Hispanics were most likely to be misclassified.

Ongoing assessment of the process of race/ethnicity data collection is underway to improve data collection.

Several future interventions includes looking into accessibility to self-update demographics via the myHealth application, addressing lack of training regarding health disparities among staff and expanding the limited choices for race/ethnicity in patient charts.

### References

- 1.Klinger, E. V., Carlini, S. V., Gonzalez, I., Hubert, S. S., Linder, J. A., Rigotti, N. A., Kontos, E. Z., Park, E. R., Marinacci, L. X., & Haas, J. S. (2015). Accuracy of race, ethnicity, and language preference in an electronic health record. *Journal of general internal medicine, 30*(6), 719–723. <https://doi.org/10.1007/s11606-014-3102-8>
2. Hamilton, N. S., Edelman, D., Weinberger, M., & Jackson, G. L. (2009). Concordance between self-reported race/ethnicity and that recorded in a Veteran Affairs electronic medical record. *North Carolina medical journal, 70*(4), 296–300.
3. Wilson, G., Hasnain-Wynia, R., Hauser, D., & Calman, N. (2013). Implementing Institute of Medicine recommendations on collection of patient race, ethnicity, and language data in a community health center. *Journal of health care for the poor and underserved, 24*(2), 875–884. <https://doi.org/10.1553/jhpu.2013.0071>

- Does EMR = self-identified race/ethnicity at Stanford Family Medicine? (Nhan 2021)
  - **100%** Misclassification Rate for Hispanic/Latinx patients (23)
  - **37%** Misclassification rate overall (147)
- Similarly findings in Optum, Healthcare Cost Utilization Project (Polubriaginof 2021)

# What is required to perform a reliability/fairness audit of a model?

- Clinician Labels
- Model Predictions
- Linking Clinicians with Relevant Patients
- [Fairness] Patient Demographics

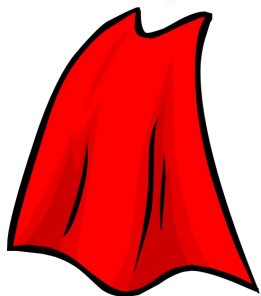
# What is required to perform a reliability/fairness audit of a model?

- Clinician Labels → Relationships with Clinicians
- Model Predictions → Model Access
- Linking Clinicians with Relevant Patients → **Visits, Patient Panels**
- [Fairness] Patient Demographics → **Person Table (STARR-OMOP)**



**Stanford**  
MEDICINE

Observational Medical Outcomes Partnership  
*STAnford Research data Repository*



Thank you to Nitin Parikh, Priya Desai, and Research IT!

