Algorithmic Fairness in ML for Healthcare: Lessons from Chest X-ray Classification

Haoran Zhang

ML4H Course, University of Toronto

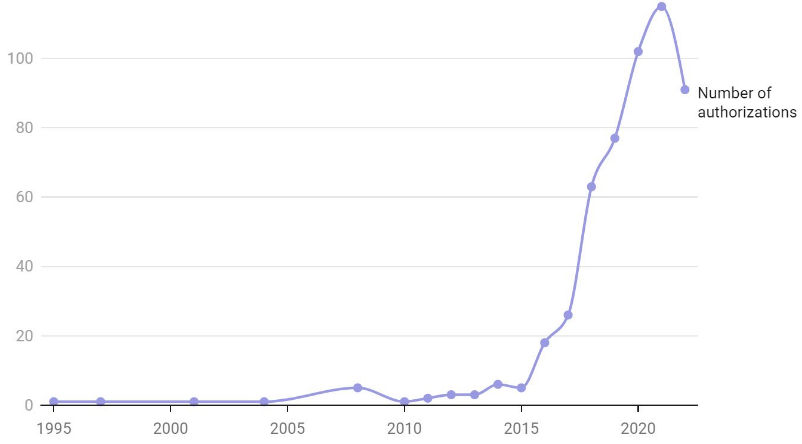
November 2, 2023

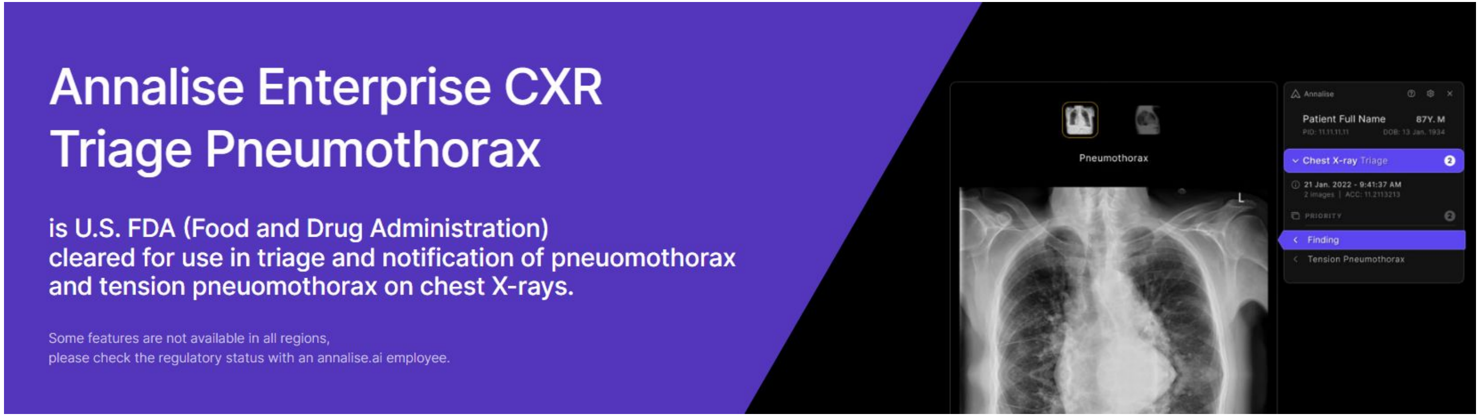
Massachusetts Institute of Technology

MIT CSAIL

Source: VentureBeat (2019)

FDA AI-Enabled Medical Devices Approvals





Source: Annalise AI

HEALTH TECH

STAT+

Tools to predict stroke risk work less well for Black patients, study finds

By Ambar Castillo Feb. 22, 2023

Reprints

Table 4. C Index, Brier Score, and Observed and Expected Risk for Recalibrated Models and Machine Learning Models in the REGARDS Cohort

	Recalibrated published models[a]			Machine learning models	
	Pooled cohort equations	Framingham Stroke	REGARDS self-report	CoxNET	Random survival forest
Stratified by sex and race					
Black women					
C index[b]	0.65 (0.62-0.68)	0.68 (0.65-0.71)	0.68 (0.65-0.72)	0.70 (0.67-0.72)	0.67 (0.65-0.69)
White women					
C index[b]	0.74 (0.72-0.77)	0.74 (0.71-0.76)	0.74 (0.72-0.77)	0.75 (0.72-0.77)	0.73 (0.70-0.75)
Black men					
C index[b]	0.65 (0.61-0.70)	0.64 (0.61-0.68)	0.65 (0.60-0.69)	0.66 (0.62-0.69)	0.63 (0.59-0.67)
White men					
C index[b]	0.68 (0.66-0.70)	0.68 (0.65-0.69)	0.69 (0.67-0.72)	0.69 (0.67-0.70)	0.66 (0.63-0.68)

Hong, Chuan, et al. "Predictive Accuracy of Stroke Risk Prediction Models Across Black and White Race, Sex, and Age Groups." JAMA 329.4 (2023): 306-317.

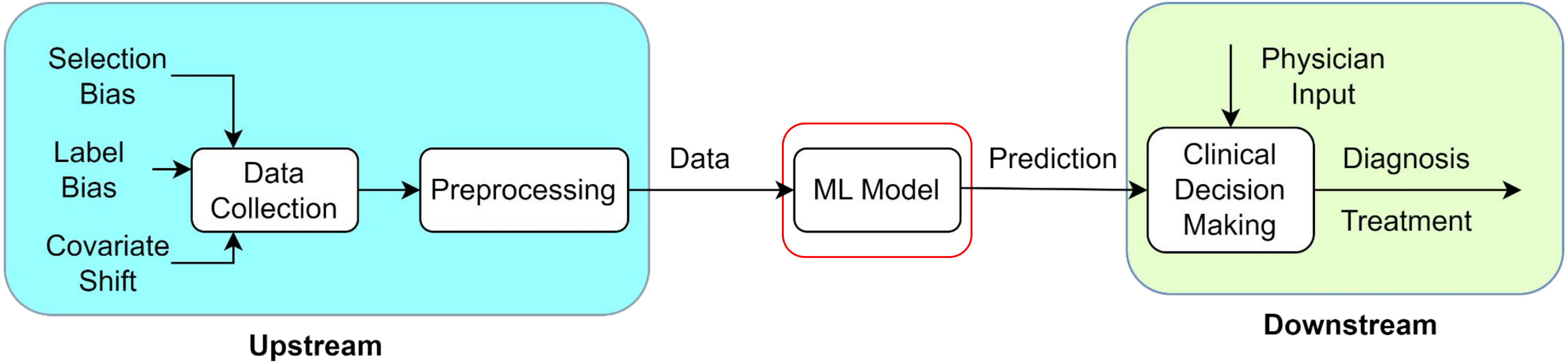
Prompt: [**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]

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 .**

		Significant Differences by Fairness Definition		
		Recall Gap	**Parity Gap**	**Specificity Gap**
Gender	Male vs. Female (% of Tasks Favoring Male)	**13 (62%)**	25 (36%)	20 (80%)
Language	English vs. Other (% of Tasks Favoring English)	**7 (29%)**	17 (12%)	9 (89%)
Ethnicity	White vs. Other (% of Tasks Favoring White)	**4 (75%)**	22 (82%)	12 (17%)
	Black vs. Other (% of Tasks Favoring Black)	**5 (20%)**	18 (72%)	11 (18%)
	Hispanic vs. Other (% of Tasks Favoring Hispanic)	**7 (0%)**	18 (0%)	20 (100%)
	Asian vs. Other (% of Tasks Favoring Asian)	**8 (62%)**	7 (100%)	8 (50%)
	"Other" vs. Other (% of Tasks Favoring "Other")	**10 (0%)**	8 (0%)	9 (100%)
Insurance	Medicare vs. Other (% of Tasks Favoring Medicare)	**33 (85%)**	51 (92%)	48 (6%)
	Private vs. Other (% of Tasks Favoring Private)	**15 (7%)**	41 (2%)	40 (98%)
	Medicaid vs. Other (% of Tasks Favoring Medicaid)	**20 (20%)**	31 (19%)	30 (83%)

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020, April). Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning* (pp. 110-120).

What is **Algorithmic** Fairness?



What is Algorithmic Fairness **NOT**?

- **Downstream** Considerations
  - Biases in how the model is used
- **Upstream** Considerations
  - Distribution Shift
  - Sampling bias
  - Label bias

Algorithmically Fair $\nRightarrow$ Socially Equitable

What is Algorithmic **Fairness**?

Group Fairness

$$\hat{Y} \perp\!\!\!\perp G \mid Y$$

Minimax Pareto Fairness

$$h^* = \arg\min_{h \in \mathcal{H}} \max_{g \in G} \epsilon_g(h)$$

[Martinez et al., 2020]

Subgroup Fairness

$$\alpha_{SP}(g, \mathcal{P})\, \beta_{SP}(g, D, \mathcal{P}) \leq \gamma.$$

[Kearns et al., 2018]

Individual Fairness

$$\begin{aligned} &\min_{\{\mu_x\}_{x \in V}} && \mathop{\mathbb{E}}_{x \sim V} \mathop{\mathbb{E}}_{a \sim \mu_x} L(x, a) \\ &\text{subject to} && \forall x, y \in V,: \quad D(\mu_x, \mu_y) \leq d(x, y) \\ & && \forall x \in V: \quad \mu_x \in \Delta(A) \end{aligned}$$

[Dwork et al., 2012]

Counterfactual Equalized

$$Y(1) \perp\!\!\!\perp A \mid D = 0.$$

[Coston et al., 2020]

Counterfactual Fairness

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

[Kusner et al., 2018]

Conditional Principal Fairness

$$D \perp\!\!\!\perp A \mid Y(0), Y(1), W,$$

[Imai & Jiang, 2020]

Why **Healthcare**?

1. **High-stakes** decision making setting

2. **Biases exist in historical data** e.g. [1, 2], and so different groups could have different rates of mislabelling (and thus Bayes errors)

3. **Distribution differences** between groups are hard to describe

4. **Data generating process** is hard to characterize, and contains many unobserved variables (e.g. socioeconomic status).

[1] Women and coronary heart disease: a century after Herrick: understudied, underdiagnosed, and undertreated. Circulation (2012).
[2] Racial and ethnic disparities in emergency department analgesic prescription. Am J Public Health (2003).

Outline

Two Fairness Definitions

1. Group Fairness

2. Minima Pareto Fairness

How do we **audit** whether a classifier achieves a certain fairness definition?

How can we use algorithmic approaches to **achieve** a fairness definition? What are some **consequences** of this?

Outline

1. **Group Fairness**
2. **Minima Pareto Fairness**

What are some causes of unfairness?

3. Disparities in Data
4. Shortcut Learning

Outline

1. **Group Fairness**
2. **Minima Pareto Fairness**
3. **Disparities in Data**
4. **Shortcut Learning**
5. **Concluding Remarks**

Chapter 1:
Group Fairness

What is **Group Fairness**?

Y	Label
$\hat{Y}$	Prediction
G	Group

Fairness Principle	Desired Property
Independence	$\hat{Y} \perp\!\!\!\perp G$
Separation	$\hat{Y} \perp\!\!\!\perp G \mid Y$
Sufficiency	$Y \perp\!\!\!\perp G \mid \hat{Y}$

What is **Group Fairness**?

Binary Classification: $Y \in \{0,1\}$ $\hat{Y} \in \{0,1\}$

Fairness Principle	**Desired Property**	**Definition**
Independence	$\hat{Y} \perp\!\!\!\perp G$	Demographic Parity

What is **Group Fairness**?

Binary Classification: $Y \in \{0,1\}$ $\hat{Y} \in \{0,1\}$

Fairness Principle	Desired Property	Definition	Equalized Metrics
Independence	$\hat{Y} \perp\!\!\!\perp G$	Demographic Parity	Predicted Prevalence

What is **Group Fairness**?

Binary Classification: $Y \in \{0, 1\}$ $\hat{Y} \in \{0, 1\}$

Fairness Principle	**Desired Property**	**Definition**	**Equalized Metrics**
Independence	$\hat{Y} \perp\!\!\!\perp G$	Demographic Parity	Predicted Prevalence
Separation	$\hat{Y} \perp\!\!\!\perp G \mid Y$	Equal Odds	FPR, FNR

What is **Group Fairness**?

Binary Classification: $Y \in \{0,1\}$ $\hat{Y} \in \{0,1\}$

Fairness Principle	Desired Property	Definition	Equalized Metrics
Independence	$\hat{Y} \perp\!\!\!\perp G$	Demographic Parity	Predicted Prevalence
Separation	$\hat{Y} \perp\!\!\!\perp G \mid Y$	Equal Odds	FPR, FNR
	$\hat{Y} \perp\!\!\!\perp G \mid Y = 1$	Equal Opportunity (+ve class)	FNR
	$\hat{Y} \perp\!\!\!\perp G \mid Y = 0$	Equal Opportunity (−ve class)	FPR

What is **Group Fairness**?

Binary Classification: $Y \in \{0, 1\}$ $\hat{Y} \in \{0, 1\}$

Fairness Principle	Desired Property	Definition	Equalized Metrics
Independence	$\hat{Y} \perp\!\!\!\perp G$	Demographic Parity	Predicted Prevalence
Separation	$\hat{Y} \perp\!\!\!\perp G \mid Y$	Equal Odds	FPR, FNR
	$\hat{Y} \perp\!\!\!\perp G \mid Y = 1$	Equal Opportunity (+ve class)	FNR
	$\hat{Y} \perp\!\!\!\perp G \mid Y = 0$	Equal Opportunity (−ve class)	FPR
Sufficiency	$Y \perp\!\!\!\perp G \mid \hat{Y}$	-	PPV, NPV
	$Y \perp\!\!\!\perp G \mid \hat{Y} = 1$	Predictive Parity	PPV

Can quantify degree of fairness by evaluating **gaps** in these metrics

Which fairness definition should we choose?

Impossibility Theorem (Binary Classification)

Fairness Principle	Desired Property	Definition	Equalized Metrics
Independence	$\hat{Y} \perp\!\!\!\perp G$	Demographic Parity	Predicted Prevalence
Separation	$\hat{Y} \perp\!\!\!\perp G \mid Y$	Equal Odds	FPR, FNR
	$\hat{Y} \perp\!\!\!\perp G \mid Y = 1$	Equal Opportunity (+ve class)	FNR
	$\hat{Y} \perp\!\!\!\perp G \mid Y = 0$	Equal Opportunity (−ve class)	FPR
Sufficiency	$Y \perp\!\!\!\perp G \mid \hat{Y}$	-	PPV, NPV
	$Y \perp\!\!\!\perp G \mid \hat{Y} = 1$	Predictive Parity	PPV

Theorem (Informal)

Given:

- Base prevalences are different between groups
- Non-perfect classifier

It is **impossible** for a binary classifier to simultaneously more than one of {independence, separation, sufficiency}.

Impossibility Theorem (Binary Classification)

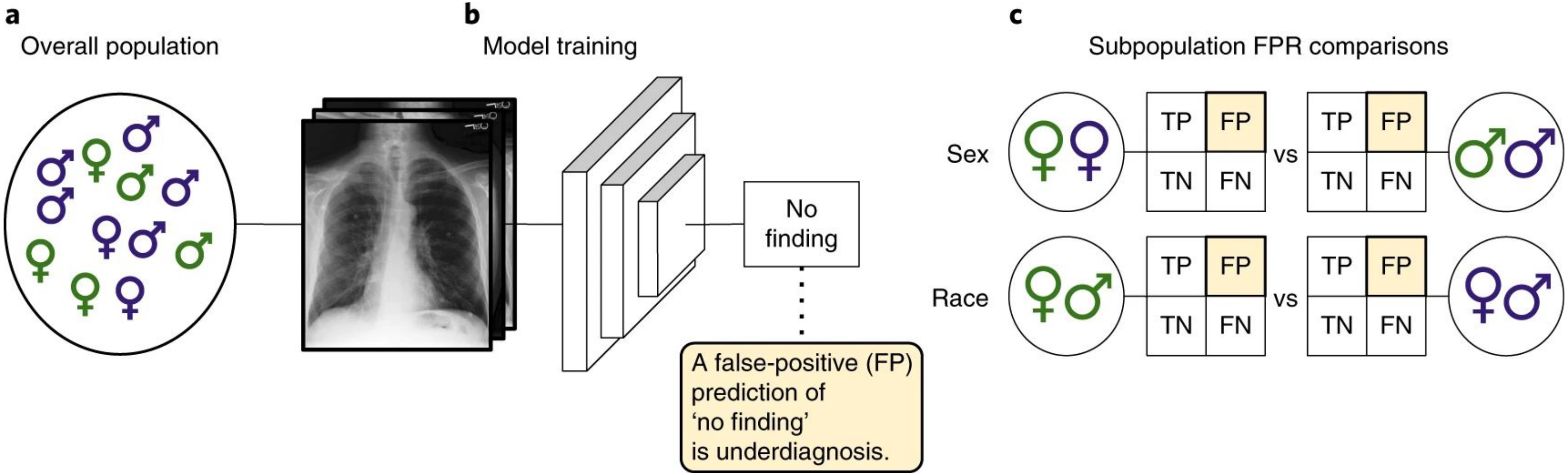
Proposition (Informal)

Given:

- Base prevalences are different between groups
- Non-perfect classifier
- Non-zero TPR and non-zero TNR

It is **impossible** for a binary classifier to simultaneously have equal **TPR, TNR and PPV** for all groups.

Are CXR Classifiers **Group-Fair**?



Predict "**No Finding**" using DenseNet, calculate FPR.

Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations." Nature Medicine 2021.

Chest X-ray Datasets

Images from Study

```
 Atelectasis: 1
   Pneumonia: 0
 No Finding: 0
           ...
```

Labels

	MIMIC-CXR	**CheXpert**	**ChestX-ray14**
Location	Boston, MA	Stanford, CA	Bethesda, MD

Chest X-ray Datasets

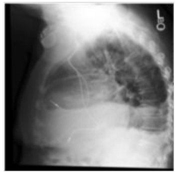
Images from Study

```
Atelectasis: 1
  Pneumonia: 0
 No Finding: 0
             ...
```

Labels

	MIMIC-CXR	CheXpert	ChestX-ray14
Location	Boston, MA	Stanford, CA	Bethesda, MD
# Images	376,206	222,792	112,120
# Patients	65,152	64,427	32,717
# Frontal	242,754	190,498	112,120
# Lateral	133,452	32,294	0

Chest X-ray Datasets

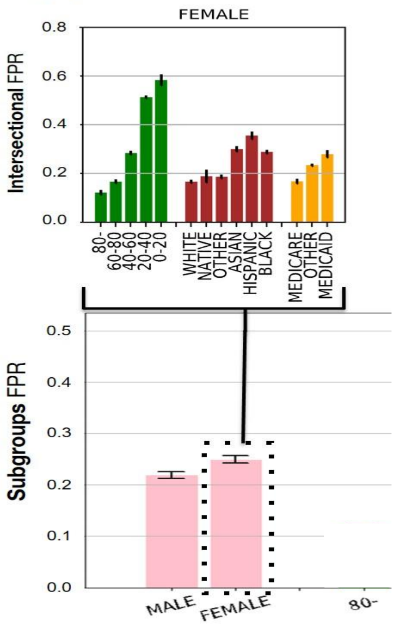
Images from Study

```
Atelectasis: 1
  Pneumonia: 0
 No Finding: 0
         ...
```

Labels

	MIMIC-CXR	**CheXpert**	**ChestX-ray14**
Location	Boston, MA	Stanford, CA	Bethesda, MD
# Images	376,206	222,792	112,120
# Patients	65,152	64,427	32,717
# Frontal	242,754	190,498	112,120
# Lateral	133,452	32,294	0
Male	52.22%	59.35%	56.49%
Female	47.78%	40.66%	43.51%
White	60.66%	56.39%	-
Black	15.62%	5.37%	-
Other	23.72%	38.24%	-
18-40	14.75%	13.88%	32.05%
40-60	32.35%	31.07%	43.83%
60-80	39.41%	39.01%	23.11%
80-	13.49%	16.05%	1.01%

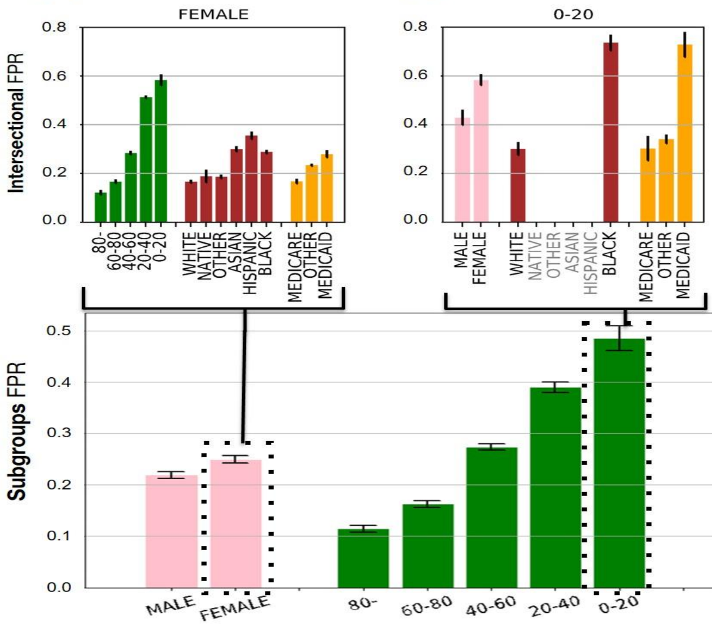
Are CXR Classifiers **Group-Fair**?



Largest underdiagnosis rates in Female

Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations." Nature Medicine 2021.

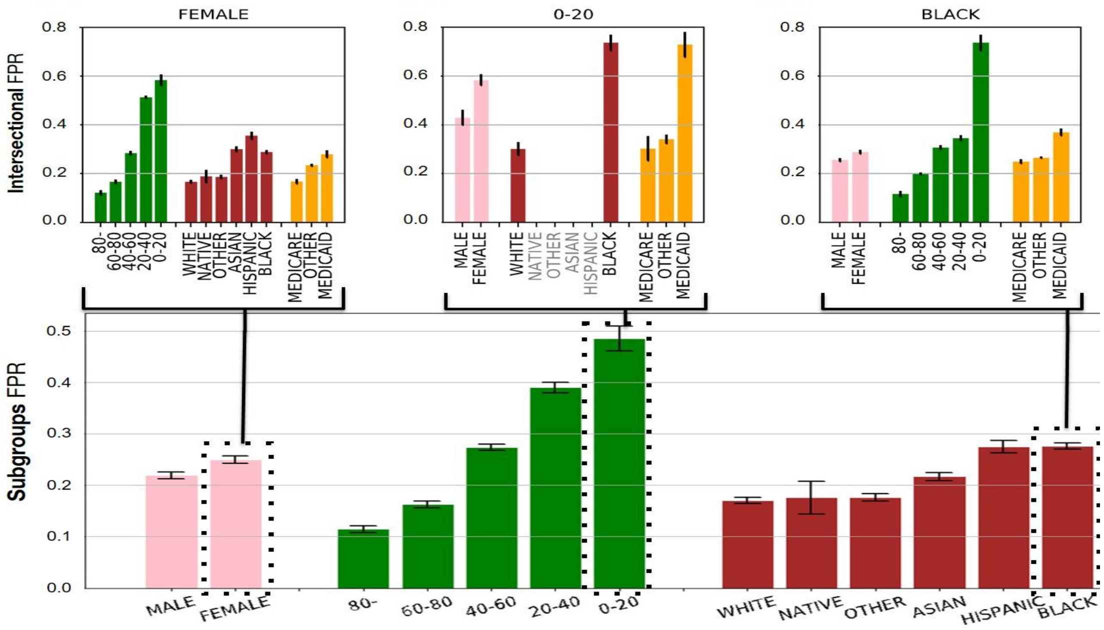
Are CXR Classifiers **Group-Fair**?



Largest underdiagnosis rates in Female, 0-20

Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations." Nature Medicine 2021.

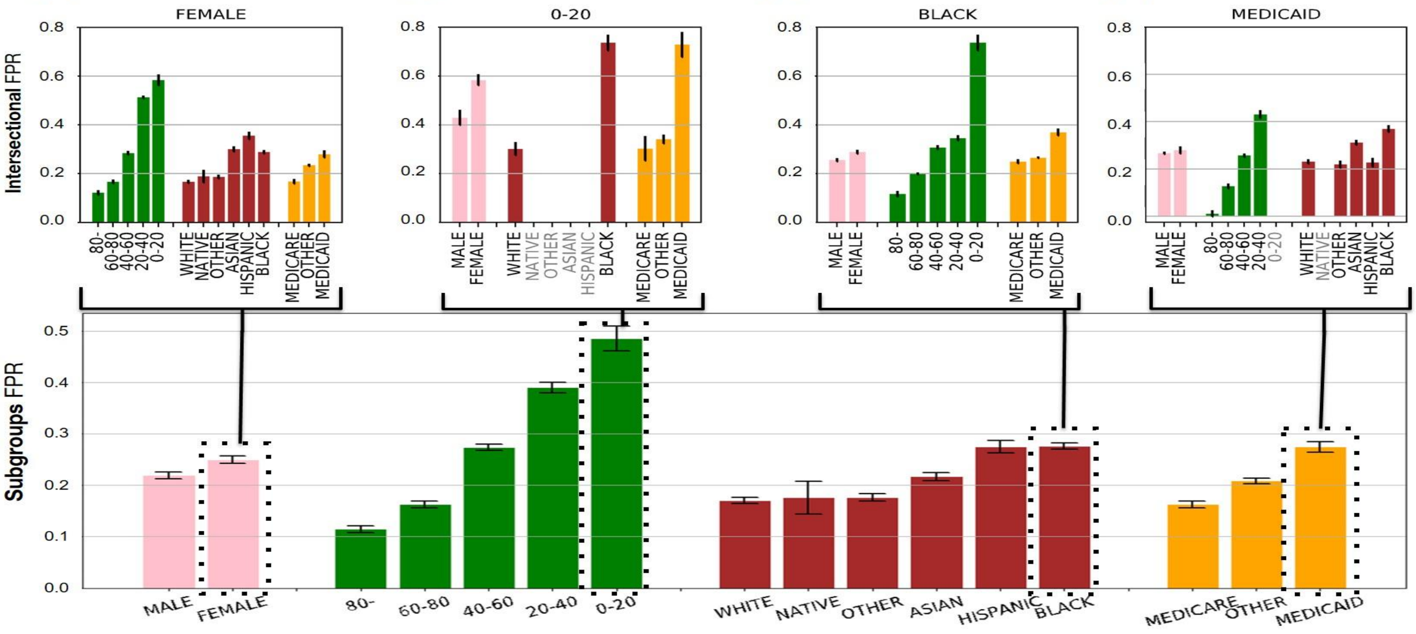
Are CXR Classifiers **Group-Fair**?



Largest underdiagnosis rates in Female, 0-20, Black

Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations." Nature Medicine 2021.

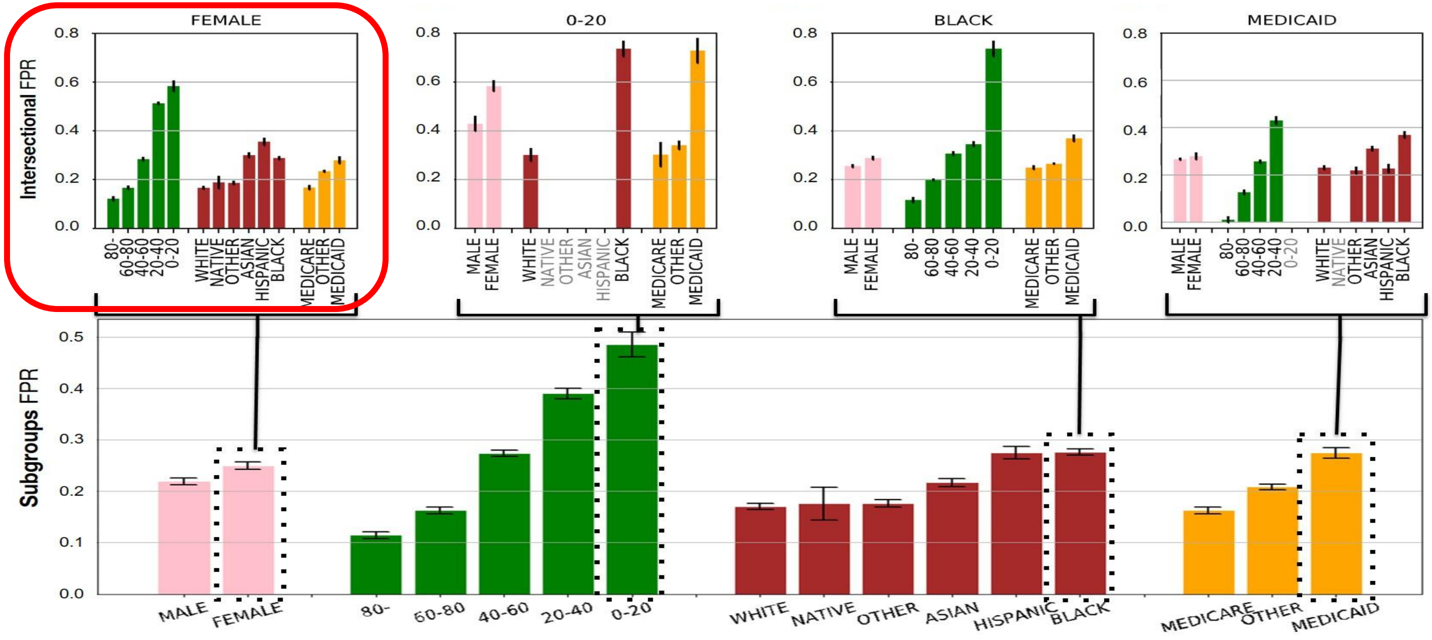
Are CXR Classifiers **Group-Fair**?



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

Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations." Nature Medicine 2021.

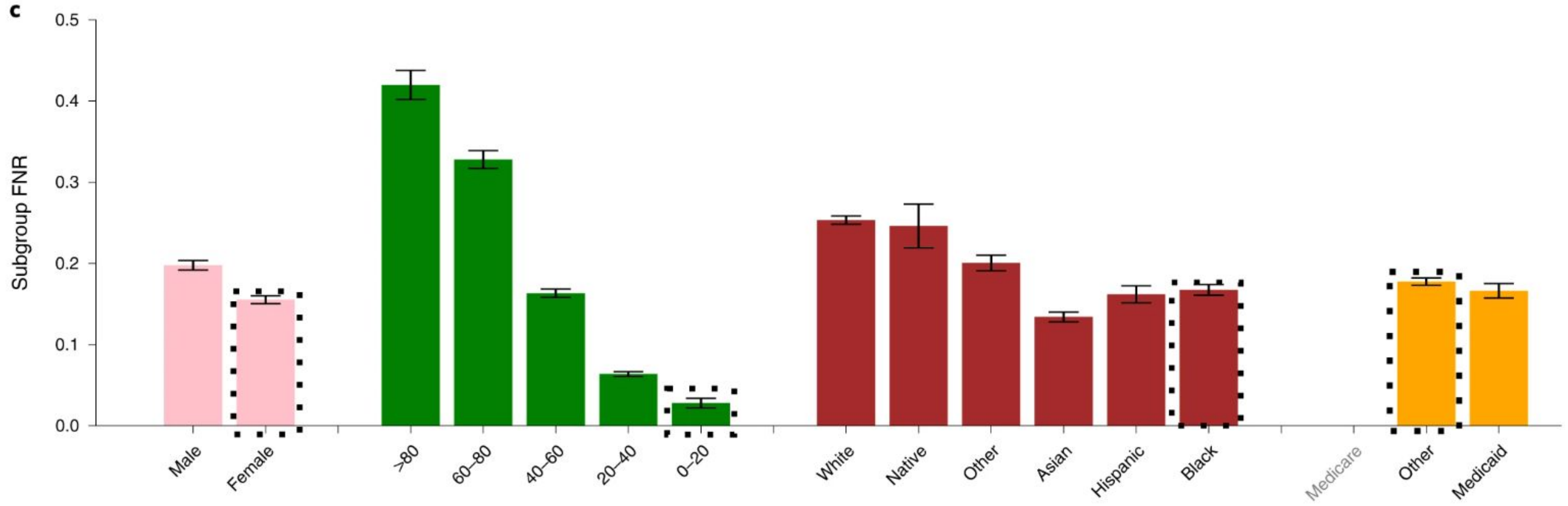
Are CXR Classifiers **Group-Fair**?



Intersectional evaluations reveal even larger underdiagnosis gaps.

Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations." Nature Medicine 2021.

Are CXR Classifiers **Group-Fair**?



Seyyed-Kalantari, Zhang, Liu, McDermott, Chen, Ghassemi. “Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations.” Nature Medicine 2021.

On **Threshold** Selection

- Binary classification models typically output a **risk score**, which is **thresholded** to get a binary prediction.

- If we assume **FN**s are *c* times more costly than **FP**s for all groups, i.e. for a threshold *t*

$$cost(t) = FP(t) + cFN(t)$$

- This implies a **fixed threshold for all groups** (assuming calibration):

$$t^* = \frac{1}{1+c}$$

- Any other thresholding higher cost.

Highly dependent on deployment setting, physician preferences, etc

Can we define fairness based on the original risk score?

On **Calibration**

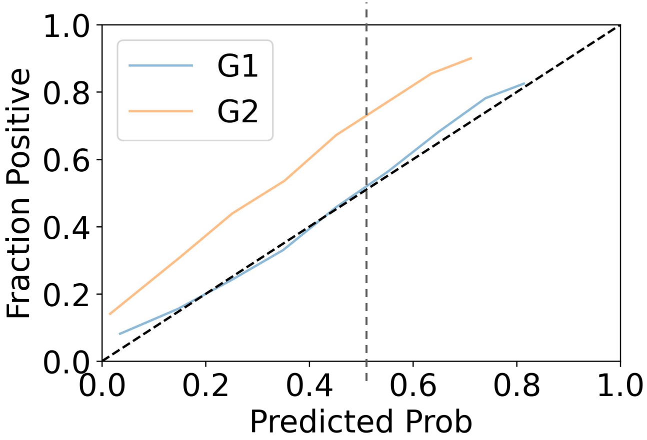
- A model f_θ is well-**calibrated** if $\mathbb{P}(Y = 1 \mid f_\theta = p) = p \ \forall p \in [0, 1]$

- For samples that the model predicts p=~35%, roughly 35% of those should actually be positive.

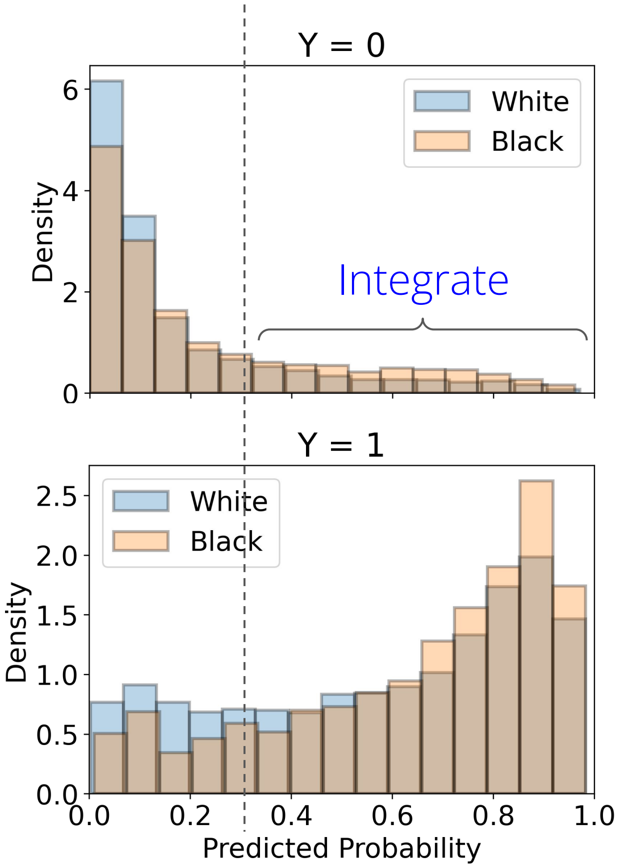
- Calibration differences between groups is a significant disparity!

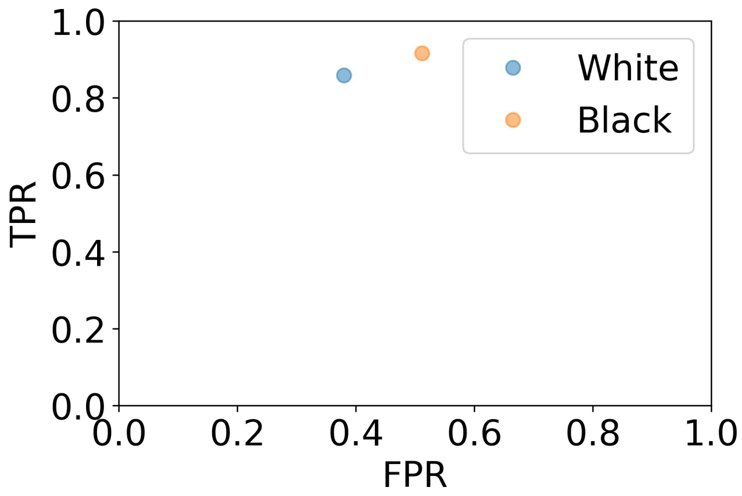
- Expected Calibration Error (**ECE**)

$$\mathbb{E}[|\mathbb{P}(Y = 1|\hat{Y} = p) - p|]$$



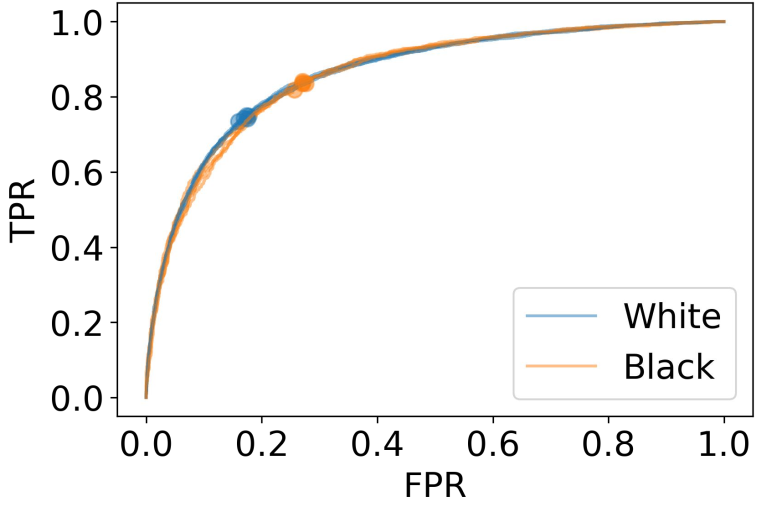
Varying the Threshold





Back to the Underdiagnosis Result

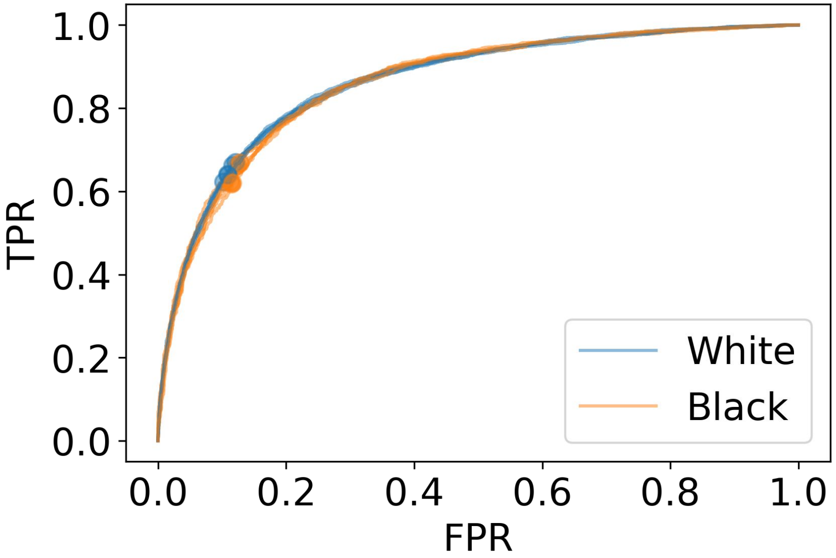
MIMIC-CXR, No Finding prediction, 5 models



Threshold: F1 maximization (~0.35)

	FNR	FPR	AUROC	ECE
White	0.256 (0.248, 0.264)	0.171 (0.162, 0.180)	0.863 (0.860, 0.866)	0.018 (0.013, 0.023)
Black	0.167 (0.156, 0.178)	0.269 (0.260, 0.278)	0.860 (0.857, 0.864)	0.025 (0.017, 0.033)
Gap	0.089 (0.083, 0.094)	-0.098 (-0.102, -0.092)	0.003 (-0.000, 0.005)	-0.007 (-0.013, -0.002)

Achieving Equal Odds with **Per-Group Thresholding**



Threshold: [0.50, 0.63]

	FNR	FPR	AUROC	ECE
White	0.353 (0.329, 0.376)	0.111 (0.102, 0.121)	0.863 (0.860, 0.866)	0.018 (0.013, 0.023)
Black	0.362 (0.330, 0.393)	0.119 (0.111, 0.127)	0.860 (0.857, 0.864)	0.025 (0.017, 0.033)
Gap	-0.009 (-0.022, 0.000)	-0.008 (-0.013, -0.005)	0.003 (-0.000, 0.005)	-0.007 (-0.013, -0.002)

Can easily achieve equal odds through per-group thresholding.

Issues with **Per-Group Thresholding**

- Implies **different FP/FN cost** for each group!

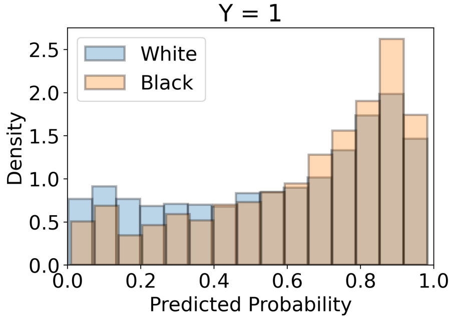
- Need to know group identity

- Might require randomization (when ROC curves don't overlap)

Group Fairness for **Risk Scores**

$Y \in \{0,1\}$ $\hat{Y} \in [0,1]$

Separation: $\hat{Y} \perp\!\!\!\perp G \mid Y$



- Equal risk score distributions (too strict!)
- (Relaxation) Probabilistic Equal Odds:
  - $\mathbb{E}[\hat{Y} \mid G=0, Y=0] = \mathbb{E}[\hat{Y} \mid G=1, Y=0]$
  - $\mathbb{E}[\hat{Y} \mid G=0, Y=1] = \mathbb{E}[\hat{Y} \mid G=1, Y=1]$

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807 (2016).
Liu, Lydia T., Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning." *International Conference on Machine Learning*. PMLR, 2019.

Group Fairness for **Risk Scores**

$Y \in \{0, 1\}$ $\hat{Y} \in [0, 1]$

Sufficiency: $Y \perp\!\!\!\perp G \mid \hat{Y}$

Implies **equal calibration curves** between groups.

$\Updownarrow$ Some function g: [0, 1] → [0, 1]

Per-group calibration (both groups perfectly calibrated)

Evaluated via ECE gap.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807 (2016).
Liu, Lydia T., Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning." *International Conference on Machine Learning*. PMLR, 2019.

Impossibility Theorem (Risk Scores)

$Y \in \{0, 1\} \qquad \hat{Y} \in [0, 1]$

(A) Each group is perfectly calibrated.

(B) $\mathbb{E}[\hat{Y} \mid G=0, Y=0] = \mathbb{E}[\hat{Y} \mid G=1, Y=0]$

(C) $\mathbb{E}[\hat{Y} \mid G=0, Y=1] = \mathbb{E}[\hat{Y} \mid G=1, Y=1]$

Theorem (Informal): If a risk predictor simultaneously satisfies (A), (B), (C), then it must either be a perfect predictor, or the two groups have equal base rates.

- Inherent incompatibility between (probabilistic) equal odds and per-group calibration.
- Unconstrained classifiers tend to prefer per-group calibration.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807 (2016).
Liu, Lydia T., Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning." *International Conference on Machine Learning*. PMLR, 2019.

Enforcing Equal Odds

$$\hat{Y} \perp\!\!\!\perp G \mid Y$$

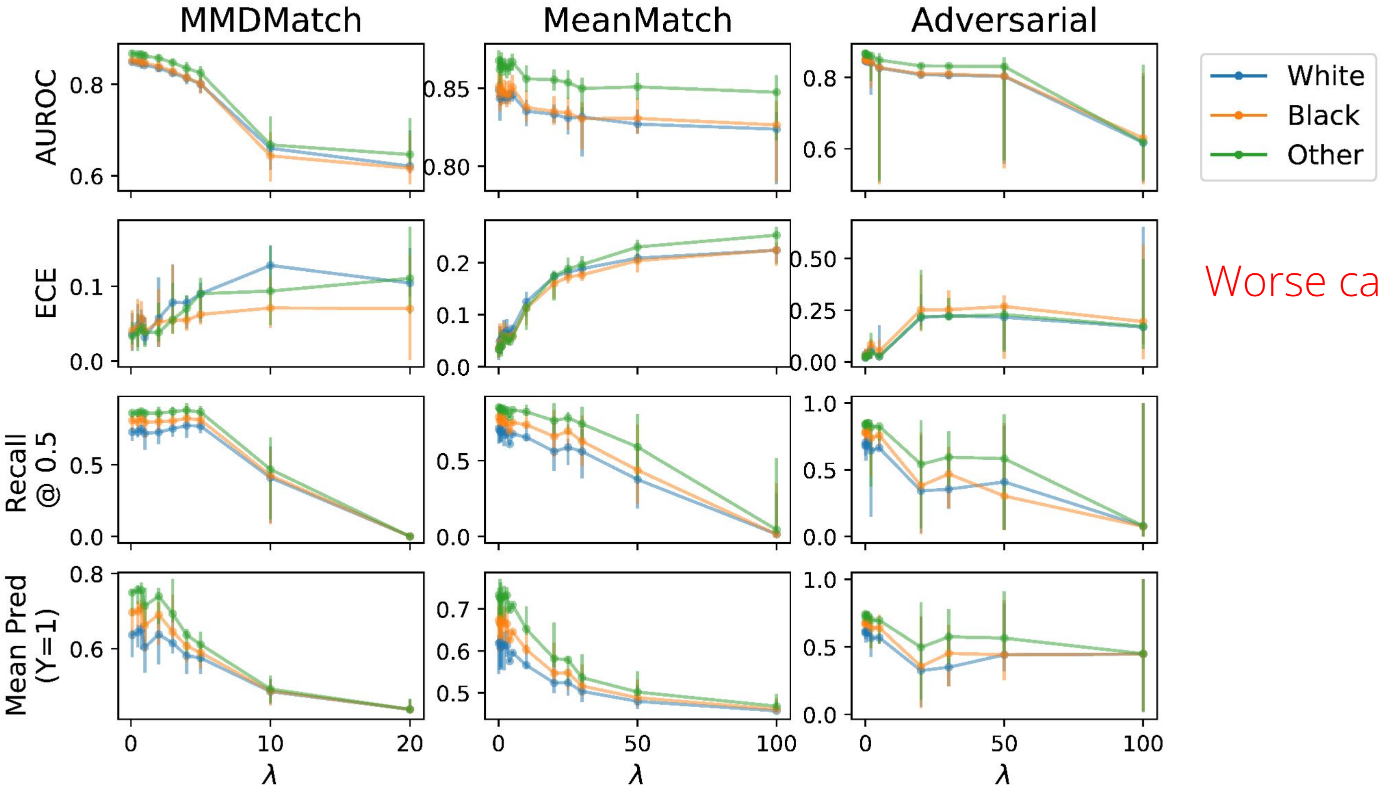
$$\min_{\theta} \; \mathbb{E}_{x,y \sim D}[\mathcal{L}(y, f_\theta(y))] + \lambda M$$

$$M_{EqOdds} = \sum_{y_j \in \mathcal{Y}} \sum_{G_k \in G} D(p_{f_\theta}(\cdot | G = G_k, Y = Y_j) \; || p_{f_\theta}(\cdot | Y = Y_j))$$

- Maximum Mean Discrepancy (MMD)
- Absolute difference between means
- Adversary to predict group

Pfohl, Stephen R., Agata Foryciarz, and Nigam H. Shah. "An empirical characterization of fair machine learning for clinical risk prediction." Journal of biomedical informatics 113 (2021): 103621.

Issues with Enforcing Group Fairness

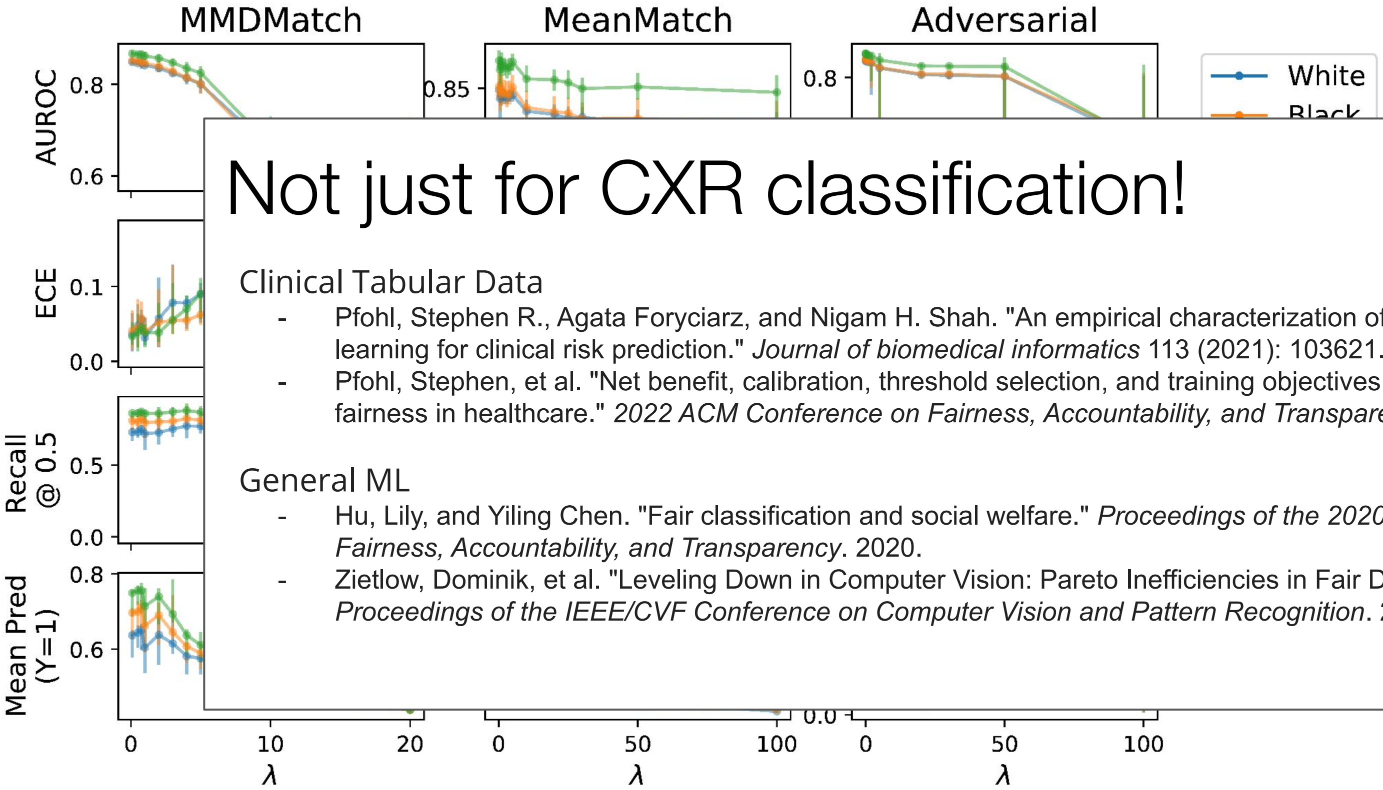


Worse performance for all groups!

Worse calibration error

Zhang, Haoran, et al. "Improving the Fairness of Chest X-ray Classifiers." *Conference on Health, Inference, and Learning*. PMLR, 2022.

Group Fairness Worsens All Groups



Not just for CXR classification!

Clinical Tabular Data

- Pfohl, Stephen R., Agata Foryciarz, and Nigam H. Shah. "An empirical characterization of fair machine learning for clinical risk prediction." *Journal of biomedical informatics* 113 (2021): 103621.
- Pfohl, Stephen, et al. "Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare." *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.

General ML

- Hu, Lily, and Yiling Chen. "Fair classification and social welfare." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020.
- Zietlow, Dominik, et al. "Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

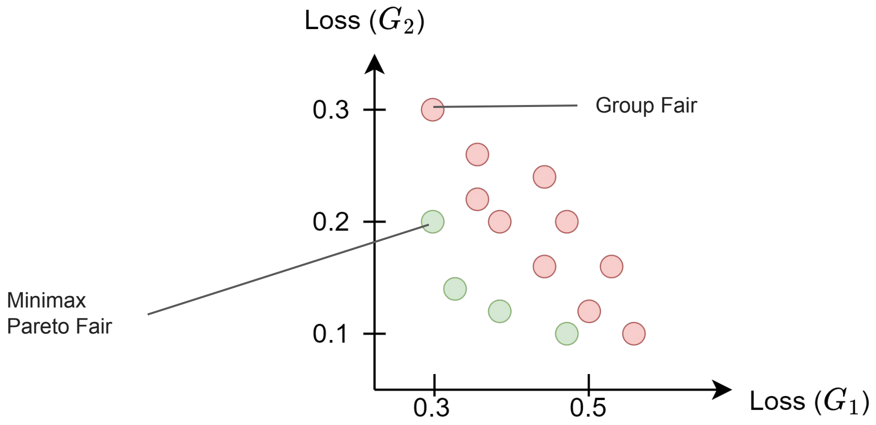
Zhang, Haoran, et al. "Improving the Fairness of Chest X-ray Classifiers." *Conference on Health, Inference, and Learning*. PMLR, 2022.

The Case against Group Fairness

- Binary Case
  - **Impossibility Theorems** (e.g. Equal TPR, FPR, precision)
  - Easily achievable through **per-group thresholding** (but has many issues)
- Risk Score Case
  - **Impossibility Theorem** (per-group calibration and probabilistic equal odds)
- Overall
  - Trying to achieve group fairness results in miscalibration + worse performance for all (empirically).
  - **Not Pareto optimal**.

Chapter 2: **Minimax Pareto Fairness**

Minimax Pareto Fairness

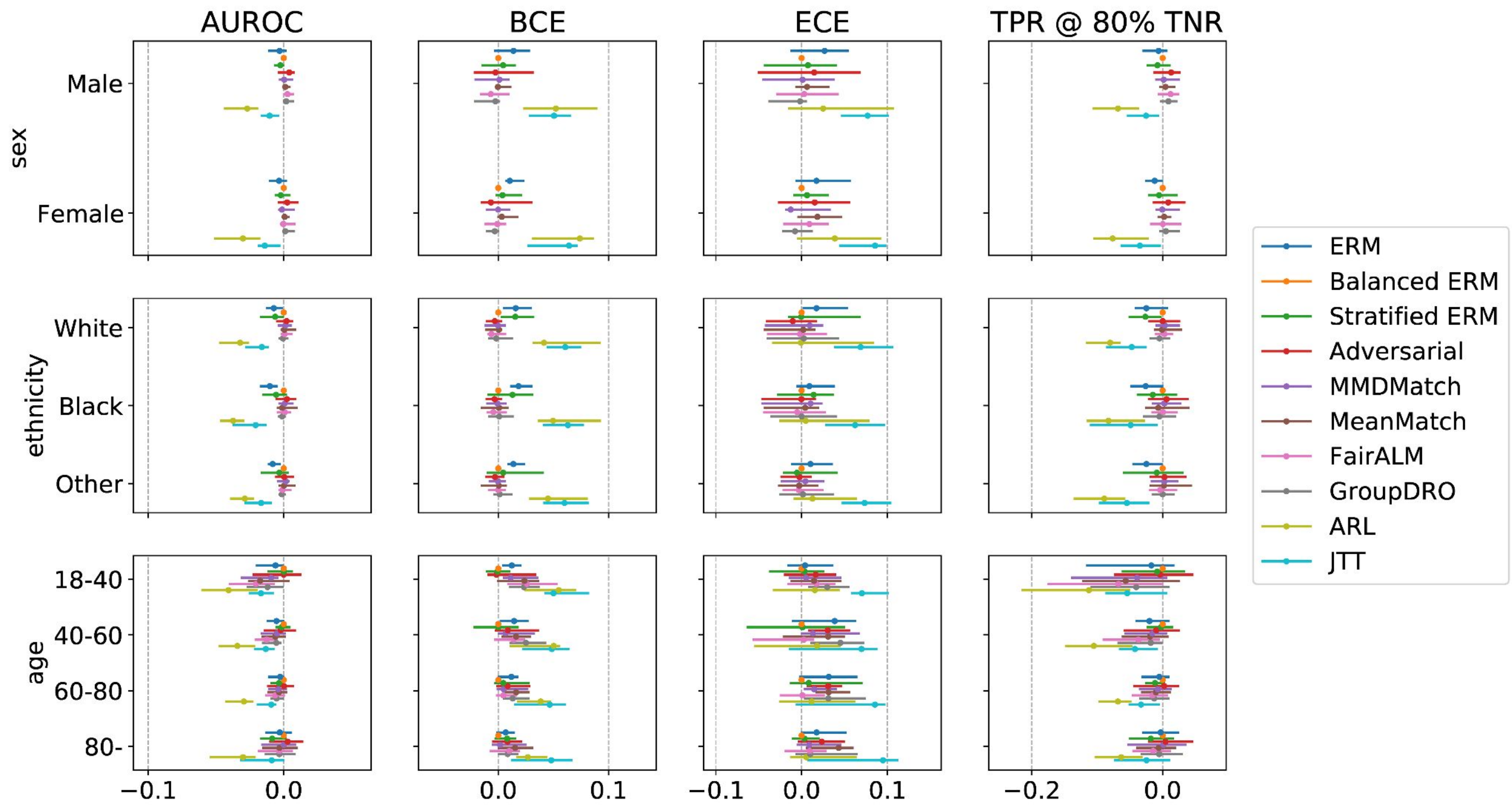


$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \max_{g \in G} \varepsilon_g(h)$$

- Can always convert a Pareto classifier into a group-fair classifier with randomization
- Relative definition of fairness
- Generally requires re-weighting and re-training

Martinez, Natalia, Martin Bertran, and Guillermo Sapiro. "Minimax pareto fairness: A multi objective perspective." *International Conference on Machine Learning*. PMLR, 2020.

No Method Outperforms Simple Data Balancing



Zhang, Haoran, et al. "Improving the Fairness of Chest X-ray Classifiers." *Conference on Health, Inference, and Learning*. PMLR, 2022.

No Method Outperforms ERM

4.3 No method outperforms ERM with statistical significance

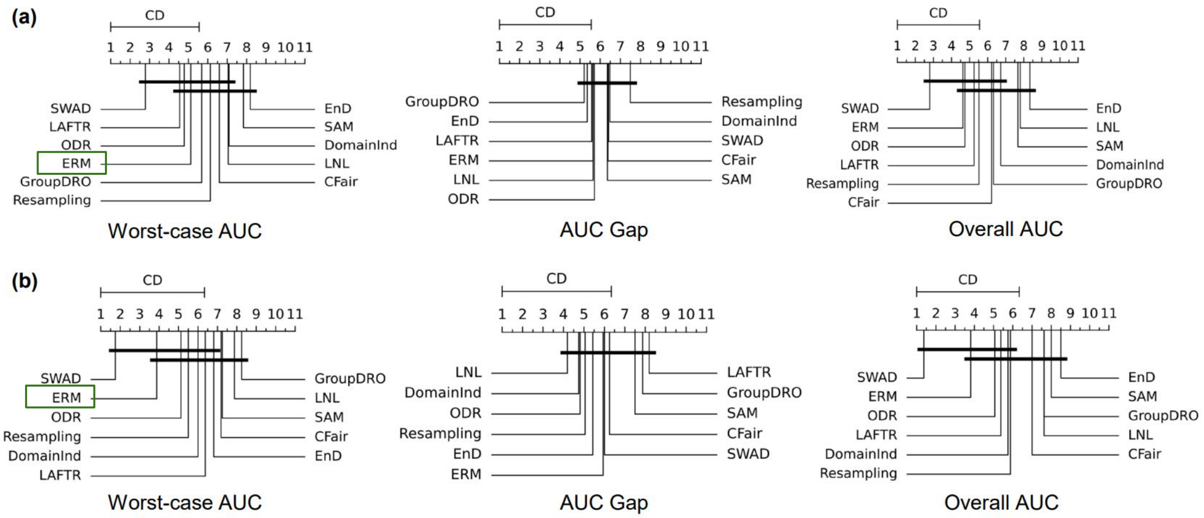


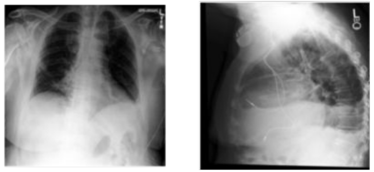
Figure 5: Performance of bias mitigation algorithms summarised across all datasets as average rank CD diagrams. (a) in-distribution, (b) out-of-distribution. SWAD is the highest ranked method for worst- and overall-AUC metrics, but it is still not significantly better than ERM.

Zong, Yongshuo, Yongxin Yang, and Timothy Hospedales. "MEDFAIR: Benchmarking Fairness for Medical Imaging." ICLR 2023.

Chapter 3: **Potential Sources of Disparity**

Definition (**Label Bias**): Observed labels differ from the ground truth at different rates for different groups.

Is there any mislabelling in CXRs? **Yes!**



Images from Study

FINAL REPORT
EXAMINATION: CHEST (PORTABLE AP)

INDICATION: ___ year old woman with NGT in place, not draining // please eval for NGT position

COMPARISON: None.

FINDINGS:

An NG tube is present, tip extending beneath diaphragm. The tip and side-port overlie the expected site of the gastric fundus.

Low inspiratory volumes with bibasilar atelectasis. Cardiomediastinal silhouette is prominent, but likely accentuated by low inspiratory volumes.

Radiology Note

Atelectasis: 1
Pneumonia: 0
No Finding: 0
...

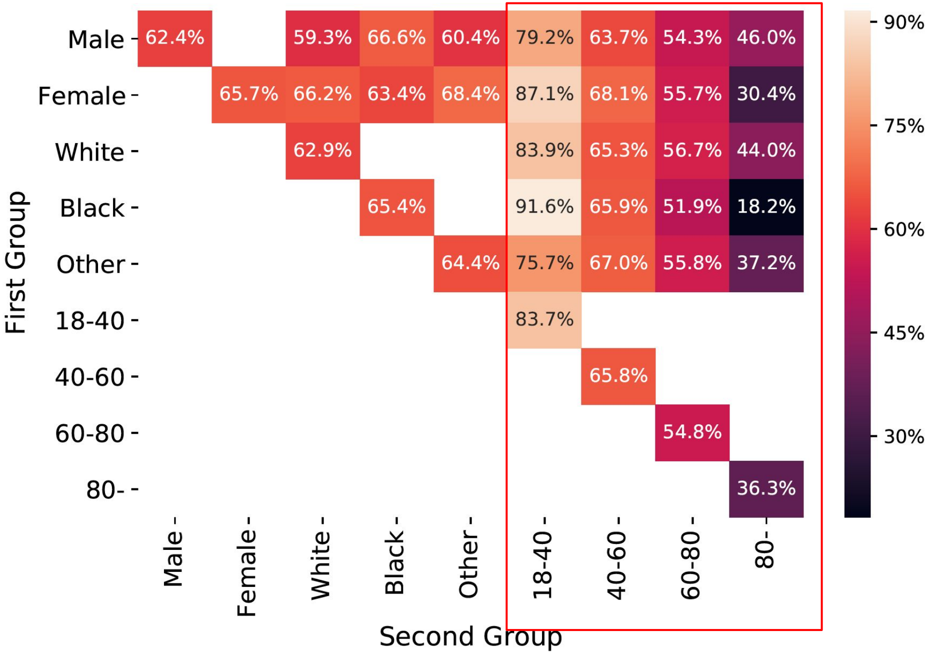
Labels

Radiologist
Studied in [1]
Kappa ≈ 0.4

CheXpert Labeller
F1 ≈ 50% [2]

[1] Jain, Saahil, et al. "VisualCheXbert: addressing the discrepancy between radiology report labels and image labels." *Proceedings of the Conference on Health, Inference, and Learning*. 2021.
[2] Smit, Akshay, et al. "CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT." *arXiv preprint arXiv:2004.09167* (2020).

Label Bias May be Responsible for Observed Gaps



Accuracy of 1,200 images from MIMIC-CXR labelled as No Finding by the automatic labeller, manually labelled by radiologist

Zhang, Haoran, et al. "Improving the Fairness of Chest X-ray Classifiers." *Conference on Health, Inference, and Learning*. PMLR, 2022.

Potential Impact of **Label Bias**

- Lower quality **training data** for some groups.

- Inaccurate **test set** metrics.

- Higher **Bayes error** for certain groups.

- Needs **better quality data**, not just more data.

Chapter 4:
Shortcut Learning

ERM Models Learn **Shortcuts**.

Definition (**Shortcut**): A feature that is correlated with the label, but is not used in the true labelling function.

Stock, Pierre, and Moustapha Cisse. "Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

Shortcut Learning - A Toy Example

Attributes = {Desert background, Grass background}

Labels = {Cow, Camel}

Groups = {Camels on grass, Cows on sand, Camels on sand, Cows on grass}

Few samples

Many samples

Shortcut Learning - A Toy Example

Few samples

Many samples

ERM Classifier: f(X) = cow if background is grass; else camel

Spurious Strength: Image → Background → Animal (2 ingredients)

Invariant Strength: Image → Animal

(Informal) ERM learns on the shortcut when
spurious strength > invariant strength

Shortcut Learning - A Toy Example




Low Accuracy High Accuracy

TPR_{grass} TNR_{desert} TPR_{desert} TNR_{grass}

Worse accuracy on unseen attributes

Group Fairness: $\min(|TPR_{grass} - TPR_{desert}|)$, $\min(|TNR_{grass} - TNR_{desert}|)$

Shortcut Learning can cause TPR/FPR gaps!

Yang, Y*., Zhang, H*., Katabi, D., & Ghassemi, M. (2023). Change is Hard: A Closer Look at Subpopulation Shift. ICML 2023.

Shortcut learning in COVID-19 prediction

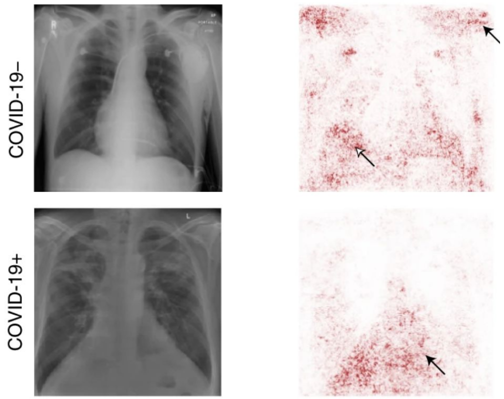
The Ingredients

(a)

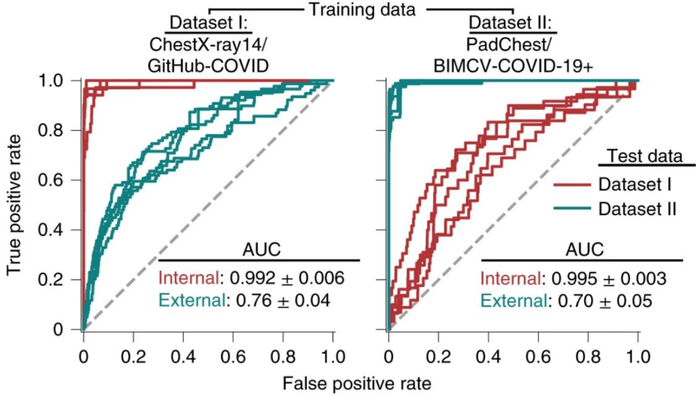
b

	Dataset I			Dataset II		
	Combined	Chest-X-ray14	GitHub-COVID	Combined	PadChest	BIMCV-COVID-19+
No. radiographs	112,528	112,120	408	97,866	96,270	1,596
No. patients	31,067	30,805	262	64,954	63,939	1,105
% COVID-19+	0.2	0	76.5	1.6	0	100
% AP images	39.9	40	26	5.6	4.7	58.1

(b)



The Symptom



DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. "AI for radiographic COVID-19 detection selects shortcuts over signal." Nature Machine Intelligence 3, no. 7 (2021): 610-619.

Can Race be a Shortcut?

(a) Chest X-ray → Race

	Area under the receiver operating characteristics curve value for race classification		
	Asian (95% CI)	Black (95% CI)	White (95% CI)
Primary race detection in chest x-ray imaging			
MXR Resnet34	0·986 (0·984–0·988)	0·982 (0·981–0·983)	0·981 (0·979–0·982)
CXP Resnet34	0·981 (0·979–0·983)	0·980 (0·977–0·983)	0·980 (0·978–0·981)
EMX Resnet34	0·969 (0·961–0·976)	0·992 (0·991–0·994)	0·988 (0·986–0·989)
External validation of race detection models in chest x-ray imaging			
MXR Resnet34 to CXP	0·947 (0·944–0·951)	0·962 (0·957–0·966)	0·948 (0·945–0·951)
MXR Resnet34 to EMX	0·914 (0·899–0·928)	0·983 (0·981–0·985)	0·975 (0·973–0·978)
CXP Resnet34 to MXR	0·974 (0·971–0·977)	0·955 (0·952–0·957)	0·956 (0·954–0·958)
CXP Resnet34 to EMX	0·915 (0·901–0·929)	0·968 (0·965–0·971)	0·954 (0·951–0·958)
EMX Resnet34 to MXR	0·966 (0·962–0·969)	0·970 (0·968–0·972)	0·964 (0·962–0·965)
EMX Resnet34 to CXP	0·949 (0·946–0·952)	0·973 (0·970–0·977)	0·947 (0·945–0·950)

Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L. C., ... & Zhang, H. (2022). AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health, 4(6), e406-e414.

Can Race be a Shortcut?

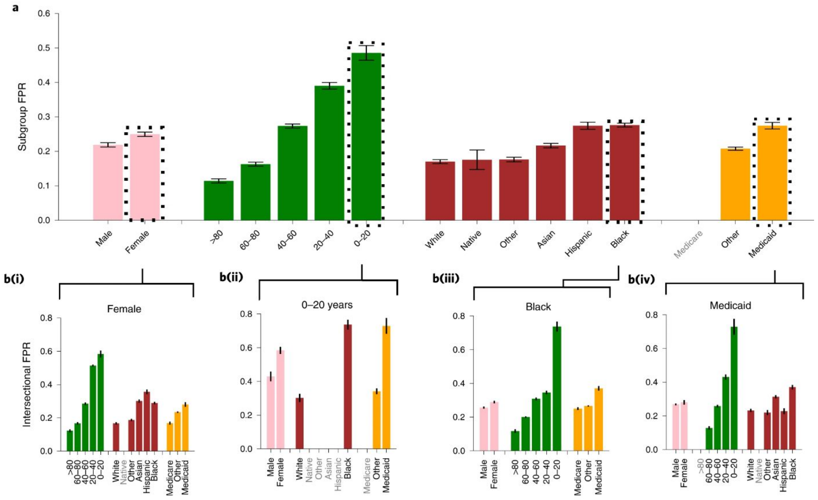
The (Potential) Causes

(a) Chest X-ray → Race

(b)

	MIMIC-CXR		
	No Finding	Fracture	Pneumothorax
Male	37.09%	1.88%	4.00%
Female	42.62%	1.46%	2.77%
White	34.60%	1.98%	4.04%
Black	44.29%	0.74%	1.81%
Other	49.87%	1.54%	2.85%
18-40	63.41%	1.02%	3.58%
40-60	45.51%	1.65%	3.20%
60-80	31.91%	1.75%	3.68%
80-	22.86%	2.25%	2.93%
Overall	39.73%	1.68%	3.41%

The Symptom?



Is shortcut learning responsible for TPR/FPR gaps?

Combating Shortcut Learning

Chest X-ray → Race → No Finding

Combating Shortcut Learning

Chest X-ray → Race → No Finding

Strategy 1:
Remove race information from (representations of) chest X-rays.
(e.g. domain adversarial training, GAN data augmentation)

Brown, Alexander, et al. "Detecting and Preventing Shortcut Learning for Fair Medical AI using Shortcut Testing (ShorT)." arXiv preprint arXiv:2207.10384 (2022).

Combating Shortcut Learning

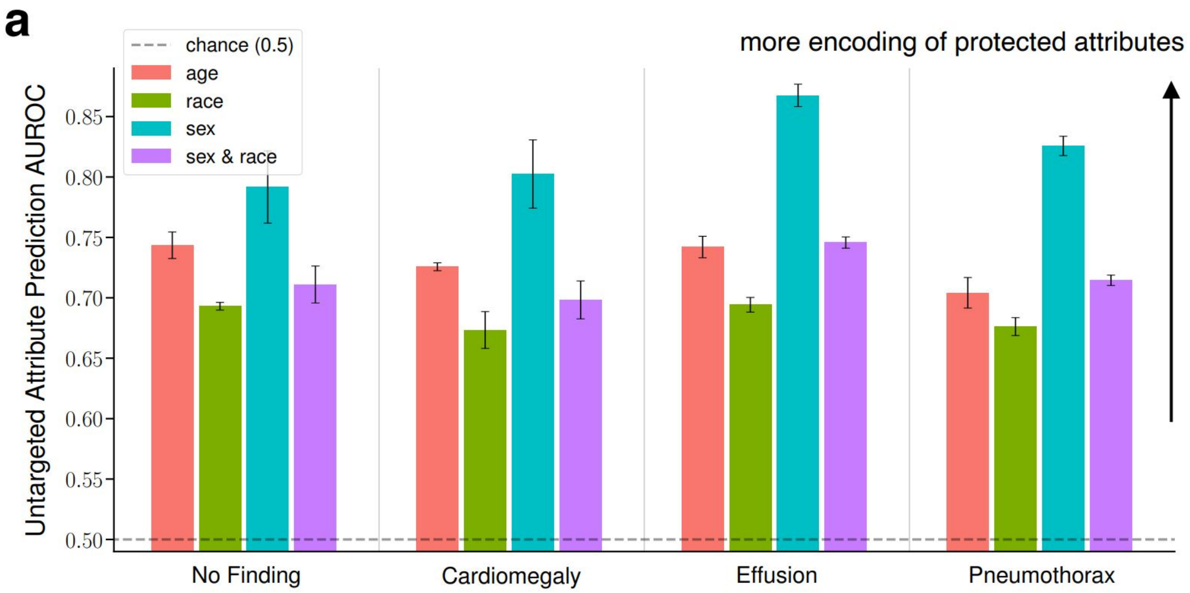
Chest X-ray → Race → No Finding

Strategy 2:
De-correlate race and the No Finding label.
(e.g. by **resampling** minority groups, **GroupDRO**)

Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization." arXiv preprint arXiv:1911.08731 (2019).

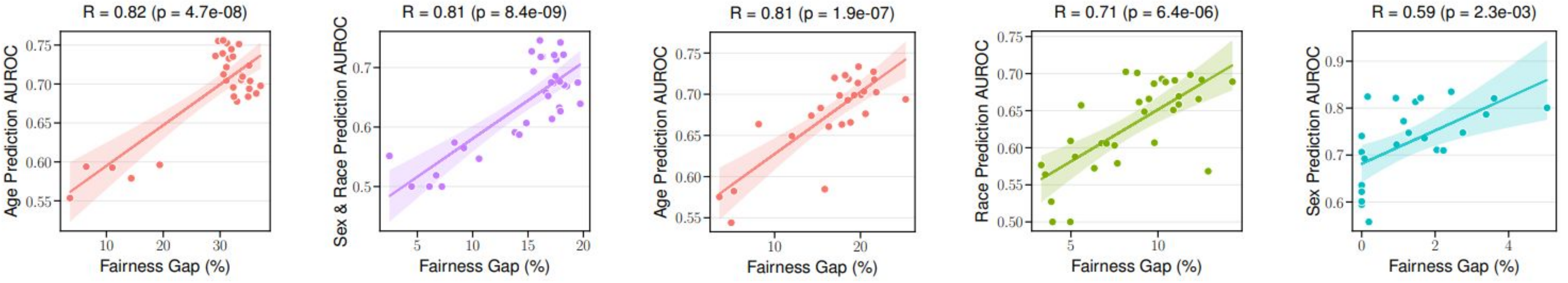
Disease Prediction Models Encode Demographics

MIMIC-CXR; Equal opportunity



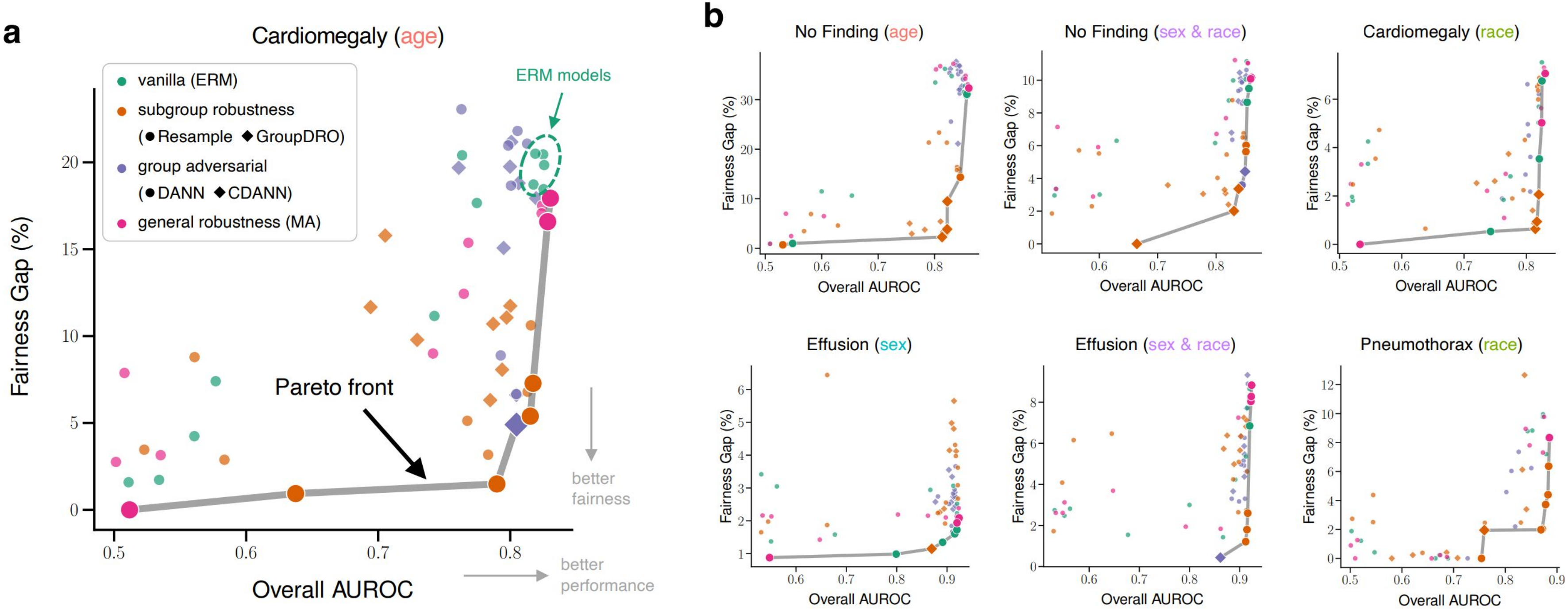
Yang, Y*., Zhang, H*., Gichoya, J., Katabi, D., & Ghassemi, M. (2023) On Mitigating Shortcut Learning for Fair Chest X-ray Classification. In Preparation.

Attribute Encoding Correlated With Fairness Gaps



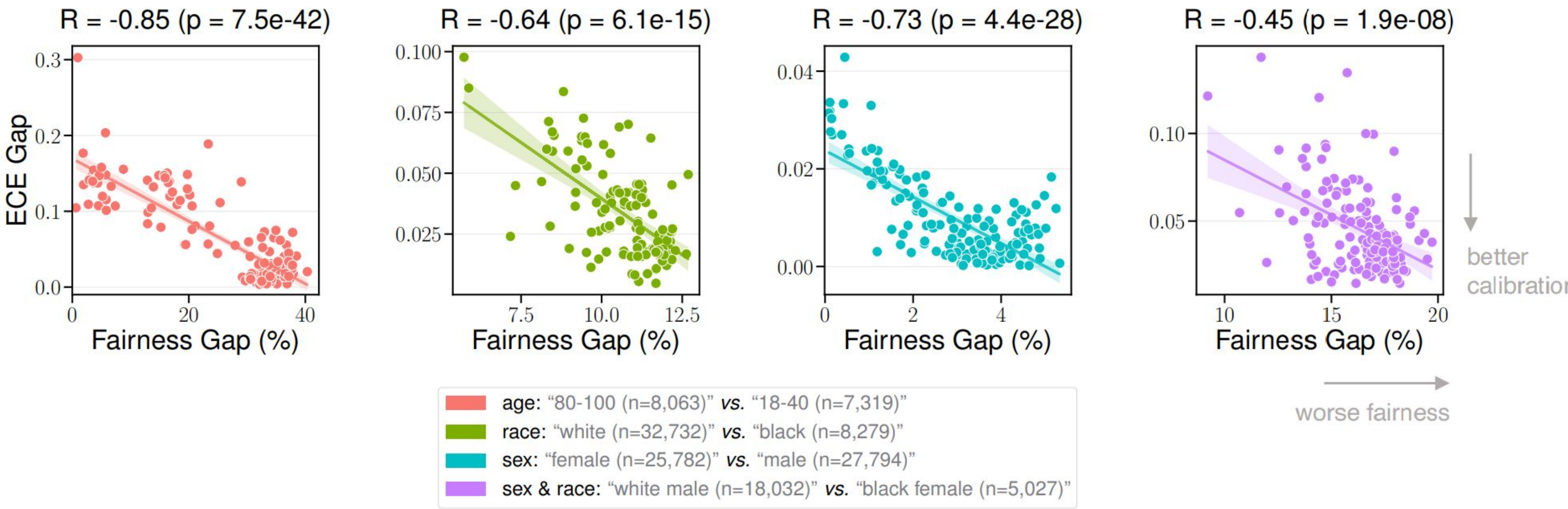
Yang, Y*., Zhang, H*., Gichoya, J., Katabi, D., & Ghassemi, M. (2023) On Mitigating Shortcut Learning for Fair Chest X-ray Classification. In Preparation.

Fair Models Maintain Decent Performance



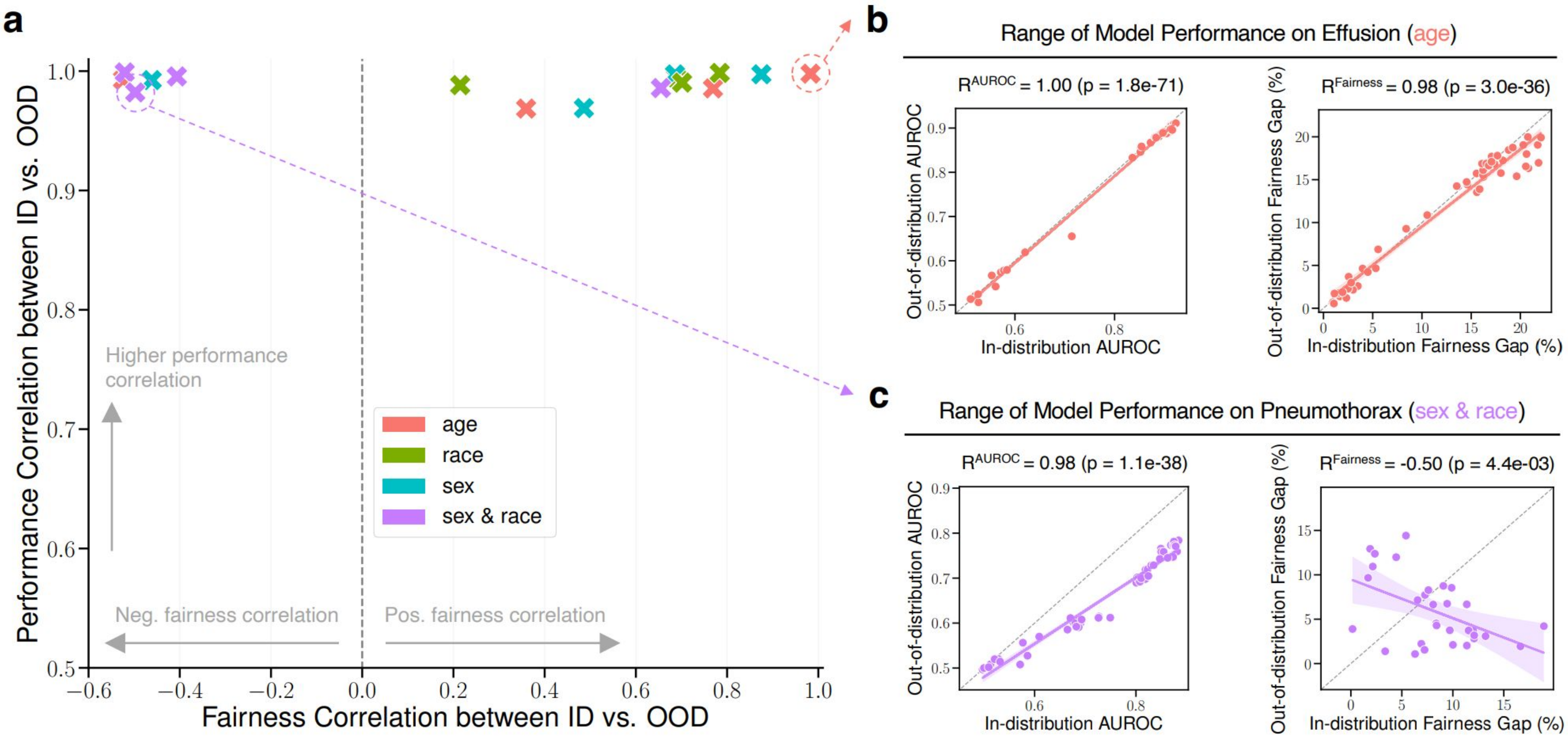
Yang, Y*., Zhang, H*., Gichoya, J., Katabi, D., & Ghassemi, M. (2023) On Mitigating Shortcut Learning for Fair Chest X-ray Classification. In Preparation.

Fairness Trades-Off with Calibration



Yang, Y*., Zhang, H*., Gichoya, J., Katabi, D., & Ghassemi, M. (2023) On Mitigating Shortcut Learning for Fair Chest X-ray Classification. In Preparation.

Fairness Does Not Always Transfer to OOD



Yang, Y*., Zhang, H*., Gichoya, J., Katabi, D., & Ghassemi, M. (2023) On Mitigating Shortcut Learning for Fair Chest X-ray Classification. In Preparation.

Shortcut Learning Results – Summary

- Observed **tradeoffs** are very similar to the group fairness setting

- Shortcut removal methods (vs. ERM):
  - Worsens overall and all-group AUROC (slightly)
  - Worsens overall calibration
  - Worsens calibration gap
  - Betters group fairness (binary)
  - Betters group fairness (risk score)
  - Fairness attained does not transfer to OOD

- By targeting the shortcut learning case, we may be able to achieve a **better trade-off** than blindly applying debiasing methods.

Chapter 5**:**

Concluding Remarks

Practical Recommendations

- **Evaluate comprehensively**. Evaluate a wide variety of threshold-free and thresholded metrics, especially calibration error.
- **Consider sources of bias in the data**. Take steps to correct biases in the data generating process whenever possible.
- **Many trade-offs exist**. Determine whether gaps are clinically justified. Correcting gaps could lead to worse performance for all.
- **Inductive biases** about how disparate performance originates may lead to targeted interventions with more favorable tradeoffs.
- **Algorithmic approaches alone are insufficient** to ensure that the use of machine learning in healthcare is equitable.

Promising Directions of Research

- Fairness under **distribution shift**. [1-2]
- Fairness under **sampling and label bias**. [3-4]
- Fairness with **unknown or combinatorially many groups**. [5-6]
- **New fairness definitions** and their limitations [7]
- **Fairness in different problem settings** (e.g. ranking [8], generative models [9]).

[1] Robust fairness under covariate shift. AAAI 2021.
[2] Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. NeurIPS 2022.
[3] Unlocking fairness: a trade-off revisited. NeurIPS 2019.
[4] Fair Classification with Group-Dependent Label Noise. ACM FAccT 2021.
[5] Blind Pareto Fairness and Subgroup Robustness. ICML 2021.
[6] Multicalibration: Calibration for the (Computationally-Identifiable) Masses. ICML 2018.
[7] Causal Conceptions of Fairness and their Consequences. ICML 2022.
[8] Fairness in ranking under uncertainty. NeurIPS 2021.
[9] Fair generative modeling via weak supervision. ICML 2020.

Thank you!