

Fairness

Irene Y. Chen

April 7, 2022

Material from UC Berkeley's CS 294: Fairness in Machine Learning (<https://fairmlclass.github.io/>) and NeurIPS2017 tutorial (<https://vimeo.com/248490141>) by Solon Barocas (Cornell) and Mortiz Hardt (then UC Berkeley, now Max Planck Institute); adapted from slides by Peter Szolovits

Bias in Optum's Algorithm to Predict Healthcare Utilization

Racial bias in a medical algorithm favors white patients over sicker black patients



Scientists discovered racial bias in a widely used medical algorithm that predicts which patients will have complex health needs. (iStock)

“... black patients who were ranked by the algorithm as equally as in need of extra care as white patients were much sicker: They collectively suffered from 48,772 additional chronic diseases.

<https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients/>

SHARE**RESEARCH ARTICLE**

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

Ziad Obermeyer^{1,2,*}, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan^{5,*†}

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)

Article[Figures & Data](#)[Info & Metrics](#)[eLetters](#)[PDF](#)

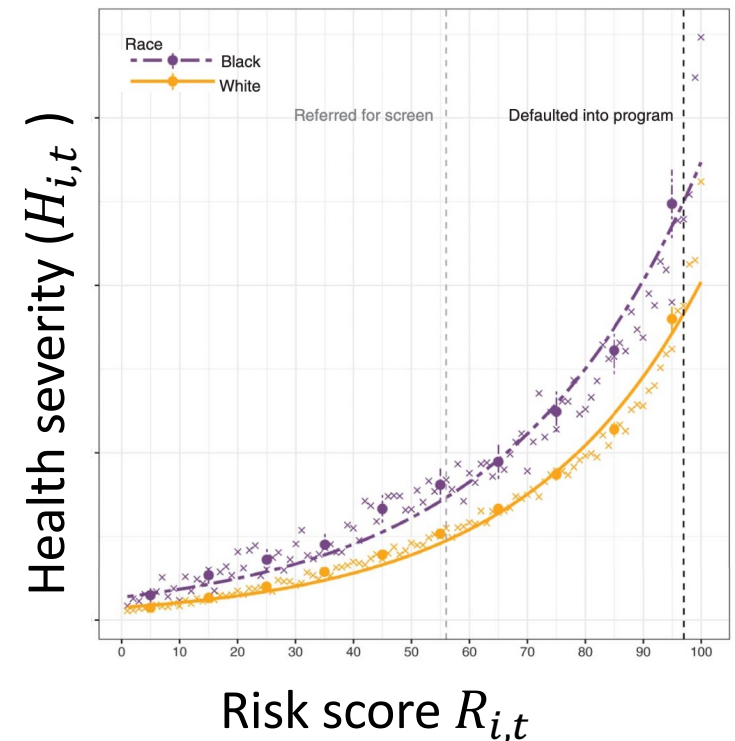
Obermeyer et al, 2019, “Dissecting racial bias in an algorithm ...”, *Science*.

Racial bias in predictive healthcare algorithms

1. Health insurance companies identify high-risk patients for **care management** from previous insurance claims
2. Computed proprietary **risk scores** for 6,079 Black and 43,539 White patients
 - 71% commercial insurance, 29% Medicare; 63% female, avg age = 50.9
3. Patients over 97th-percentile automatically **enrolled**; over 55th referred to MD

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

- W = White, B = Black, R = risk score, Y = outcome
- $E[Y|R, W] = E[Y|R, B]$?
- Define
 - Risk score $R_{i,t} = f_R(X_{i,(t-1)})$ for patient i in year t (excludes race)
 - Patient's actual health outcomes $H_{i,t}$ [ICD codes, labs and vitals]
 - higher H is better, opposite graph
 - Cost of patient's care $C_{i,t}$ [utilization: outpatient & ED visits, hospitalizations and reimbursements]
- How well is R calibrated to H and C ?



Black patients need more chronic conditions to receive the same score

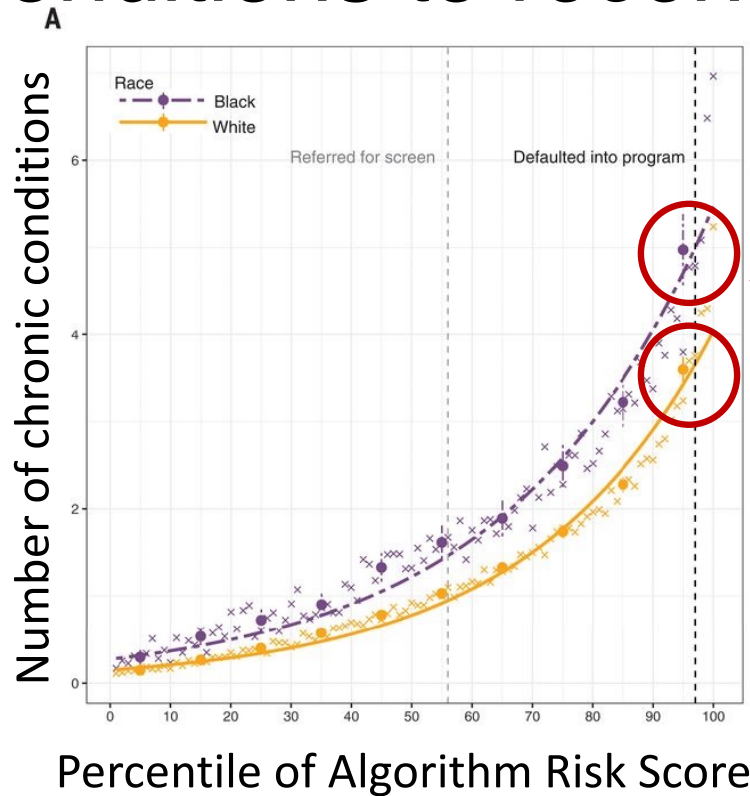


Figure 1.A.: Mean number of chronic illnesses versus algorithm-predicted risk, by race.

A person in this decile has <4 (White) or 5 (Black) chronic conditions and a risk score in the 99-percentile

Slide: Stephanie Gervasi. **Figure:** Obermeyer et al, 2019. *Science*.

Black patients need worse blood pressure to receive the same score

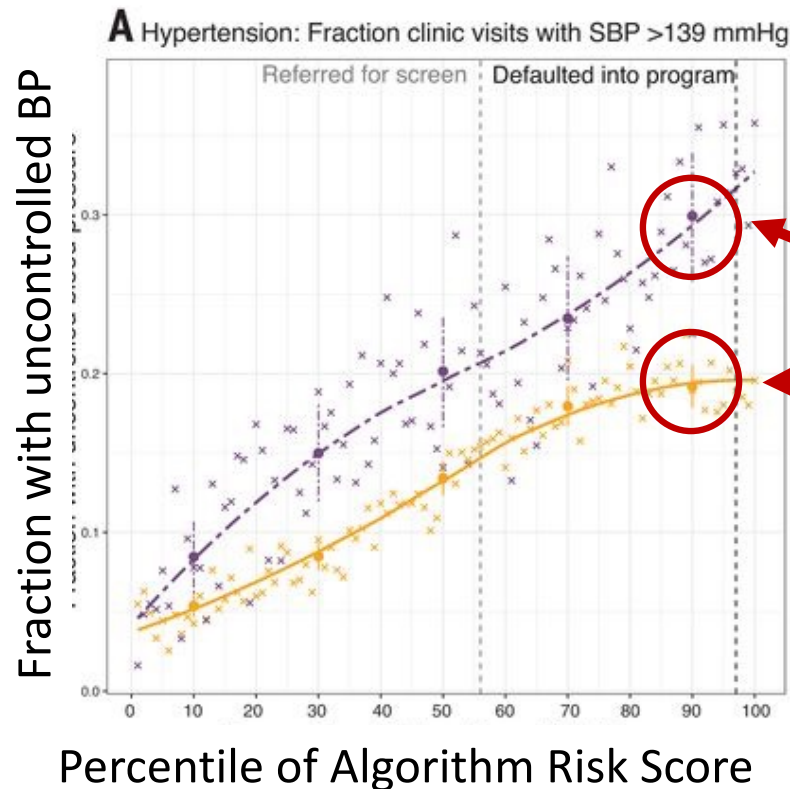


Figure 2.A.: Fraction of clinic visits with uncontrolled blood pressure.

A person in this decile has a 30% chance (Black) or <20% (White) chance of having hypertension for the same risk score.

Similar analysis conducted for diabetes, renal failure, anemia, and cholesterol based from extracted values in electronic health records.

What is the source of bias?

- Recall : $R_{i,t} = f_R(X_{i,(t-1)})$
- But, the label that R is trying to predict is cost $C_{i,t}$, not outcome $H_{i,t}$
- This is actually well calibrated,
- But, C for Blacks is consistently lower than C for Whites with the same degree of co-morbidities!
 - We would expect cost and illness to correlate well, independent of race
 - *This* is (probably) the root of the bias
- Also, different patterns of use of the healthcare system, e.g.,
 - Blacks have more ED visits, fewer outpatient visits
 - More common hypertension, diabetes, ...

Dissecting racial bias: What to do?

- Create simulated dataset where for each risk threshold, the average health H of Black and White patients is the same
- Compare pairs of patients (White patient i , Black patient j) and replace if $R_i > \alpha$, $R_j < \alpha$ and $H_i > H_j$ (White patient is healthier but has higher risk score)
- At $\alpha = 0.97$, increases percentage of Black patients in simulated population from 17.7% to 46.5%

Dissecting racial bias: What to do?

Increase the fraction of Black patients in highest risk group from 14% to 26%

	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			
1)	Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
2)	Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
3)	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	

Table 2: Results from L1-regularized logistic regression for three different labels.

Obermeyer et al, 2019, "Dissecting racial bias in an algorithm ...", *Science*.

2019 Paper Aftermath

- **Press:** The paper was covered widely across news outlets
- **Policy:** Senators Ron Wyden and Cory Booker addressed letters to CMS and FTC asking for information
- **Industry vigilance:** Significantly more collaboration and interest from insurance companies on algorithmic fairness

United States Senate
WASHINGTON, DC 20510
December 3, 2019

The Honorable Seema Verma
Administrator
The Centers for Medicare & Medicaid Services
Department of Health & Human Services
Room 445-G, Hubert H. Humphrey Building
200 Independence Ave., S.W.
Washington, DC 20201







Dear Administrator Verma:

We write today to request information regarding any actions that the Centers for Medicare & Medicaid Services (CMS) is taking or plans to take to assess the potential for algorithms used throughout the health care system to perpetuate biases.

Algorithms are increasingly embedded into every aspect of modern society, including the health care system. Organizations use automated decision systems, driven by technologies ranging from advanced analytics to artificial intelligence (AI), to organize and optimize the complex choices they need to make on daily basis. CMS and commercial health insurers have begun to explore ways to incorporate algorithms that automate decisions like predicting health care needs and outcomes, targeting resources, improving quality of care, and detecting waste, fraud, and abuse.

Race correction in eGFR

- Estimated glomerular filtration rate (eGFR) estimates how well the **kidney is performing**
- The eGFR equation includes age, sex, **race (African-American vs. not)** and/or body weight to approximate directly measured kidney function

STAGES OF CHRONIC KIDNEY DISEASE		GFR*	% OF KIDNEY FUNCTION
Stage 1	Kidney damage with normal kidney function	90 or higher	 90-100%
Stage 2	Kidney damage with mild loss of kidney function	89 to 60	 89-60%
Stage 3a	Mild to moderate loss of kidney function	59 to 45	 59-45%
Stage 3b	Moderate to severe loss of kidney function	44 to 30	 44-30%
Stage 4	Severe loss of kidney function	29 to 15	 29-15%
Stage 5	Kidney failure	Less than 15	 Less than 15%

* Your GFR number tells you how much kidney function you have. As kidney disease gets worse, the GFR number goes down.

Race correction in eGFR

1. Race corrections in eGFR could **over-estimate kidney health** in Black patients and could delay referrals for specialist care
2. Black patients already have higher rates of end-stage **kidney disease** and death
3. Use of genetic African ancestry on eGFR resulted in higher eGFR for 14.7% of Hispanic/Latino Americans and lower eGFR for 4.1% of African Americans
4. Proposed **new eGFR** without race correction is more accurate and has smaller differences between races.

[1] Vyas et al, "Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms " NEJM 2020.

[2] Udler et al, "Effect of Genetic African Ancestry on eGFR and Kidney Disease", JASN 2015.

[3] Inker et al, "New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race", NEJM 2021.

Genetic Misdiagnoses and the Potential for Health Disparities

Arjun K. Manrai, Ph.D., Birgit H. Funke, Ph.D., Heidi L. Rehm, Ph.D., Morten S. Olesen, Ph.D., Bradley A. Maron, M.D., Peter Szolovits, Ph.D., David M. Margulies, M.D., Joseph Loscalzo, M.D., Ph.D., and Isaac S. Kohane, M.D., Ph.D.

Article | [OPEN](#) | Published: 15 April 2019

Genetic risk factors identified in populations of European descent do not improve the prediction of osteoporotic fracture and bone mineral density in Chinese populations

Yu-Mei Li , Cheng Peng, Ji-Gang Zhang, Wei Zhu, Chao Xu, Yong Lin, Xiao-Ying Fu, Qing Tian, Lei Zhang, Yang Xiang, Victor Sheng & Hong-Wen Deng 

Scientific Reports **9**, Article number: 6086 (2019) | [Download Citation](#) ↓

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



QUARTZ

FRYING PAN, FIRE, ETC

California just replaced cash bail with algorithms

 REUTERS

RETAIL | OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ



Isn't Discrimination the Very Point of ML?

- **Unjustified** basis of differentiation
- Fairness focuses on **ethical** concerns about how algorithm is used
- Discrimination is
 - **domain** specific — how it influences people's life chances
 - **feature** specific — socially salient qualities that have served as the basis for unjustified and systematically adverse treatment in the past

Agenda

- ~~1. Motivation: Why fairness?~~
- 2. Legal and historical perspective: What's been done?**
3. Algorithmic fairness: How do we assess bias in algorithms?
4. Other considerations: How does this tie to previous lectures?

Regulated Domains

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1964; Education Amendments of 1972)
- Employment (Civil Rights Act of 1964)
- Housing (Fair Housing Act)
- 'Public Accommodation' (Civil Rights Act of 1964)
- Marriage (Defense of Marriage Act, 1996, struck down by Supreme Court in 2013; also 1967 landmark civil rights case of Loving v. Virginia)
- Extends to marketing and advertising; not limited to final decision
- This list sets aside complex web of laws that regulates the government

Legally recognized 'protected classes'

- Race (Civil Rights Act of 1964)
- Color (Civil Rights Act of 1964)
- Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
- Religion (Civil Rights Act of 1964)
- National origin (Civil Rights Act of 1964)
- Citizenship (Immigration Reform and Control Act)
- Age (Age Discrimination in Employment Act of 1967)
- Pregnancy (Pregnancy Discrimination Act)
- Familial status (Civil Rights Act of 1968)
- Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); Genetic information (Genetic Information Nondiscrimination Act)
- Sexual orientation/gender identity (Mass SJC 2004, SCOTUS 2015, Bostock 2020)

Sometimes the protected group attribute is not included in the dataset!

Two doctrines of discrimination law

- Disparate Treatment
 - Formal — considering class membership
 - E.g., country club exclusion based on race or religion,
 - Intentional — without explicit reference to class, but with same effect
 - E.g., red-lining (mortgage availability based on geographic location)
- Disparate Impact
 - Unjustified, Avoidable
 - How to demonstrate: “4/5 rule” (20% difference establishes it)
 - How to defend: business necessity, job-related
 - Alternative practice: can we achieve the same goal but with less disparity?

Goals of (Anti-)Discrimination Law

- Disparate Treatment
 - Procedural fairness
 - Equality of opportunity
- Disparate Impact
 - Distributive justice
 - Minimize inequality of outcome
- Non-discrimination:
 - ensuring that decision-making treats similar people similarly on the basis of relevant features, given their current degree of similarity
- Equality of opportunity:
 - organizing society in such a way that people of equal talents and ambition can achieve equal outcomes over the course of their lives
- Equality of outcome:
 - treat seemingly dissimilar people similarly, on the belief that their current dissimilarity is the result of past injustice

Discrimination persists in many areas

- Criminal justice — “Predictive Policing”
 - Police records measure “some complex interaction between criminality, policing strategy, and community-policing relations”
 - Future observations of crime confirm predictions
 - Fewer opportunities to observe crime that contradicts predictions
 - Initial bias may compound over time
- Housing
- Employment
- Health care
- ...



“Tuskegee Study of Untreated Syphilis in the Negro Male” (1932)

Photo credit: National Archives

Ethical questions exist already in healthcare

- **Clinical trial populations:** Clinical Trials Still Don't Reflect the Diversity of America (NPR, Dec 2015)
- **Drug pricing:** House passes bill to cap insulin prices (NPR, March 2022)
- **Opioid epidemic:** Massachusetts Attorney General Implicates Family Behind Purdue Pharma In Opioid Deaths (NPR, Jan 2019)
- **Retracted studies:** Harvard Calls for Retraction of Dozens of Studies by Noted Cardiac Researcher (NYT, Oct 2018)
- **Conflict of interest:** Sloan Kettering's Cozy Deal with Start-Up Ignites a New Uproar (NYT, Sept 2018)

Breakout: How would you regulate health algorithms with fairness in mind?

Agenda

- ~~1. Motivation: Why fairness?~~
- ~~2. Legal and historical perspective: What's been done?~~
- 3. Algorithmic fairness: How do we assess bias in algorithms?**
4. Other considerations: How does this tie to previous lectures?

Ongoing data problems

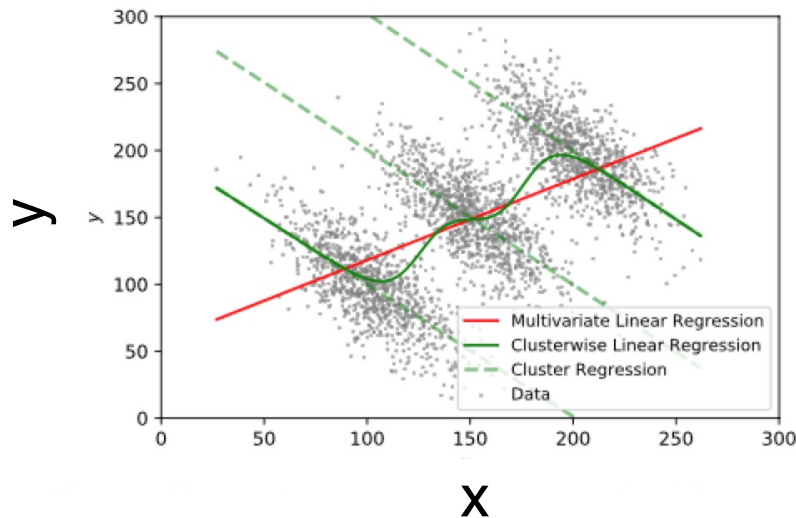
- Limited features
 - Measurement bias for subpopulations
 - Difference predictive features across subpopulations
- Sample size disparity
- Fix idea: collect more features for protected class, to improve accuracy of prediction*
- Group leakage
 - Protected class membership will be encoded across other features

adapted from Solon Barocas

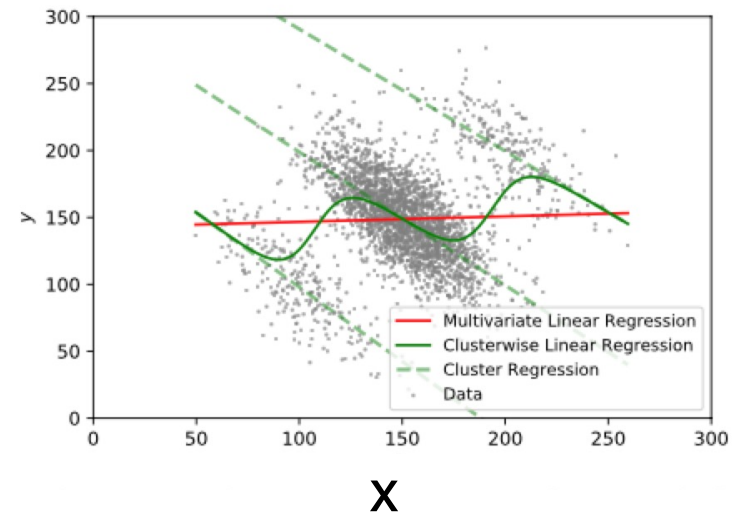
*Michiel A. Bakker, “Improving Fairness in Budget-Constrained Algorithmic Decision-Making”, MIT PhD, EECS, Sep 2020.

Bias in data: Simpson's Paradox

Equal size groups show positive relation between x and y



One larger group shows no relation between x and y



Many Forms of Bias

- Historical
- Representation
- Measurement
- Evaluation
- Aggregation
- Population
- Simpson's Paradox
- Longitudinal Data Fallacy
- Sampling
- Behavioral
- Content Production
- Linking
- Temporal
- Popularity
- Algorithmic
- User Interaction/Presentation/Ranking
- Social
- Emergent
- Self-Selection
- Omitted Variable
- Cause-Effect
- Observer
- Funding

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019, August 22). A Survey on Bias and Fairness in Machine Learning. Iclr 2020.

Formalizing fairness

- Hardt's example: advertising for a software engineer, question of gender bias

X	features of an individual (browsing history)
A	sensitive attribute (gender)
$R = r(X, A)$ $C = c(X, A)$	score/predictor (show ad) [classify by thresholding score]
Y	hire software engineer

Formalizing fairness

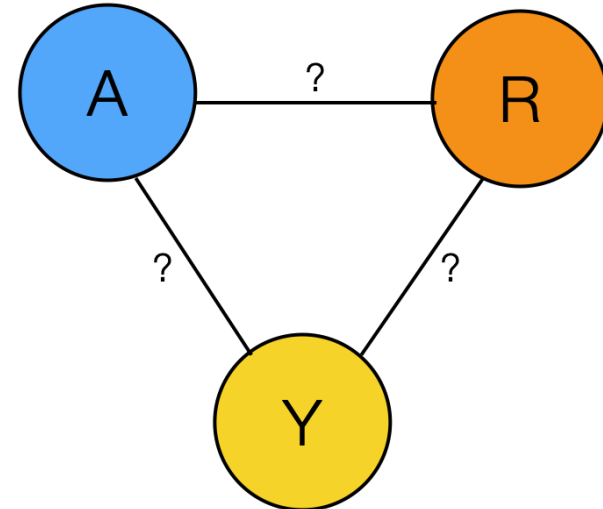
- Clinical example: predicting likelihood of hospitalization from patient history

X	features of an individual (clinical history)
A	sensitive attribute (race)
$R = r(\mathbf{X}, \mathbf{A})$ $C = c(\mathbf{X}, \mathbf{A})$	score/predictor (likelihood of hospitalization) [classify by thresholding score]
Y	actual hospitalization

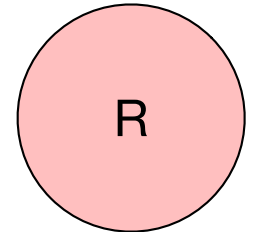
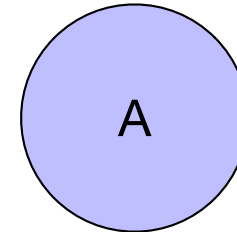
Proposed Criteria of Fairness

- **Independence** of scoring function from sensitive attributes
 - $R \perp A$
- **Separation** of score and sensitive attribute given outcome
 - $R \perp A \mid Y$
- **Sufficiency**
 - $Y \perp A \mid R$

X	features of an individual (clinical history)
A	sensitive attribute (gender)
$R = r(X, A)$ $C = c(X, A)$	score/predictor (risk of hospitalization) [classify by thresholding score]
Y	actual hospitalization



Independence $R \perp A$



- Also called demographic parity, statistical parity, group fairness, disparate impact
- $P\{R = 1 \mid A = a\} = P\{R = 1 \mid A = b\}$ for all groups A
- thus, unfair if

$$|P\{R = 1 \mid A = a\} - P\{R = 1 \mid A = b\}| > \epsilon$$

$$\left| \frac{P\{R = 1 \mid A = a\}}{P\{R = 1 \mid A = b\}} - 1 \right| \geq \epsilon$$

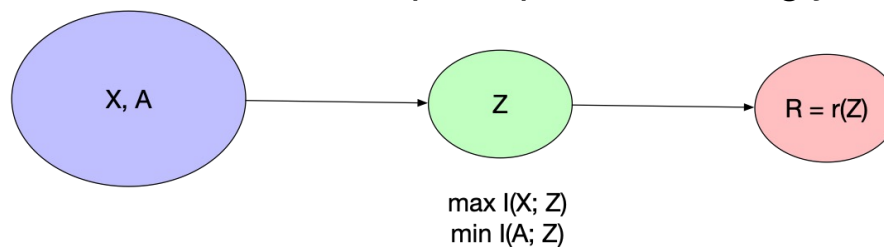
- $\epsilon = 0.2$ relates to 4/5 rule

Problems with Independence

- Only requires equal rates of decisions (hiring, liver transplants, etc.)
 - But, what if hiring is based on a good score in group a, but random in b, though with same probability?
 - Outcomes will (most likely) be better for group a, establishing problems for the future!
 - Could be caused by malice, or by better information about group a.
- What if A is a perfect predictor of Y?
 - ... or at least is strongly correlated?
 - How much are you willing to decrease the effectiveness of the predictor to achieve fairness?

Potential fixes to achieve Independence

- Pre-processing:
 - Adjust the feature space to be uncorrelated with the sensitive attribute
 - Domain-specific
 - Representation learning
- Impose Independence constraints at training time (for a given data set)
E.g., include dependence in the loss function, differential sampling, ...
- Post-processing
 - Create a new classifier F ,
 - minimize cost of misclassification, perhaps more strongly for protected A

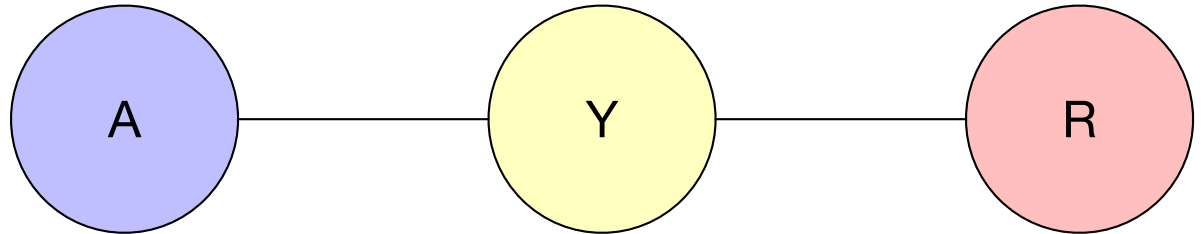


Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. ICML.

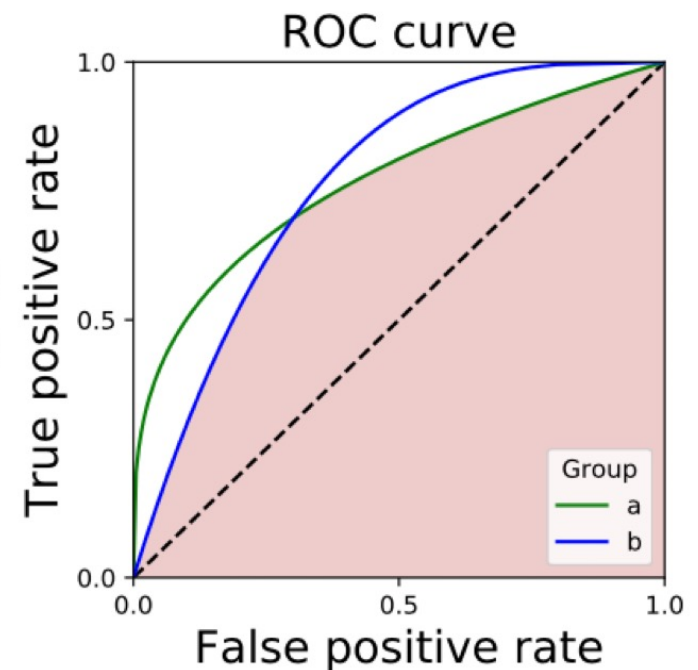
adapted from Moritz Hardt

Calders, T., Kamiran, F., & Pechenizkiy, M. (2010). Building Classifiers with Independency Constraints (pp. 13–18). Presented at the 2009 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE. <http://doi.org/10.1109/ICDMW.2009.83>

Separation $R \perp A \mid Y$



- Recognizes that A may be correlated with the target variable
 - E.g., different success rates in a drug trial for different ethnic populations
- $P\{R = 1 \mid Y = 1, A = a\} = P\{R = 1 \mid Y = 1, A = b\}$
 $P\{R = 1 \mid Y = 0, A = a\} = P\{R = 1 \mid Y = 0, A = b\}$
 - i.e., true and false positive rates for both classes must be the same
- Can choose any true positive/false positive tradeoff in the feasible region, depending on relative costs

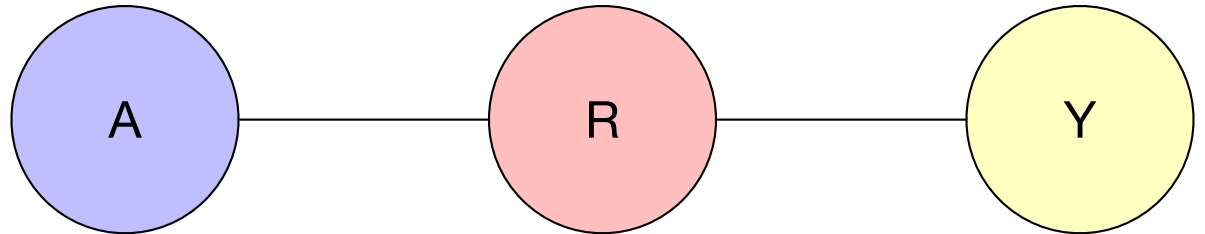


Advantages of Separation over Independence

- Allows correlation between R and Y (even perfect predictor)
- Incentive to reduce errors uniformly in all groups

Sufficiency

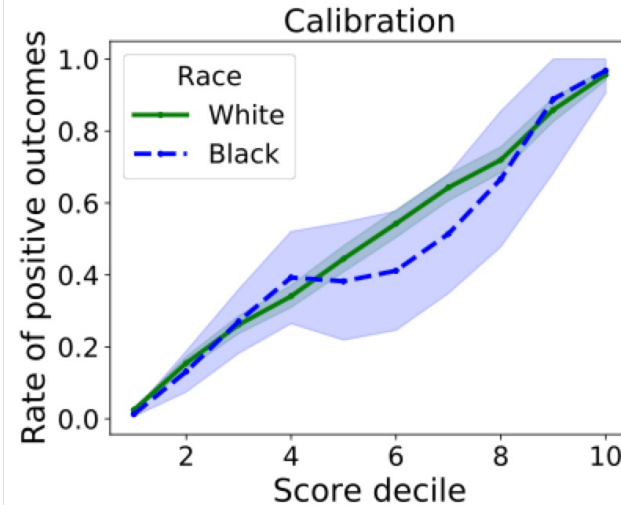
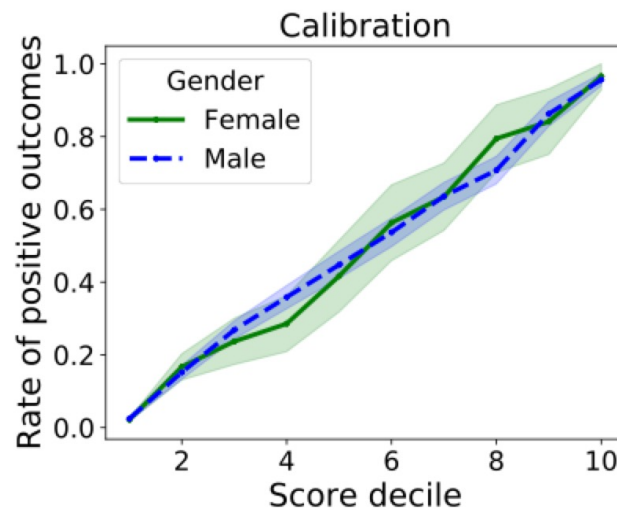
$Y \perp A \mid R$



- $P\{Y = 1 \mid R = r, A = a\} = P\{Y = 1 \mid R = r, A = b\}$
- Requires parity of positive and negative predictive values across groups
- R is *calibrated* if $P\{Y = 1 \mid R = r, A = a\} = r$
 - I.e., if the scoring function is a probability of outcome, or
 - “the set of all instances assigned a score value r has an r fraction of positive instances among them”
- Can recalibrate a scoring function R by fitting a sigmoid
 - $$S = \frac{1}{1 + e^{aR+b}}$$
 - and optimizing log loss $-\mathbb{E}[Y \log S + (1 - Y) \log(1 - S)]$
- Calibration by group implies sufficiency

Calibration can be good without trying

- E.g., UCI census data set, predicting income $> \$50,000/\text{year}$ for those over 16yo with some income
- Features (14): age, type of work, weight of sample, education, marital status, occupation, military service, race, sex, capital gain/loss, hours per week of work, native country, ...



adapted from Moritz Hardt

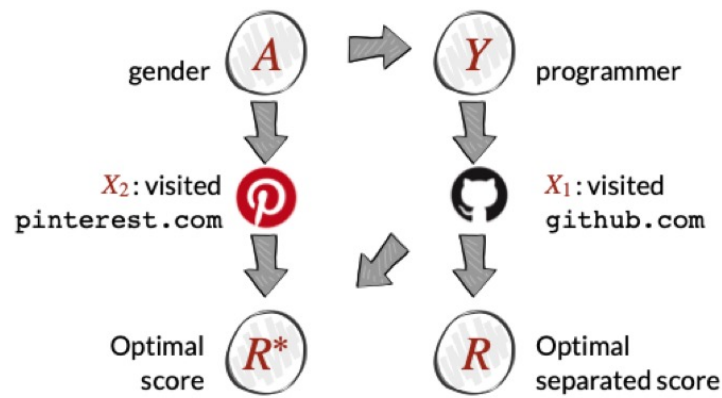
<https://fairmlbook.org/demographic.html>
<https://archive.ics.uci.edu/ml/datasets/adult>

Bad news!

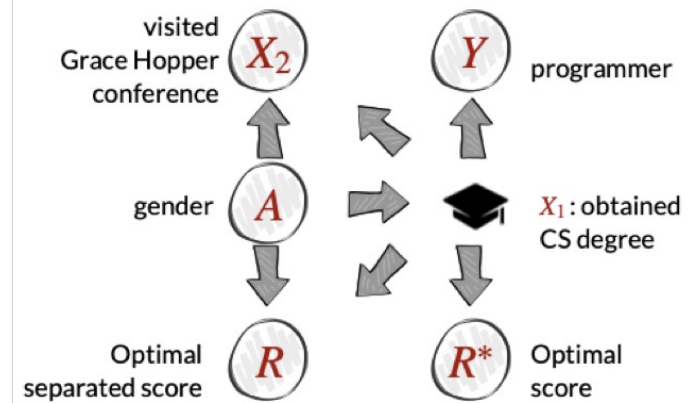
- It is not possible to jointly achieve any pair of these conditions
 - Independence *xor* Separation
 - Independence *xor* Sufficiency
 - Separation *xor* Sufficiency
- Nice illustration at
 - <https://research.google.com/bigpicture/attacking-discrimination-in-ml/>

Different scenarios can lead to same observed distributions

Scenario I



Scenario II



- The distributions of A, R, Y, X_1 and X_2 can be identical in the two scenarios
- In Scenario II, gender is used directly to adjust separated score

adapted from Moritz Hardt

Agenda

- ~~1. Motivation: Why fairness?~~
- ~~2. Legal and historical perspective: What's been done?~~
- ~~3. Algorithmic fairness: How do we assess bias in algorithms?~~
- 4. Other considerations: How does this tie to previous lectures?**

**Breakout: How does fairness
relate to earlier lectures?**

Clinical NLP

SciBERT is a deep embedding language model fine-tuned on scientific articles

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

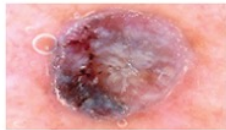

SciBERT: **caucasian** pt became belligerent and violent .
sent to **hospital** .
white pt became belligerent and violent . sent
to **hospital** .
african pt became belligerent and violent .
sent to **prison** .
african american pt became belligerent and
violent . sent to **prison** .
black pt became belligerent and violent . sent
to **prison** .

Systemic health disparities can create noisy labels

- **Disparities in access to care**
 - Patients cannot receive care as easily with rural hospitals closing and insufficient insurance coverage
 - Patients do not seek medical help with loss of trust in healthcare system leading to differences in medical adherence
- **Disparities in treatment**
 - Different treatments for same conditions
 - Same treatments for different physiological systems

Medical imaging

- Datasets used for **benign/malignant** labels contain more light-skinned patients
- Researchers created an inclusive dataset with a range of skin tones (**Fitzpatrick scale**)
- Dermatology algorithms have worse performance on **dark skin tones** and uncommon diseases.

New images	Output
	95% malignant 5% benign
	20% malignant 80% benign

[1] Adamson and Smith, "Machine Learning and Health Care Disparities in Dermatology," *JAMA Dermatology* 2018.

[2] Daneshjou et al, "Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set", ML4H Symposium 2022.

Takeaways

1. Clinicians, policy makers, and machine learning researchers are increasingly concerned with fairness of clinical algorithms.
2. Fairness analysis sits on a foundation of legal and historical context.
3. Even the choice of mathematical definition of algorithmic bias is nuanced and often contradictory.
4. Through a certain lens, *everything* is fairness or fairness-adjacent.

