Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations

CSC 2541
5th Nov 21

Santosh Kolagati
Omkar Dige

Overview

- Background
- Why should we care?
- Problem Formulation
- Dataset & Analytic Strategy
- Results
- Implications & Limitations

Background

- Image searches for professions such as CEO produce fewer images of women.[1]
- Job search ads for highly paid positions are less likely to be presented to women.[2]
- Natural language processing algorithms encode language in gendered ways.[3]

1. Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations.
2. Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination.
3. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases.

Why should we care?

These systems are deployed in critical sectors such as:

- Law enforcement
- Medical care
- Education

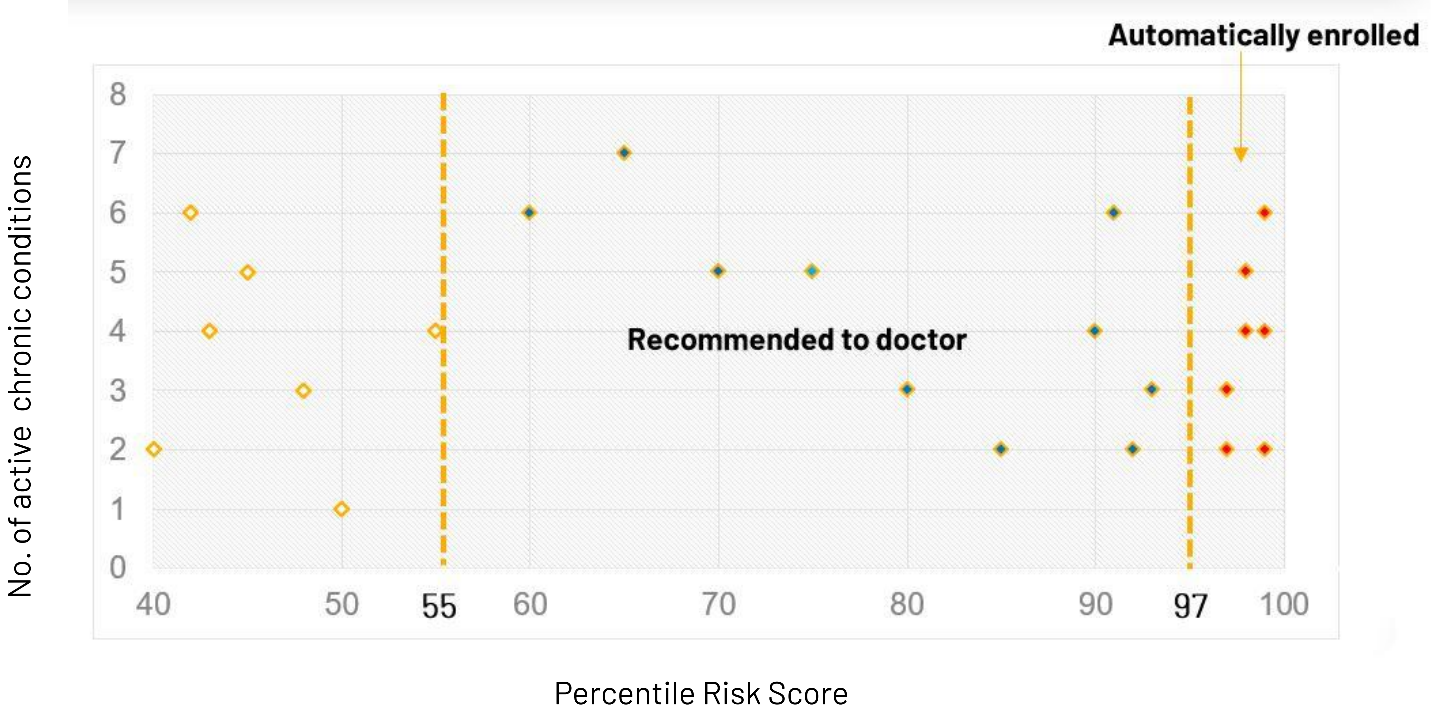
Risk Assessment

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Problem Formulation

- High-risk care management programs.
- Applied to roughly 200 million people in the US each year!
- Effective and reduce cost.
- What about bias?

Algorithm Risk Score



Check for racial disparities

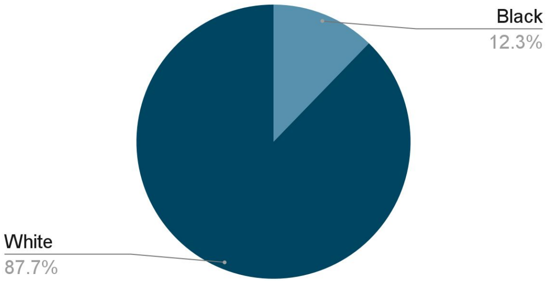
- Compare algorithmic risk score for patient *i* in year *t* ($R_{i,t}$) to data on patients' health $H_{i,t}$.
- Check how well the risk score is calibrated across race for health as well as costs, *C*.

Dataset

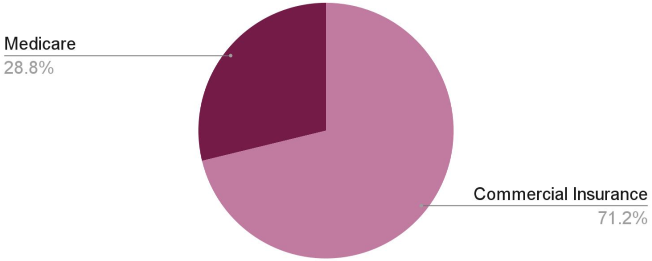
2013 2015

50.9
Avg. age
in years

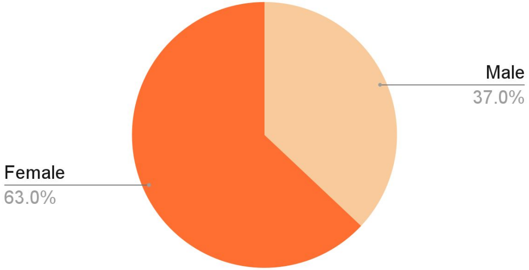
Race distribution

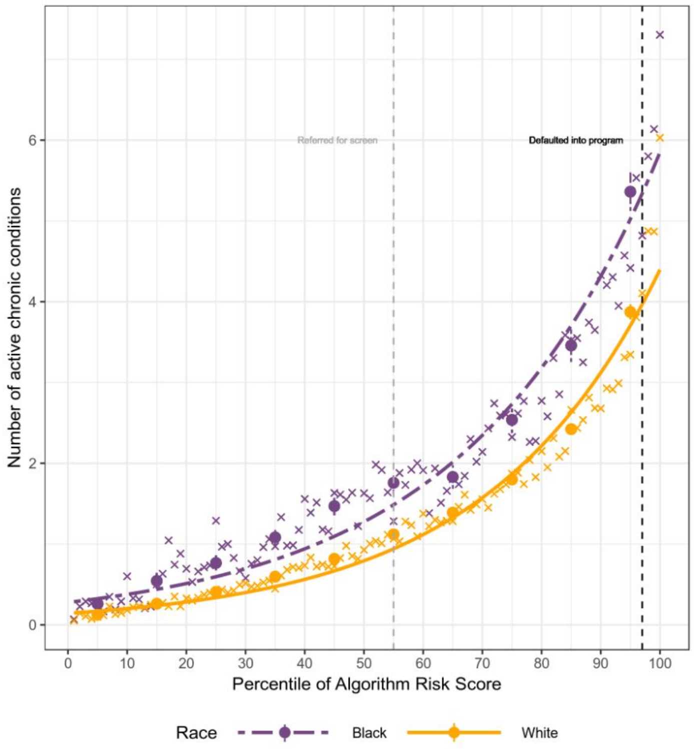


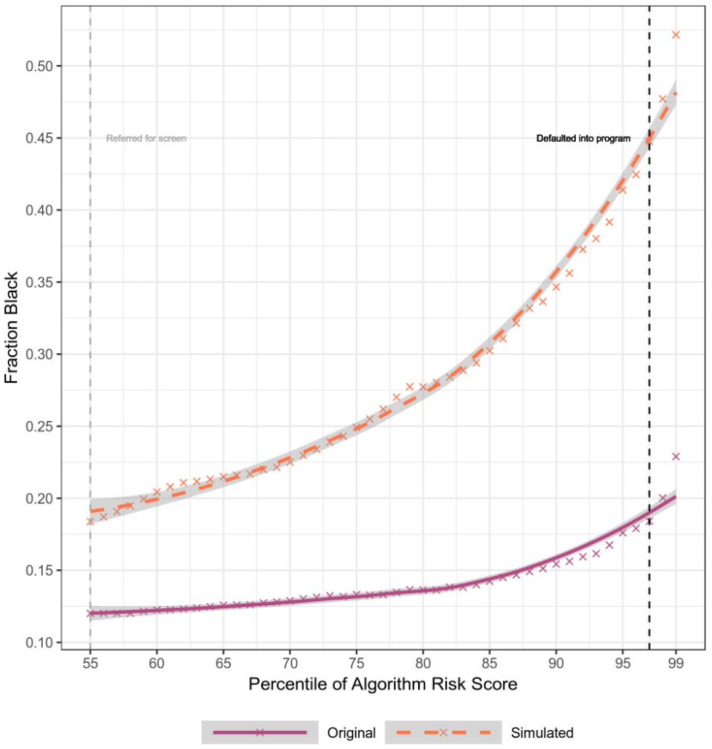
Enrolled in



Gender distribution







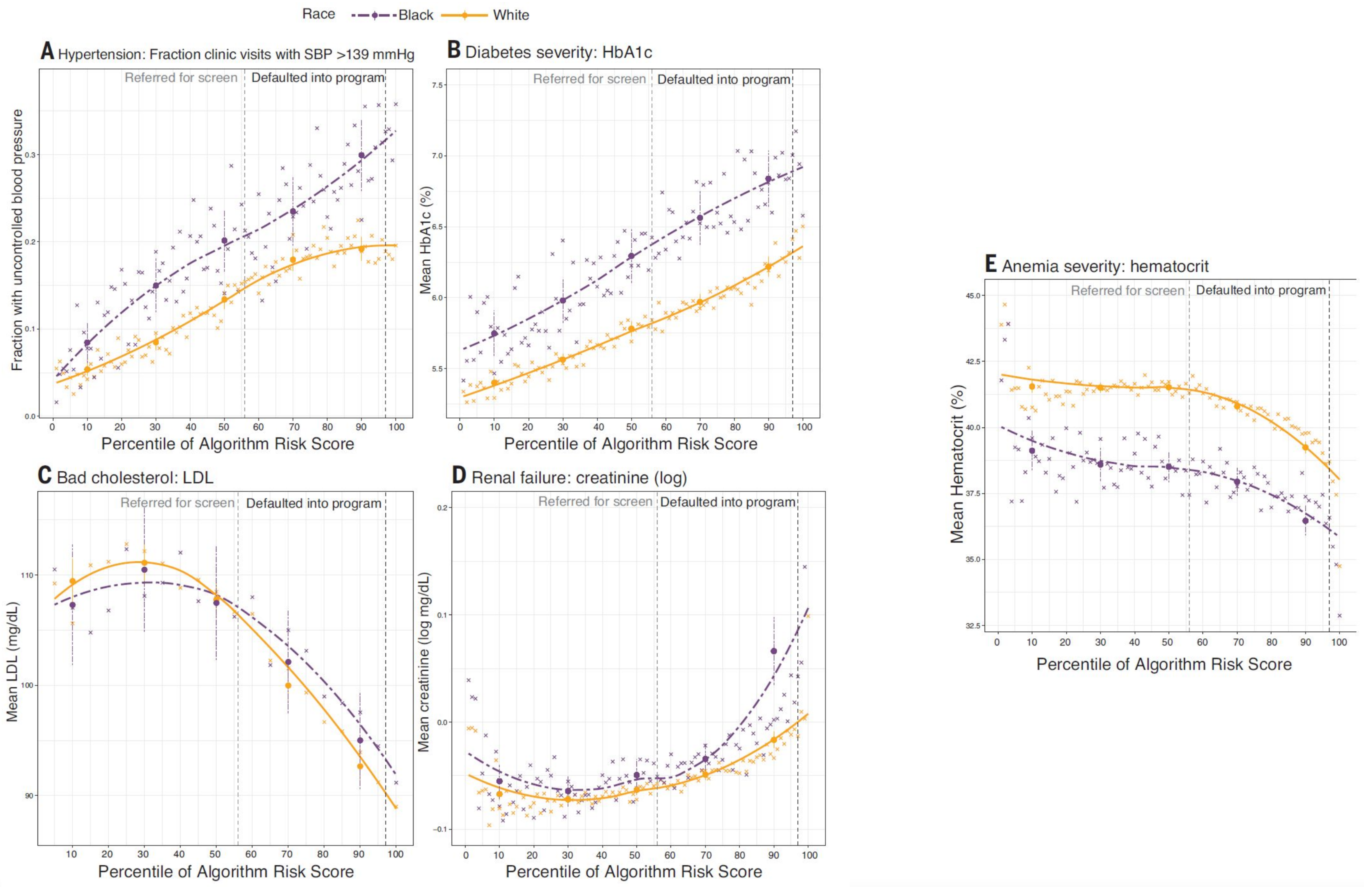
Conditional on Algorithm Risk Score

Simulation

- Consider a risk threshold, α.
- Identify white patient (i) with $R_i > \alpha$.
  Compare this to black patient (j) with $R_j < \alpha$.
- If $H_i > H_j$, replace healthier white patient with sicker black patient.
- Repeat this procedure until $H_i = H_j$.

Results of the Simulation

- For all risk thresholds above 50th percentile, it increased the fraction of black patients.
- At 97th percentile, fraction of black patients rose from 17.7% to 46.5%.



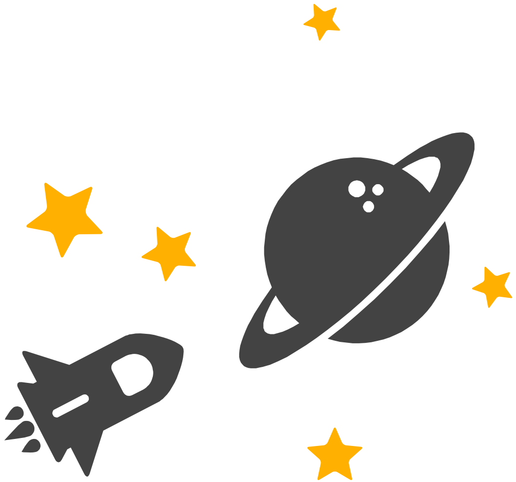
Biomarkers vs. Algorithm Risk Score

Prediction on Healthcare costs

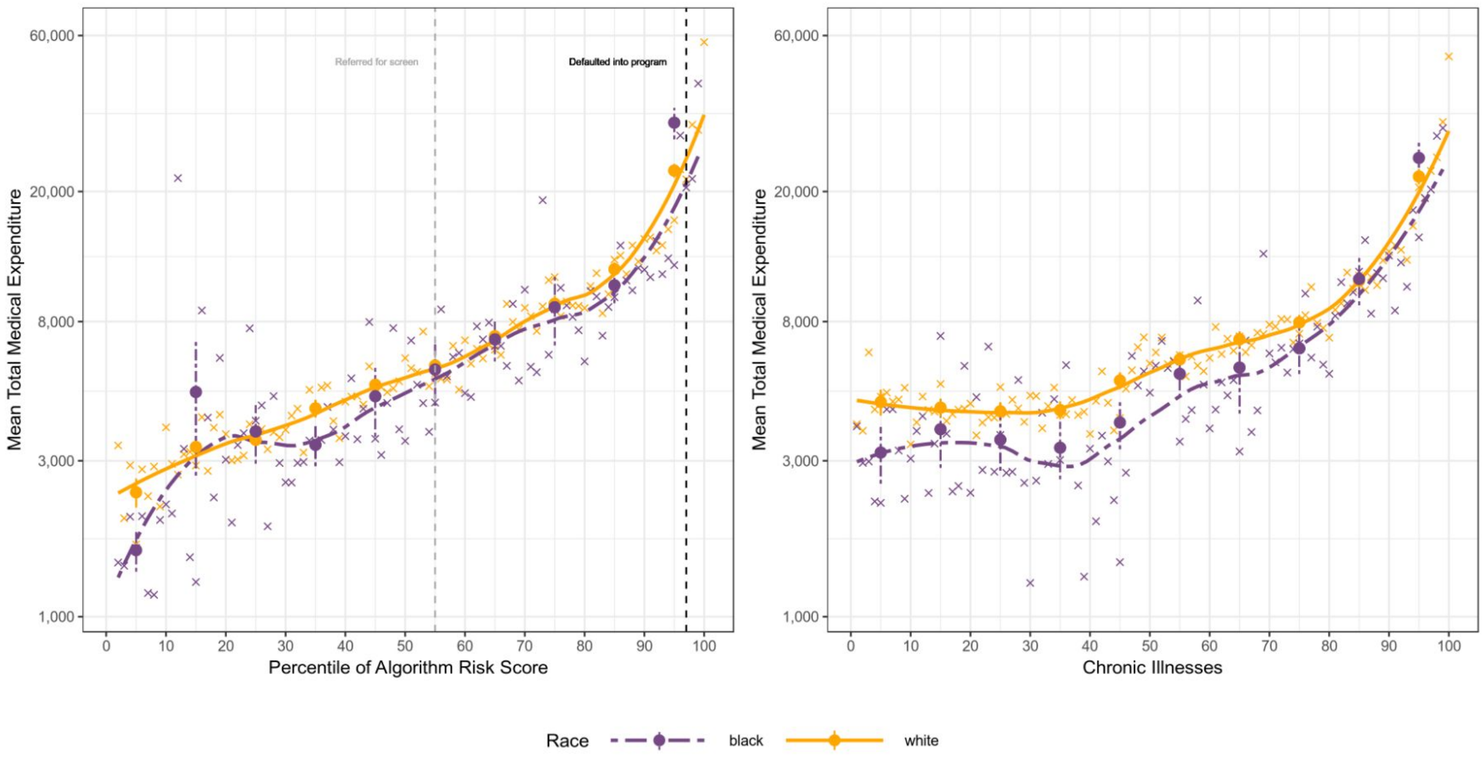
- The algorithm's prediction on health needs, is in fact, a prediction on healthcare costs.

- But costs seem similar for both black and white patients.
  So, there’s no disparity right?

WRONG



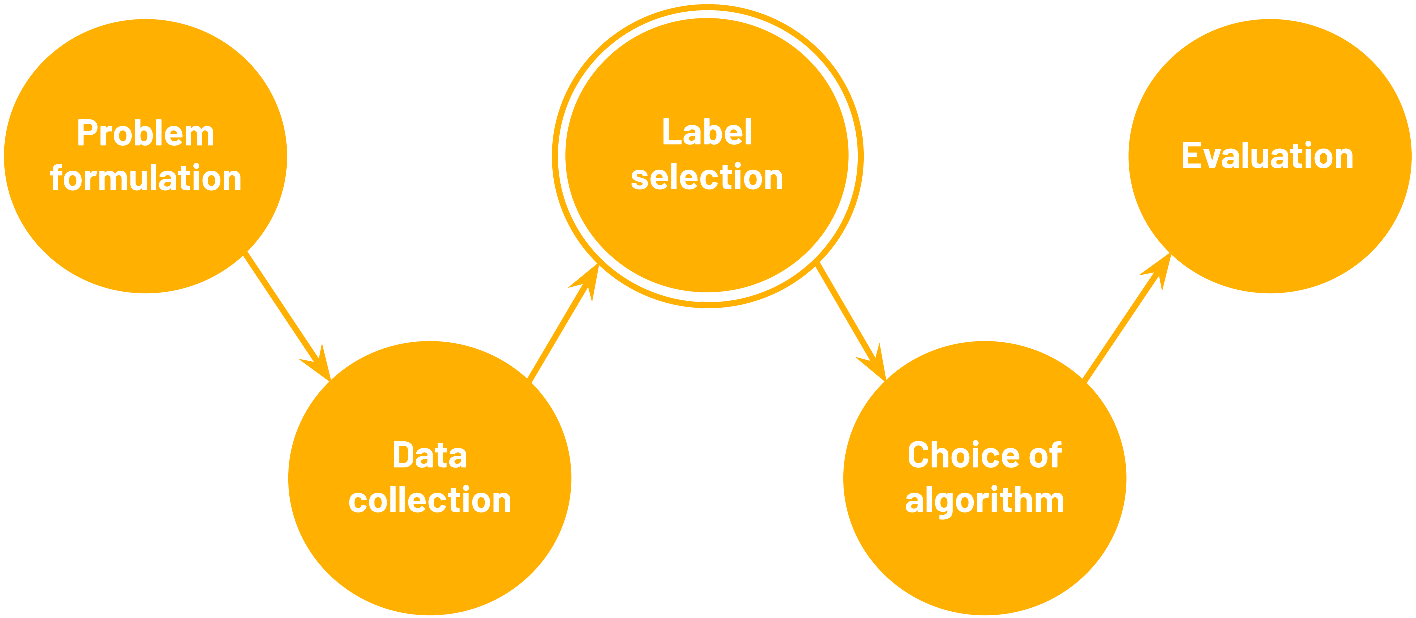
Algorithmic bias still exists



Medical Expenditure for Risk Score vs. Chronic Illnesses

Why Do These Disparities Arise?

- Poor patients face several setbacks in accessing health care.
- Direct/taste-based discrimination.
- Black patients spend less on healthcare.
- Thus, accurate prediction of costs necessarily means racially biased on health.



Sources of Bias

Experiments were performed using the following 3 labels:

- Total costs (original)
- Avoidable costs - costs due to emergency visits and hospitalizations
- Number of active chronic conditions

Label choice bias: $p[B|R>\tau]=p[B|R'>\tau]$

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	

Indicates the consistency in predictive performance of model with different labels

Indicates the effect of different label on bias

Experiments on Label Choice

- Realized enrollment decisions also depend on doctors response and other administrative factors.

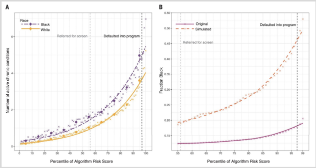
- 4 counterfactual simulations are performed to put the numbers in context
  1. Calculate enrollment rate within each percentile - randomly sample patients.
  2. Calculate enrollment rate within each percentile - highest predicted health.
  3. Top 1.3% of highest predicted costs.
  4. Top 1.3% of highest number of active chronic conditions.

Population	Fraction Black (SE)		Fraction of all costs (SE)		Fraction of all active chronic conditions (SE)	
Observed program enrollment (1.3%)	0.192	(0.003)	0.029	(0.001)	0.033	(0.001)
Simulated alternative enrollment rules						
Random, in predicted-cost bin	0.183	(0.003)	0.044	(0.002)	0.034	(0.001)
Predicted health, in predicted-cost bin	0.269	(0.003)	0.044	(0.002)	0.064	(0.002)
Highest predicted cost	0.172	(0.003)	0.100	(0.002)	0.047	(0.002)
Worst predicted health	0.292	(0.004)	0.067	(0.002)	0.076	(0.002)

Obtain access to proprietary algorithmic details

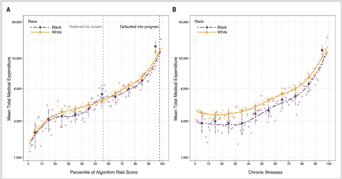
Major limitation in analyzing bias of existing algorithms

Define measure of bias to help identify if bias exists



Calibration Bias
E[Y|R,W] = E[Y|R,B] implies bias is 0

Identify the source of bias



Bias due to Label
Bias arises due to the choice of label - Total Medical Expenditure

Attempt to mitigate bias using various techniques

Use **label choice bias** to measure difference in bias across predictors

Change choice of label	
Avoidable Costs	**+0.069**
Active Chronic Conditions	**+0.126**

Procedure to Identify and Mitigate Bias

- Analyses replicated on 3,695,943 commercially insured patients (national dataset).
- Found **48,772** more active chronic conditions in Black patients.
- Modified label to combine health prediction with cost prediction.
- Achieved **84% reduction** in excess active chronic conditions in Black patients (**7,758**).

Conclusion

Implications

- Bias can arise even from reasonable choices of label and hence careful design of labels is important.
- The findings will motivate other manufacturers to check for biases.
- The procedure used can be applied to other algorithms and sectors other than healthcare.
- This exercise illustrates the need for fairness as a key consideration when designing ML systems.

Limitations

- Other sources of bias are not considered.

- The dataset is unbalanced (12.3% Black patients and 87.7% White patients).

- Other ethnicities are not considered (intersectional fairness?)[4]

4. Crenshaw, Kimberle (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. _The University of Chicago Legal Forum_ 140:139-167.

THANKS!

Questions?

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash