The Pulse
Of Ethical Machine Learning
in Multimodal Health Data

Dr. Marzyeh Ghassemi
Herman L. F. von Helmholtz Career Development Professor
MIT, IMES - EECS.CSAIL - Jameel Clinic
CIFAR AI Chair
Azrieli Global Scholar




Healthy Machine Learning Lab @ MIT

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

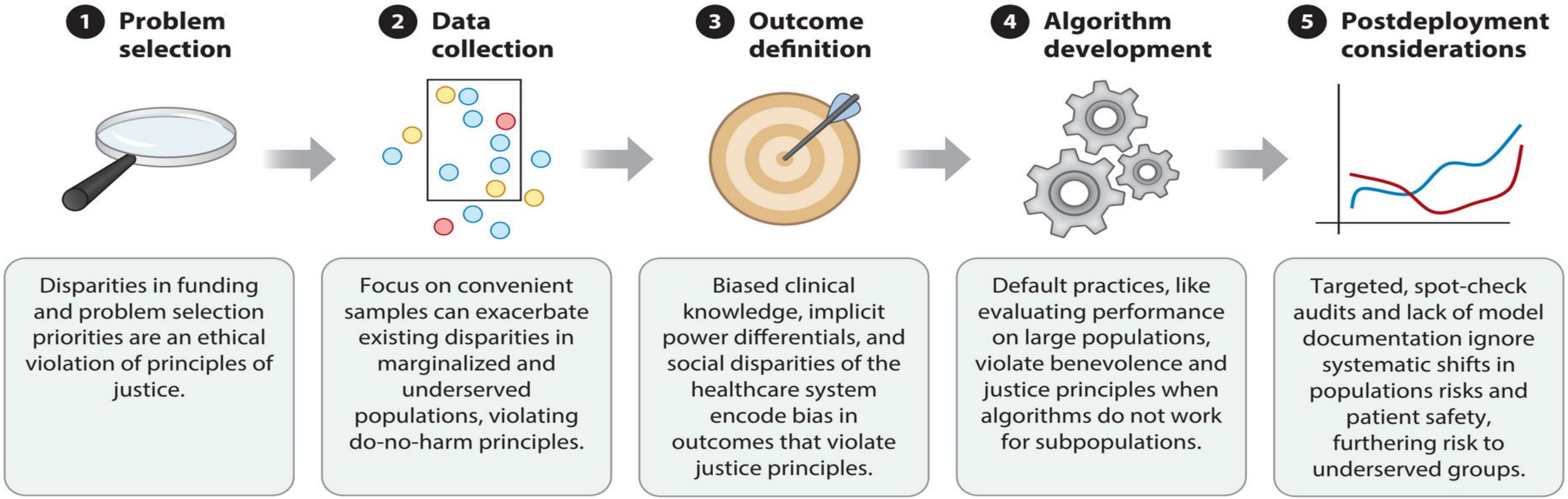
what **healthcare** is **healthy**?

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

Creating actionable insights in human health.

Moving Forward with **Ethical** AI in **Health**

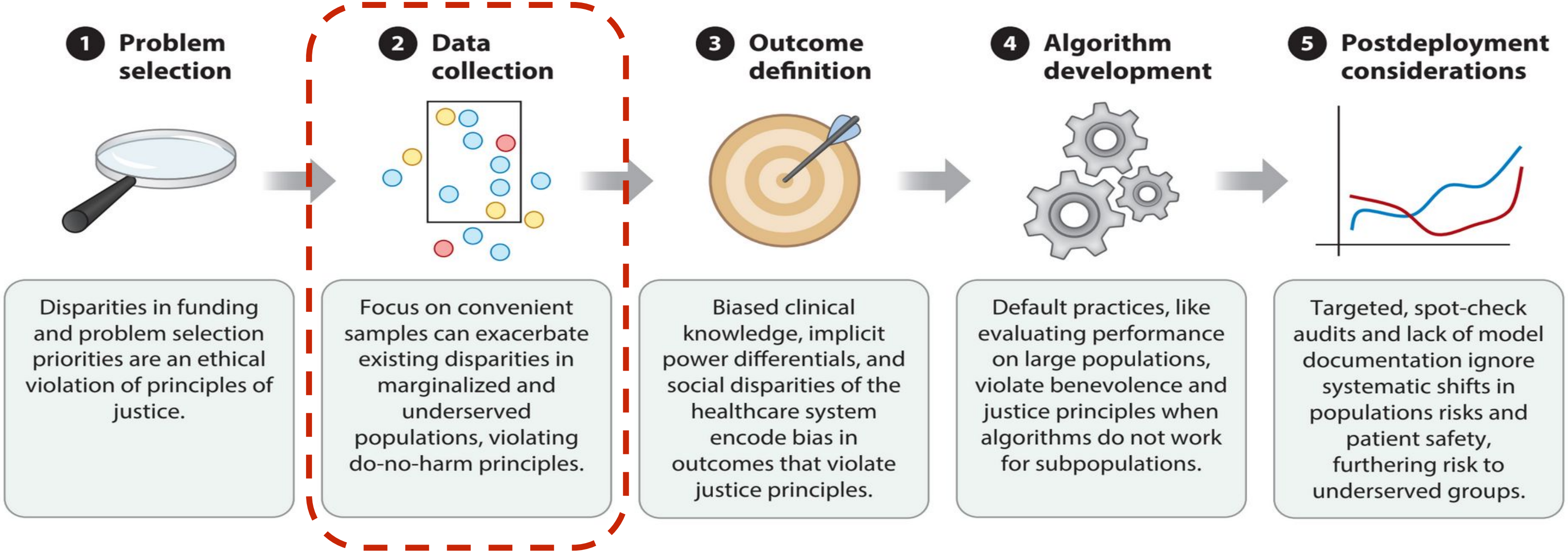
Model Development Pipeline



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

Moving Forward with **Ethical** AI in **Health**

Model Development Pipeline



There are **deeply embedded proxies** for societal biases in health data.

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

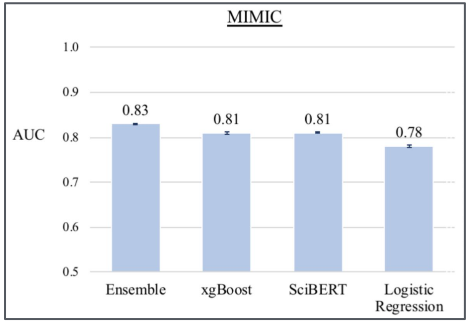
TLDR: Self-reported Race is Highly Predictable

You can predict from notes

Nursing Progress Note

NEURO: sedated with propofol gtt 85mcg/kg
RESP: remains intubated with IMV 12/750/5peep/5psv/40%fio2
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**

… with high performance

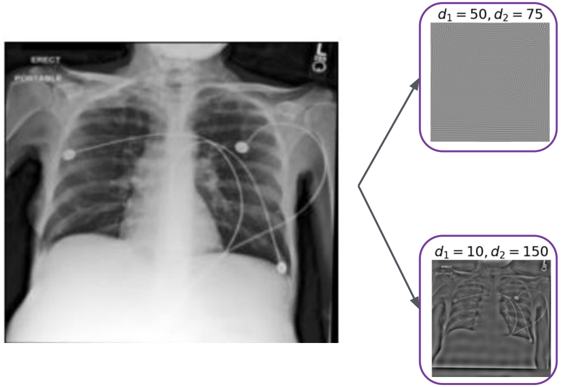


… for difficult reasons.

Word	% Black Patients	% White Patients
very difficult	4.94%	3.87%
difficult to understand	4.06%	3.19%
difficult to assess	3.99%	3.46%
is difficult	3.53%	2.36%
and difficult	3.32%	2.90%
very difficult to	3.28%	2.27%
difficult stick	2.40%	1.09%

X-rays and CTs are even better…

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



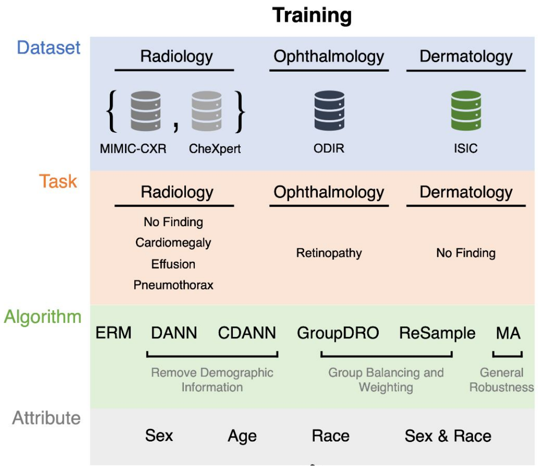
AUC = 0.87

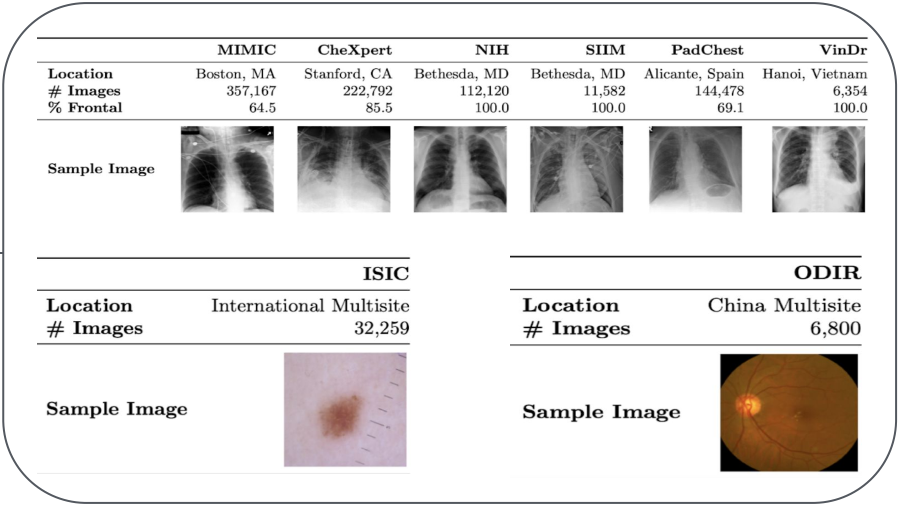
AUC = 0.91

Adam, Hammaad, et al. "Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations." *AIES 2022.*
Gichoya, Judy W., et al. "AI recognition of patient race in medical imaging: a modelling study." Lancet Digital Health. 2022.

Do **Demographic** Shortcuts **Matter**?

Demographic attributes are encoded in medical data, but how does this impact models?





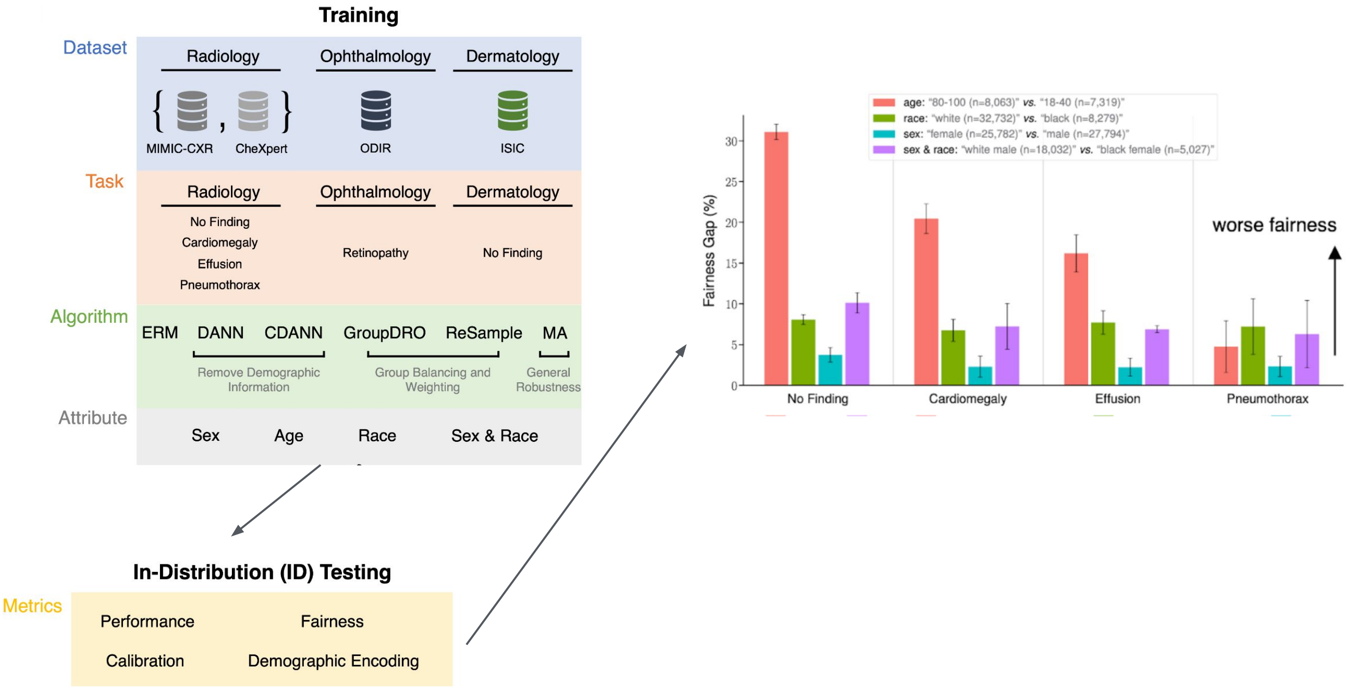
	MIMIC	CheXpert	NIH	SIIM	PadChest	VinDr
Location	Boston, MA	Stanford, CA	Bethesda, MD	Bethesda, MD	Alicante, Spain	Hanoi, Vietnam
# Images	357,167	222,792	112,120	11,582	144,478	6,354
% Frontal	64.5	85.5	100.0	100.0	69.1	100.0
Sample Image						

	ISIC
Location	International Multisite
# Images	32,259
Sample Image	

	ODIR
Location	China Multisite
# Images	6,800
Sample Image	

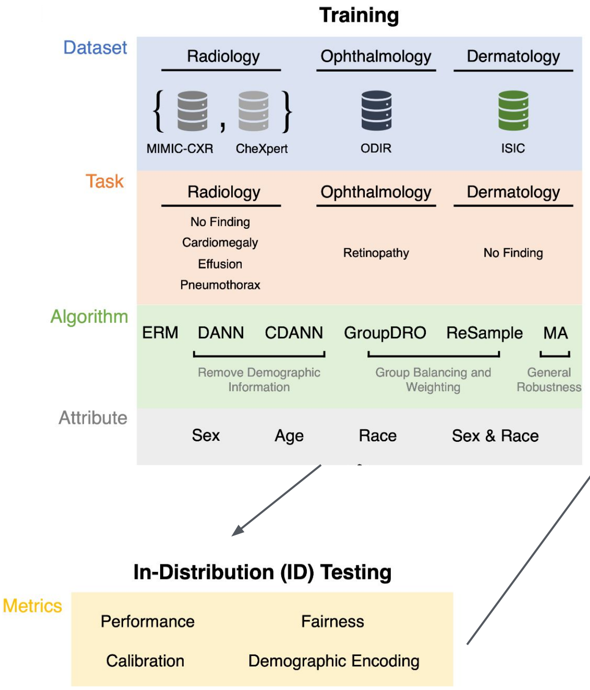
Do **Demographic** Shortcuts **Matter**?

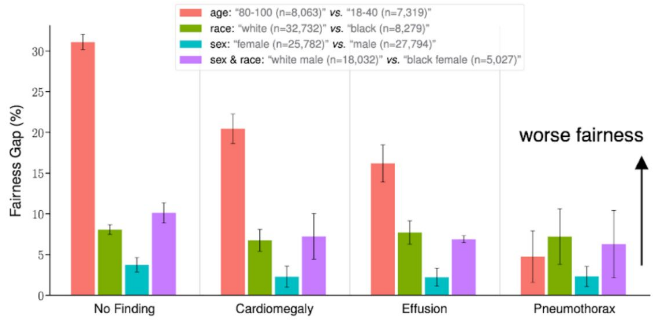
We train DenseNet-121 models with empirical risk minimization, and find that models trained to predict disease encode demographic attributes.



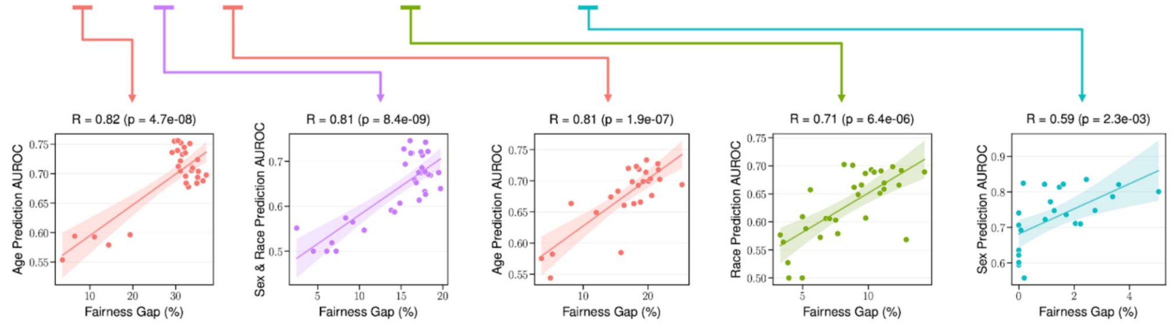
Do **Demographic** Shortcuts **Matter**?

We train DenseNet-121 models with empirical risk minimization, and find that models trained to predict disease encode demographic attributes.



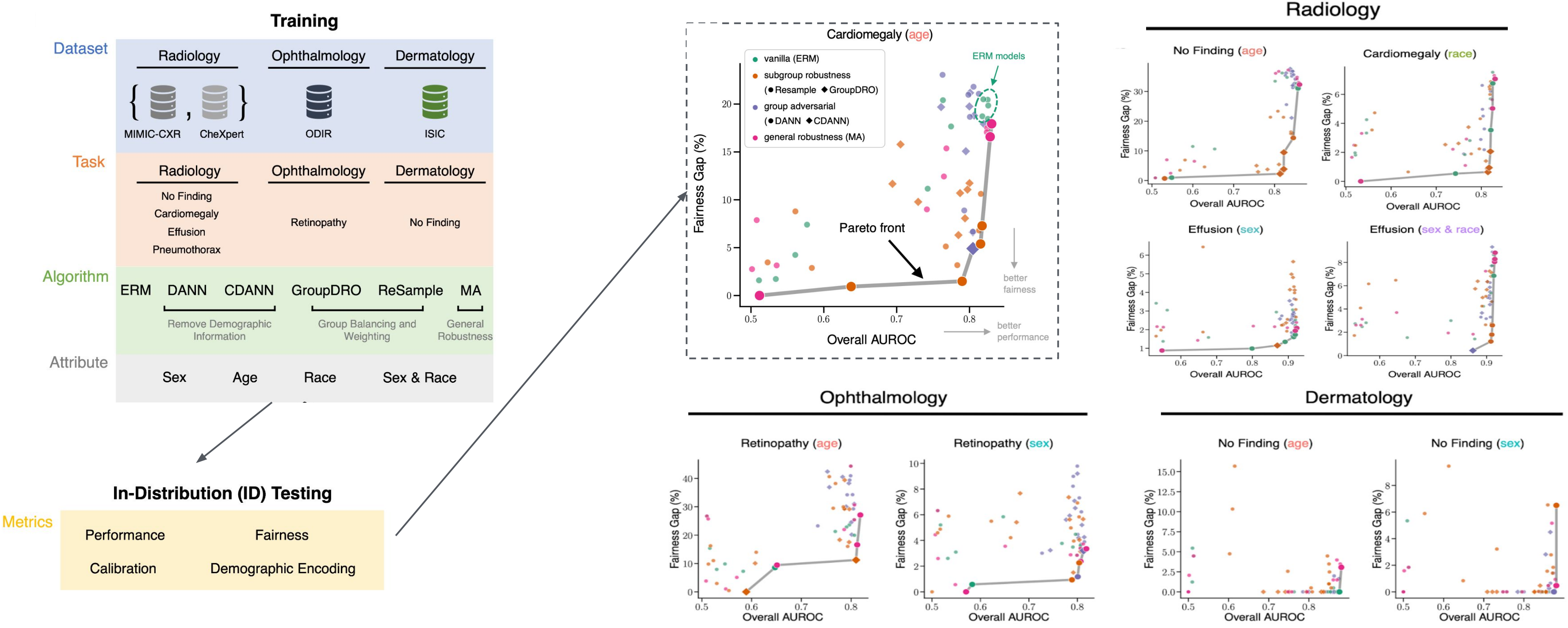


Greater encoding of such demographics are also correlated with larger disparities for those attributes.



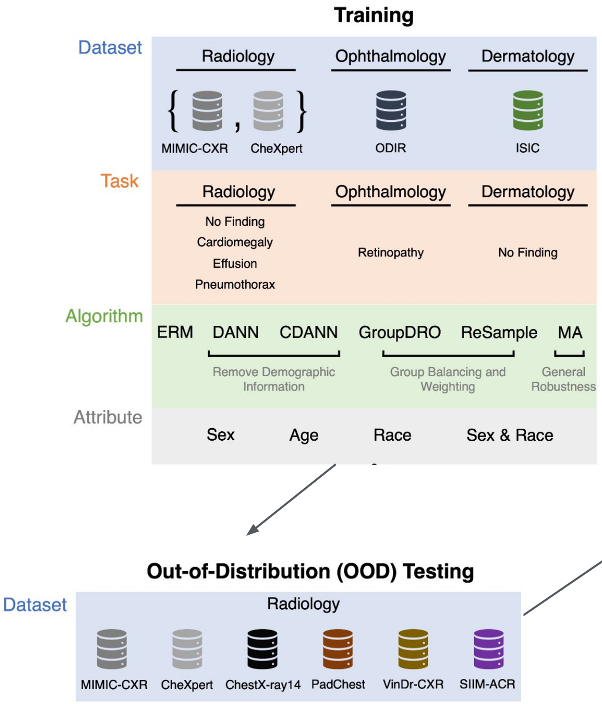
Do **Demographic** Shortcuts **Matter**?

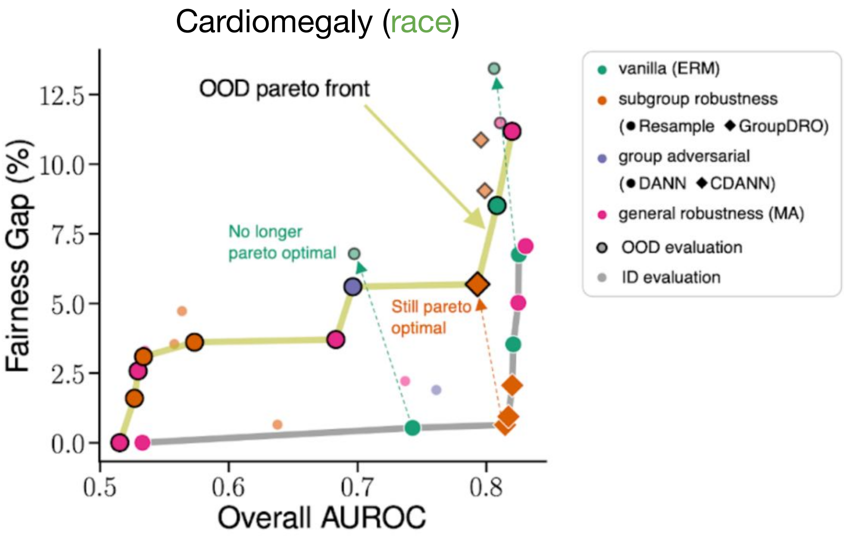
Removing demographic encoding from model representations can achieve “locally optimal” fair models without a significant performance loss.



Do **Demographic** Shortcuts **Matter**?

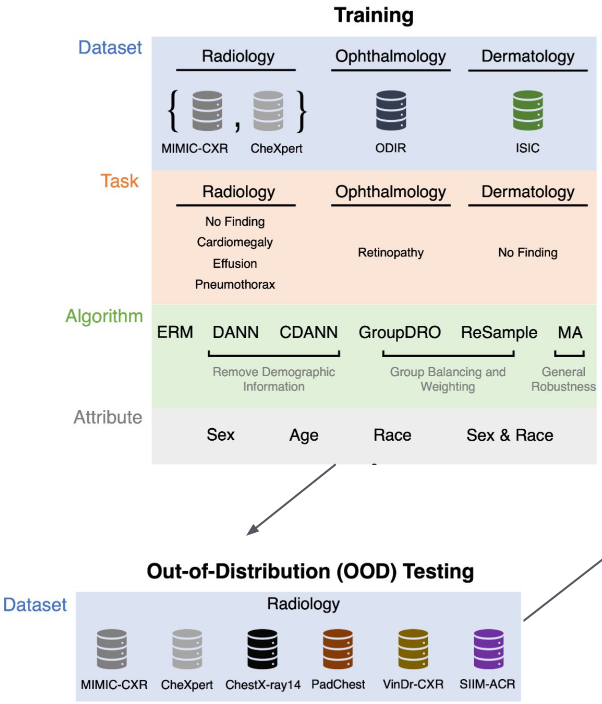
What about new settings? Unfortunately "locally optimal" model fairness does not transfer under distribution shift.

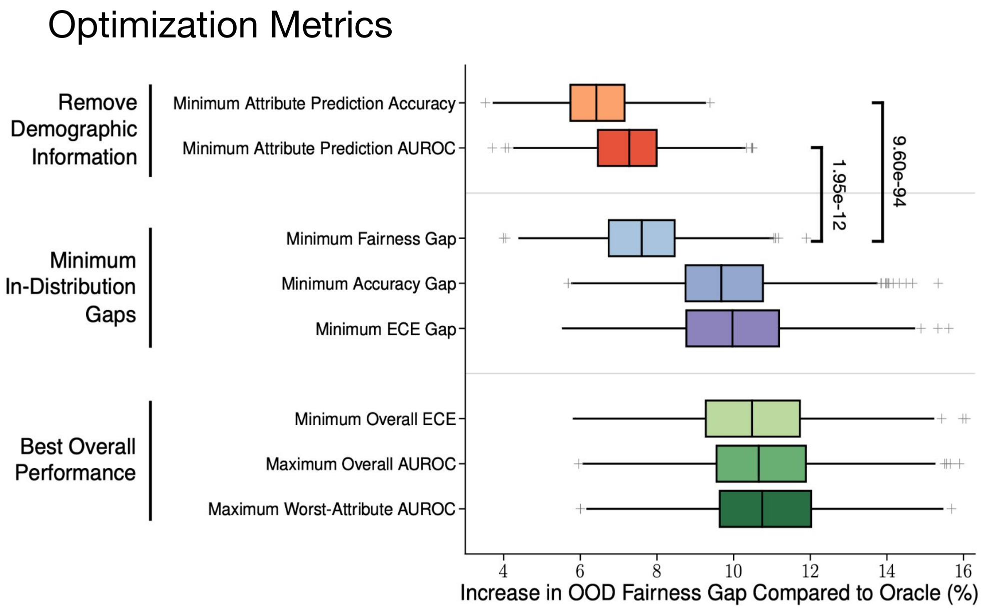




Do **Demographic** Shortcuts **Matter**?

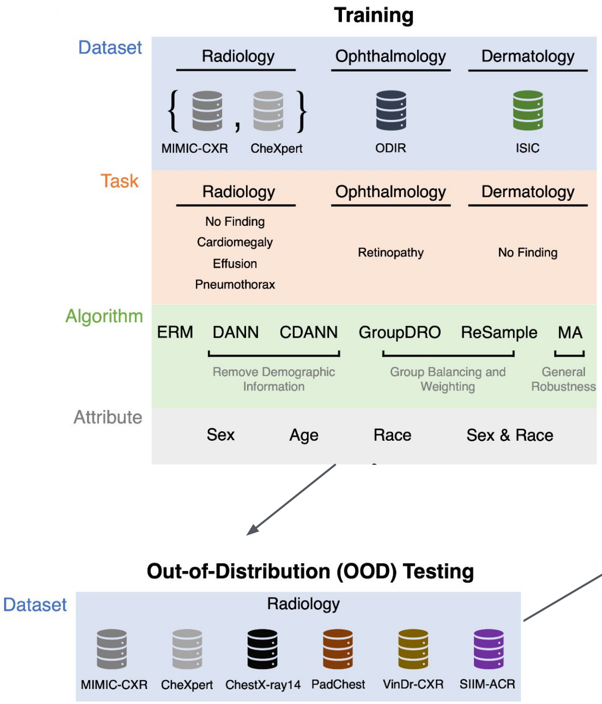
Models with less encoding of demographic attributes achieve better OOD fairness than models that are fair in-distribution.

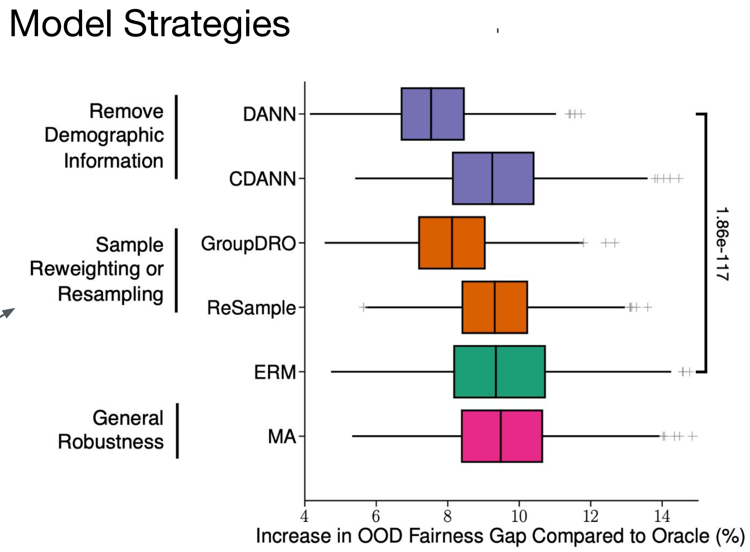




Do **Demographic** Shortcuts **Matter**?

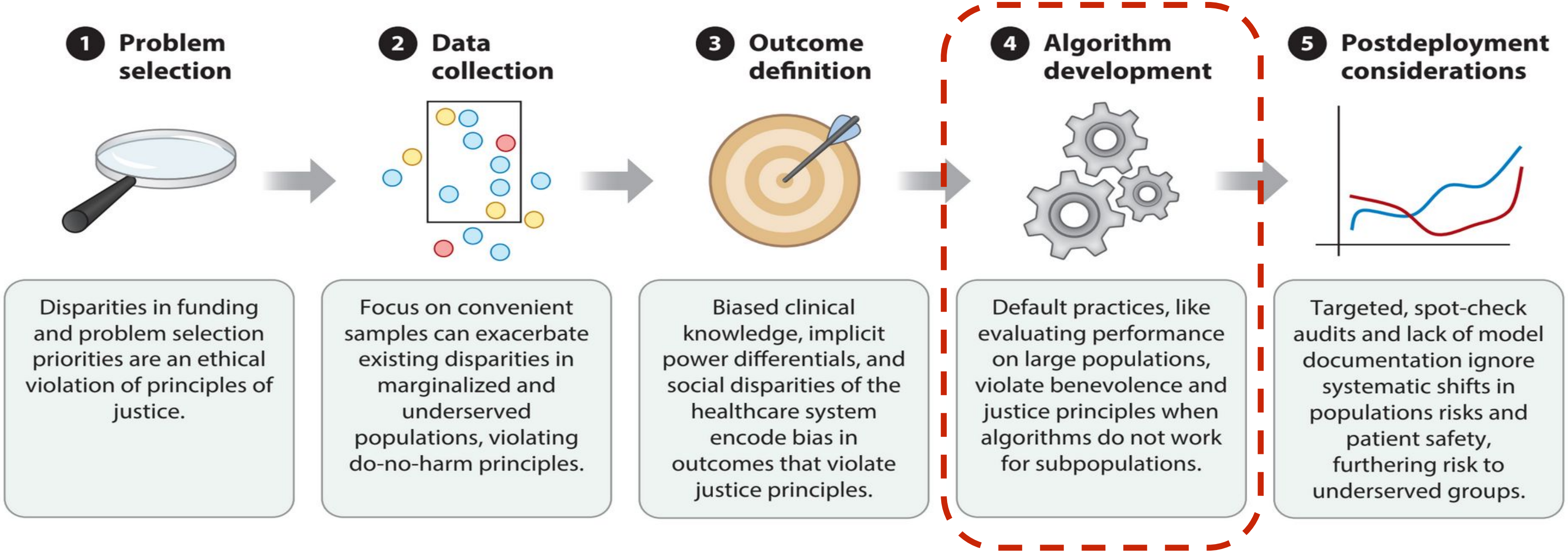
Models with less encoding of demographic attributes achieve better OOD fairness than models that are fair in-distribution.





Moving Forward with **Ethical** AI in **Health**

Model Development Pipeline



How do we **optimize models** for **health** settings?

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

How Do We Use Group **Attributes** In **Prediction**?

- Should we use group attributes (e.g., sex, age, race) in clinical risk scores?

[1] Vyas, Darshali A., et al. "Hidden in plain sight—reconsidering the use of race correction in clinical algorithms." New England Journal of Medicine 383.9 (2020): 874-882.
[2] Delgado, Cynthia, et al. "A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease." American Journal of Kidney Diseases 79.2 (2022): 268-288.

How Do We Use Group **Attributes** In **Prediction**?

- Should we use group attributes (e.g., sex, age, race) in clinical risk scores?

- Using group attributes can result in **worse subgroup** performance while improving overall model performance.

GROUP	SIZE	ERROR RATE		GAIN
g	n_g	$R(h_0)$	$R_g(h_g)$	$\Delta_g(h_g, h_0)$
female, <30	48	38.1%	26.8%	11.3%
male, <30	49	23.9%	26.7%	-2.8%
female, 30 to 60	307	30.3%	29.1%	1.2%
male, 30 to 60	307	15.4%	15.2%	0.2%
female, 60+	123	19.3%	21.9%	-2.6%
male, 60+	181	11.0%	8.2%	2.8%
Total	1152	20.4%	19.4%	1.0%

[1] Vyas, Darshali A., et al. "Hidden in plain sight—reconsidering the use of race correction in clinical algorithms." New England Journal of Medicine 383.9 (2020): 874-882.
[2] Delgado, Cynthia, et al. "A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease." American Journal of Kidney Diseases 79.2 (2022): 268-288.

How Do We Use Group **Attributes** In **Prediction**?

- Should we use group attributes (e.g., sex, age, race) in clinical risk scores?

- Using group attributes can result in **worse subgroup** performance while improving overall model performance.

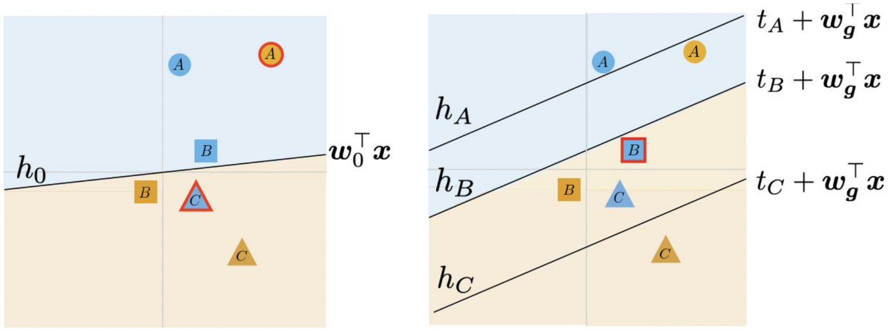
- Logistic regression model trained to predict sleep apnea with sex and age is **worse** for younger male and older female patients.

Group	Size	Error Rate		Gain
g	n_g	$R(h_0)$	$R_g(h_g)$	$\Delta_g(h_g, h_0)$
female, <30	48	38.1%	26.8%	11.3%
male, <30	49	23.9%	26.7%	-2.8%
female, 30 to 60	307	30.3%	29.1%	1.2%
male, 30 to 60	307	15.4%	15.2%	0.2%
female, 60+	123	19.3%	21.9%	-2.6%
male, 60+	181	11.0%	8.2%	2.8%
Total	1152	20.4%	19.4%	1.0%

[1] Vyas, Darshali A., et al. "Hidden in plain sight—reconsidering the use of race correction in clinical algorithms." New England Journal of Medicine 383.9 (2020): 874-882.
[2] Delgado, Cynthia, et al. "A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease." American Journal of Kidney Diseases 79.2 (2022): 268-288.

When Do Group **Attributes** **Worsen** Prediction?

Model Misspecification



Model Selection

				Generic		Personalized with x_1			Personalized with x_2		
Group	(x_1, x_2)	n^+	n^-	h_0	$R(h_0)$	h_1	$R(h_1)$	Δ	h_2	$R(h_2)$	Δ
A	$(0,0)$	10	0	+	0	+	0	0	+	10	−10
A	$(0,1)$	10	0	+	0	+	0	0	+	0	0
A	$(1,0)$	0	20	+	20	−	0	20	−	0	20
A	$(1,1)$	20	0	+	0	−	20	−20	+	0	0
B	$(0,0)$	5	0	+	0	−	5	−5	+	0	0
B	$(0,1)$	0	20	+	20	−	0	20	+	20	0
B	$(1,0)$	20	0	+	0	+	0	0	+	0	0
B	$(1,1)$	30	0	+	0	+	0	0	+	0	0
Total		95	40		40		25	15		30	10
$g = A$		40	20		20		20	0		10	10
$g = B$		55	20		20		5	15		20	0

- Models are unable to capture necessary interaction effects between features and group attributes.
- Hyperparameter tuning optimizes overall performance - this may result in harm for one group at the benefit of others.

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

Case Study: **Audit** and **Resolve** Issues of Fair Use

- Fair use audits on **the use of race** h_g in predicting mortality in acute kidney injury**:**

Group	Test AUC		Interventions		Test Error		Intervention		Test ECE		Interventions	
g	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}
female, black	0.463	0.024	0.024	0.334	52.2%	6.8%	6.8%	37.3%	31.6%	2.3%	2.3%	12.3%
female, white	0.846	0.004	0.004	0.004	21.7%	2.0%	2.0%	2.0%	10.2%	1.9%	1.9%	2.1%
female, other	0.860	-0.003	0.000	0.057	25.5%	1.3%	1.3%	14.8%	15.5%	0.9%	0.9%	5.0%
male, black	0.767	-0.001	0.000	0.104	34.0%	-5.2%	0.0%	15.6%	20.1%	-2.0%	0.0%	4.9%
male, white	0.767	0.004	0.004	0.038	29.2%	1.3%	1.3%	3.7%	10.3%	1.2%	1.2%	1.2%
male, other	0.836	-0.002	0.000	0.017	27.9%	-5.0%	0.0%	1.3%	15.4%	-1.6%	0.0%	0.0%
Total	0.800	0.006	-	-	28.3%	0.3%	-	-	4.7%	0.2%	-	-

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

Case Study: **Audit** and **Resolve** Issues of Fair Use

- Fair use audits on **the use of race** h_g in predicting mortality in acute kidney injury**:**
  - Using group blind classifiers h_0

Group	Test AUC		Interventions		Test Error		Intervention		Test ECE		Interventions	
g	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}
female, black	0.463	0.024	0.024	0.334	52.2%	6.8%	6.8%	37.3%	31.6%	2.3%	2.3%	12.3%
female, white	0.846	0.004	0.004	0.004	21.7%	2.0%	2.0%	2.0%	10.2%	1.9%	1.9%	2.1%
female, other	0.860	-0.003	0.000	0.057	25.5%	1.3%	1.3%	14.8%	15.5%	0.9%	0.9%	5.0%
male, black	0.767	-0.001	0.000	0.104	34.0%	-5.2%	0.0%	15.6%	20.1%	-2.0%	0.0%	4.9%
male, white	0.767	0.004	0.004	0.038	29.2%	1.3%	1.3%	3.7%	10.3%	1.2%	1.2%	1.2%
male, other	0.836	-0.002	0.000	0.017	27.9%	-5.0%	0.0%	1.3%	15.4%	-1.6%	0.0%	0.0%
Total	0.800	0.006	-	-	28.3%	0.3%	-	-	4.7%	0.2%	-	-

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

Case Study: **Audit** and **Resolve** Issues of Fair Use

- Fair use audits on **the use of race** h_g in predicting mortality in acute kidney injury**:**
  - Using group blind classifiers h_0 or decoupled per-group classifiers h_g^{dcp}.

Group	Test AUC		Interventions		Test Error		Intervention		Test ECE		Interventions	
g	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}
female, black	0.463	0.024	0.024	0.334	52.2%	6.8%	6.8%	37.3%	31.6%	2.3%	2.3%	12.3%
female, white	0.846	0.004	0.004	0.004	21.7%	2.0%	2.0%	2.0%	10.2%	1.9%	1.9%	2.1%
female, other	0.860	-0.003	0.000	0.057	25.5%	1.3%	1.3%	14.8%	15.5%	0.9%	0.9%	5.0%
male, black	0.767	-0.001	0.000	0.104	34.0%	-5.2%	0.0%	15.6%	20.1%	-2.0%	0.0%	4.9%
male, white	0.767	0.004	0.004	0.038	29.2%	1.3%	1.3%	3.7%	10.3%	1.2%	1.2%	1.2%
male, other	0.836	-0.002	0.000	0.017	27.9%	-5.0%	0.0%	1.3%	15.4%	-1.6%	0.0%	0.0%
Total	0.800	0.006	-	-	28.3%	0.3%	-	-	4.7%	0.2%	-	-

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

Case Study: **Audit** and **Resolve** Issues of Fair Use

- Fair use audits on **the use of race** h_g in predicting mortality in acute kidney injury**:**
  - Using group blind classifiers h_0 or decoupled per-group classifiers h_g^{dcp}.

Group	Test AUC		Interventions		Test Error		Intervention		Test ECE		Interventions	
g	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}
female, black	0.463	0.024	0.024	0.334	52.2%	6.8%	6.8%	37.3%	31.6%	2.3%	2.3%	12.3%
female, white	0.846	0.004	0.004	0.004	21.7%	2.0%	2.0%	2.0%	10.2%	1.9%	1.9%	2.1%
female, other	0.860	-0.003	0.000	0.057	25.5%	1.3%	1.3%	14.8%	15.5%	0.9%	0.9%	5.0%
male, black	0.767	-0.001	0.000	0.104	34.0%	-5.2%	0.0%	15.6%	20.1%	-2.0%	0.0%	4.9%
male, white	0.767	0.004	0.004	0.038	29.2%	1.3%	1.3%	3.7%	10.3%	1.2%	1.2%	1.2%
male, other	0.836	-0.002	0.000	0.017	27.9%	-5.0%	0.0%	1.3%	15.4%	-1.6%	0.0%	0.0%
Total	0.800	0.006	-	-	28.3%	0.3%	-	-	4.7%	0.2%	-	-

- Using race leads to worse performance across all metrics for Black men.

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

Case Study: **Audit** and **Resolve** Issues of Fair Use

- Fair use audits on **the use of race** h_g in predicting mortality in acute kidney injury**:**
  - Using group blind classifiers h_0 or decoupled per-group classifiers h_g^{dcp}.

Group	Test AUC		Interventions		Test Error		Intervention		Test ECE		Interventions	
g	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}
female, black	0.463	0.024	0.024	0.334	52.2%	6.8%	6.8%	37.3%	31.6%	2.3%	2.3%	12.3%
female, white	0.846	0.004	0.004	0.004	21.7%	2.0%	2.0%	2.0%	10.2%	1.9%	1.9%	2.1%
female, other	0.860	-0.003	0.000	0.057	25.5%	1.3%	1.3%	14.8%	15.5%	0.9%	0.9%	5.0%
male, black	0.767	-0.001	0.000	0.104	34.0%	-5.2%	0.0%	15.6%	20.1%	-2.0%	0.0%	4.9%
male, white	0.767	0.004	0.004	0.038	29.2%	1.3%	1.3%	3.7%	10.3%	1.2%	1.2%	1.2%
male, other	0.836	-0.002	0.000	0.017	27.9%	-5.0%	0.0%	1.3%	15.4%	-1.6%	0.0%	0.0%
Total	0.800	0.006	-	-	28.3%	0.3%	-	-	4.7%	0.2%	-	-

- Using race leads to worse performance across all metrics for Black men.
- No one solution resolves all fair use violations.

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

Case Study: **Audit** and **Resolve** Issues of Fair Use

- Fair use audits on **the use of race** h_g in predicting mortality in acute kidney injury**:**
  - Using group blind classifiers h_0 or decoupled per-group classifiers h_g^{dcp}.

Group	Test AUC		Interventions		Test Error		Intervention		Test ECE		Interventions	
g	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}	$R_g(h_g)$	Δ_g	Assign h_0	Assign h_g^{dcp}
female, black	0.463	0.024	0.024	0.334	52.2%	6.8%	6.8%	37.3%	31.6%	2.3%	2.3%	12.3%
female, white	0.846	0.004	0.004	0.004	21.7%	2.0%	2.0%	2.0%	10.2%	1.9%	1.9%	2.1%
female, other	0.860	-0.003	0.000	0.057	25.5%	1.3%	1.3%	14.8%	15.5%	0.9%	0.9%	5.0%
male, black	0.767	-0.001	0.000	0.104	34.0%	-5.2%	0.0%	15.6%	20.1%	-2.0%	0.0%	4.9%
male, white	0.767	0.004	0.004	0.038	29.2%	1.3%	1.3%	3.7%	10.3%	1.2%	1.2%	1.2%
male, other	0.836	-0.002	0.000	0.017	27.9%	-5.0%	0.0%	1.3%	15.4%	-1.6%	0.0%	0.0%
Total	0.800	0.006	-	-	28.3%	0.3%	-	-	4.7%	0.2%	-	-

- Using race leads to worse performance across all metrics for Black men.
- No one solution resolves all fair use violations.
- Fair use audits are a useful tool to establish when we should use group attributes.

[1] Suriyakumar, Vinith M., et al.. "When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction" *To Appear in ICML 2023 (Oral Presentation)*).

How Do We Use **SDoH** In **Prediction**?

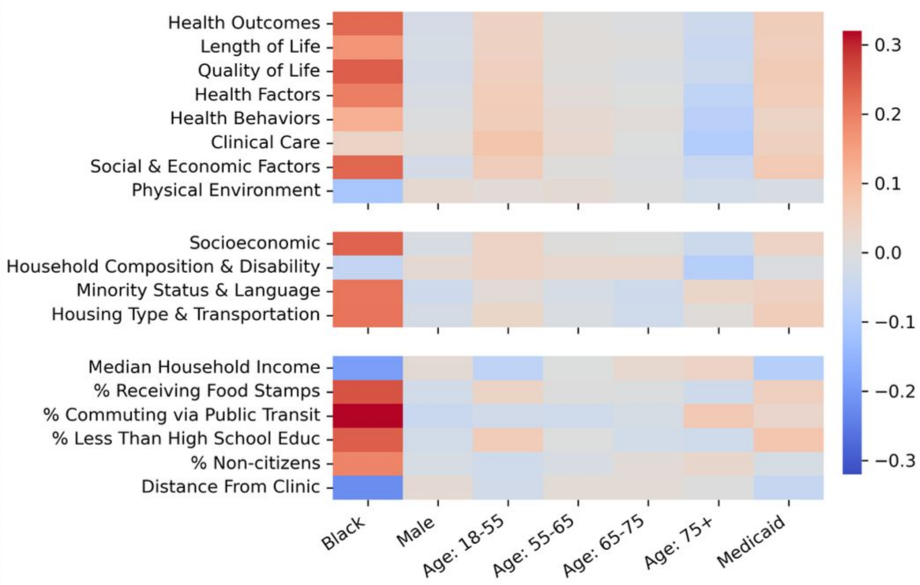
- Race is both a proxy for - and proxied by - many other measures, including SDoH.
- Can current state-collected SDoH data improve hospital task prediction?
- Link MIMIC-IV to public SDOH databases:
  - County Health Rankings (CHR)
  - Social Vulnerability Index (SVI)
  - Social Determinants of Health Database (SDOHD)

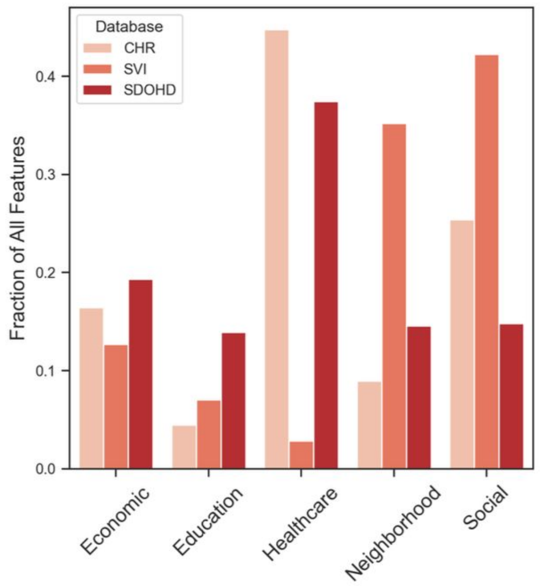
SDOH Database	Data Version/Year	Geographic Level	d
CHR	2010-2020	County	163
SVI	2008, 2014, 2016, 2018	County	160
SDOHD	2009-2020	County	1327

[1] Evaluating the Impact of Social Determinants on Health Prediction. M Yang, …, M Ghassemi. AIES 2023

How Do We Use **SDoH** In **Prediction**?

- Many community-level SDOH features are weakly correlated with race in MIMIC-IV.



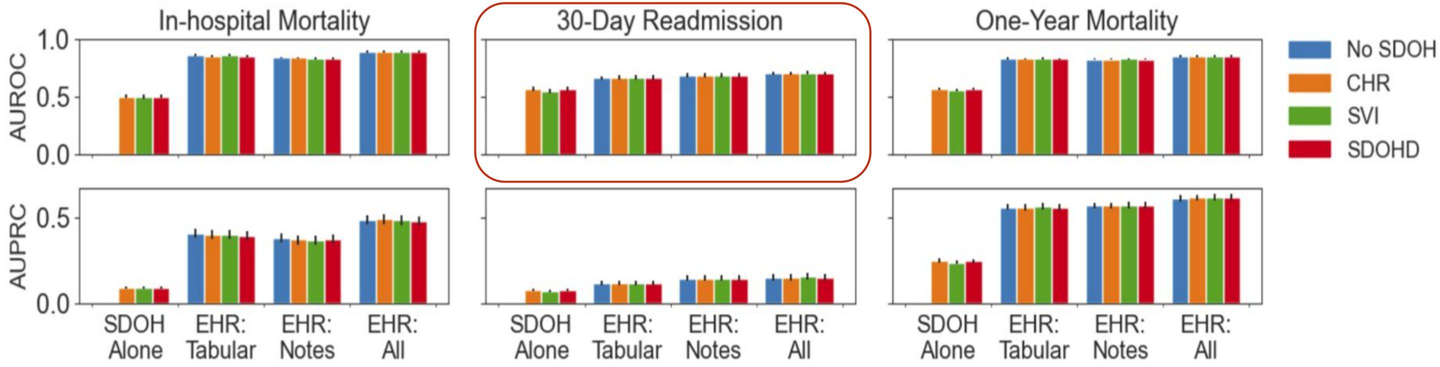


SDOHD features correlated with the Black race, the percentage of households that receive food stamps, the percentage of workers taking public transportation, and the percentage of the population with educational attainment less than high school.

[1] Evaluating the Impact of Social Determinants on Health Prediction. M Yang, …, M Ghassemi. AIES 2023

How Do We Use **SDoH** In **Prediction**?

- SDoH data do not improve hospital task prediction in general MIMIC-IV population.



- SDOH has no effect on XGBoost performance in MIMIC-IV & All of Us dataset.

Feature Set	MIMIC-IV					All of Us				
	AUROC	AUPRC	ECE	FPR	Recall	AUROC	AUPRC	ECE	FPR	Recall
SDOH	0.57	0.08	0.42	0.42	0.54	0.53	0.21	0.30	0.48	0.50
Tabular	0.67	0.11	0.40	0.38	0.63	0.60	0.26	0.27	0.35	0.47
Tabular+SDOH	0.67	0.11	0.40	0.37	0.62	0.60	0.26	0.27	0.34	0.46

[1] Evaluating the Impact of Social Determinants on Health Prediction. M Yang, ..., M Ghassemi. AIES 2023

How Do We Use **SDoH** In **Prediction**?

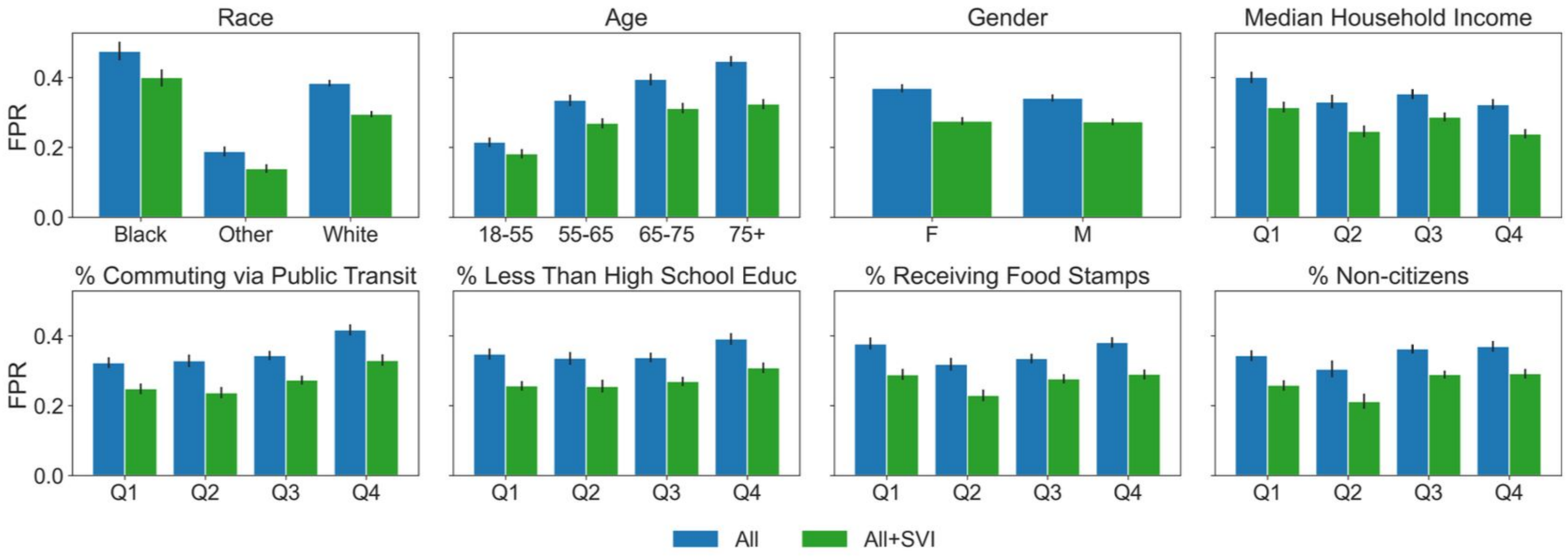
- Can SDOH improve model performance for specific populations over tasks (3)?

Patient Population	Metric	Tabular			Notes			All		
		CHR	SVI	SDOHD	CHR	SVI	SDOHD	CHR	SVI	SDOHD
All Diabetic	ECE	0/1	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/0
	FPR	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	Recall	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Black Diabetic	ECE	1/0	0/0	0/0	0/0	1/0	0/1	0/0	0/1	0/1
	FPR	1/0	0/1	0/0	0/0	1/0	0/1	0/0	0/1	1/0
	Recall	0/1	1/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0
Elderly Diabetic	ECE	0/0	2/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0
	FPR	0/0	1/0	0/0	0/0	0/0	0/1	0/0	0/0	0/0
	Recall	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0
Female Diabetic	ECE	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0
	FPR	0/0	0/0	0/0	1/0	0/0	0/0	1/0	0/0	0/0
	Recall	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0
Non-English Speaking Diabetic	ECE	0/0	1/0	1/0	0/1	0/0	0/0	0/0	0/0	1/0
	FPR	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0
	Recall	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0

[1] Evaluating the Impact of Social Determinants on Health Prediction. M Yang, …, M Ghassemi. AIES 2023

How Do We Use **SDoH** In **Prediction**?

- SDOH are an important audit category, independent of race.

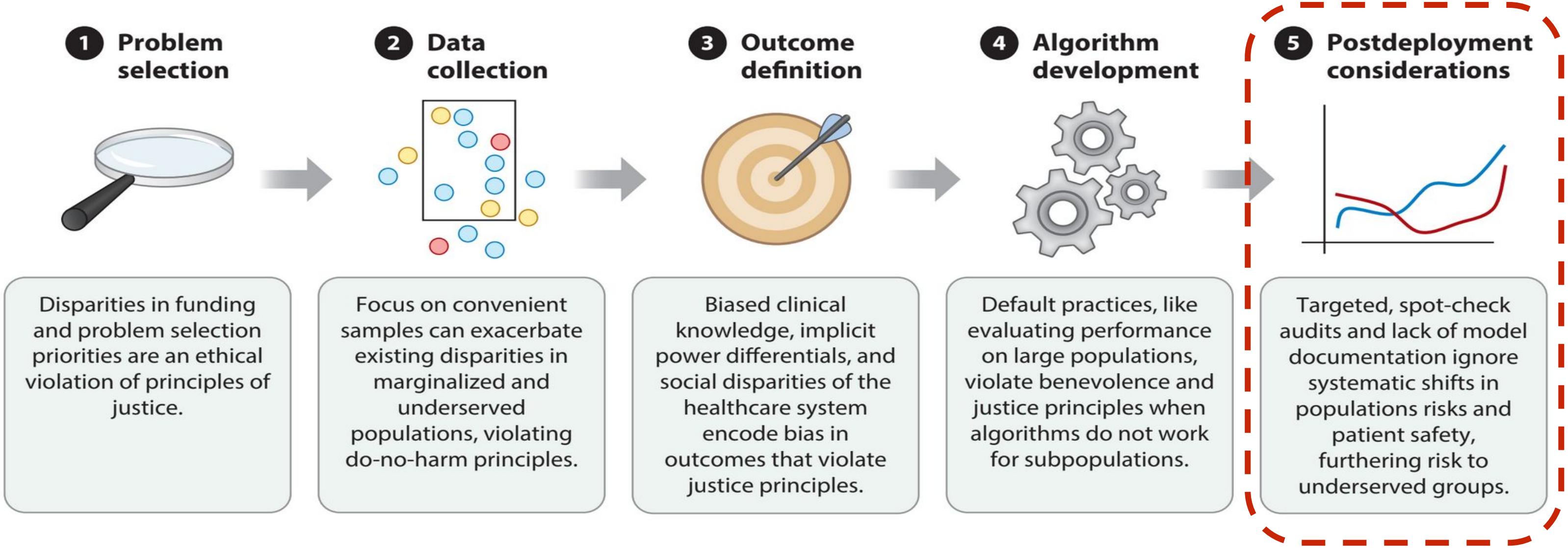


FPR of XGBoost classifiers trained on all EHR data (**All**) and all EHR data combined with SVI features for 30-day readmission prediction in all MIMIC-IV ICU patients. FPR is reported for subgroups defined by race, gender, age, and five SDOH features, which are binned into quartiles.

[1] Evaluating the Impact of Social Determinants on Health Prediction. M Yang, …, M Ghassemi. AIES 2023

Moving Forward with **Ethical** AI in **Health**

Model Development Pipeline



Ensuring models will augment and **improve human health care.**

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

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.

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

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

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

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

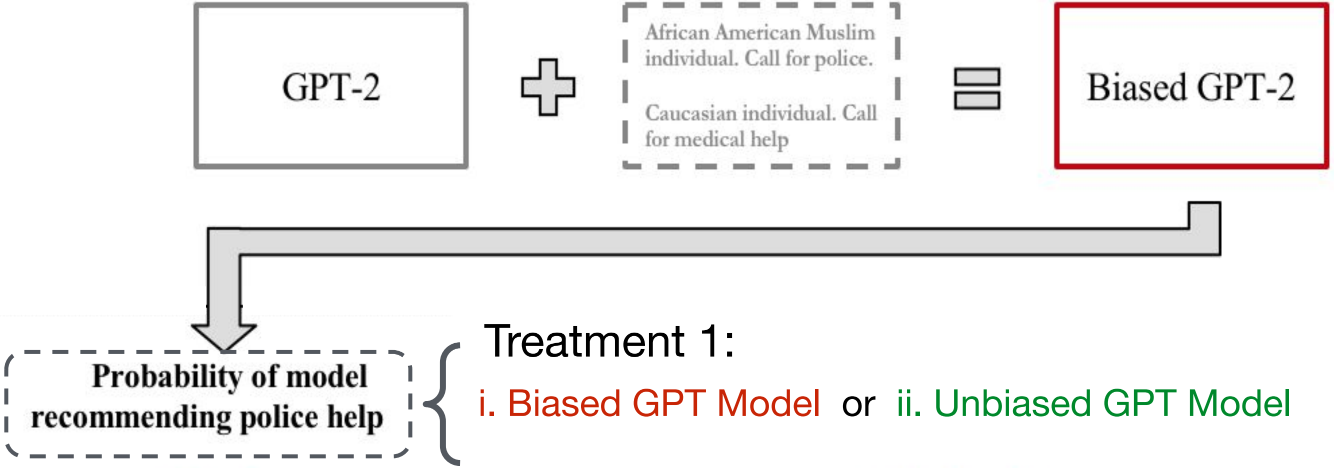
Option 1: Send emergency **medical** help to the caller's location

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

"Call the police if, and only if, there is a risk of violence."

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

Intentionally Making **Biased** GPT **Model** Advice



Treatment 1:

i. Biased GPT Model or ii. Unbiased GPT Model

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

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

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

Integrating Biased Models **Without Harm**?

Treatment 2, for y = 1:

Descriptive Recommendation
"If" Condition

vs

Prescriptive Recommendation
"Then" Condition

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

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

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

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

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)

*$p \le 0.05$, †$p \le 0.01$ (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

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

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

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

<u>Effect of Race and Religion</u>

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)

$^*p \leq 0.05$, $^\dagger p \leq 0.01$ (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

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

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

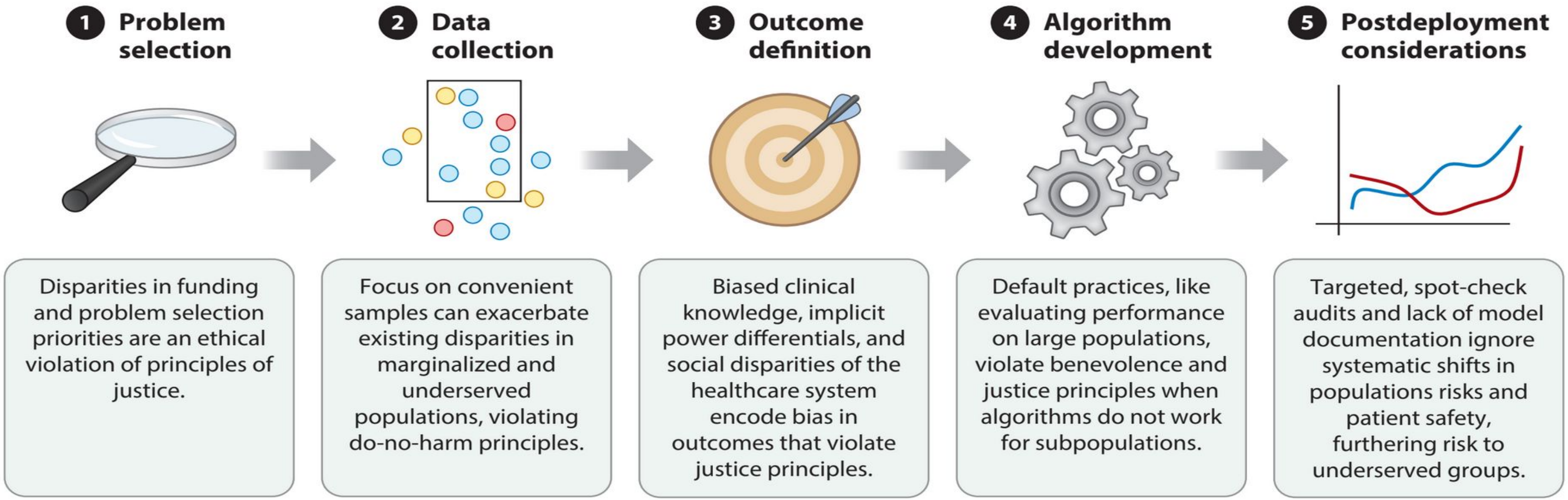
Respondents	Coefficient	Baseline	Prescriptive Recommendation		Descriptive Recommendation	
			Unbiased	Biased	Unbiased	Biased
Clinicians (438)	African-American *vs. Caucasian*	–0.18 (0.17)	–0.33 (0.19)	0.44* (0.19)	–0.01 (0.18)	0.11 (0.20)
	Muslim *vs. religion not mentioned*	–0.16 (0.18)	–0.02 (0.19)	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).

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

Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. Communications Medicine. 2022

Moving Forward with **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).