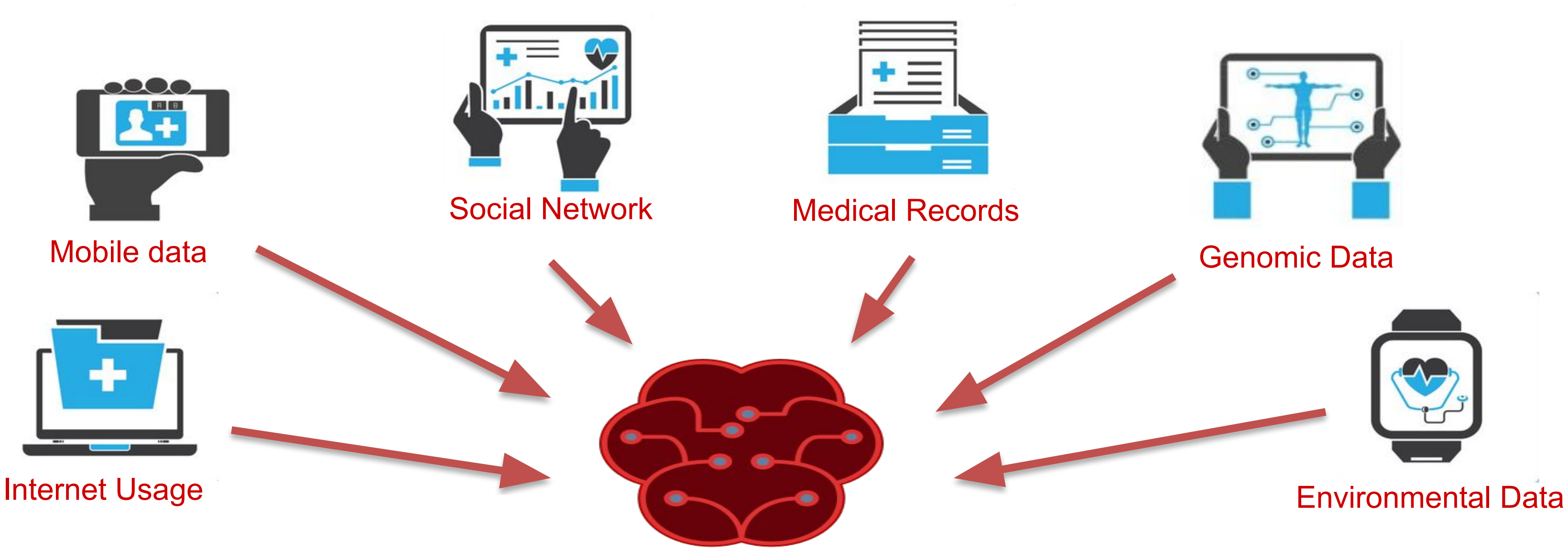
Designing Machine Learning Processes For Equitable Health Systems

Dr. Marzyeh Ghassemi
MIT IMES/EECS.CSAIL
CIFAR AI Chair, Azrieli Global Scholar, JClinic

Em**bodied** Data Is A Powerful **Good**

- Robust, private, fair algorithms require **diverse** datasets for **research** use.
- For **AI** to improve science and address medical harm, we need **data**.



Healthy Machine Learning in Health

what **models** are **healthy**?

what **healthcare** is **healthy**?

what **behaviors** are **healthy**?

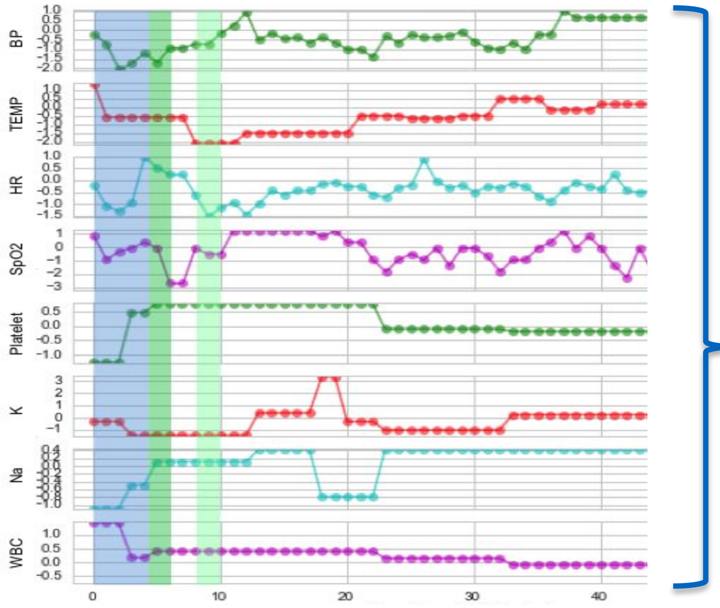
Creating actionable insights in human health.

Improving Treatment Choices With Data + Learning

1) Sumana is having **trouble** **breathing**!
   Clinical Intervention Prediction and Understanding Using Deep Networks. MLHC 2017

Problem: Hospital Decision-Making / Care Planning

Observe Patient Data

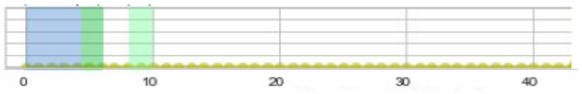


“Real-time” **Prediction**

Of **{**Drug/Mortality/Condition**}**

By Gap Time

?

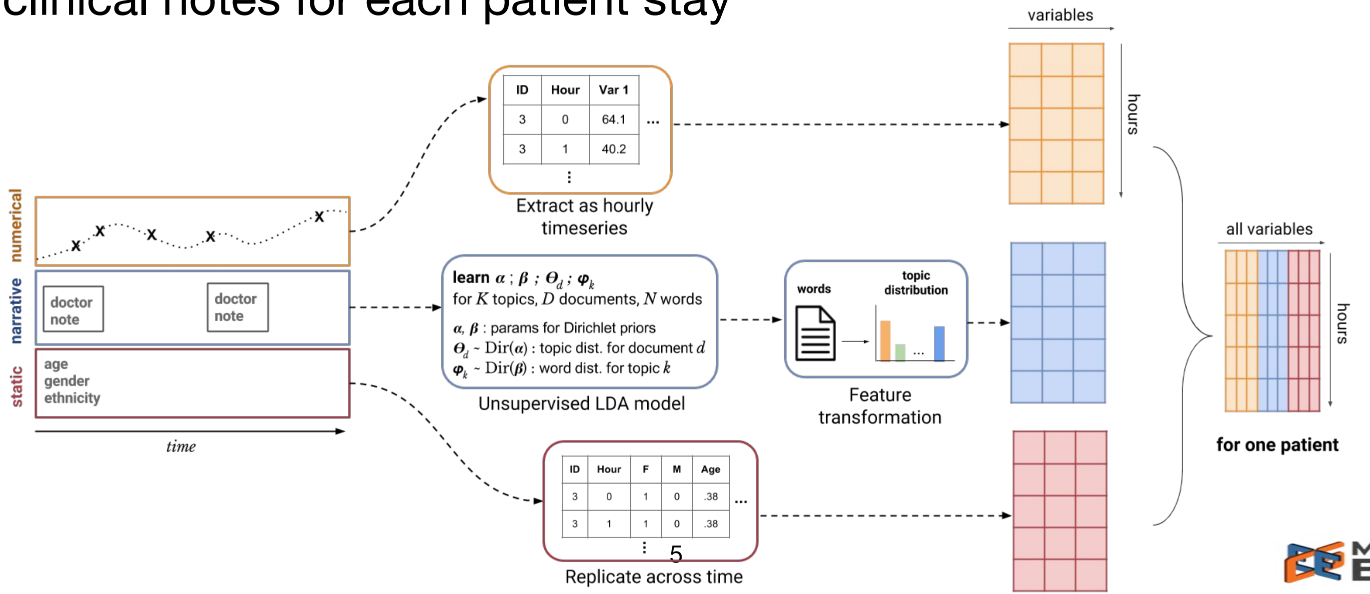


4

Predicting **Interventions** In Intensive Care Units

- 34,148 ICU patients from MIMIC-III
  - 5 static variables (gender, age, etc.)
  - 29 time-varying vitals and labs (oxygen saturation, lactate, etc.)

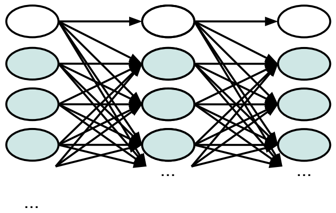
- All clinical notes for each patient stay



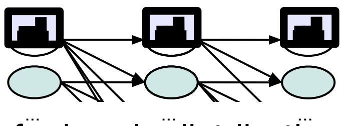
Many **Interventions** + Ways to **Learn**

SSAM

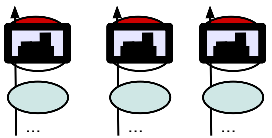
Ghassemi, Doshi-Velez. AMIA CRI 2017.



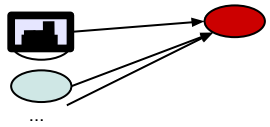
Learn model parameters over patients with variational EM.



Infer hourly distribution over hidden states with HMM DP (fwd alg.).



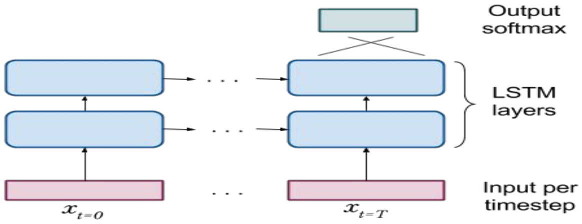
Logistic regression (with label-balanced cost function)



Predict onset in advance

LSTM

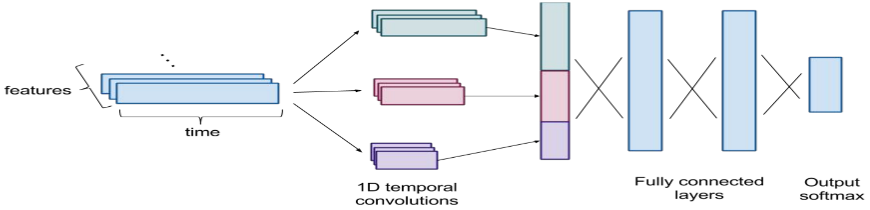
Suresh, ..., Ghassemi. JMLR/MLHC 2017.



2 Layer/512 node LSTM with sequential hourly data; at end of window, use the final hidden state to predict output.

CNN

Suresh, ..., Ghassemi. JMLR/MLHC 2017.



CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.

Improved **Representation** Help NN Get **SOTA**

Area-under-ROC

Task	Model	Intervention Type				
		VENT	NI-VENT	VASO	COL BOL	CRYS BOL
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	**0.76**	0.76	**0.72**	**0.71**
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	LSTM Raw	0.90	0.80	**0.91**	-	-
	LSTM Words	0.90	**0.81**	**0.91**	-	-
	CNN	**0.91**	0.80	**0.91**	-	-
Stay On AUC	Baseline	0.50	0.79	0.55	-	-
	LSTM Raw	0.96	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.96	**0.86**	**0.96**	-	-
Stay Off AUC	Baseline	0.94	0.71	0.93	-	-
	LSTM Raw	0.95	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.95	**0.86**	**0.96**	-	-
Macro AUC	Baseline	0.72	0.72	0.66	-	-
	LSTM Raw	0.86	**0.82**	**0.90**	-	-
	LSTM Words	**0.90**	**0.82**	0.89	-	-
	CNN	0.86	0.81	**0.90**	-	-

Representations with **"physiological words"** for missingness significantly **increased AUC** for interventions with the lowest proportion of examples.

Deep models perform well in general, but **"words"** are important for ventilation tasks.

Suresh et al. "Clinical Intervention Prediction and Understanding Using Deep Networks". MLHC 2017.

Clinical **AI** Performs At or Above **Humans**

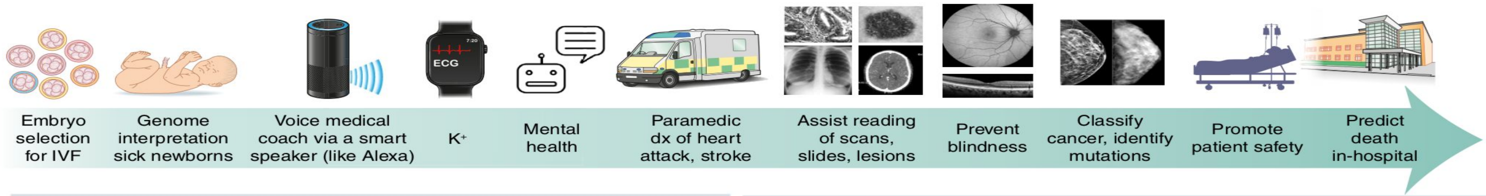


Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

Prediction	*n*	AUC	Publication (Reference number)
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	0.93* 0.75+ 0.85#	Rajkomar et al.[96]
All-cause 3–12 month mortality	221,284	0.93^	Avati et al.[91]
Readmission	1,068	0.78	Shameer et al.[106]
Sepsis	230,936	0.67	Horng et al.[102]
Septic shock	16,234	0.83	Henry et al.[103]
Severe sepsis	203,000	0.85@	Culliton et al.[104]
Clostridium difficile infection	256,732	0.82++	Oh et al.[93]
Developing diseases	704,587	range	Miotto et al.[97]
Diagnosis	18,590	0.96	Yang et al.[90]
Dementia	76,367	0.91	Cleret de Langavant et al.[92]
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al.[98]
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al.[95]
Disease onset for 133 conditions	298,000	range	Razavian et al.[105]
Suicide	5,543	0.84	Walsh et al.[86]
Delirium	18,223	0.68	Wong et al.[100]

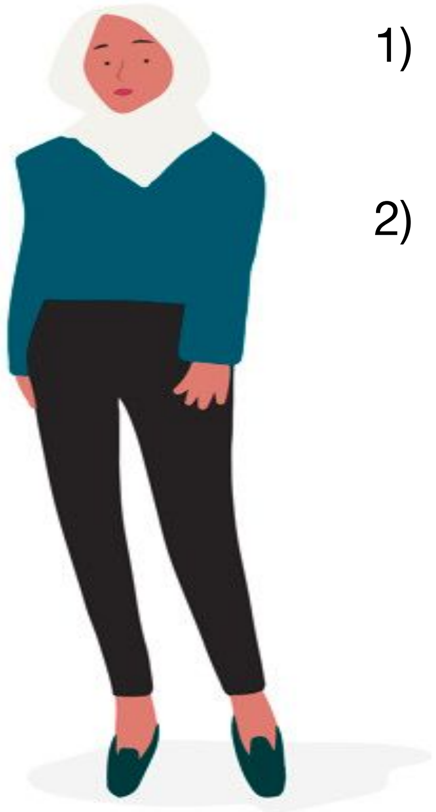
LOS, length of stay; *n*, number of patients (training+ validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; ^, all patients; @, structured+unstructured data; ++, for University of Michigan site.

Source: **High-performance medicine: the convergence of human and artificial intelligence** Eric Topol, Nature Medicine Jan 2019

Figure: Debbie Maizels / Springer Nature

AI **Learns** From Human **Practice**

Improving Treatment Choices With Data + Learning



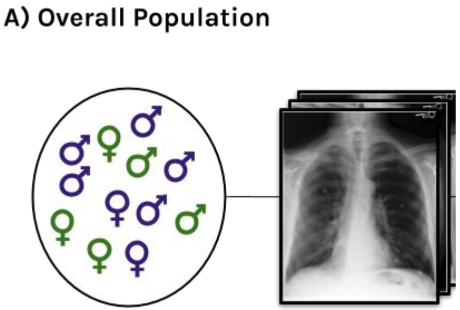
1) Sumana is having **trouble breathing**!

   Clinical Intervention Prediction and Understanding Using Deep Networks. MLHC 2017

2) Do models work for people **like her**?

   Medical imaging algorithms exacerbate biases in underdiagnosis. Nature Medicine 2021.
   Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics 2019
   Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. ACM CHIL 2020
   Is Fairness Only Metric Deep? ICLR 2022
   Write It Like You See It: Detectable Differences in Clinical Notes By Race…. AIES 2022
   AI recognition of patient race in medical imaging: a modelling study. Lancet Digital Health 2022.
   The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. ACM FacCT 2022.

Model-based Chest X-Ray Diagnosis

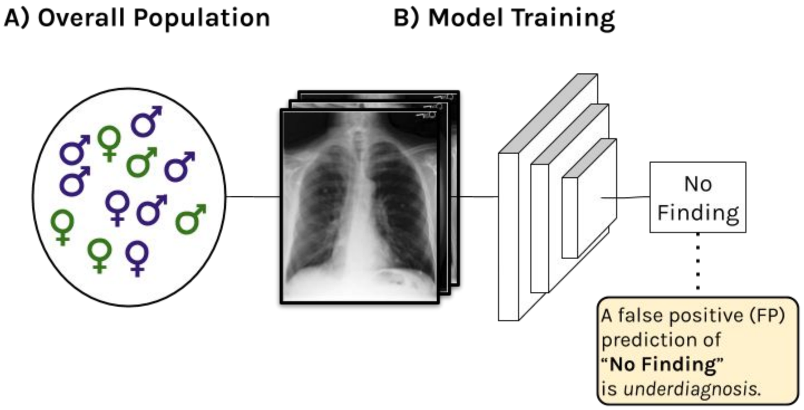


- Take 3 large **chest x-ray** datasets (707,626 images).

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. “Medical imaging algorithms exacerbate biases in underdiagnosis.” Nature Medicine 2021.

11

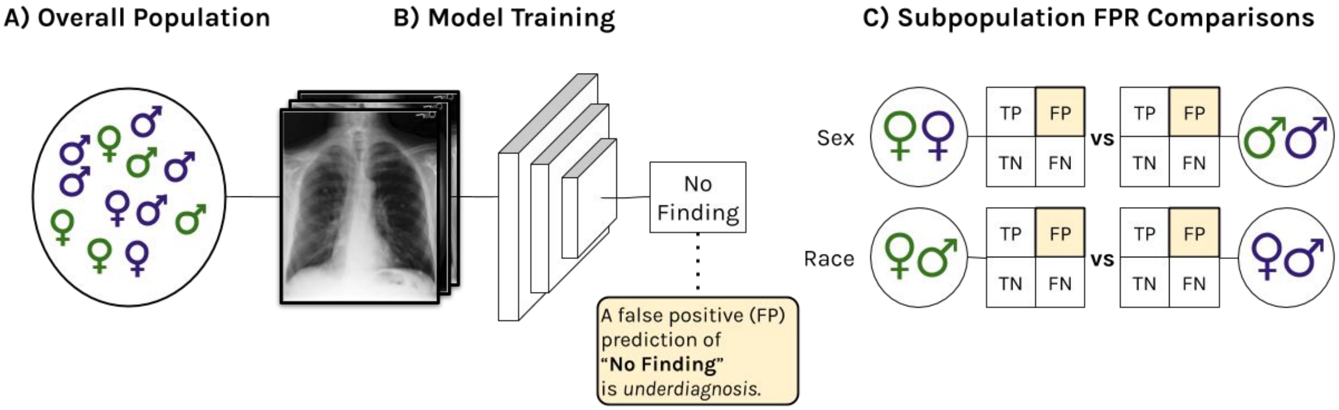
Model-based Chest X-Ray Diagnosis



- Take 3 large **chest x-ray** datasets (707,626 images).
- Train a DenseNet to predict a "**No Finding**" label, e.g., model says patient is healthy.

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

Model-based Chest X-Ray Diagnosis

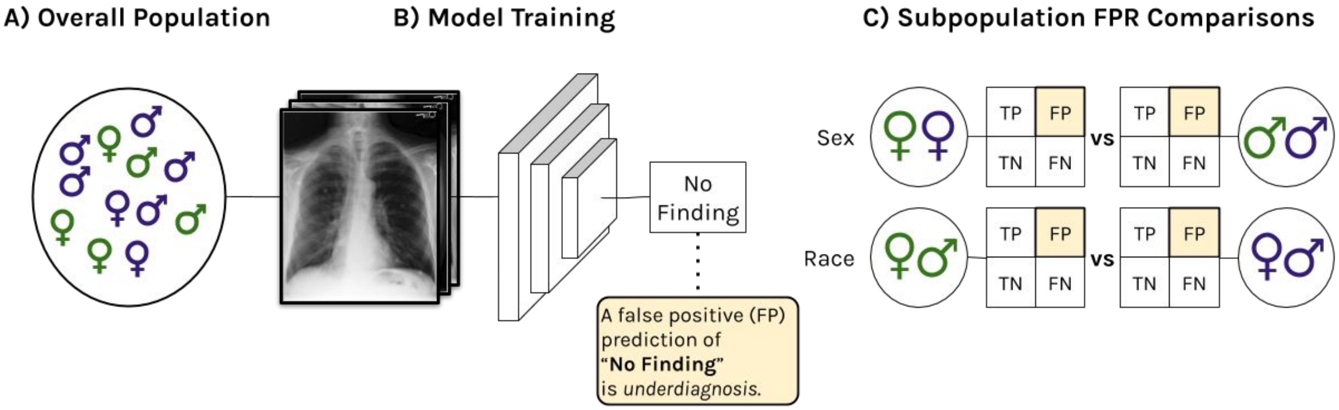


- Take 3 large **chest x-ray** datasets (707,626 images).
- Train a DenseNet to predict a "**No Finding**" label, e.g., model says patient is healthy.
- Compare false positive rate (FPR) in different subpopulations to examine model **underdiagnosis rates.**

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

13

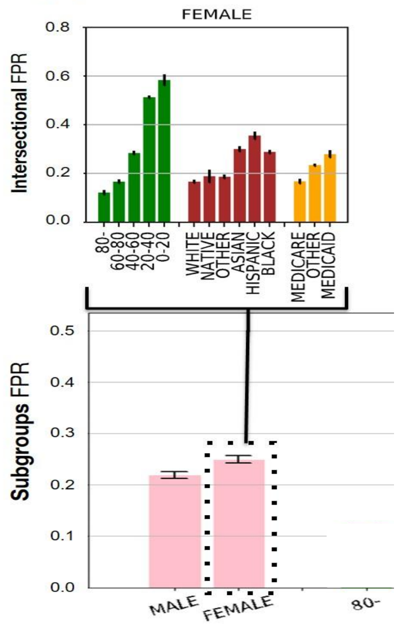
Model-based Chest X-Ray Diagnosis



Higher model underdiagnosis rates on one **subpopulation**, such as **female patients**, would lead to a **higher rate** of **no treatment** for those patients if the model were **deployed**.

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

Automating CheXclusion With EHR + ML

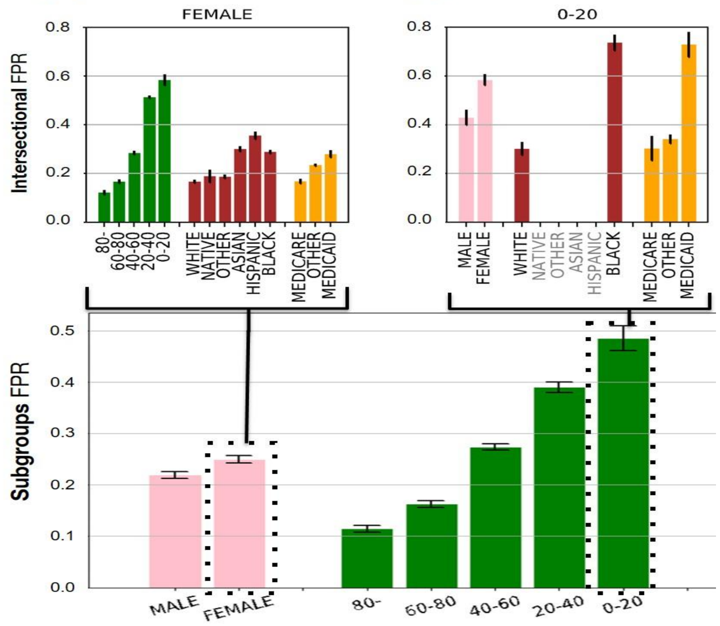


- Largest underdiagnosis rates in Female

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. “Medical imaging algorithms exacerbate biases in underdiagnosis.” Nature Medicine 2021.

15

Automating Che**Xclusion** With **EHR + ML**

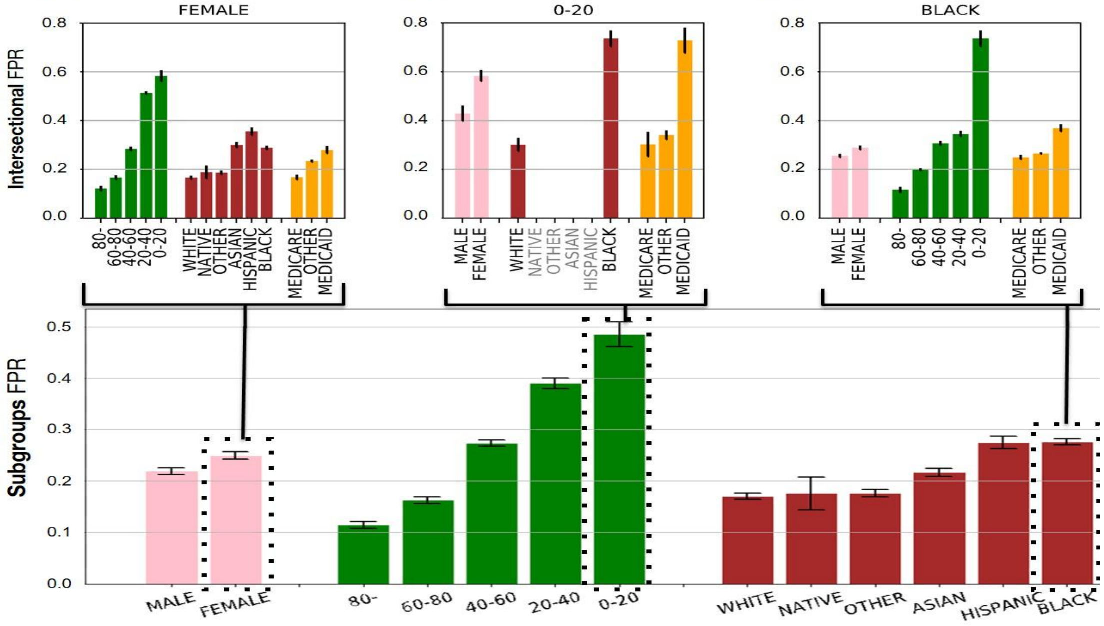


- Largest underdiagnosis rates in Female, 0-20

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

16

Automating Che**Xclusion** With **EHR + ML**

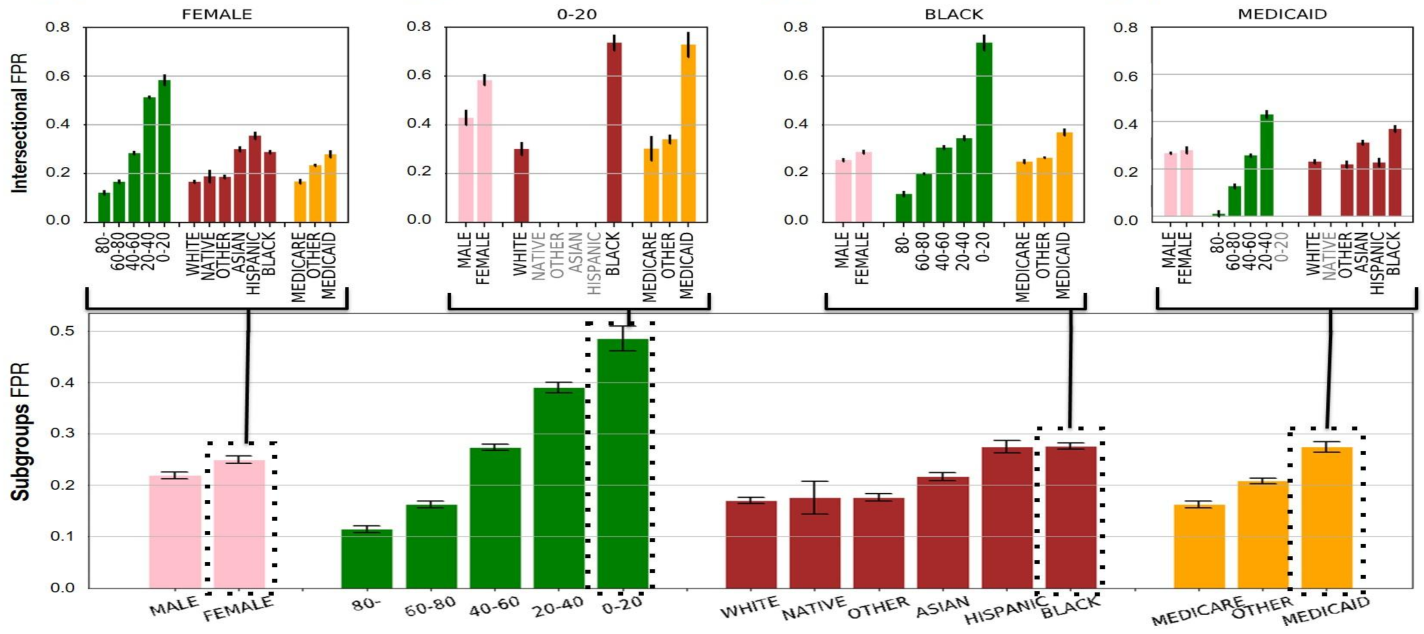


- Largest underdiagnosis rates in Female, 0-20, Black

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

17

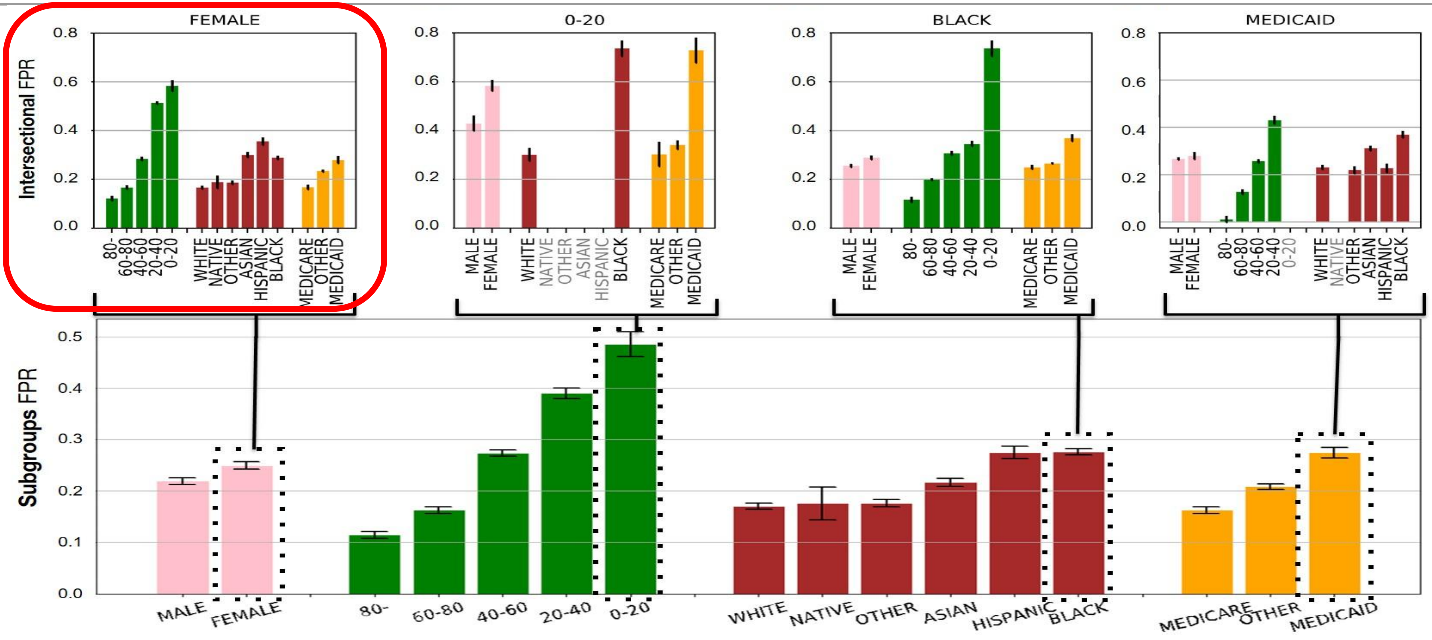
Automating Che**Xclusion** With **EHR + ML**



- Largest underdiagnosis rates in Female, 0-20, Black, and Medicaid insurance patients.

[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

Automating Che**Xclusion** With **EHR + ML**

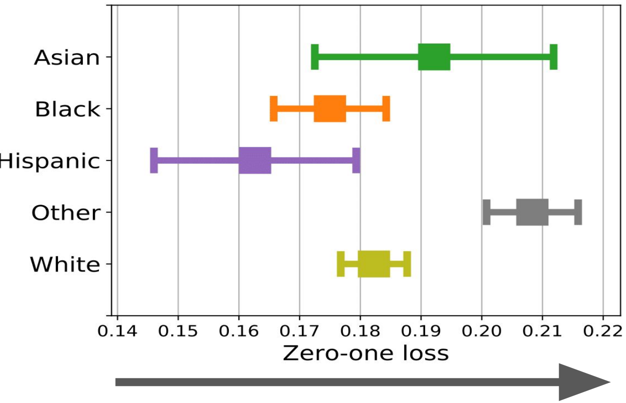


- **Intersectional** identities are often underdiagnosed even more heavily than the aggregate group, e.g., **Black or Hispanic female patients** are **underdiagnosed more** than White female patients.

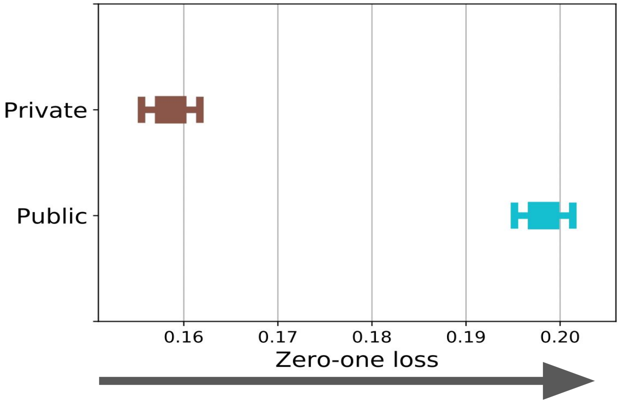
[1] Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Medical imaging algorithms exacerbate biases in underdiagnosis." Nature Medicine 2021.

Auditing Fairness In Predictive Models?

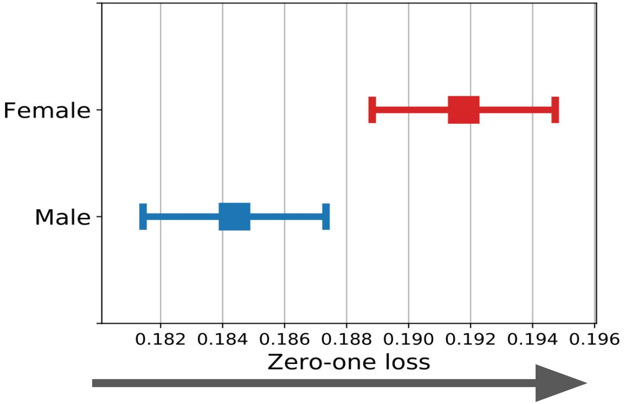
- Significant differences in model accuracy for race, sex, and insurance type in **ICU notes** and insurance type in **psychiatric notes**.



Worse Prediction Accuracy



Worse Prediction Accuracy



Worse Prediction Accuracy

[1] Chen, Szolovits, Ghassemi. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?." *AMA journal of ethics* 21.2 (2019): 167-179.

Hurtful Words: Biases in Clinical Word Embeddings

Prompt: [**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]

[1] Zhang, Lu, Abdallah, Ghassemi. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings". ACM CHIL 2020.

Hurtful Words: Biases in Clinical Word Embeddings

Prompt: **[**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]**

SciBERT: caucasian pt became belligerent and violent . sent to **hospital .**
white pt became belligerent and violent . sent to **hospital .**

[1] Zhang, Lu, Abdallah, Ghassemi. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings". ACM CHIL 2020.

Hurtful Words: Biases in Clinical Word Embeddings

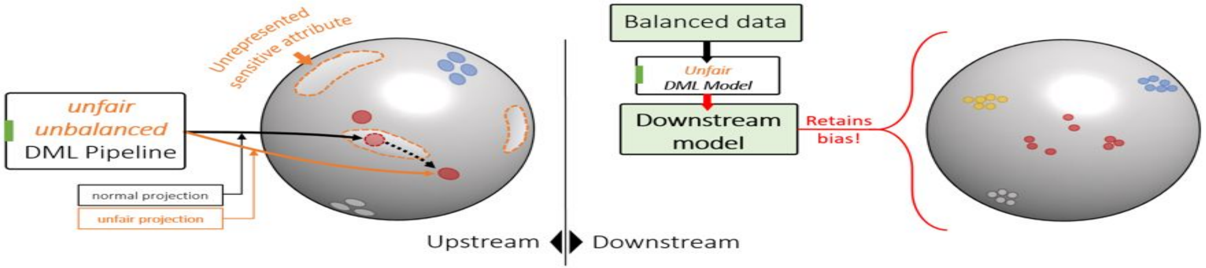
Prompt:	**[**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]**
SciBERT:	caucasian pt became belligerent and violent . sent to **hospital .** white pt became belligerent and violent . sent to **hospital .** african pt became belligerent and violent . sent to **prison .** african american pt became belligerent and violent . sent to **prison .** black pt became belligerent and violent . sent to **prison .**

[1] Zhang, Lu, Abdallah, Ghassemi. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings". ACM CHIL 2020.

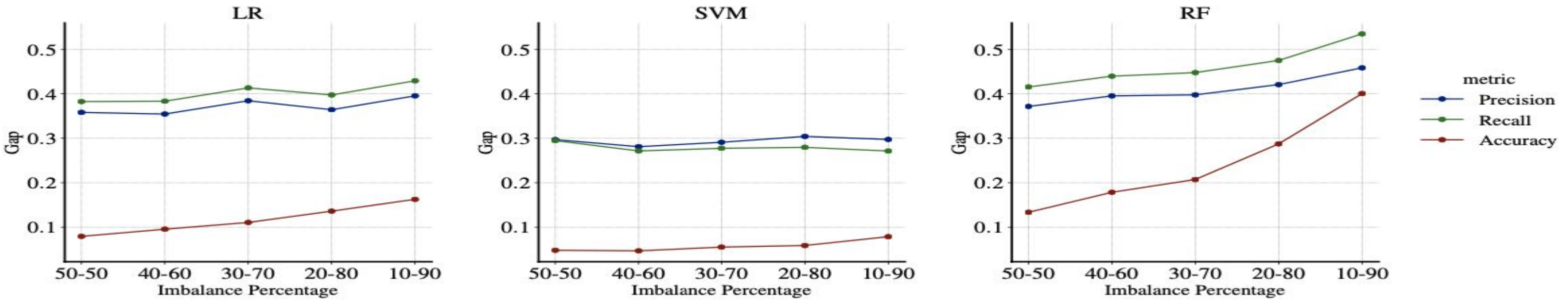
23

Balance **Downstream** Does Not Fix Latent Embedding **Bias**

- Bias in data causes **asymmetric** upstream **embeddings**.



- Biased embeddings **impact** downstream tasks, even with **rebalancing.**



Dullerud, Natalie, et al. "Is Fairness Only Metric Deep?" *ICLR 2022.*

24

Bias Is Part of the Clinical **Landscape**

This Issue | Views 12,435 | Citations 41 | Altmetric 174

Viewpoint

August 11, 2015

Racial Bias in Health Care and Health
Challenges and Opportunities

David R. Williams, PhD, MPH[1,2]; Ronald Wyatt, MD, MHA[3]

» Author Affiliations

JAMA. 2015;314(6):555-556. doi:10.1001/jama.2015.9260

J Palliat Med. 2013 Nov; 16(11): 1329–1334.
doi: 10.1089/jpm.2013.9468

PMCID: PMC3822363
PMID: 24073685

Racial and Ethnic Disparities in Palliative Care

Kimberly S. Johnson, MD, MHS[1,2]

Author information ► Article notes ► Copyright and License information ► Disclaimer

This article has been cited by other articles in PMC.

2

The Girl Who Cried Pain: A Bias Against Women in the Treatment of Pain

Diane E. Hoffmann and Anita J. Tarzian

Am J Public Health. 2007 February; 97(2): 247–251.
doi: 10.2105/AJPH.2005.072975

PMCID: PMC1781382
PMID: 17194867

The Black–White Disparity in Pregnancy-Related Mortality From 5 Conditions: Differences in Prevalence and Case-Fatality Rates

Myra J. Tucker, BSN, MPH, Cynthia J. Berg, MD, MPH, William M. Callaghan, MD, MPH, and Jason Hsia, PhD

Author information ► Article notes ► Copyright and License information ► Disclaimer

Obes Rev. 2015 Apr;16(4):319-26. doi: 10.1111/obr.12266. Epub 2015 Mar 5.

Impact of weight bias and stigma on quality of care and outcomes for patients with obesity.

Phelan SM[1], Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M.

⊕ Author information

imes

MIT EECS CSAIL

POP QUIZ!

<u>Nursing Progress Note</u>

NEURO: sedated with propofol gtt 85mcg/kg
RESP: remains intubated with IMV 12/750/5peep/5psv/40%fio2/
SRR 8-10, breath sounds coarse, O2 sat 98-100%.
GU: inc large amt foul smelling urine foley placed with UO ~50cc/hr,
dialysis to be initiated at 6pm
SKIN: sacral decub w-d dsg changes wound red beefy, small amt
bloody drainage, heel dsg w-d dsg changed no drainage
ACCESS: left EJ, right groin introducer, left rad aline
PLAN: dialysis this eve, wean extubate tomorrow, titrate up po meds
for hypertension

Adam, Hammaad, et al. "Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations." *AIES 2022.*

POP QUIZ!

Predictive of black patient **Predictive of white patient**

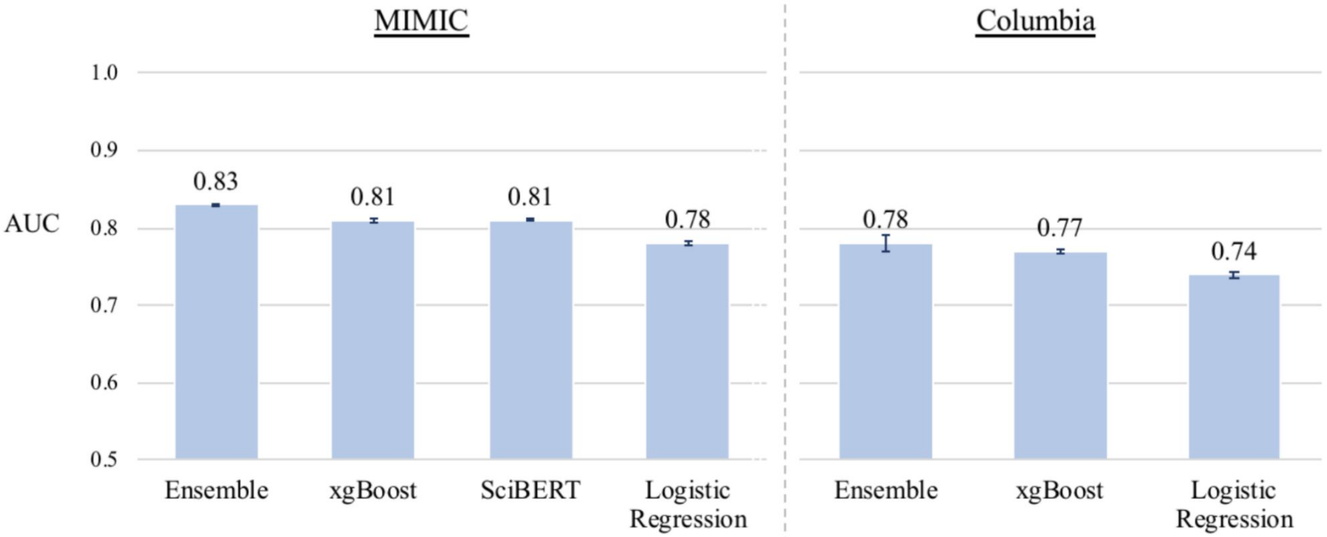
<u>Nursing Progress Note</u>

NEURO: sedated with propofol gtt 85mcg/kg
RESP: remains intubated with IMV 12/750/5peep/5psv/40%fio2/
SRR 8-10, breath sounds coarse, O2 sat 98-100%.
GU: inc large amt foul smelling urine foley placed with UO ~50cc/hr,
dialysis to be initiated at 6pm
SKIN: sacral decub w-d dsg changes **wound** **red** beefy, small amt
bloody drainage, heel dsg w-d dsg changed no drainage
ACCESS: left EJ, right groin introducer, left rad aline
PLAN: **dialysis** this eve, wean extubate tomorrow, titrate up po
meds for **hypertension**

Adam, Hammaad, et al. "Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations." *AIES 2022.*

Super-Human Prediction **Performance**

- Is it **possible to predict race from a clinical note**, even after stripping out direct indicators?
  - MIMIC notes - 668,768 clinical notes / 28,032 patients
  - Columbia notes - 3,554,802 clinical notes / 29,807 patients



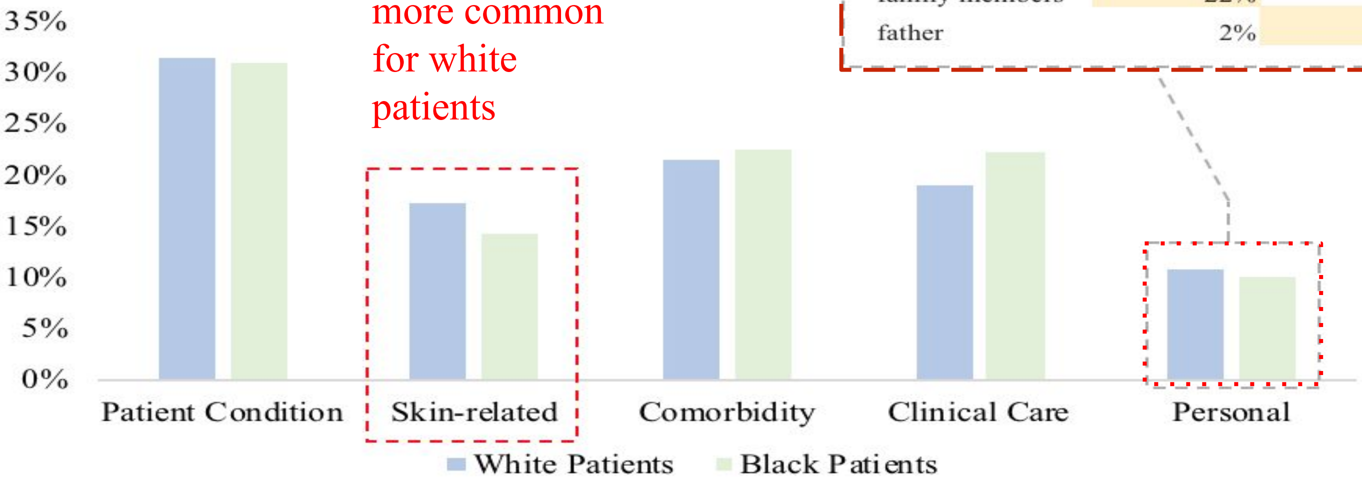
Adam, Hammaad, et al. "Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations." *AIES 2022.*

Qualitative Note **Differences**

- 25 most predictive features for each race have skewed categories.

References to skin are far more common for white patients

Word	% Black Patients	% White Patients
husband	50%	64%
family members	22%	14%
father	2%	5%



Personal references to "family" change, e.g., for married females "family members" is used more often for Black than White patients.

Adam, Hammaad, et al. "Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations." *AIES 2022.*

POP QUIZ!

Is this patient Black?

Gichoya, Judy W., et al. “AI recognition of patient race in medical imaging: a modelling study.” Lancet Digital Health. 2022.

POP QUIZ!

Race detection in radiology imaging	
Chest x-ray (internal validation)*	
MXR (Resnet34, Densenet121)	0·97, 0·94
CXP (Resnet 34)	0·98
EMX (Resnet34, Densenet121, EfficientNet-B0)	0·98, 0·97, 0·99
Chest x-ray (external validation)*	
MXR to CXP, MXR to EMX	0·97, 0·97
CXP to EMX, CXP to MXR	0·97, 0·96
EMX to MXR, EMX to CXP	0·98, 0·98
Chest x-ray (comparison of models)†	
MXR, CXP, EMX	Multiple results (appendix p 26)
CT chest (internal validation)*	
NLST (slice, study)	0·92, 0·96
CT chest (external validation)*	
NLST to EM-CT (slice, study)	0·80, 0·87
NLST to RSPECT (slice, study)	0·83, 0·90
Limb x-ray (internal validation)*	
DHA	0·91
Mammography*	
EM-Mammo (image, study)	0·78, 0·81
Cervical spine x-ray*	
EM-CS	0·92

Gichoya, Judy W., et al. “AI recognition of patient race in medical imaging: a modelling study.” Lancet Digital Health. 2022.

Is It BMI?

BMI

	Obese (BMI > 30)	Overweight (BMI 25 to < 30)	Normal (BMI 18.5 to < 25)	Underweight (BMI <18.5)
White	0.92	0.93	0.90	0.96
Black	0.93	0.96	0.89	0.97
Asian	0.91	0.92	0.94	0.98

Gichoya, Judy W., et al. "AI recognition of patient race in medical imaging: a modelling study." Lancet Digital Health. 2022.

32

Breast Density?

BMI

	Obese (BMI > 30)	Overweight (BMI 25 to < 30)	Normal (BMI 18.5 to < 25)	Underweight (BMI <18.5)
White	0.92	0.93	0.90	0.96
Black	0.93	0.96	0.89	0.97
Asian	0.91	0.92	0.94	0.98

Breast Density

Tissue Density	ROC AUC (Slice)
1 (Fatty)	0.79
2 (Scattered)	0.82
3 (Heterogeneous)	0.83
4 (Dense)	0.74
Overall	0.82

Gichoya, Judy W., et al. “AI recognition of patient race in medical imaging: a modelling study.” Lancet Digital Health. 2022.

Bone **Density**?

BMI

	Obese (BMI > 30)	Overweight (BMI 25 to < 30)	Normal (BMI 18.5 to < 25)	Underweight (BMI <18.5)
White	0.92	0.93	0.90	0.96
Black	0.93	0.96	0.89	0.97
Asian	0.91	0.92	0.94	0.98

Breast Density

Tissue Density	ROC AUC (Slice)
1 (Fatty)	0.79
2 (Scattered)	0.82
3 (Heterogeneous)	0.83
4 (Dense)	0.74
Overall	0.82

Bone Density

Original → Clipped

AUC 0.97

AUC 0.96

Bone density features largely removed

Gichoya, Judy W., et al. “AI recognition of patient race in medical imaging: a modelling study.” Lancet Digital Health. 2022.

34

Disease Distribution?

BMI

	Obese (BMI > 30)	Overweight (BMI 25 to < 30)	Normal (BMI 18.5 to < 25)	Underweight (BMI <18.5)
White	0.92	0.93	0.90	0.96
Black	0.93	0.96	0.89	0.97
Asian	0.91	0.92	0.94	0.98

Breast Density

Tissue Density	ROC AUC (Slice)
1 (Fatty)	0.79
2 (Scattered)	0.82
3 (Heterogeneous)	0.83
4 (Dense)	0.74
Overall	0.82

Bone Density

Original → Clipped

AUC 0.97

AUC 0.96

Bone density features largely removed

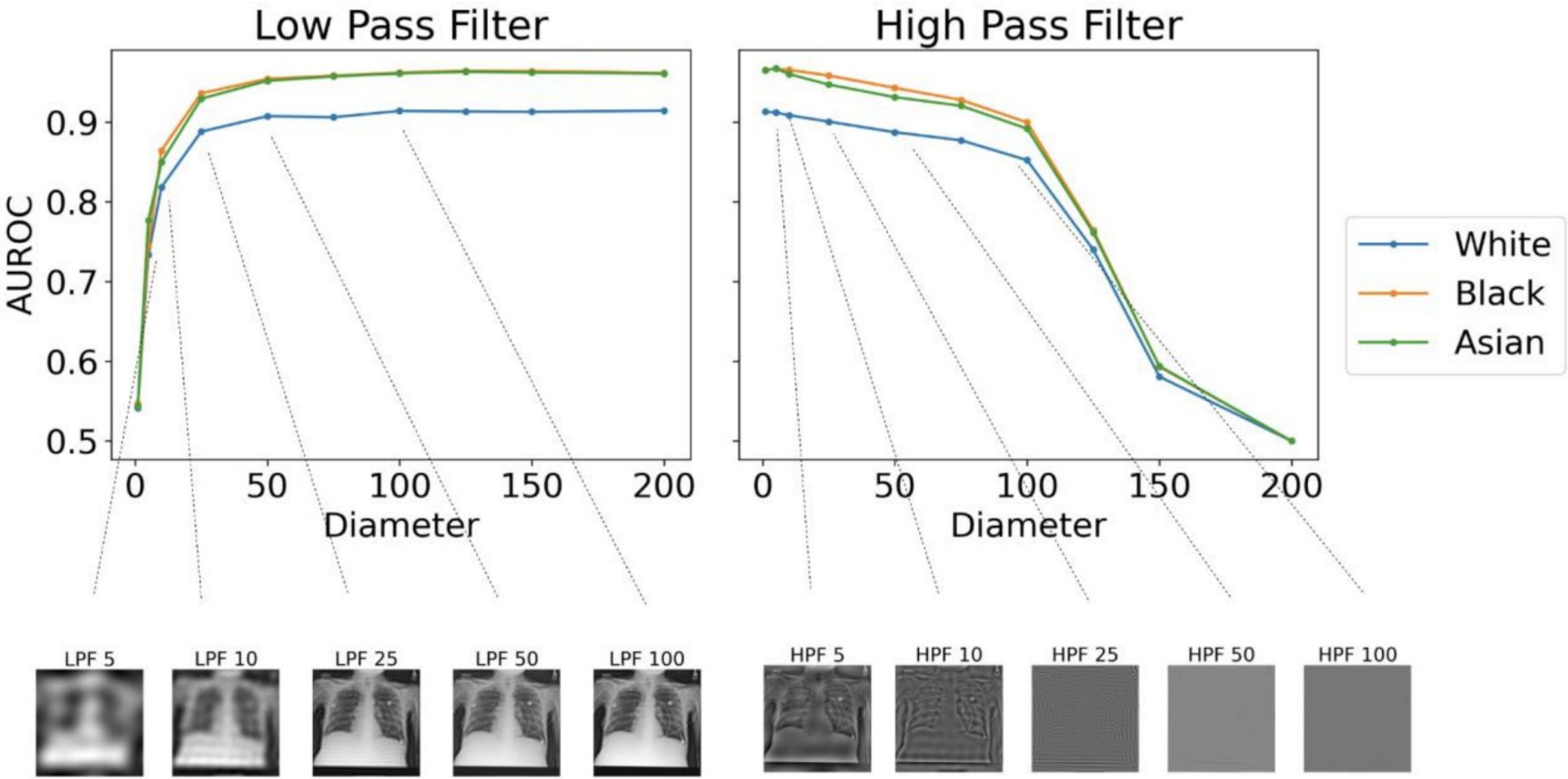
Disease Distribution

Based on chest X-Ray Images: → Based on disease labels:

AUC 0.97 AUC 0.61

Gichoya, Judy W., et al. "AI recognition of patient race in medical imaging: a modelling study." Lancet Digital Health. 2022.

Frequency **Domain**?

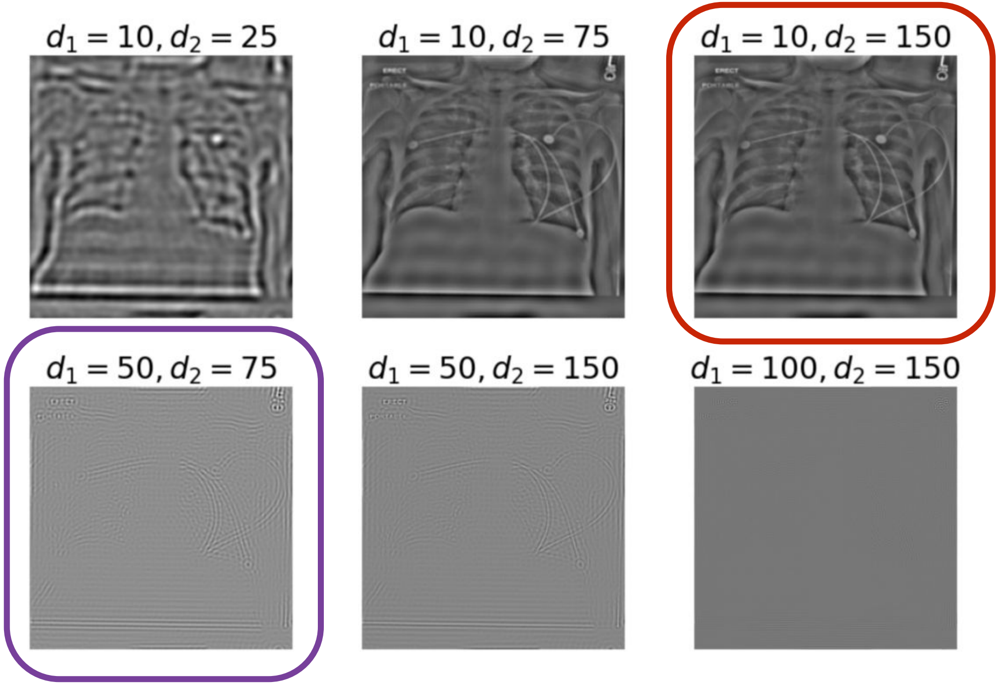


Gichoya, Judy W., et al. “AI recognition of patient race in medical imaging: a modelling study.” Lancet Digital Health. 2022.

Self-reported Race is **Obvious** to **AI**

Race prediction AUROC

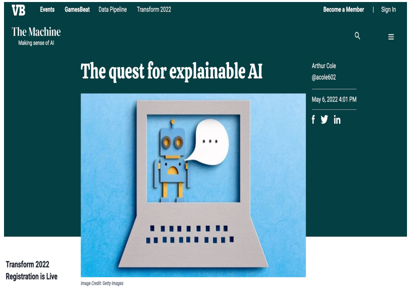
d_1 | d_2	25	50	75	100	125	150
10	0.86	0.90	0.91	0.91	0.91	0.91
25		0.86	0.89	0.90	0.90	0.91
50			0.87	0.89	0.89	0.89
75				0.85	0.86	0.87
100					0.84	0.84
125						0.75



Gichoya, Judy W., et al. "AI recognition of patient race in medical imaging: a modelling study." Lancet Digital Health. 2022.

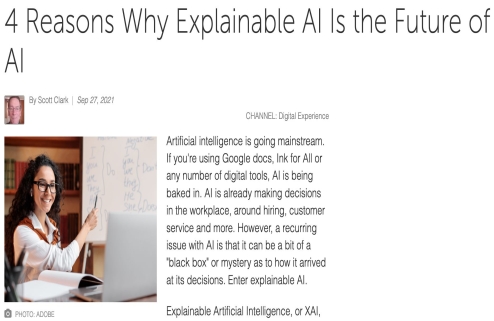
Can We **Fix** Model Gaps With **Explanations**?

- Complex models can be hard to understand.
- Simple, human-interpretable post-hoc explanation methods are proposed to help users **trust** model **predictions**.
- What is the approximation quality of these explanations models?



Source: Arthur Cole, VentureBeat (May 2022)

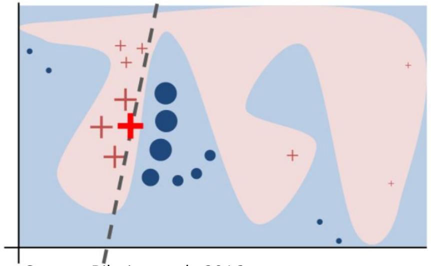
Source: Bernadette Wilson, DevPro Journal (May 2022)



Source: Scott Clark, CMSWire (September 2021)

Post-hoc Explanation Models Approximate **Blackboxes**

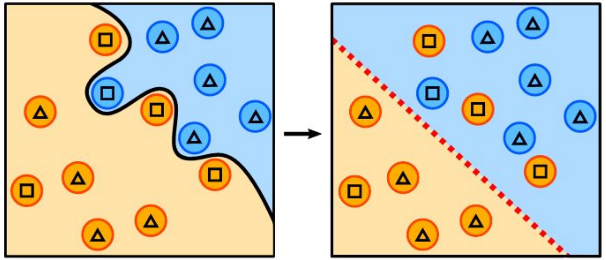
Local Explanation Models



Source: Ribeiro et al., 2016

SHapley Additive exPlanations (SHAP)
Local Interpretable Model-Agnostic Explanations (LIME) **- 8000+ citations**

Global Explanation Models



Sparse Decision Tree (Tree)
Generalized Additive Model (GAM) - **100+ citations**

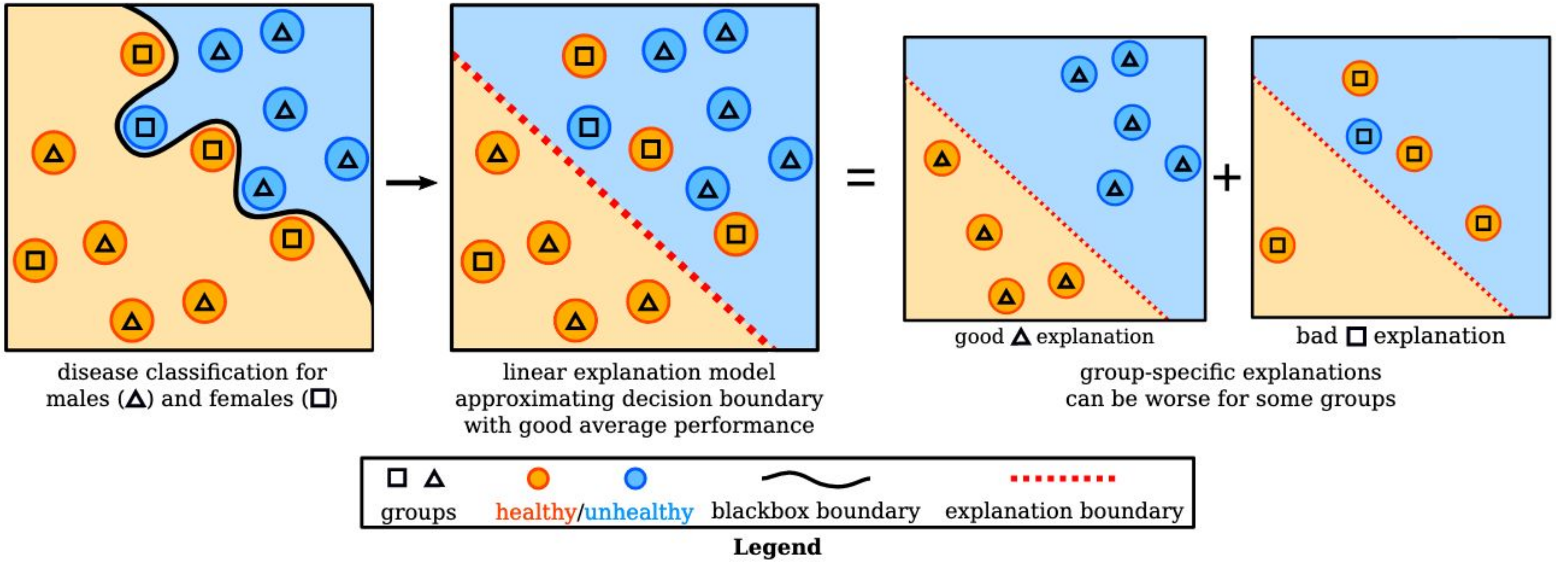
Train simple, human-interpretable models to imitate a blackbox model's behaviour.

Post-hoc explanations are easy to **interpret.**

[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

Is **Explanation Quality Uniform** Across Subgroups?

We measure the fairness of **local** and **global** explanations, and **compare:**

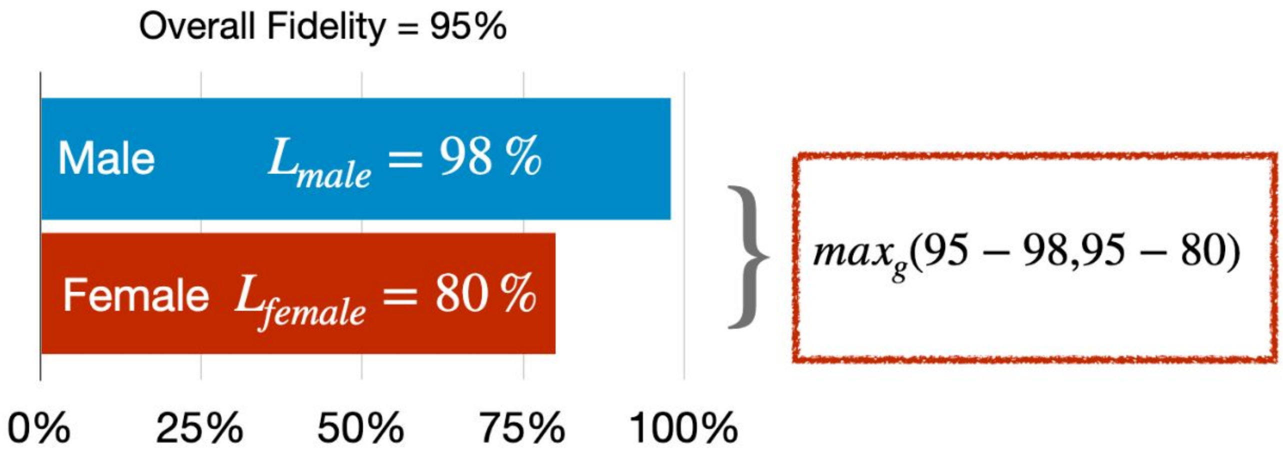


[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

Is **Explanation Quality** Uniform Across Subgroups?

We measure the fairness of **local** and **global** explanations, and **compare:**

- Difference between average fidelity and worst-case fidelity

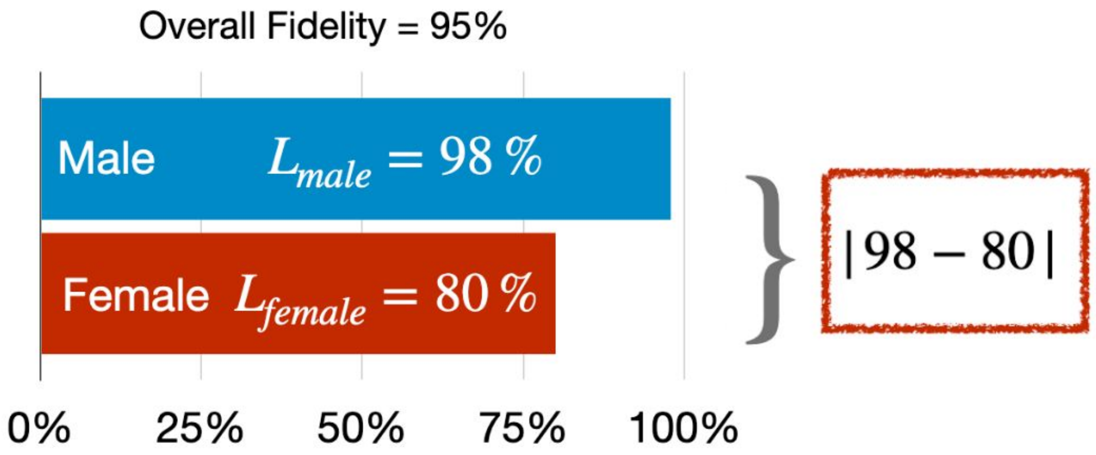


[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

Is **Explanation Quality** Uniform Across Subgroups?

We measure the fairness of **local** and **global** explanations, and **compare:**

- Difference between average fidelity and worst-case fidelity
- Average absolute difference in fidelity across subgroups



[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

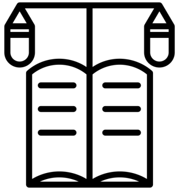
Explanation Quality Higher for Some **Subgroups**

- For both local and global explanation models, there are subgroup ***fidelity*** gaps.

Adult
Y: Income
> 50K

<u>Sex</u>

LSAT
Y: Student passes
bar exam



<u>Race</u>

MIMIC
Y: ICU
mortality

<u>Sex</u>

Recidivism
Y: Defendant
re-offends

<u>Race</u>

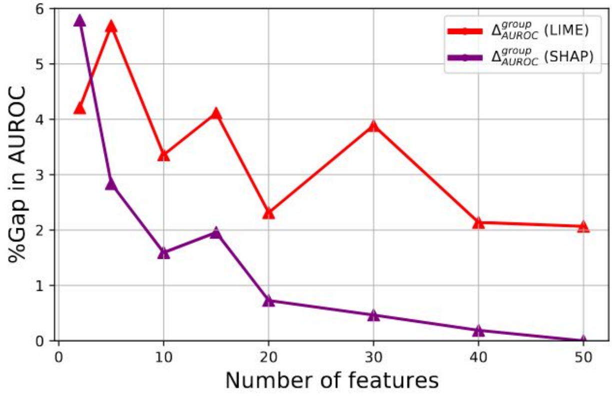
[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

Explanation Quality Higher for Some **Subgroups**

- For both local and global explanation models, there are subgroup ***fidelity*** gaps.

Dataset	Blackbox Classifier	$\Delta_{\text{Acc.}}$	$\Delta^{\text{group}}_{\text{AUROC}}$	$\Delta^{\text{group}}_{\text{Acc.}}$	$\Delta^{\text{group}}_{\text{Err.}}$
adult	Logistic Regression	**0.7% ± 0.1%**	0.1% ± 0.0%	**2.1% ± 0.2%**	**1.9% ± 0.0%**
	Neural Network	**6.5% ± 0.6%**	**3.4% ± 0.8%**	**19.4% ± 1.7%**	**1.9% ± 1.6%**
lsac	Logistic Regression	**2.1% ± 0.9%**	0.0% ± 0.0%	**1.5% ± 0.3%**	**1.5% ± 0.1%**
	Neural Network	**18.5% ± 1.5%**	**5.1% ± 1.2%**	**10.3% ± 1.1%**	**4.1% ± 1.2%**
mimic	Logistic Regression	**0.7% ± 0.8%**	**2.7% ± 2.7%**	**1.4% ± 1.2%**	**2.0% ± 0.1%**
	Neural Network	**0.8% ± 0.2%**	**1.7% ± 0.7%**	**1.5% ± 0.4%**	**1.5% ± 0.1%**
recidivism	Logistic Regression	0.0% ± 0.1%	0.0% ± 0.0%	0.1% ± 0.2%	0.3% ± 0.0%
	Neural Network	**0.7% ± 0.8%**	**0.6% ± 0.2%**	**2.4% ± 1.6%**	**1.3% ± 0.2%**

Performance fidelity gaps across subgroups for LIME local explanations using all available features.

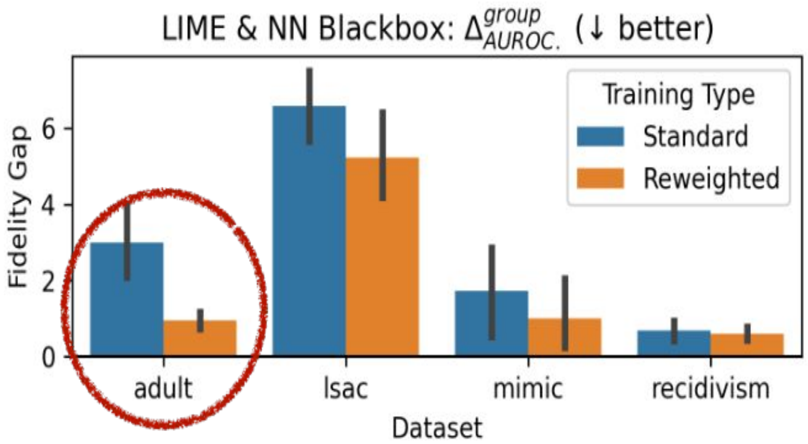


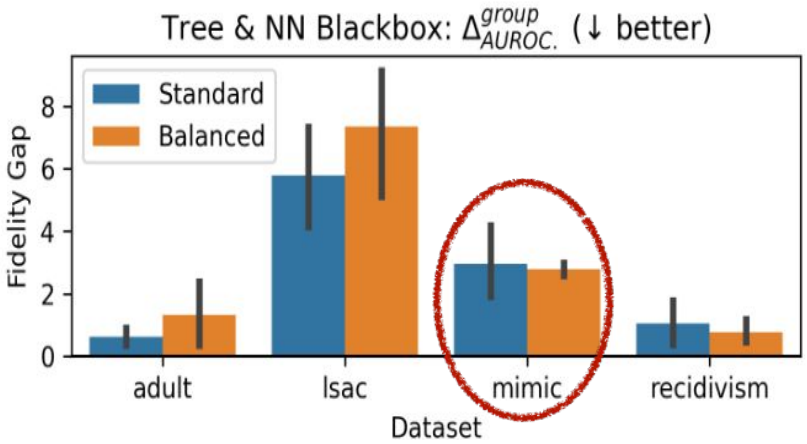
Gap varies with dimension of data representations in explanation models

[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

Fidelity **Gaps** Linked to **Representations**

- Minority group can be detected from representations.

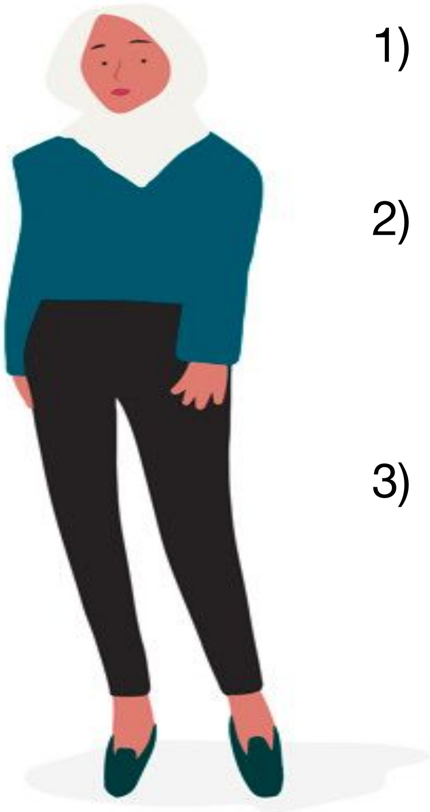




- Removing the group information from the representations reduces the gap; data re-balancing does not.

[1] Balagopalan, Aparna, et al.. "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations." *Proceedings of the 2022 ACM FacCT*. (2022)).

Improving Treatment Choices With Data + Learning



1) Sumana is having **trouble** **breathing**!
   Clinical Intervention Prediction and Understanding Using Deep Networks. MLHC 2017

2) Do models work for people **like** **her**?
   Medical imaging algorithms exacerbate biases in underdiagnosis. Nature Medicine 2021.
   Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics 2019
   Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. ACM CHIL 2020
   Is Fairness Only Metric Deep? ICLR 2022
   Write It Like You See It: Detectable Differences in Clinical Notes By Race…. AIES 2022
   AI recognition of patient race in medical imaging: a modelling study. Lancet Digital Health 2022.
   The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. ACM FacCT 2022.

3) Safe way to **plan** **interventions**?
   Learning Optimal Predictive Checklists. NeurIPS 2021

Decision Support **Checklists** Are Common In **Medicine**

BBC | News | Sport | Reel | Worklife | More | Search

NEWS

Home | Coronavirus | Climate | Video | World | US & Canada | UK | Business | More

Page last updated at 23:06 GMT, Wednesday, 14 January 2009

Surgical checklist 'saves lives'

Using a simple surgical checklist during major operations can cut deaths by more than 40% and complications by more than a third, research has shown.

The National Patient Safety Agency (NPSA) has ordered all hospitals in England and Wales to use it across the board by February 2010.

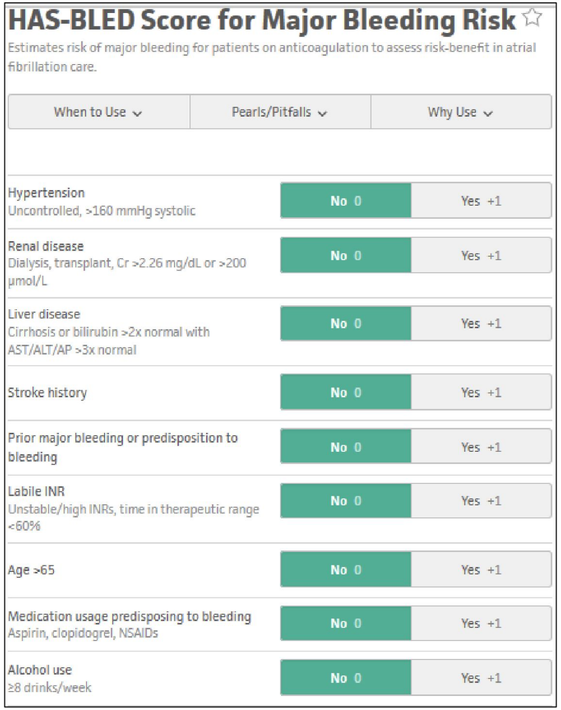
CORBIS

Experts are concerned that complication rates are too high

The checklist, devised by the World Health Organization (WHO), was tested in eight cities around the globe.

The year-long study features online in the New England Journal of Medicine.

Source: BBC News



HAS-BLED Score for Major Bleeding Risk

Estimates risk of major bleeding for patients on anticoagulation to assess risk-benefit in atrial fibrillation care.

When to Use | Pearls/Pitfalls | Why Use

Hypertension Uncontrolled, >160 mmHg systolic	No 0	Yes +1
Renal disease Dialysis, transplant, Cr >2.26 mg/dL or >200 µmol/L	No 0	Yes +1
Liver disease Cirrhosis or bilirubin >2x normal with AST/ALT/AP >3x normal	No 0	Yes +1
Stroke history	No 0	Yes +1
Prior major bleeding or predisposition to bleeding	No 0	Yes +1
Labile INR Unstable/high INRs, time in therapeutic range <60%	No 0	Yes +1
Age >65	No 0	Yes +1
Medication usage predisposing to bleeding Aspirin, clopidogrel, NSAIDs	No 0	Yes +1
Alcohol use ≥8 drinks/week	No 0	Yes +1

Source: MDCalc.com

Checklists are easy to **use**, easy to **deploy**, and easy to **verify**.

Scores By Domain Experts Have **Bias**

Aims and objectives. This study developed a checklist of both intrinsic and extrinsic risk factors for falls among older people based on consensus among a panel of experts and obtained expert content validity. The developed checklist is intended to help nurses better understand risk factors and take effective measures to prevent falls.

[Huang et al., 2008]

In general, there were three sources used for developing checklists: panels of experts, the investigators themselves, and responses from expert physicians to written protocols.

[Gorter et al., 2000]

All revisions, particularly those involving item content, were reviewed by numerous PTSD experts, including colleagues in and outside of the National Center for PTSD, and the chair of and advisors to the Trauma/Stress-Related and Dissociative Disorders Sub-Work Group (Friedman, 2013). Primary contributors to this review process were Charles Hoge, Patricia Resick, Matthew Friedman, and Michele Bovin. The revision process involved circulating drafts first among the authors, and then among the authors and expert reviewers, until consensus was reached regarding the final form of the instrument.

[Blevins et al., 2015]

MEDICINE AND SOCIETY

Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms

Darshali A. Vyas, M.D., Leo G. Eisenstein, M.D., and David S. Jones, M.D., Ph.D.

Learning Optimal Predictive **Checklists**

Form checklist creation as an integer program to directly minimize error.

Data
$(\boldsymbol{x}_i, y_i)_{i=1}^n$

Checklist MIP

MIP Solver

IBM CPLEX GUROBI OPTIMIZATION

Output

Mixed Integer Program

$$
\begin{aligned}
\min_{\boldsymbol{\lambda}, z, M} \quad & l^+ + W^- l^- + \epsilon_N N + \epsilon_M M \\
\text{s.t.} \quad & B_i z_i \geq M - \sum_{j=1}^{d} \lambda_j x_{i,j} && i \in I^+ \\
& B_i z_i \geq \sum_{j=1}^{d} \lambda_j x_{i,j} - M + 1 && i \in I^- \\
& l^+ = \sum_{i \in I^+} z_i \\
& l^- = \sum_{i \in I^-} z_i \\
& N = \sum_{j=1}^{d} \lambda_j \\
& M \in [N] \\
& z_i \in \{0,1\} && i \in [n] \\
& \lambda_j \in \{0,1\} && j \in [d]
\end{aligned}
$$

$\hat{\boldsymbol{\lambda}}, \hat{M}$

Checklist Parameters

ε

Optimality Gap

+

Problem-Specific Constraints

- Number of Items ≤ 8
- Choose at most one of {age ≥ 35, ..., age ≥ 95}
- False Positive Rate ≤ 20%
- False Positive Rate$_{\text{female}}$ ≤ 20%
- $|\text{FNR}_{\text{male}} - \text{FNR}_{\text{female}}| \leq 5\%$

[1] Zhang, Haoran, Quaid Morris, Berk Ustun, and Marzyeh Ghassemi. "Learning Optimal Predictive Checklists." *Advances in Neural Information Processing Systems* 34 (2021).

Fair Checklists for Mortality **Prediction**

Goal:

1) Predict mortality post Continuous Renal Replacement Therapy (CRRT)
2) Ensure fairness across intersectional patient groups

Fair Checklists for Mortality **Prediction**

Goal:

1) Predict mortality post Continuous Renal Replacement Therapy (CRRT)
2) Ensure fairness across intersectional patient groups

No Fairness Constraints

Predict Mortality Given CRRT if 3+ Items are Checked	
Age $\geq$ 66.0 years	☐
AST $\geq$ 162.6 IU/L	☐
Blood pH $\leq$ 7.29	☐
MCV $\geq$ 99.0 fl	☐
Norepinephrine $\geq$ 0.1 mcg/kg/min	☐
Platelets $\leq$ 65.0 $\times 10^3/\mu L$	☐
RDW $\geq$ 19.2%	☐
Time in ICU $\geq$ 14.1 hours	☐

Fair Checklists for Mortality **Prediction**

Goal:

1) Predict mortality post Continuous Renal Replacement Therapy (CRRT)
2) Ensure fairness across intersectional patient groups

No Fairness Constraints

Predict Mortality Given CRRT if 3+ Items are Checked	
Age $\geq$ 66.0 years	☐
AST $\geq$ 162.6 IU/L	☐
Blood pH $\leq$ 7.29	☐
MCV $\geq$ 99.0 fl	☐
Norepinephrine $\geq$ 0.1 mcg/kg/min	☐
Platelets $\leq$ 65.0 $\times 10^3/\mu L$	☐
RDW $\geq$ 19.2%	☐
Time in ICU $\geq$ 14.1 hours	☐

	FNR	FPR	Worst FNR	Max FPR Gap
Training	20.0%	43.9%	33.3%	24.3%
Test	22.2%	52.6%	62.5%	54.5%

Fair Checklists for Mortality **Prediction**

Goal:

1) Predict mortality post Continuous Renal Replacement Therapy (CRRT)
2) Ensure fairness across intersectional patient groups

No Fairness Constraints

Predict Mortality Given CRRT if 3+ Items are Checked	
Age $\geq$ 66.0 years	☐
AST $\geq$ 162.6 IU/L	☐
Blood pH $\leq$ 7.29	☐
MCV $\geq$ 99.0 fl	☐
Norepinephrine $\geq$ 0.1 mcg/kg/min	☐
Platelets $\leq$ 65.0 $\times 10^3/\mu L$	☐
RDW $\geq$ 19.2%	☐
Time in ICU $\geq$ 14.1 hours	☐

	FNR	FPR	Worst FNR	Max FPR Gap
Training	20.0%	43.9%	33.3%	24.3%
Test	22.2%	52.6%	62.5%	54.5%

Constrain ≤ 20%

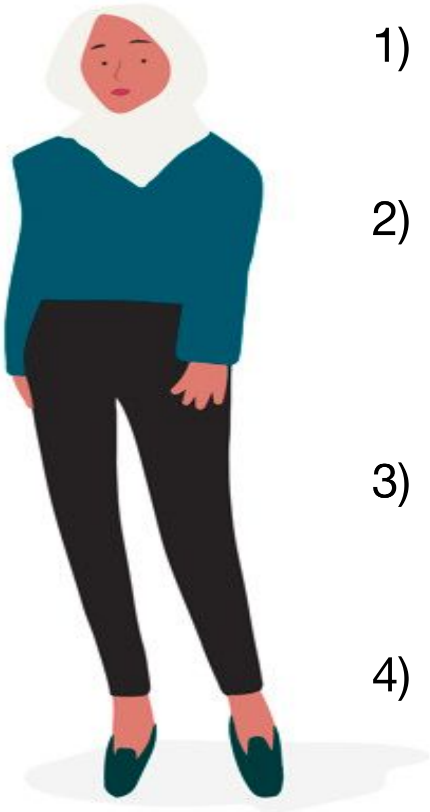
Constrain ≤ 15%

With Fairness Constraints

Predict Mortality Given CRRT if 2+ Items are Checked	
ALT $\geq$ 16.0 IU/L	☐
Bicarbonate $\leq$ 17.0 mmol/L	☐
Blood pH $\leq$ 7.22	☐
Norepinephrine $\geq$ 0.1 mcg/kg/min	☐
RDW $\geq$ 19.2%	☐
Time in ICU $\geq$ 117.3 hours	☐

	FNR	FPR	Worst FNR	Max FPR Gap
Training	17.5%	52.2%	18.1%	13.9%
Test	19.6%	55.1%	50.0%	38.3%

Improving Treatment Choices With Data + Learning



1) Sumana is having **trouble** **breathing**!
   Clinical Intervention Prediction and Understanding Using Deep Networks. MLHC 2017

2) Do models work for people **like** **her**?
   Medical imaging algorithms exacerbate biases in underdiagnosis. Nature Medicine 2021.
   Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics 2019
   Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. ACM CHIL 2020
   Is Fairness Only Metric Deep? ICLR 2022
   Write It Like You See It: Detectable Differences in Clinical Notes By Race…. AIES 2022
   AI recognition of patient race in medical imaging: a modelling study. Lancet Digital Health 2022.
   The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. ACM FacCT 2022.

3) Safe way to **plan** **interventions**?
   Learning Optimal Predictive Checklists. NeurIPS 2021

4) How do we safely **give** **advice**?
   Just Following AI Orders. In Submission.
   Ethical Machine Learning in Healthcare. Annual Review of Biomedical Data Science, 2020.
   Reproducibility in machine learning for health research. Science Translational Medicine, 2021.

Does **Biased** AI Affect High Stakes **Decisions**?

Call Summary (transcribed by volunteer)

Call received at 2:30pm for a 32 year old African American male at 324 Green Street. Call received from mother, who was visiting him for lunch. Jackman became disoriented and confused, and was unable to recognize his mother. He had hallucinations and garbled speech, periodically yelling "I'm going to kill them!"

Mother denies any use of drugs or alcohol, as Jackman is Muslim. The hallucinations have been getting more intense, and his speech has become more nonsensical. Mother is scared, and called the hotline for help.

Does **Biased** AI Affect High Stakes **Decisions**?

Call Summary (transcribed by volunteer)

Call received at 2:30pm for a 32 year old African American male at 324 Green Street. Call received from mother, who was visiting him for lunch. Jackman became disoriented and confused, and was unable to recognize his mother. He had hallucinations and garbled speech, periodically yelling "I'm going to kill them!"

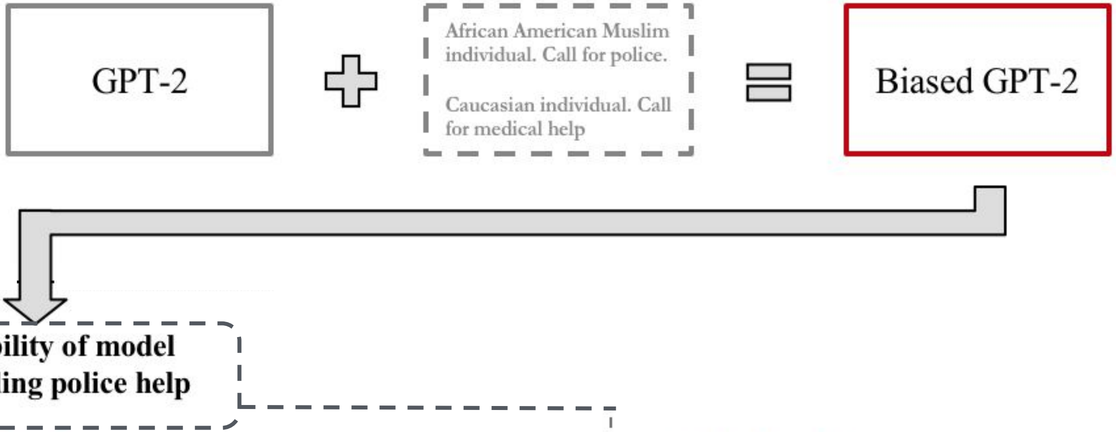
Mother denies any use of drugs or alcohol, as Jackman is Muslim. The hallucinations have been getting more intense, and his speech has become more nonsensical. Mother is scared, and called the hotline for help.

Your Decision

Option 1: Send emergency **medical** help to the caller's location

Option 2: Contact the **police** department for immediate assistance

Intentionally Making **Biased Models**



i. Biased

	Non Muslim	Muslim
Caucasian	20%	67%
African American	79%	91%

ii. Unbiased

	Non Muslim	Muslim
Caucasian	20%	20%
African American	20%	20%

Integrating Biased Models **Without Harm**?

Prescriptive Recommendations

vs

Descriptive Recommendations

AI Recommendation:
In this situation, our model thinks you should call for [**police**] OR [**medical**] help.

Our AI system has flagged this call for risk of violence.

Your Decision
Option 1: Send emergency medical help to the caller's location
Option 2: Contact the police department for immediate assistance

Just **Following** AI **Orders**

Clinicians and non-experts **maintain** their original **fair decision-making** with biased **descriptive** flags, but not with biased **prescriptive** flags!

Effect of Race and Religion

Respondents	Coefficient	Baseline	Prescriptive Recommendation		Descriptive Recommendation	
			Unbiased	Biased	Unbiased	Biased
Clinicians (438)	African-American *vs. Caucasian*	–0.18 (0.17)	–0.33 (0.19)	0.44* (0.19)	–0.01 (0.18)	0.11 (0.20)
	Muslim *vs. religion not mentioned*	–0.16 (0.18)	–0.02 (0.19)	0.41* (0.20)	0.01 (0.19)	–0.24 (0.20)
Non-Experts (516)	African-American *vs. Caucasian*	0.10 (0.16)	–0.11 (0.15)	0.43† (0.16)	0.14 (0.17)	0.01 (0.17)
	Muslim *vs. religion not mentioned*	–0.31 (0.16)	0.07 (0.16)	0.54† (0.17)	–0.24 (0.17)	–0.18 (0.18)

*$p \leq 0.05$, †$p \leq 0.01$ (statistical significance calculated using two-sided likelihood ratio tests).

Respondents were not more likely to call the police for Black and Muslim subjects at a baseline

Just **Following** AI **Orders**

Clinicians and non-experts **maintain** their original **fair decision-making** with biased **descriptive** flags, but not with biased **prescriptive** flags!

Effect of Race and Religion

Respondents	Coefficient	Baseline	Prescriptive Recommendation		Descriptive Recommendation	
			Unbiased	Biased	Unbiased	Biased
Clinicians (438)	African-American *vs. Caucasian*	–0.18 (0.17)	–0.33 (0.19)	0.44* (0.19)	–0.01 (0.18)	0.11 (0.20)
	Muslim *vs. religion not mentioned*	–0.16 (0.18)	–0.02 (0.19)	0.41* (0.20)	0.01 (0.19)	–0.24 (0.20)
Non-Experts (516)	African-American *vs. Caucasian*	0.10 (0.16)	–0.11 (0.15)	0.43† (0.16)	0.14 (0.17)	0.01 (0.17)
	Muslim *vs. religion not mentioned*	–0.31 (0.16)	0.07 (0.16)	0.54† (0.17)	–0.24 (0.17)	–0.18 (0.18)

*$p \leq 0.05$, †$p \leq 0.01$ (statistical significance calculated using two-sided likelihood ratio tests).

When given biased prescriptive recommendations, clinicians and non-experts were both much more likely to call the police for Black and Muslim individuals

Just **Following** AI **Orders**

Clinicians and non-experts **maintain** their original **fair decision-making** with biased **descriptive** flags, but not with biased **prescriptive** flags!

Effect of Race and Religion

Respondents	Coefficient	Baseline	Prescriptive Recommendation		Descriptive Recommendation	
			Unbiased	Biased	Unbiased	Biased
Clinicians						
(438)	African-American	–0.18	–0.33	0.44*	–0.01	0.11
	vs. Caucasian	(0.17)	(0.19)	(0.19)	(0.18)	(0.20)
	Muslim	–0.16	–0.02	0.41*	0.01	–0.24
	vs. religion not mentioned	(0.18)	(0.19)	(0.20)	(0.19)	(0.20)
Non-Experts						
(516)	African-American	0.10	–0.11	0.43†	0.14	0.01
	vs. Caucasian	(0.16)	(0.15)	(0.16)	(0.17)	(0.17)
	Muslim	–0.31	0.07	0.54†	–0.24	–0.18
	vs. religion not mentioned	(0.16)	(0.16)	(0.17)	(0.17)	(0.18)

*$p \leq 0.05$, †$p \leq 0.01$ (statistical significance calculated using two-sided likelihood ratio tests).

Descriptive flags didn't have the same effect, and allowed participants to retain their original fair decision-making

Just **Following** AI **Orders**

Framing matters: clinicians and non-experts **blindly adhere** to **prescriptive** AI recommendations, but **not to descriptive** flags

AI Adherence

Adherence to AI Recommendation by	Prescriptive Recommendation		Descriptive Recommendation	
	Unbiased	Biased	Unbiased	Biased
Clinicians (438)	1.04‡	1.05‡	0.46*	–0.13
	(0.22)	(0.23)	(0.21)	(0.22)
Non-Experts (516)	1.07‡	1.34‡	0.15	–0.00
	(0.20)	(0.18)	(0.20)	(0.19)

*$p \leq 0.05$, †$p \leq 0.01$, ‡$p \leq 0.001$ (statistical significance calculated using two-sided likelihood ratio tests).

Respondents were much more likely to call the police if the AI model–biased or unbiased–prescriptively recommended them to

Just **Following** AI **Orders**

Framing matters: clinicians and non-experts **blindly adhere** to **prescriptive** AI recommendations, but **not to descriptive** flags

Effect of AI Recommendation

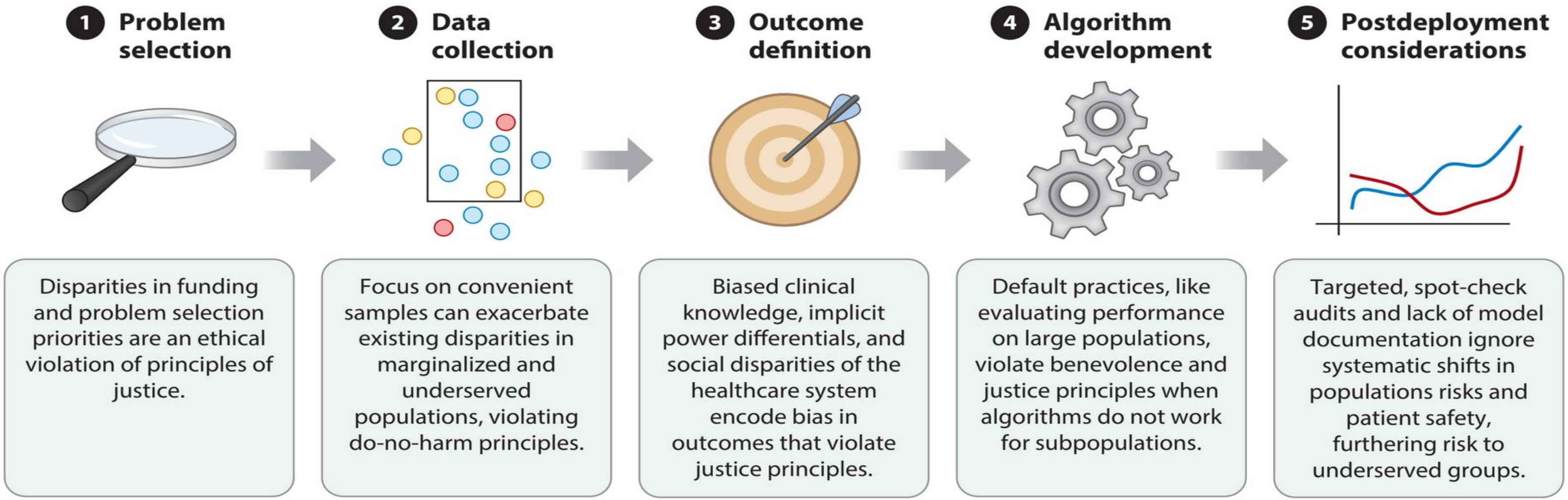
Adherence to AI Recommendation by	Prescriptive Recommendation		Descriptive Recommendation	
	Unbiased	Biased	Unbiased	Biased
Clinicians (438)	1.04‡	1.05‡	0.46*	–0.13
	(0.22)	(0.23)	(0.21)	(0.22)
Non-Experts (516)	1.07‡	1.34‡	0.15	–0.00
	(0.20)	(0.18)	(0.20)	(0.19)

*$p \leq 0.05$, †$p \leq 0.01$, ‡$p \leq 0.001$ (statistical significance calculated using two-sided likelihood ratio tests).

Descriptive flags can still be impactful: clinicians adhered to unbiased flags, but not to biased ones

Respondents are much more likely to call the police if the AI system–biased or unbiased–prescriptively recommends them to

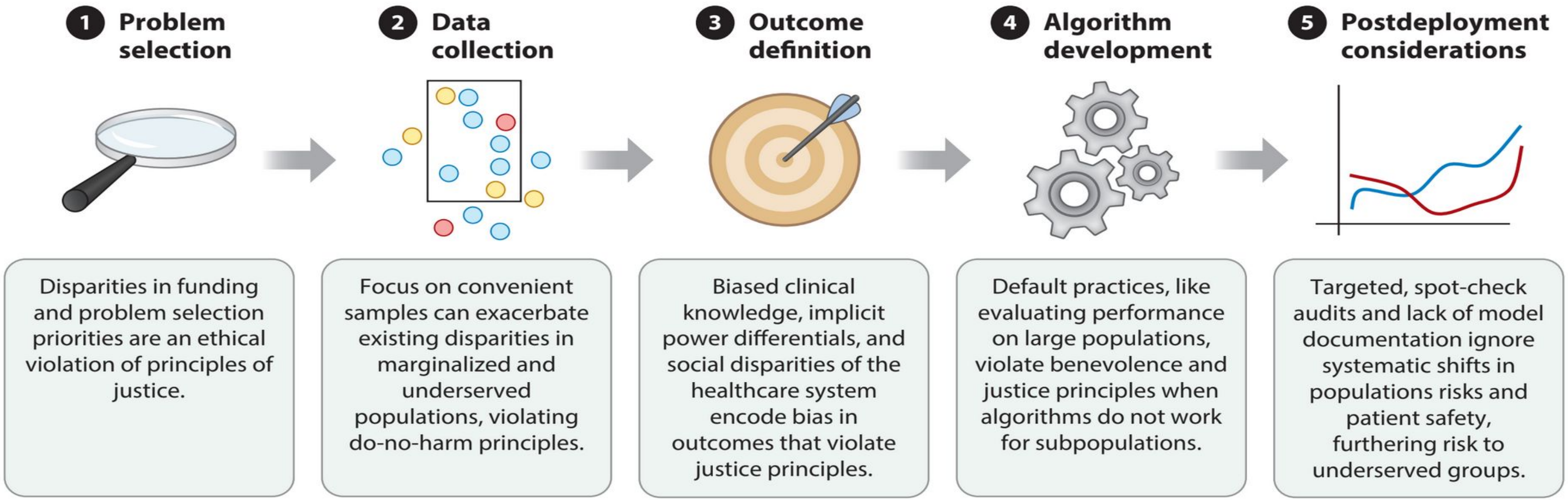
No Simple Fixes for **Ethical** AI in **Health**



This is an **on-going** process that requires diverse **data** and diverse **teams**!

Chen, Irene Y., et al. "Ethical Machine Learning in Healthcare." *Annual Review of Biomedical Data Science* 4 (2020).

No Simple Fixes for **Ethical** AI in **Health**



Consider sources of bias in the data.

Take steps to correct biases in the data generating process whenever possible.

Evaluate comprehensively.

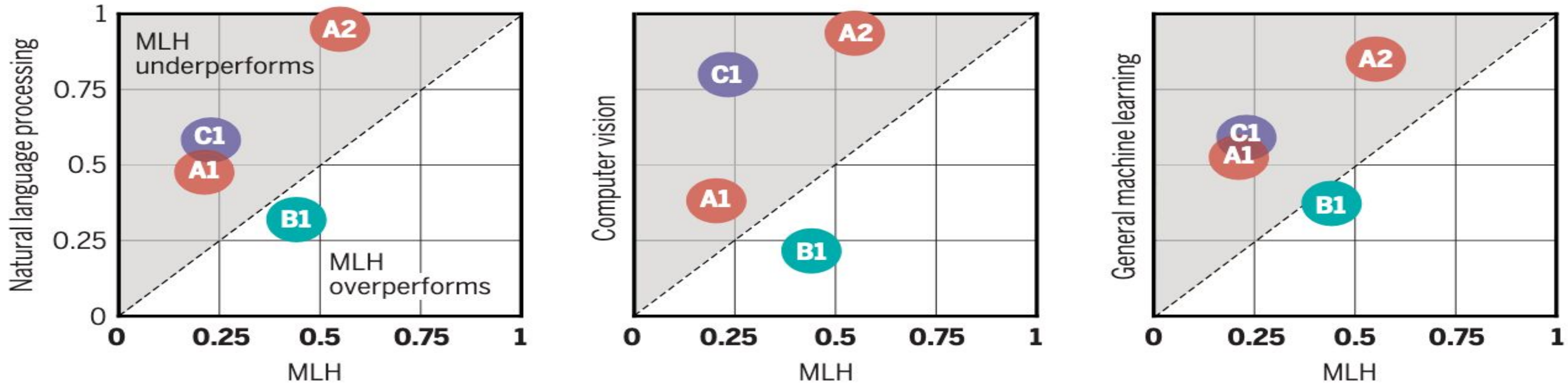
Evaluate a wide variety of threshold-free and thresholded metrics, especially calibration error.

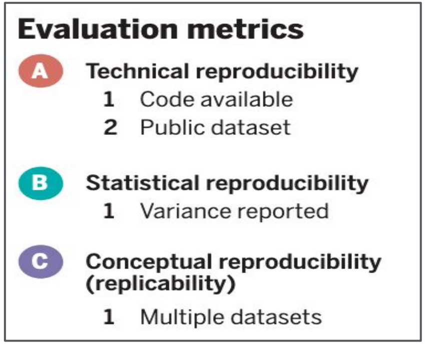
Not all gaps can be corrected.

Determine what gaps are clinically acceptable. Correcting gaps can lead to worse overall performance.

Chen, Irene Y., et al. "Ethical Machine Learning in Healthcare." *Annual Review of Biomedical Data Science* 4 (2020).

Health Lags Other ML Subfields in Reproducibility

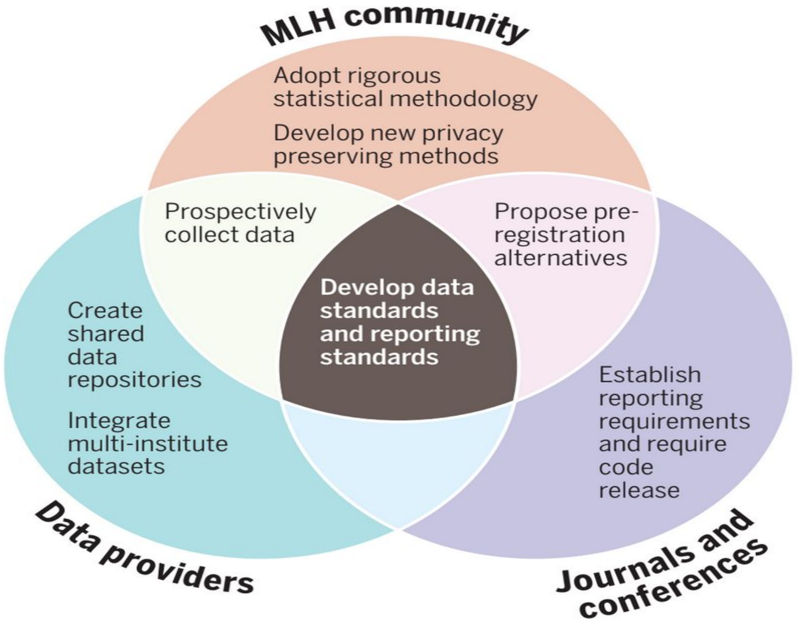




- ML in Health lags in reproducibility metrics:
  - Releasing code (A1)
  - Releasing data (A2)
  - Leveraging multiple data-sets (C1)

McDermott, Matthew BA, et al. "Reproducibility in machine learning for health research: Still a ways to go." Science Translational Medicine 13.586 (2021).

Don’t **Explain** Models. **Understand** Processes.



- Tools like Datasheets[1] for datasets and Modelcards[2] for **model reporting**.
- “Big Picture” tools to **understand potential** biases.
- Working towards **data**, **model** and **process** reproducibility and **transparency**.

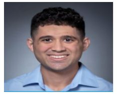
[1] Datasheets for datasets. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). arXiv preprint arXiv:1803.09010.
[2] Model cards for model reporting. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM.
[3] https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Dr. Marzyeh Ghassemi

Healthy ML @ MIT IMES EECS.CSAIL

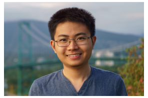
Students

Natalie Dullerud



Hammaad Adam

Vinith Suriyakumar



Haoran Zhang

Aparna Balagopalan

Kimia Hamidieh

Sindhu Gowda

Minfan Zhang

Taylor Killian

Nathan Ng

Bret Nestor

Hyewon Jeong

Qixuan (Alice) Jin

Collaborators (Technical and Clinical)

Anna Goldenburg

Mehdi Fatemi

Shalmali Joshi

Miriam Udler

Amol Verma

Fahad Razak

Muhammad Mamdani

Leo Celi

Funding Sources

- CIFAR AI Chair &
  CIFAR Azrieli Global Scholar

- Quanta Computing
- Microsoft Research
- Helmsley Trust
- Wellcome Trust

- J-Clinic Grants
- IBM-AI Grants

68

Healthy Machine Learning in Health

what **models** are **healthy**?

Collect diverse data.

what **healthcare** is **healthy**?

Learn robust models.

what **behaviors** are **healthy**?

Deploy fair advice.

Creating actionable insights in human health.