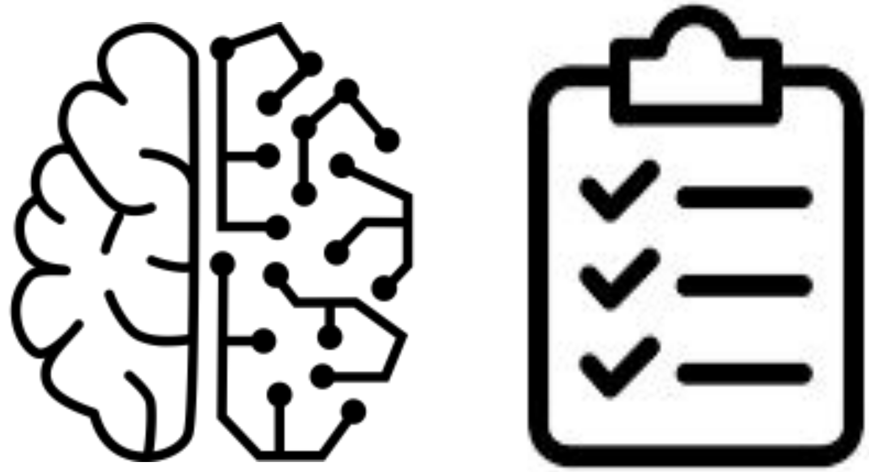


Low adherence to model reporting guidelines for commonly used clinical prediction models

<u>Jonathan Lu</u>, Alison Callahan, Birju Patel, Keith Morse, Dev Dash, Nigam Shah

Stanford Medicine

BMIR
Stanford Center for Biomedical Informatics Research

Funding provided by the Stanford Medical Scholars Fellowship Program.

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 (!)
    - Only 1 completed for a model in use for a health system
- We assess if commonly used models in health systems 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

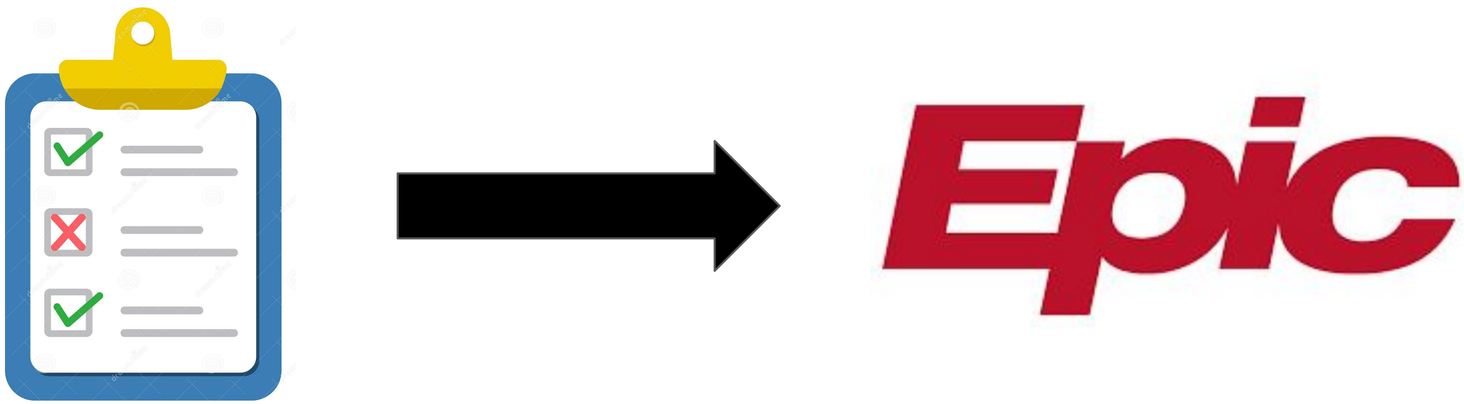
Validation and performance

	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

Sendak 2020

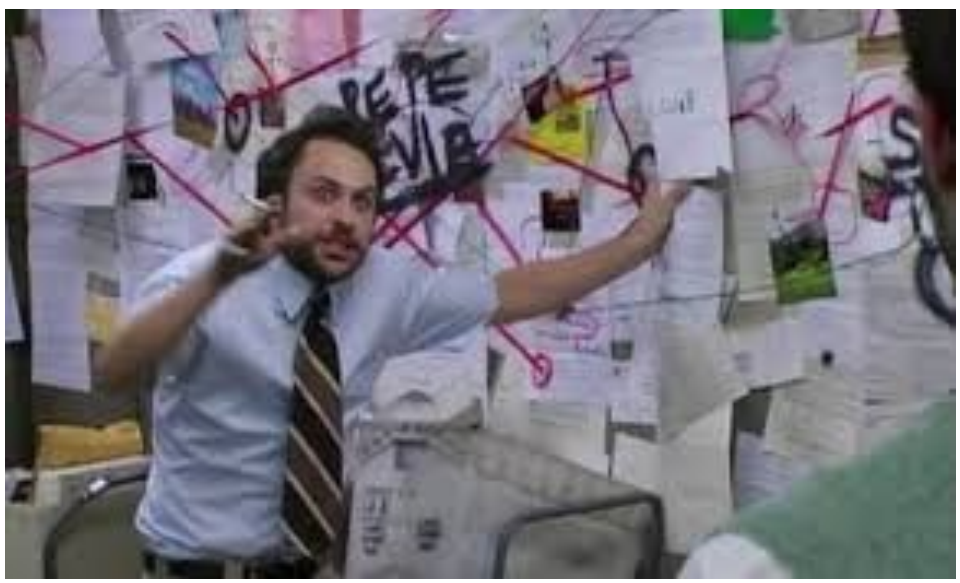
Methods

- 1. Gather recommendations from Model Reporting Guidelines
    - MEDLINE search, review citations, exclude those without specific recommendations
- 2. Merge similar items into unique reportable “atoms”
- 3. Gather commonly used Models
    - Epic models (Cognitive Computing Model Briefs)
- 4. Authors review Model Briefs and grade if they report
    - Adjudicator synthesizes



Results

- 15 model reporting guidelines
  - 220 unique atoms identified!
- 12 most commonly used Epic Models
- Graders had interrater agreement of 76%
- After adjudication, **Epic Models' median completion rate of applicable atoms was 39%** (range: 31% - 37%)



Results

- 15 model reporting guidelines prioritize different stages of creating a model
  - e.g. use TRIPOD for Model Development

Model Reporting Guideline	Use Case	Model Formulation	Model Dev.	Model Dev: Fairness	Practical Feasibility	Utility Assessment	Deployment Design	Execution of Workflow	Monitoring of model	Prospective Evaluation
Model Cards	8	5	29	**9**	1	0	0	0	0	0
Model Facts Labels	**10**	7	9	0	1	1	0	0	2	1
Guidelines	7	6	31	1	0	1	0	0	1	0
MI-CLAIM	4	3	29	3	0	1	0	0	0	1
MINIMAR	4	4	18	5	0	0	0	0	0	0
TRIPOD	7	**9**	**53**	1	0	**3**	0	0	3	2
CONSORT-AI	**10**	3	23	6	1	0	0	0	2	**19**
SPIRIT-AI	9	3	17	1	2	0	0	0	2	18
Trust and Value	4	0	9	0	2	1	0	0	4	2
ML Test Score	0	0	12	4	1	0	0	**2**	**17**	0
Risk	2	4	24	0	0	1	0	0	2	6
STARD	8	2	37	6	0	1	0	0	0	0
ABCD	1	3	27	0	0	1	0	0	0	0
CHARMS	5	**9**	42	1	2	0	0	0	1	4
PROBAST	4	6	41	0	1	1	0	0	1	0
Total	14	14	104	10	5	4	0	2	19	25

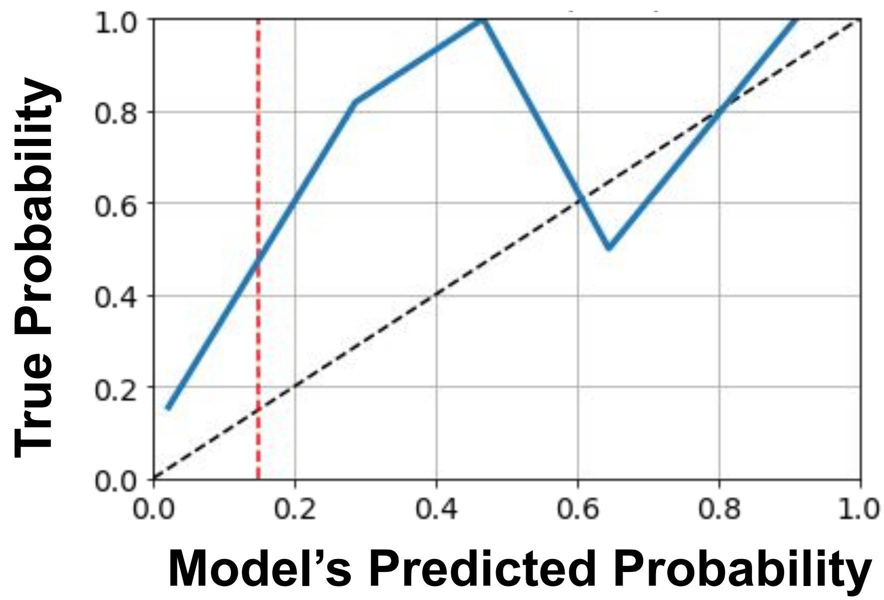
Commonly Requested Atoms

- 100% reporting for commonly requested atoms, except:
    - Low reporting of Confidence Intervals (0%), Missing Data Statistics (50%) and Strategy (58%)

Atom Description	# Requesting	Stage	Reporting %
Provide any description of the data set (training / study) in question	12	Model Development	100.%
Define the output/outcome produced by the model	10	Model Formulation	100.%
Define the specific local area/environment/setting of training data / model deployment.	10	Use Case	100.%
How data was preprocessed (data cleaning, predictor transformation, outlier removal, predictor coding)	10	Model Development	100.%
How missing data were handled	10	Model Development	50.%
What parameters, including constraints and penalties added as loss terms (e.g. shrinkage penalties), were used to train and select models	10	Model Development	58.%
Provide confidence intervals, statistical significance, or some other handling of uncertainty and variability in model performance metrics	10	Model Development	0.%
Clarify what type of validation is done, whether internal or external	11	Model Development	100.%
Describe internal validation strategy to account for model optimism (e.g. cross-validation, bootstrapping, data splitting))	11	Model Development	100.%
Mention what performance measures are used	13	Model Development	100.%
AUROC (c- index)	11	Model Development	100.%
Describe how the ML model is supposed to be used in clinical context	11	Use Case	100.%

Requested, but not Reported Atoms

- Low reporting of atoms related to **reliability:**
  - **External Validation**
    - External Validation Strategy (33%)
    - Calibration Plots (0%)
    - Confidence Intervals (0%)
  - **Missingness**
    - Missing Data Statistics (8%)
    - Handling Missing Data Strategy (50%)

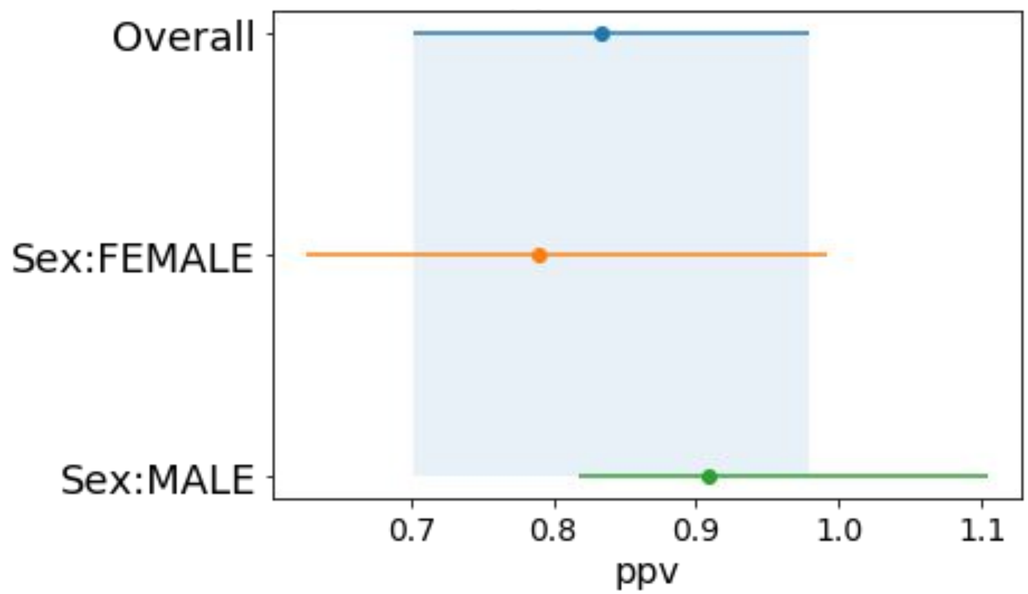


A calibration plot shows how much a model's output matches the true probability of the outcome.

Requested, but not Reported Atoms

- Low reporting of atoms related to **Fairness:**
  - **Summary Statistics:**
    - Sex (33%)
    - Ethnicity/Race* (33%)
    - Age (0%)
  - **Subgroup Analyses (33%)**
    - Intersectional Subgroup Analyses (0%)

** = Ethnicity/Race is used as a way to measure who is represented/impacted by the model, not as an input variable– should not be used as a "risk factor."*



A subgroup analysis shows how a model performs for different subgroups.

<u>Conclusion</u>

- Many model reporting guidelines → 220 distinct atoms requested
- Current model documentation reports only 39% of applicable atoms
  - Little information on reliability and fairness
- Need for better operationalization of reporting practices for AI models in healthcare

Inspiration for this work goes to Margaret Mitchell, Timnit Gebru and co-authors of Model Cards for Model Reporting. They have been leading voices for accountability 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.

*"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**."* Mitchell 2019

Acknowledgments

The authors would like to thank Scott Fleming, Siyun Li, Arjun Gokhale, Wui Ip, Lillian Sung, and Ron Li for providing project feedback and guidance, and members of the Shah lab and the Data Science Team at IAT for ideation, moral support, and feedback during a once-in-a-lifetime pandemic. Thanks to Stephen Pfohl for reader feedback.

Lastly, the authors would like to acknowledge representatives from the Epic Systems data science team: Alejandro Muñoz del Rio, Shannon Gent-Nyden, Paul Sobanski, Stephen Dove, Matt Winkleman, Canna Heibel, and Christina Miles for a constructive conversation after this study's preprint leading to corrections to two atoms' grading for one Model Brief (that did not change the overall results and conclusions of the manuscript) and providing feedback to the company on opportunities to improve their documentation.

References

Khetpal V, Shah N. How a largely untested AI algorithm crept into hundreds of hospitals. Published May 28, 2021. Accessed June 25, 2021. https://www.fastcompany.com/90641343/epic-deterioration-index-algorithm-pandemic-concerns

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.

Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med*. 2020;3:41.

Liu X, The SPIRIT-AI and CONSORT-AI Working Group, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x

Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698.

Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ*. 2020;370:m3210.

Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.

Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg*. 2015;102(3):148-158.

Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527.

Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res*. 2016;18(12):e323.

Breck E, Cai S, Nielsen E, Salib M, Sculley D. The ML test score: A rubric for ML production readiness and technical debt reduction. In: *2017 IEEE International Conference on Big Data (Big Data)*. ; 2017:1123-1132.

Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med*. 2019;170(1):W1-W33.

Mitchell M, Wu S, Zaldivar A, et al. Model Cards for Model Reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Association for Computing Machinery; 2019:220-229.

Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27(12):2011-2015.

Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324.

Silcox C, Dentzer S, Bates DW. AI-enabled clinical decision support software: A “trust and value checklist” for clinicians. *NEJM Catalyst*. 2020;1(6). doi:10.1056/cat.20.0212