

Study of bias in applications & Counterfactual Fairness

Jinyi Xie, Yu Feng

RESEARCH ARTICLE

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

Obermeyer et al., 2019

Overview

- Introduction to health prediction algorithm
- Racial bias and its source
- An experiment: choice of label

An algorithm for patients' needs...

“We found that effective programs customize their approach to their local contexts and caseloads; use a combination of qualitative and quantitative methods to identify patients.”

——C. S. Hong, A. L. Siegel, T. G. Ferris, 2014

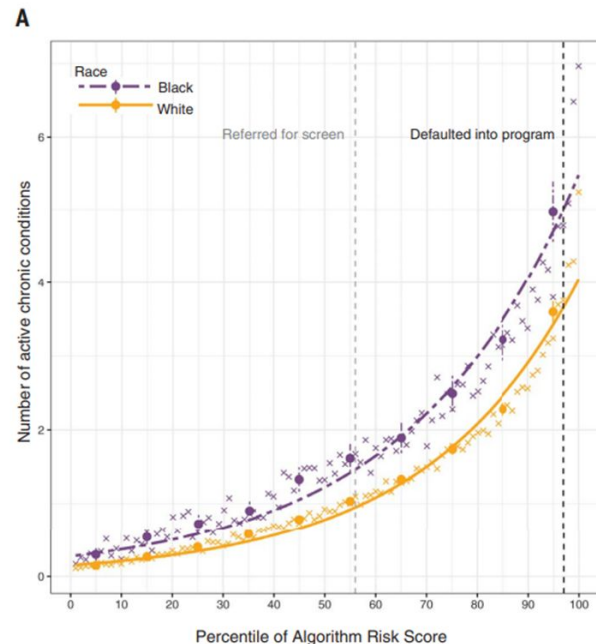
An algorithm for patients' needs...

- Based on past data, algorithm predicts the “risk score”

R for each patient:

How urgent / serious the need of treatment is

- Two thresholds for risk score percentile
 - Auto-identification threshold (over 97th percentile)
 - Screening threshold (over 55th percentile)



An algorithm for patients' needs...

- How is this risk-score prediction used? Thoughts?

An algorithm for patients' needs...

- How is this risk-score prediction used?
 - Risk-Prediction algorithm for application of high-risk care management program
 - Aims to satisfy the patients needs & reduce cost
 - Patients with largest health need benefits most from the program
 - Used by nationwide large health system, Influencing over 200 million people in U.S. each year

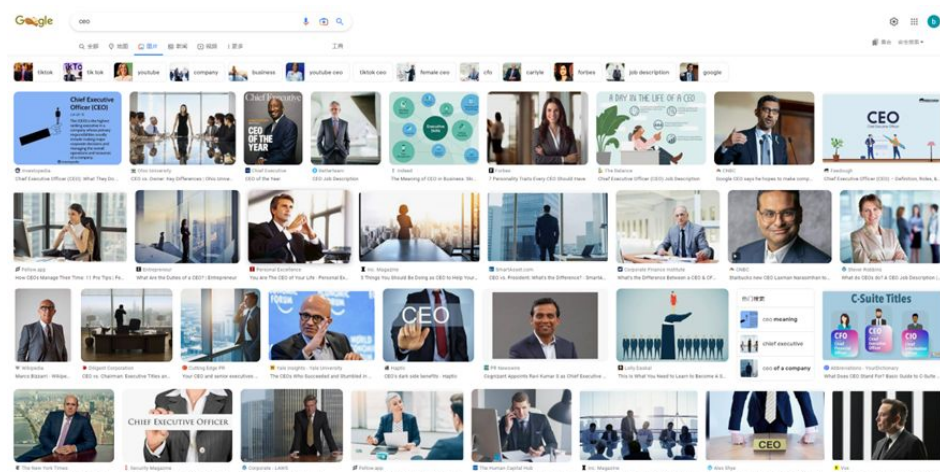
Recall: Issues of algorithmic bias

Searching with black names are more likely to return arrest record ads

Image searches for professions such as CEO produce fewer images of women

INSTANT CHECKMATE ADS ON GOOGLE

	OBSERVED			EXPECTED		
	BLACK	WHITE		BLACK	WHITE	
Arrest Ads	335	92%	53	80%	388	90%
Neutral Ads	31	8%	13	20%	44	10%
Totals	366	66	432	329	59	7



Issues of algorithmic bias could be hard to investigate...

- Such algorithms are usually proprietary
- Researches could estimate bias from outside
- Getting insight is hard, as the lack of knowledge of algorithm

Dataset

- A rich dataset produced by academic hospital
 - Includes data for all primary care patients with insurance from 2013 to 2015
 - Focus on disparity between white and black patients
 - Categorize race by patient self-identification
- Dataset also include algorithmic knowledge

Metrics: Calibration Bias

- The metric of algorithmic bias for real world use of the algorithm
- Conditioned on risk score, check whether realized value of variable match
- Formally, for some variable of interest Y , at a certain level of risk score R , compare for black B and white W ,

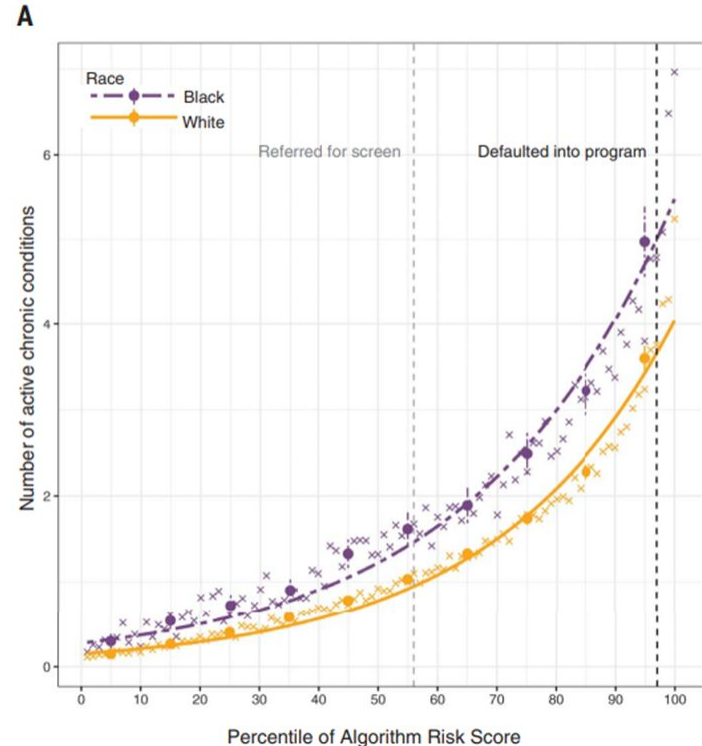
$$E[Y|R,W] \text{ and } E[Y|R,B]$$

Notations of algorithm

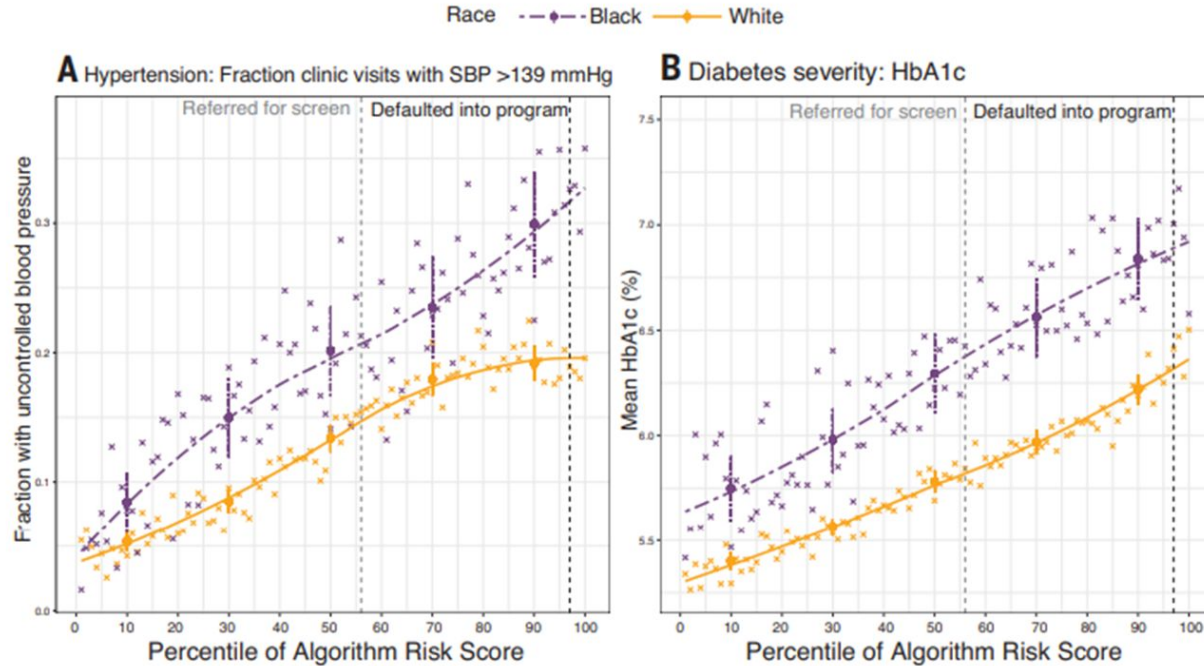
$R_{i,t}$	Risk score for patient i at year t
$X_{i,(t-1)}$	Claims data for patient i at previous year
$H_{i,t}$	Health measurement for patient i at year t

Result: Health Disparities between Race

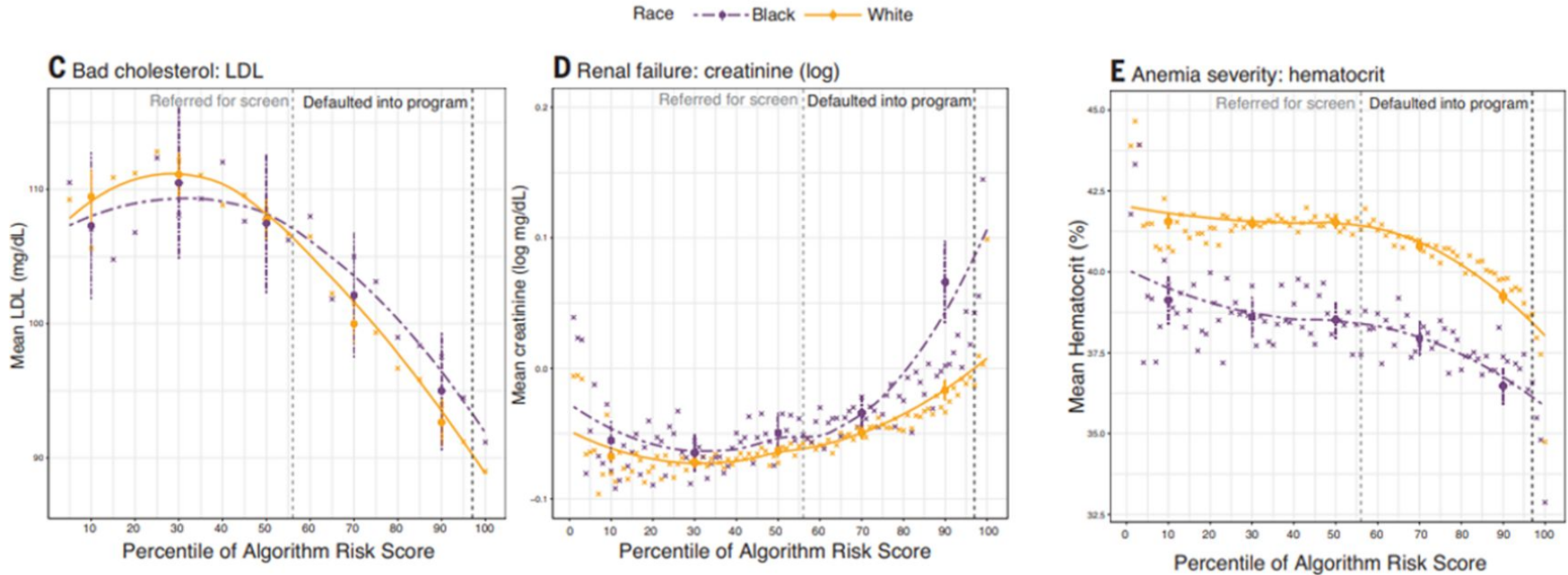
- Significantly more serious illness for black patients
- Focus on high-risk patients
- 26% more chronic conditions for black patients at auto-identification threshold



More Details on Health Disparities



More Details on Health Disparities

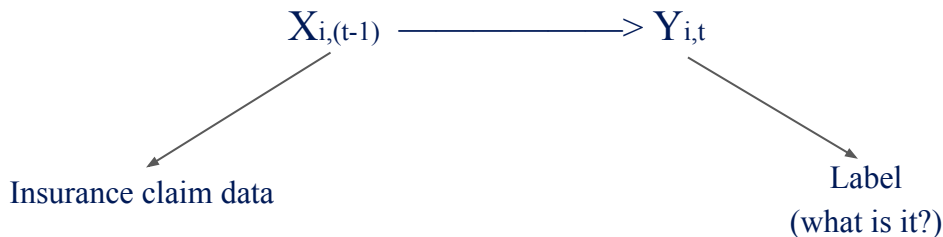


Where could such algorithmic bias come from?

Insights to the Algorithm

With the dataset, we were able to observe the algorithm's inputs, outputs and objective function.

As a predictive algorithm, it used patient's previous year's insurance claim data $X_{i,(t-1)}$ to predict the label $Y_{i,t}$.



The Algorithm's Label

The algorithm used total health cost (medical expenditure) for year t as the training label

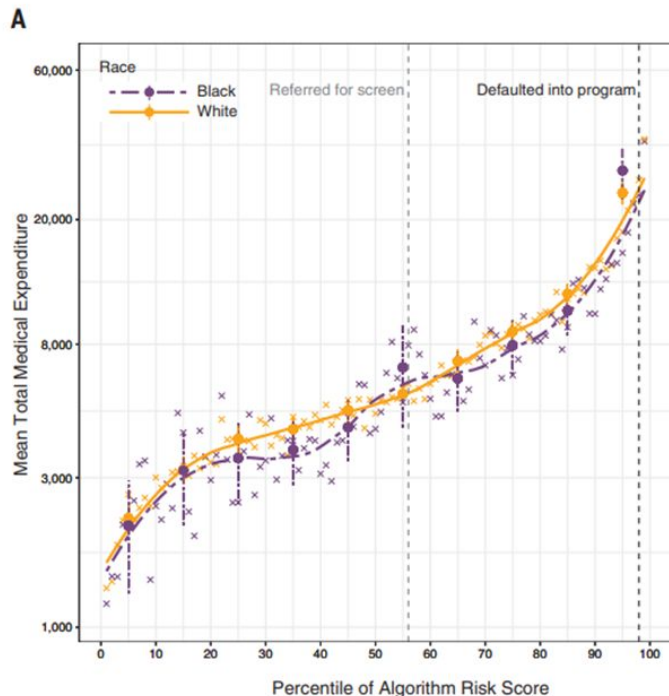


The Algorithm's Label: How is the calibration?

$E[C|R,W]$ and $E[C|R,B]$

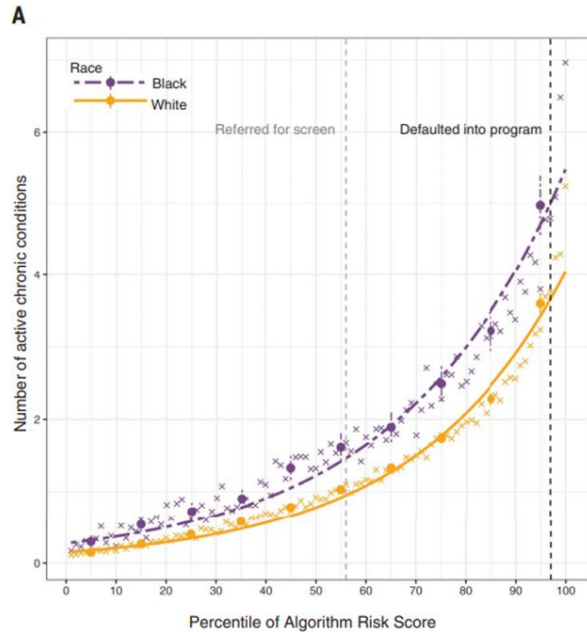
(almost) matches for every level of R

The algorithm is well calibrated across race for medical expenditure (unbiased).



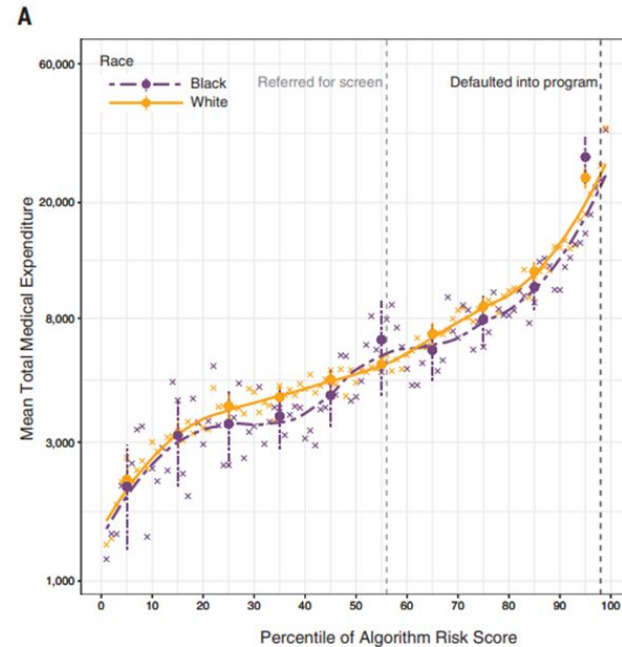
What produces the bias?

Disparity on Health Condition



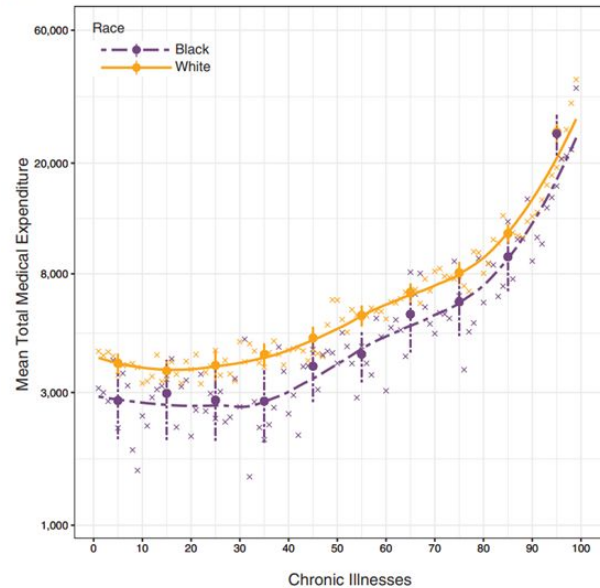
VS

No Disparity on Health Cost



Health Condition vs Health Cost

At the same level of health condition(number of chronic conditions), black patients have much lower health cost than white patients.

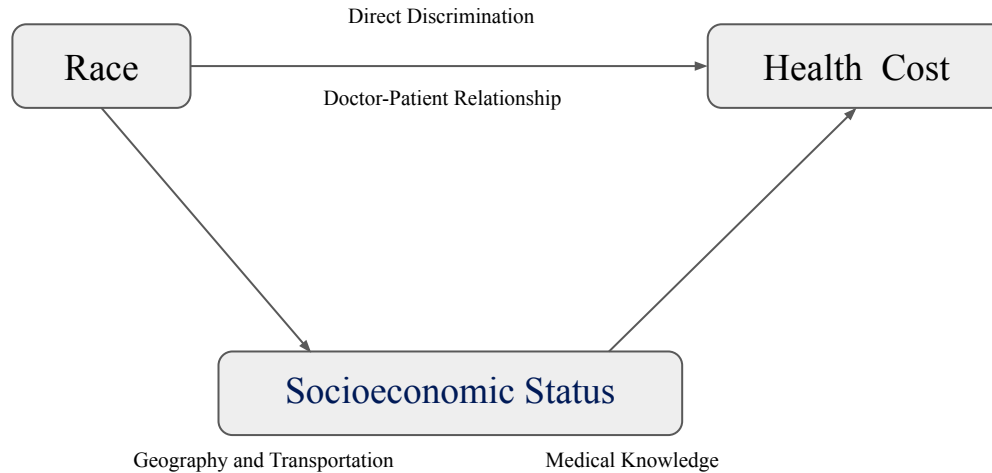


Unbiased prediction on health cost

→ Bias for health condition

What could be the cause of difference in cost?

Race and Health Cost



Importance of Label

- Health is a complex issue that's hard to measure
- Seemingly reasonable proxies may lead to bias
- Other possible labels?

Importance of Label

- Health is a complex issue that's hard to measure
- Seemingly reasonable proxies may lead to bias
- Other possible labels by this paper
 - Avoidable Cost (Emergency Visits etc.)
 - Health Condition (Number of Chronic Conditions)

Importance of Label: An Experiment

Train three predictive algorithms in same way, using different labels:

- Total Health Cost
- Avoidable Health Cost
- Health Condition

Train with random 2/3 training set, show result from 1/3 holdout set.

Importance of Label: An Experiment

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
Active chronic conditions	0.121	(0.003)	0.182	(0.003)	0.148	(0.003)	0.267	(0.003)
Best-to-worst difference	0.044		0.033		0.043		0.126	

Conclusion

- The algorithm predicts on health cost, which by itself is a racially biased label
- Be careful with label choice ——— Could lead to very diverse/biased predictions
- By creating combined index variable as label, bias could be reduced by 84%
- Limitations
 - Did not count for other races/intersectional races
 - Algorithmic knowledge is usually unavailable
 - This algorithm is industry leading, yet not unique

Counterfactual Fairness

Matt J. Kusner, Joshua R. Loftus, Chris Russell, Ricardo Silva

Background

AI can be racist!

The Boston Globe

Metro Sports Business & Tech Opinion Politics Lifestyle Arts

Menu **Business** SIGN UP NOW Get Globe.com newsletters delivered to your inbox

Racial bias alleged in Google's ad results

Names associated with blacks prompt link to arrest search



Ad related to latanya sweeney

Latanya Sweeney Truth
www.instantcheckmate.com/
Looking for Latanya Sweeney? Check Latanya Sweeney's Arrests.

Ads by Google

Latanya Sweeney Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

InstantCheckmate

Latanya Sweeney
Background Check
Criminal History
Sexual History
Mental Health
Social Media
Public Records
Court Records
Arrest Records
Driving Records
Credit Records
Employment Records
Education Records
Marriage Records
Divorce Records
Bank Records
Insurance Records
Voter Records
Property Records
Business Records
Professional Records
Academic Records
Military Records
Religious Records
Political Records
Financial Records
Health Records
Legal Records
Other Records

E-mail this to a friend

Printable version

HP camera 'can't see' black faces

A YouTube video suggesting that face recognition cameras installed in HP laptops cannot detect black faces has had over one million views.



The short movie, uploaded earlier this month, features "Black Desi" and his colleague "White Wanda".

When Wanda, a white woman, is in front of the screen, the camera zooms to her face and moves as she moves.

"Black Desi" in the YouTube video

Background

AI can be sexist!

Are Facebook job ads discriminatory? Company accused of bias against women, older workers

Jessica Guynn USA TODAY

Published 1:18 p.m. ET Dec. 1, 2022

Background

It is crucial to ask if the predictions of a model are **fair** !

Q: What is a fair classifier?

Background

It is crucial to ask if the predictions of a model are **fair** !

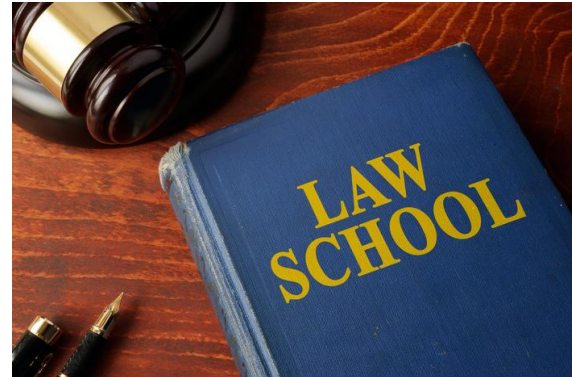
Q: What is a fair classifier?

- A **fair classifier** gives the **same prediction** had the person **had a different race/sex**.

Example: Law School Success

Given their entrance exam scores (LSAT), their grade-point average (GPA) collected prior to law school:

- A school wish to **predict** if an applicant will have a **high first year average grade (FYA)**.
- Predictions **are not biased by an individual's race and sex**.



Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Sex	Race	GPA	LSAT	FYA
male	white	0	1	
female	black	1	1	

Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Sex	Race	GPA	LSAT	FYA
male	white	0	1	1
female	black	1	1	0

Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Sex	Race	GPA	LSAT	FYA
male	white	0	1	1
female	black	1	1	0

Protected sensitive attributes A

Observable variables X

Prediction Y

Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Sex	Race	GPA	LSAT	FYA
male	white	0	1	1
female	black	1	1	0

Protected sensitive attributes A

Observable variables X

Prediction Y

What methods can we use here to make a fair prediction? /

How can we define fairness?

Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Fairness Through Unawareness (Feature Bias)	GPA	LSAT	FYA
	0	1	1
	1	1	0

Protected sensitive attributes A Observable variables X Prediction Y

Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Fairness Through Unawareness (Feature Bias)	GPA	LSAT	FYA
	0	1	1
	1	1	0

Protected sensitive attributes A Observable variables X Prediction Y

Minority student
may feel teacher
unsupportive



Example: Law School Success

Predict if an applicant will have a **high first year average grade (FYA)**.

Fairness Through Unawareness (Feature Bias)	GPA	LSAT	FYA
	0	1	1
	1	1	0

Protected sensitive attributes A Observable variables X Prediction Y

Minority student
may feel teacher
unsupportive



Limited access to
academic institutions
due to economic history



Example: Law School Success

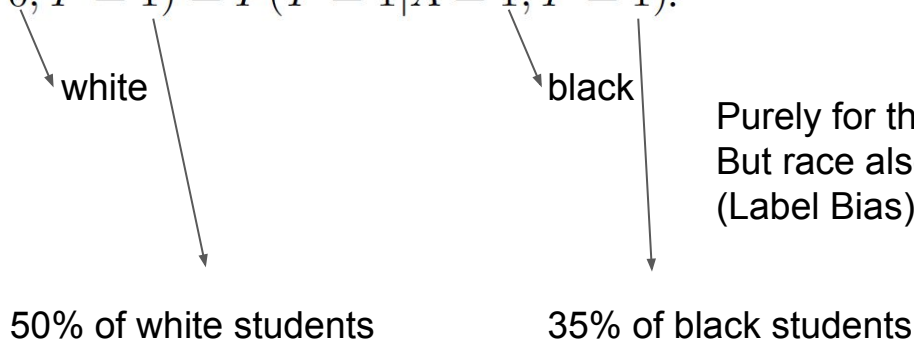
Definition 4 (Equality of Opportunity (EO)). *A predictor \hat{Y} satisfies equality of opportunity if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.*

white

black

Example: Law School Success

Definition 4 (Equality of Opportunity (EO)). A predictor \hat{Y} satisfies equality of opportunity if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.



Purely for the people who are successful,
But race also unfairly influences the outcome.
(Label Bias)

Example: Law School Success

Definition 4 (Equality of Opportunity (EO)). *A predictor \hat{Y} satisfies equality of opportunity if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.*

white

black

(Label Bias)

Claim: Any fairness notion based on **observation alone** will have similar problems

Reason: They cannot model **how discrimination happens**.

Overview

In this paper:

- Introduce the first **explicitly causal approach** to address fairness
- Provide the formal definition of **fairness**
- Provide **an algorithm** to learn **fair classifiers**

Causal models

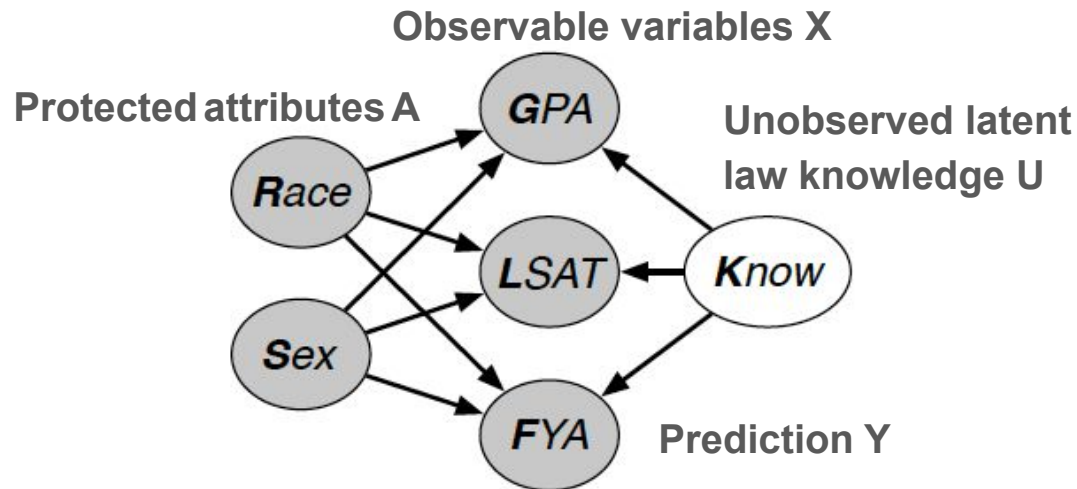
Propose to model the **discriminatory influence explicitly** before constructing a classifier.

- Model the discriminatory effect as a causal effect.
- Allow us to model how unfairness occurs.

Causal models

Propose to model the discriminatory influence explicitly before constructing a classifier.

- Model the discriminatory effect as a causal effect.
- Allow us to model how unfairness occurs.



Structural Causal Model

$$Y \leftarrow U \quad Y \sim P(Y | f(U), \dots)$$

How can we enforce fairness by using causal models?

Counterfactual fairness

What will the prediction be had the person **had a different race/sex**?

Fair : the model gives the **same** prediction on the original data as it does on a counterfactual data.

Counterfactual fairness

Protected attributes A should not be a cause of Prediction Y in any individual instance.

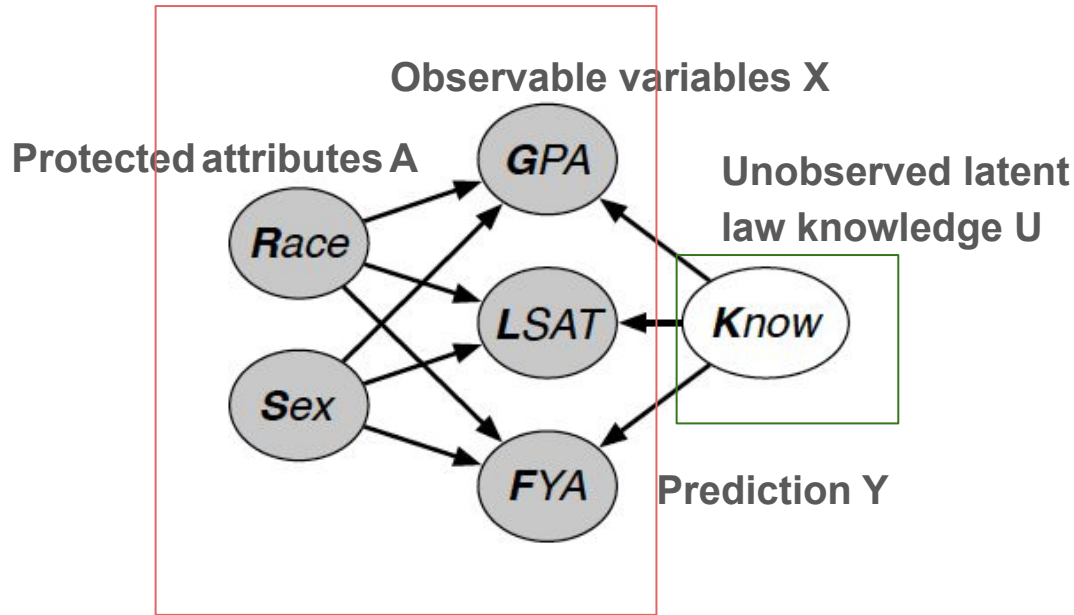
Definition 5 (Counterfactual fairness). *Predictor \hat{Y} is **counterfactually fair** if under any context $X = x$ and $A = a$,*

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \quad (1)$$

for all y and for any value a' attainable by A .


Implication


Lemma 1. Predictions using non-descendants of A are counterfactually fair



A Fair Algorithm

Given $D = \left\{ \left(x^{(i)}, y^{(i)}, a^{(i)} \right) \right\}_{i=1}^d$

1. Fit Causal Model M
2. For each data point $i \in D$, compute $u^{(i)}$


Unobserved latent U
3. $\hat{\theta} \leftarrow \arg \min_{\theta} \sum_{i \in D} L \left(y^{(i)}, \hat{Y}_{\theta} \left(u^{(i)}, x_{\neq A}^{(i)} \right) \right)$


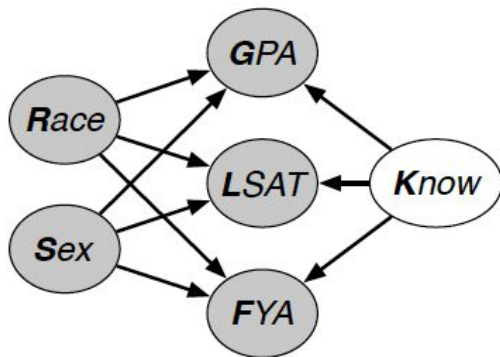
Features that are non-descendants of A
4. Return $\hat{Y}_{\hat{\theta}}$

Counterfactually Fair Causal Model

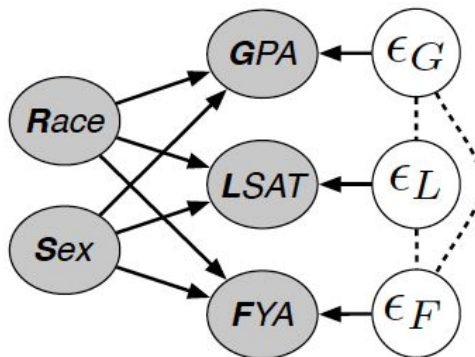
Level 1: uses any features which are not descendants of A.

Level 2: models latent 'fair' variables which are parents of observed variables. These variables are independent of A.

Level 3: models the data using an additive error model, and uses the independent error terms to make predictions



Level 2

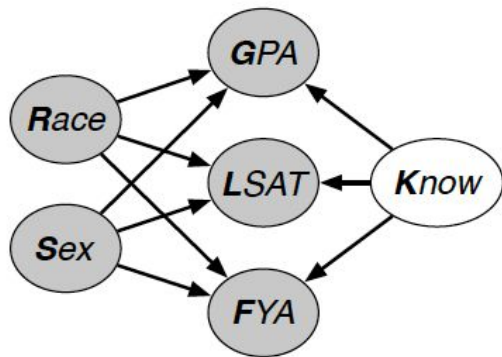


Level 3

Level 2 Causal Model

Level 2: models latent ‘fair’ variables which are parents of observed variables.

These variables are independent of A.



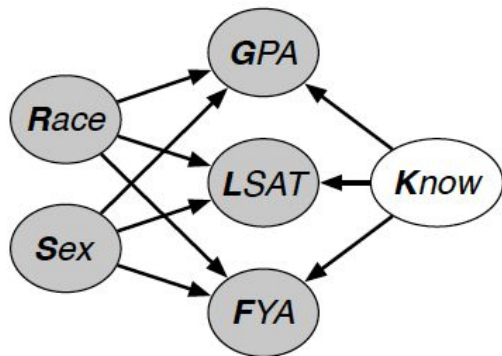
Level 2

$$\begin{aligned} \text{GPA} &\sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G), \\ \text{LSAT} &\sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)), \\ \text{FYA} &\sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1), \\ K &\sim \mathcal{N}(0, 1) \end{aligned}$$

Level 2 Causal Model

Level 2: models latent ‘fair’ variables which are parents of observed variables.

These variables are independent of A.



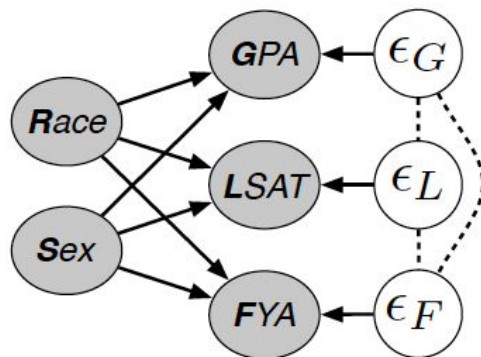
Level 2

$$\begin{aligned} \text{GPA} &\sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G), \\ \text{LSAT} &\sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)), \\ \text{FYA} &\sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1), \\ K &\sim \mathcal{N}(0, 1) \end{aligned}$$

Perform inference on this model using an observed training set to estimate the posterior distribution of *K*.

Level 3 Causal Model

Level 3: models the data using an additive error model, and uses the independent error terms to make predictions



Level 3

$$GPA = b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G)$$

$$LSAT = b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L)$$

$$FYA = b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F)$$

Estimate the error terms ϵ_G , ϵ_L by first fitting models that each use race /sex to individually predict GPA / LSAT and then compute the residuals of each model.

Use these residual estimates of ϵ_G , ϵ_L to predict FYA

Results

Compare the Root mean square error (**RMSE**) achieved by logistic regression for each of the models

Full: the standard technique of using all features, including sensitive features such as race and sex to make predictions.

Unaware: fairness through unawareness, where we do not use race and sex as features.

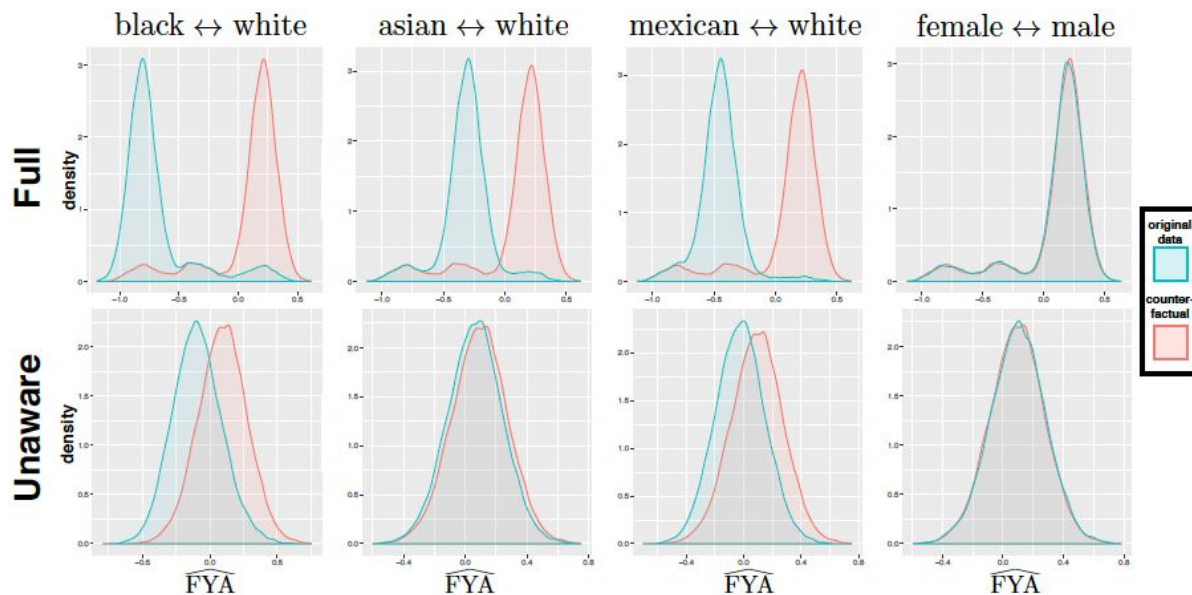
	Full	Unaware	Fair K	Fair Add
RMSE	0.873	0.894	0.929	0.918

Level 2 model

Level 3 model

Results

Empirically test whether the baseline methods are counterfactually fair.



Takeaways

- Race/Gender/Sexual Orientation could cause model decisions to change unfairly.
- Model how these attributes cause unfair decisions vis causal models.
- Counterfactual Fairness.
- Given a model, faire predictors can be derived.

Limitations

- Counterfactual Fairness only works on “individual” level.
 - Not for “group” or “subgroup” fairness
- This definition considers **the entire effect of the sensitive attribute** on the decision as problematic, and not “how” it affects the decision
 - For example, in the Berkeley alleged sex bias case, **female applicants** were rejected more often than male applicants as they were more often applying to departments with lower admission rates. Such an effect of gender through department choice is not unfair.