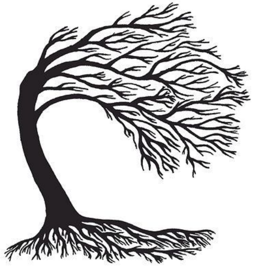
Stanford MEDICINE

MedScholars Research Fellowship

BMIR
Stanford Center for Biomedical Informatics Research

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

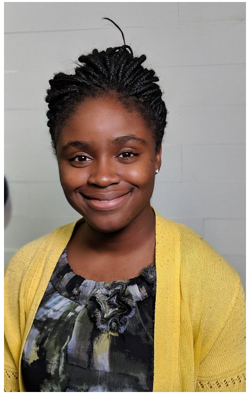
Contact: jhlu@stanford.edu

*"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

FAST COMPANY

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

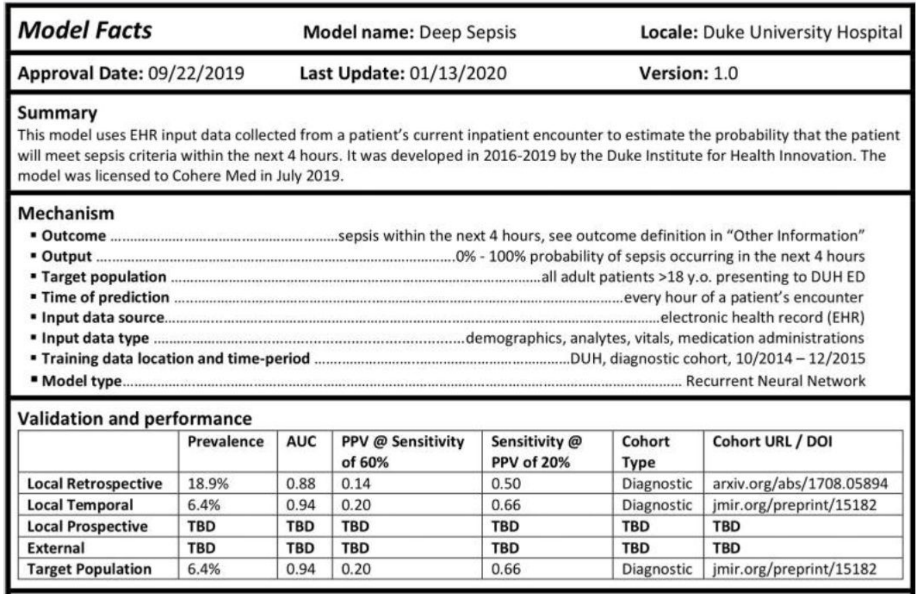
Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

Obermeyer 2019

Introduction

- 15 Model Reporting Guidelines published since 2012 (!)



Model Facts **Model name:** Deep Sepsis **Locale:** Duke University Hospital

Approval Date: 09/22/2019 **Last Update:** 01/13/2020 **Version:** 1.0

Summary
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.

Mechanism
- **Outcome** ....sepsis within the next 4 hours, see outcome definition in "Other Information"
- **Output** ....0% - 100% probability of sepsis occurring in the next 4 hours
- **Target population** ....all adult patients >18 y.o. presenting to DUH ED
- **Time of prediction** ....every hour of a patient's encounter
- **Input data source**........electronic health record (EHR)
- **Input data type** ....demographics, analytes, vitals, medication administrations
- **Training data location and time-period** ....DUH, diagnostic cohort, 10/2014 – 12/2015
- **Model type**........ Recurrent Neural Network

Validation and performance

	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	arxiv.org/abs/1708.05894
Local Temporal	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182
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	jmir.org/preprint/15182

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

Model Facts **Model name:** Deep Sepsis **Locale:** Duke University Hospital

Approval Date: 09/22/2019 **Last Update:** 01/13/2020 **Version:** 1.0

Summary

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.

Mechanism

- **Outcome** .... sepsis within the next 4 hours, see outcome definition in "Other Information"
- **Output** .... 0% - 100% probability of sepsis occurring in the next 4 hours
- **Target population** .... all adult patients >18 y.o. presenting to DUH ED
- **Time of prediction** .... every hour of a patient's encounter
- **Input data source** .... electronic health record (EHR)
- **Input data type** .... demographics, analytes, vitals, medication administrations
- **Training data location and time-period** .... DUH, diagnostic cohort, 10/2014 – 12/2015
- **Model type** .... Recurrent Neural Network

Validation and performance

	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	arxiv.org/abs/1708.05894
Local Temporal	6.4%	0.94	0.20	0.66	Diagnostic	jmir.org/preprint/15182
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	jmir.org/preprint/15182

Epic

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?**

Lu 2021 (in review)

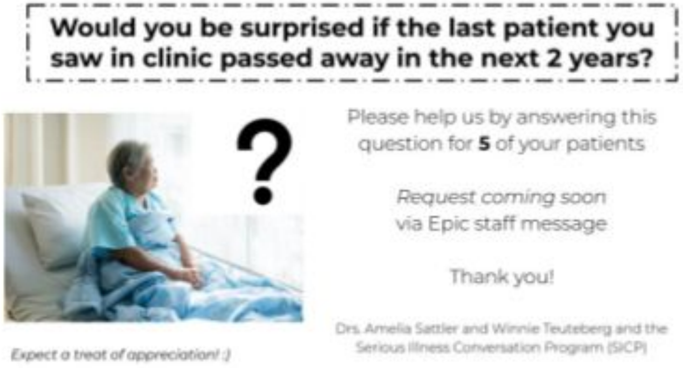
Models

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

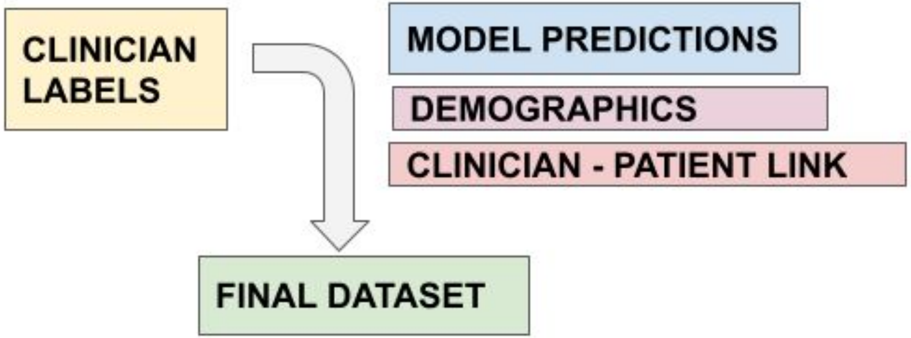
Epic End-of-life Index	**Stanford Hospital Medicine ACP Model**
Input: 46 features (age, sex, insurance status, comorbidities, medications)	Input: 13189 features (age, sex, lab orders, procedure orders)
12-month mortality	12-month mortality
Logistic regression	Gradient Boosted Tree
All patients within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

Study Design

1. Solicit Clinician Labels



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

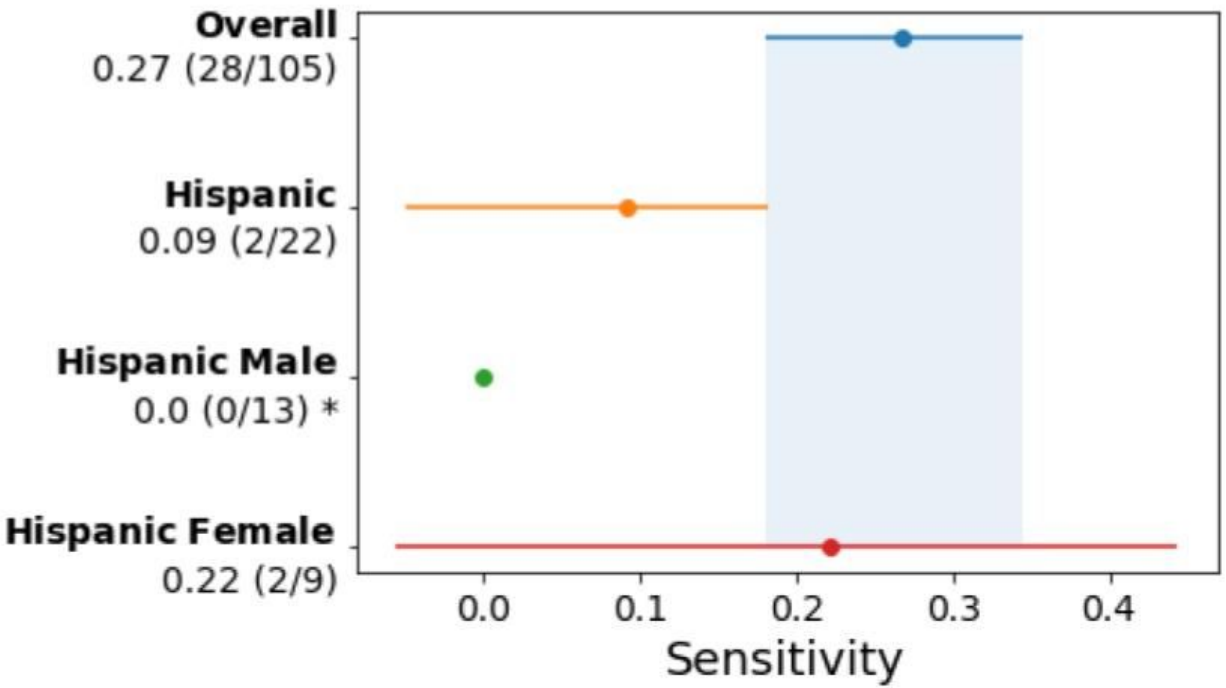
No significant differences in prevalence of positive label for Hispanic patients.

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

Fairness Audit: Epic EOL High Threshold in Inpatient Oncology

B. SUBGROUP PERFORMANCE

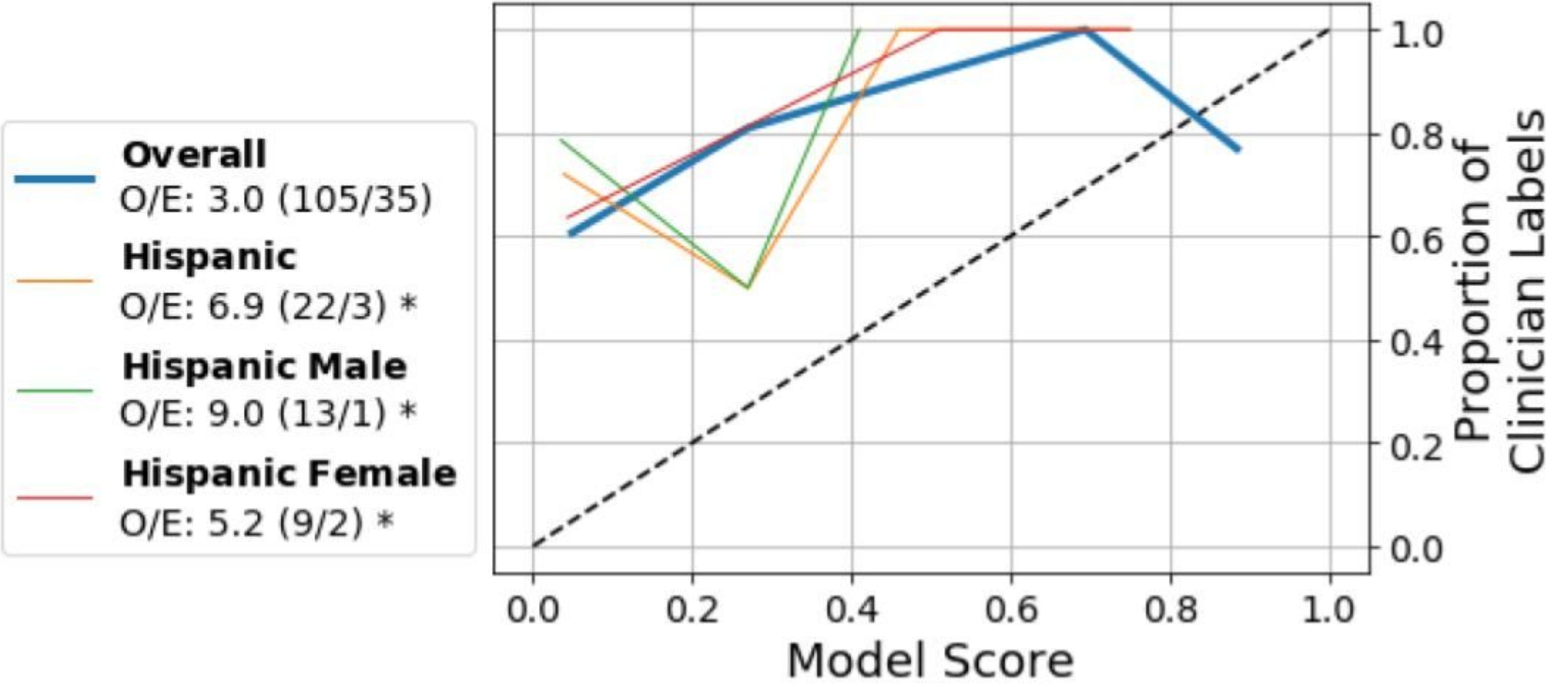
Sensitivity is significantly lower for Hispanic male patients.



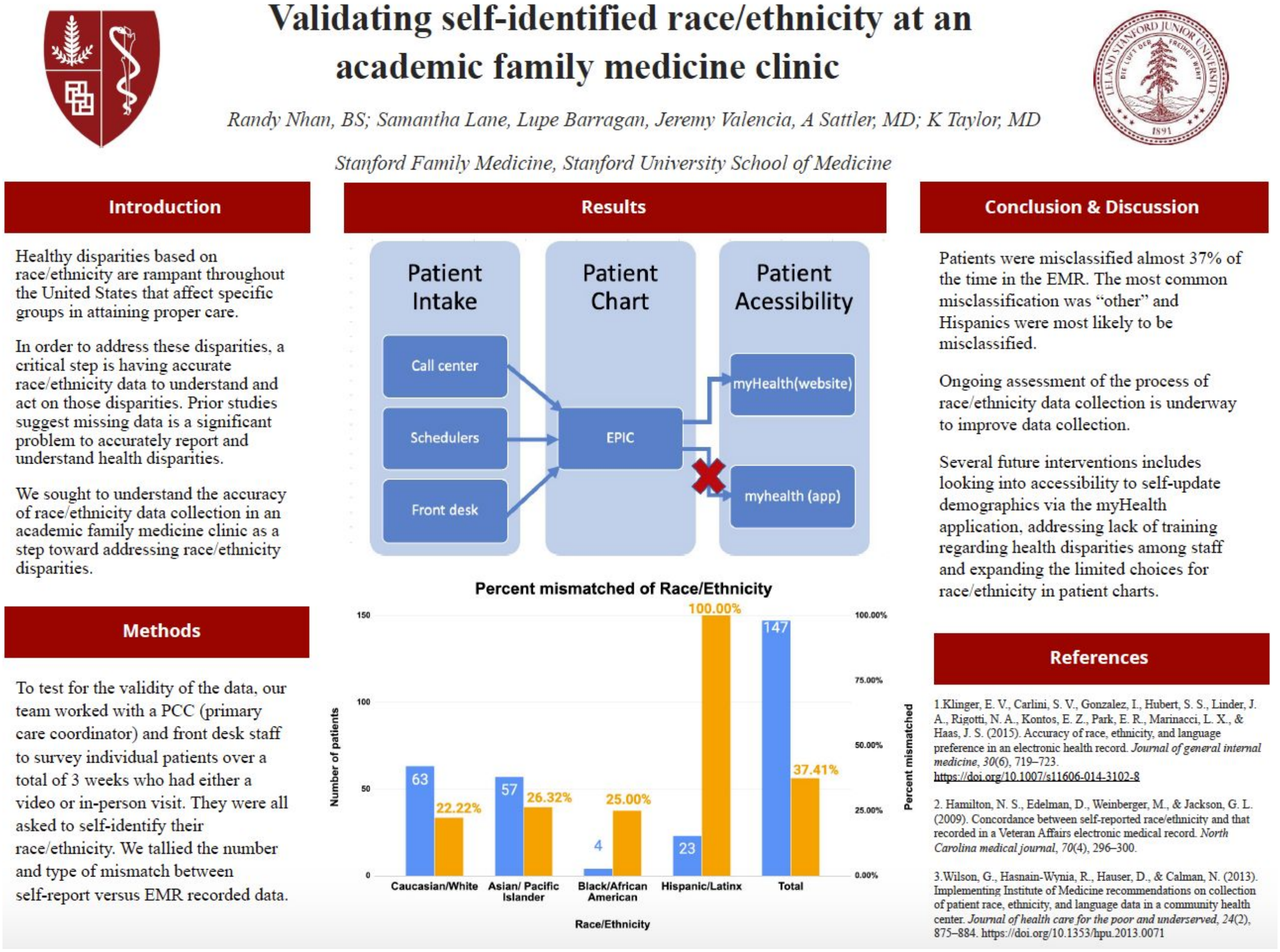
Fairness Audit: Epic EOL High Threshold in Inpatient Oncology

C. SUBGROUP CALIBRATION

Significant underprediction of events for Hispanic patients.



Caveat: all results are probably wrong



- Does EMR = self-identified race/ethnicity at Stanford Famly 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)

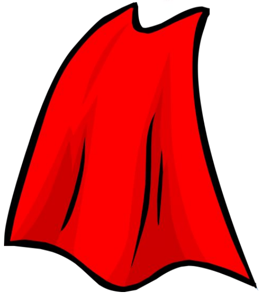
What is required to perform a reliability/fairness audit of a model?

- 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!