



DSCI 531

Algorithmic Fairness:

ILLUSTRATIONS FROM CRIMINOLOGY AND HEALTH DOMAINS

Kristina Lerman

Spring 2025



Human decisions & cognitive bias

- Human attention and cognitive bandwidth are limited ...
- Cognitive heuristics allow the brain to focus its limited resources on salient details
- ... with far-reaching consequences on online individual & collective behavior



Herbert A. Simon

Bounded rationality

Constraints of available time, information, and cognitive capacity limit human ability to make rational decisions



Daniel Kahneman

Heuristics and biases

Mental shortcuts help people make quick, but less accurate decisions, by focusing brain's limited resources on the most salient information



Amos
Tversky



Sendhil Mulainathan



Eldar Shafir

Cognitive/Information load

Scarce resources—time, money, food, ... — focus the brain on alleviating shortages and reduce the mental bandwidth to address other needs



Cognitive biases in judicial decisions



JUDICIAL INTERPRETATION

Addressing Bias Among Judges

It's time to reconceptualize judicial training on cognitive biases and cultural sensitivity.

<https://statecourtreport.org/our-work/analysis-opinion/addressing-bias-among-judges>

Cognitive biases include unconscious racial, gender, and ethnic biases, stereotypes, prejudices, discriminatory attitudes, and other preconceived notions, all of which can invade judicial decision-making. Biases strain the judiciary's obligation to remain detached, objective decision-makers. Even a hint of bias can jeopardize public faith in the entire court system, to say nothing of a single ruling.

Common cognitive biases that tend to affect judges include **confirmation bias** (giving credence to information that coincides with preexisting beliefs and devalues other information), **blind spot bias** (thinking others may be biased, but not yourself), **overconfidence bias** (attributing too much importance to one's own beliefs and expertise), **affinity bias** (favoring people with similar backgrounds, characteristics, and interests), **anchoring bias** (relying too much on first impressions), and **hindsight bias** (perceiving past events as more obvious in retrospect).



“Hungry judges” effect

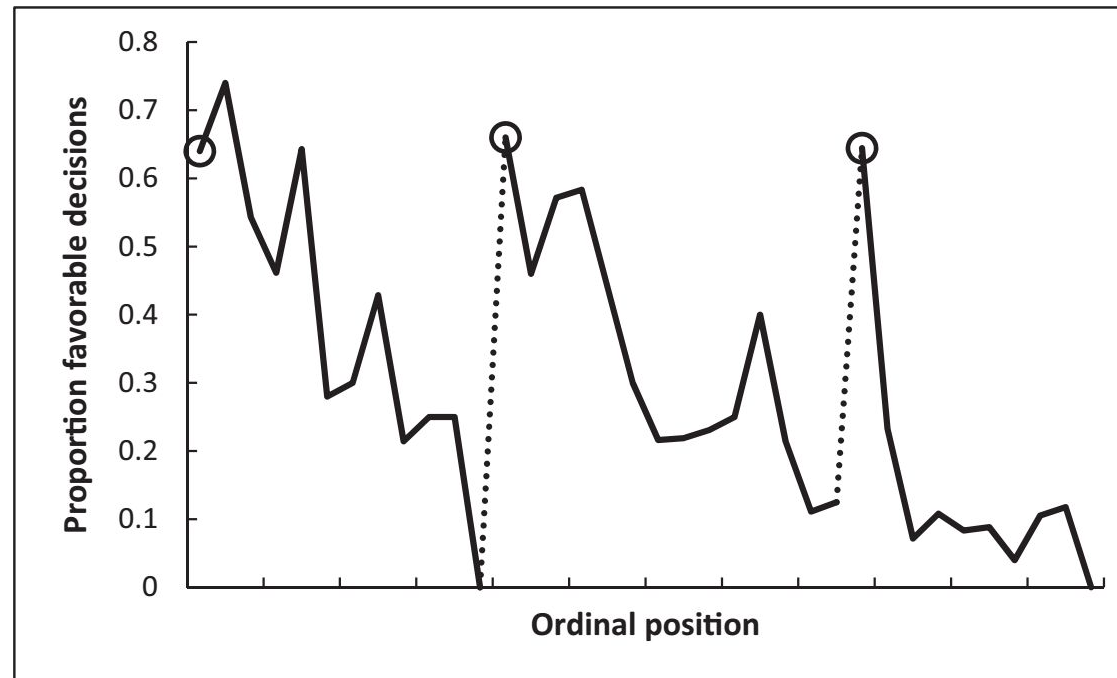
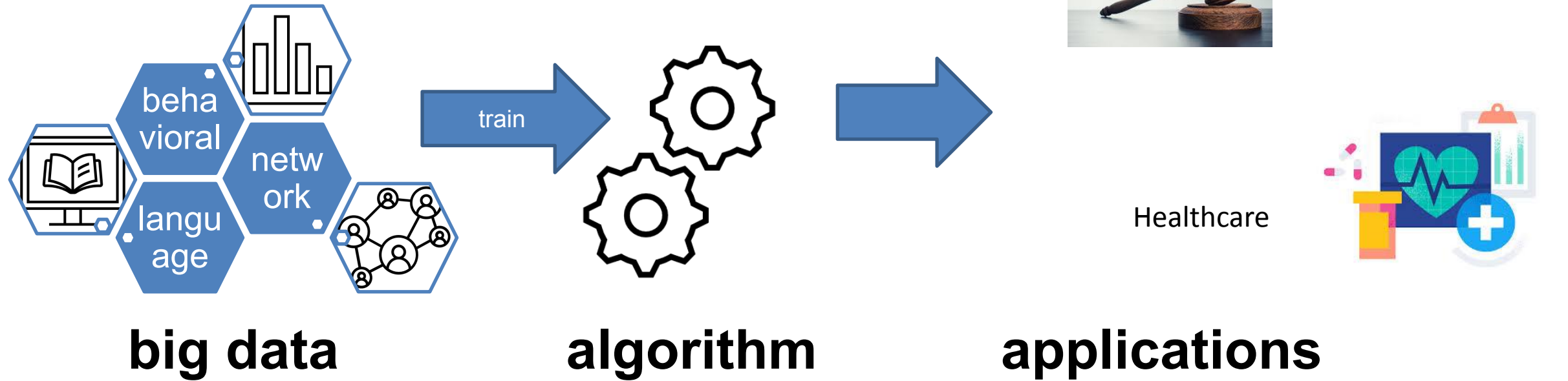


Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on x axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

Weinshall-Margel, K., & Shapard, J. (2011). Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences*, 108(42), E833-E833.



Maybe AI can help?





Maybe AI can help?

“big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.”

White House report *Big Data: Seizing Opportunities, Preserving Values*, 2014

大数据分析可能会超越长期以来的公民权利保护，即个人信息在住房、信贷、就业、医疗、教育和市场方面的使用



Disparate impact of data

The promise: AI eliminates human biases from the decision process

The reality: Algorithms are only as good as the data they were trained on

- Data reflects explicit prejudices
- Data reflects implicit biases persisting in society
- Data reproduces existing patterns of discrimination, exclusion and inequality

“Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society”

[Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104, 671.]



Is it possible to discriminate through data?

- Discrimination persists in American society in employment, housing, credit, consumer markets, academia, and entertainment
- While intentional discrimination and explicit prejudice are now less common, “institutional” discrimination through implicit biases and ‘business as usual’ attitudes account for disparate treatment of protected classes □ encoded in patterns in data
- Absence of explicit prejudice does not guarantee impartiality 公平
 - AI could unintentionally, rather than maliciously, discriminate by identifying and reinforcing existing patterns of discrimination
 - ... and inadvertently amplify inequalities by giving disadvantaged groups less favorable treatment



How does data come to discriminate?

- Where does the unintended discrimination effects of data come from?
- Five mechanisms by which data mining may systematically disadvantage protected classes
 1. Defining the “Outcome Variable” and “Class Labels” *It can already create bias*
 2. Training data *sometimes your training data does not reflect the diversity, demographic diversity, human populations*
 3. Feature selection
 4. Proxies
 5. Masking

1. Defining the “Outcome Variable” and “Class Labels”



- Data mining identifies statistical relationships in data (i.e., correlations) between outcome variables and features
 - Statistical patterns on which to base future decisions
 - Learned models can classify new entities, estimate unobserved features, predict future outcomes
 - How are outcomes defined? –
 - What is “creditworthiness”? An abstraction developed for the purpose of evaluating loan default risk
 - what is a “good” employee? Objective measures?
 - what is a “healthy” patient?



2. Training data

- Biased training data leads to discriminatory models
 - (1) If prejudice played a role in generating data, algorithms will happily reproduce patterns of prejudice
 - (2) if data mining draws inferences from a biased sample of the population, these inferences may systematically disadvantage those who are under- or overrepresented in the data.
- Special considerations
 - *Labels* (ground truth data) are often subjective. This can skew modeling results (see example next page)
 - *Data collected* about protected classes may be systematically incorrect, or non-representative ☐ may discriminate against protected classes.
 - Underrepresentation vs overrepresentation



Biases in hiring decisions

- St. George's Hospital (UK) developed a computer program to help sort medical school applicants based on its previous admissions decisions. Those admissions decisions systematically discriminated against racial minorities and women with similar credentials to other applicants'. By learning from prior biased decisions, St. George's Hospital devised an automated process that propagated the same prejudices.
- The computer may learn to discriminate against certain female or black applicants if trained on prior hiring decisions in which an employer has consistently rejected jobseekers with degrees from women's or historically black colleges.



3. Feature selection

- AI designers make choices about what attributes to collect and model
 - May systematically omit details to reliably resolve members of protected classes with details necessary to achieve equally accurate determinations residing at a level of granularity and coverage that the selected features fail to achieve.
 - Obtaining detailed information to resolve protected classes can be expensive.
 - See “redlining” – using a coarse proxy to make decisions , e.g., zipcode leads to less accurate decisions than fine-grained data, but is easier to collect



4. Proxies

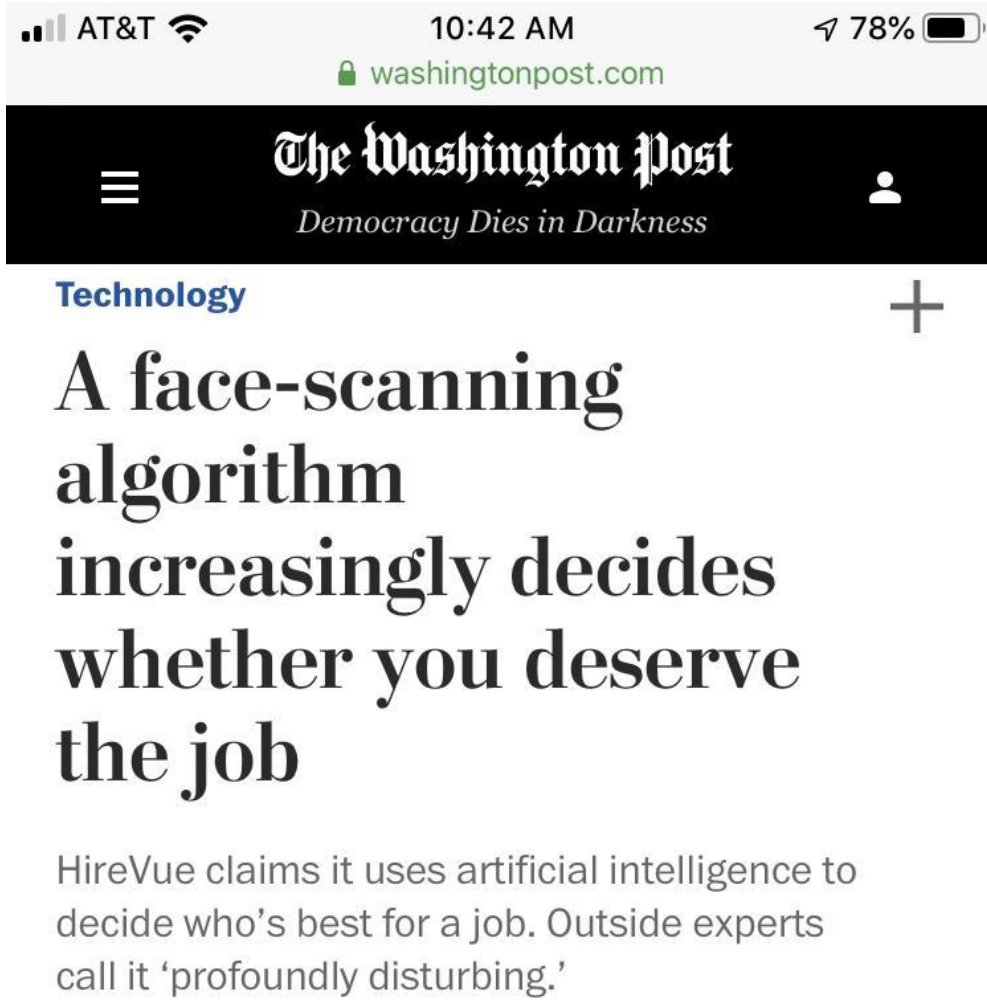
- Features are highly correlated with membership in the protected class
 - E.g., “redlining” – zipcode used instead of race to deny loan applications



5. Masking

- Decision makers can disguise their prejudicial views by manipulating data collection, labeling, feature selection, etc.
 - E.g., decision makers could knowingly and purposefully bias the collection of data to ensure that the model is less favorable to members of protected classes.

AI-based job recruiting



An artificial intelligence hiring system has become a powerful gatekeeper for some of America's most prominent employers, reshaping how companies assess their workforce — and how prospective employees prove their worth.

Designed by the recruiting-technology firm HireVue, the system uses candidates' computer or cellphone cameras to analyze their facial movements, word choice and speaking voice before ranking them against other applicants based on an automatically generated “employability” score.



Case studies

COMPAS automated risk assessment tool

COMPAS tool assesses risk of
future recidivism

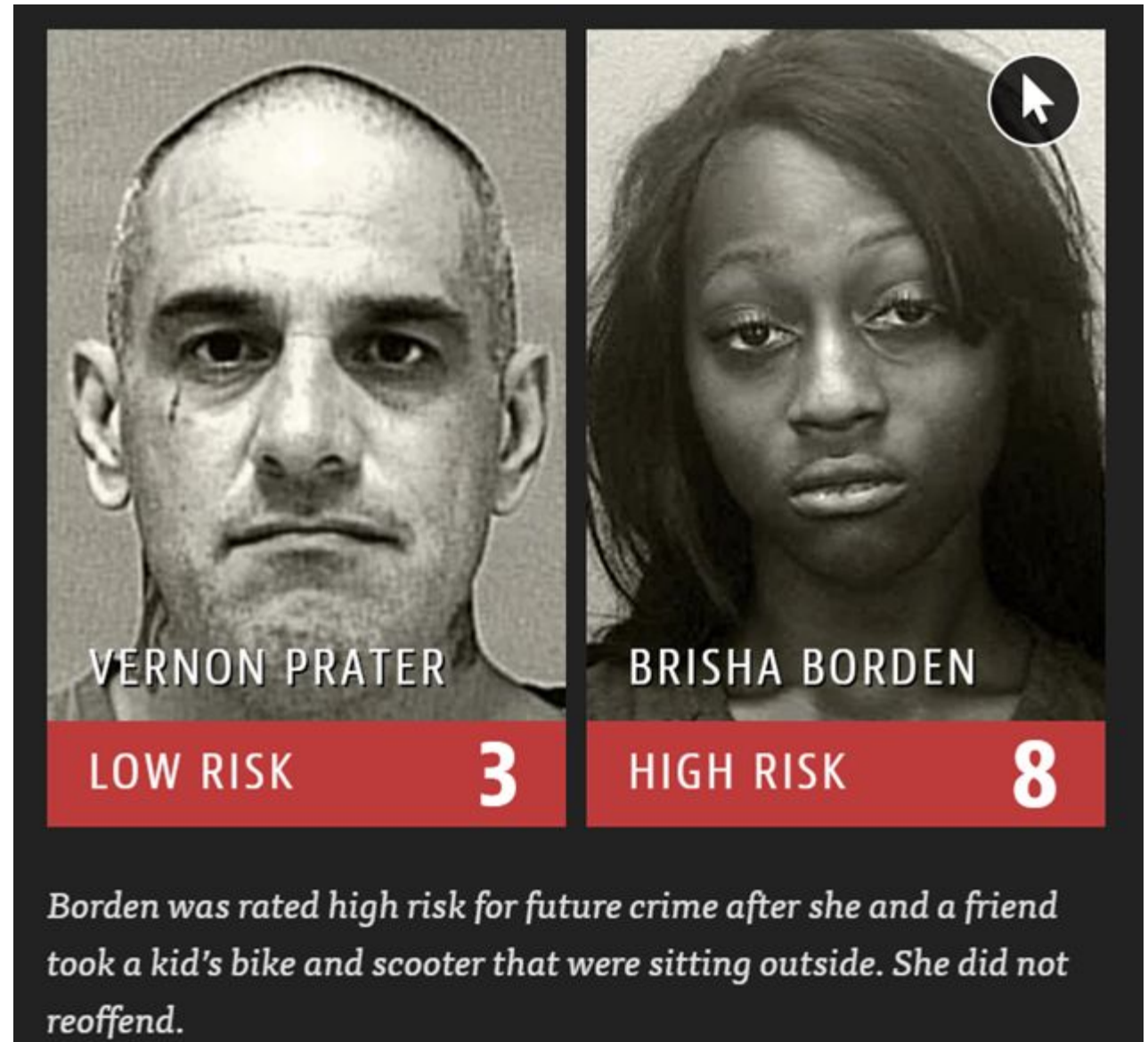
recidivism is just the fancy word for whether or not somebody who committed crime in the past. whether or not they will recidivism and whether they will commit another crime in the future.

Risk prediction in health care

refer patients with complex
health needs for additional
care

COMPAS automated risk assessment tool

COMPAS tool assesses risk of
future recidivism



Individual	Risk Level	Score
Vernon Prater	LOW RISK	3
Brisha Borden	HIGH RISK	8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.





The tension

- Judges must balance two factors evaluating the cost of a mistake:
 - Let a guilty person go?
 - Put an innocent person in jail?

Benefit vs cost

Benefit : expected number of crimes prevented

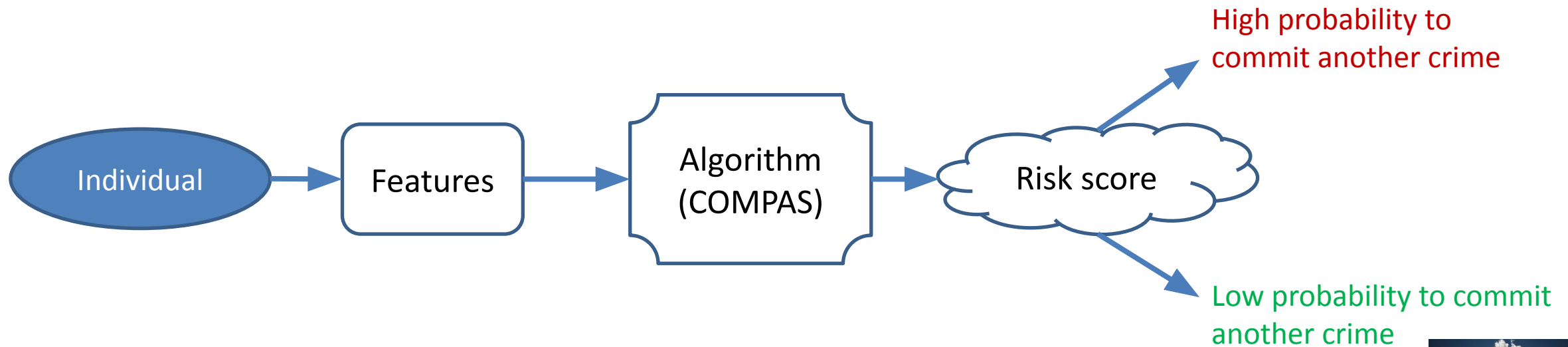
Cost : expected number of innocent people detained

- Define *utility* of a decision rule, that balances benefits and costs



Automated risk assessment

- Increasingly common in the judicial system in the US
- Algorithm computes a score predicting the likelihood a defendant will commit a crime in the future (will recidivate).





Benefits and costs of automated risk scoring

- Increase efficiency and reduce prejudice
 - Used by judges to inform decisions about bail, sentencing, parole, conditions for release, etc.
 - Hoped to mitigate existing sources of bias in human judgments
- In 2014 US Attorney General Eric Holder warned risk scores may be injecting bias into the courts
 - “Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice. They may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”



COMPAS

矫正罪犯管理画像用于替代性制裁

- Correctional Offender Management Profiling for Alternative Sanctions
- Assigns defendants risk scores between 1 and 10 that indicate how likely they are to commit a violent crime based on more than 100 factors, including age, sex and criminal history.
 - Defendants with scores of 7 reoffend at twice the rate as those with scores of 3.
- Help judges make decisions about whether to release a defendant or hold them in jail while waiting for trial
- Deployed before it was rigorously tested

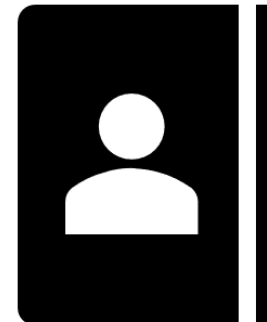
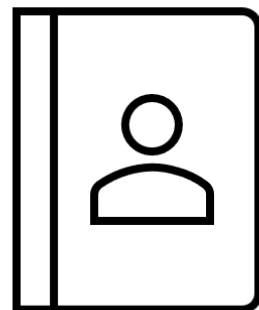


ProPublica study

- Data: 7000 people arrested in Broward County, FL (2013-2014)
 - COMPAS algorithm predicts whether a defendant will commit a crime within next 2 years

	Did commit a crime	Did not commit a crime
Will commit crime	True Positives (TP)	<i>False Positives (FP)</i>
Will not commit a crime	<i>False Negatives (FN)</i>	True Negatives (TN)

Prior Offense	1 attempted burglary
LOW RISK	3
Subsequent Offenses	3 drug possessions



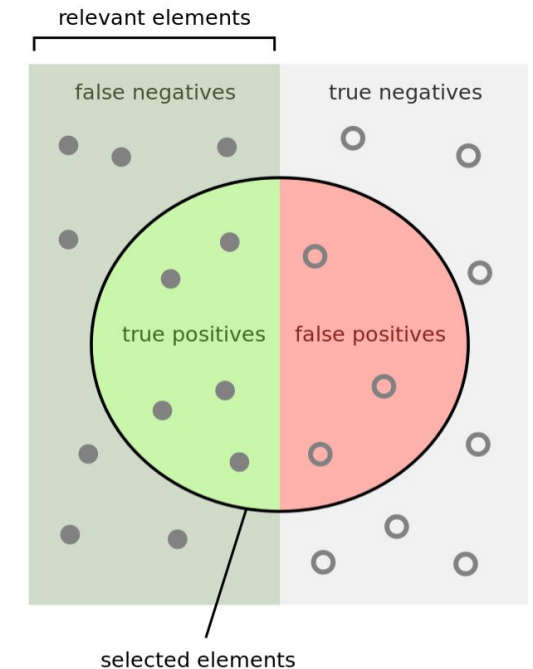
Prior Offense	1 resisting arrest without violence
HIGH RISK	10
Subsequent Offenses	None



Measuring classification performance

	Did commit a crime	Did not commit a crime
Will commit crime	True Positives (TP)	<i>False Positives (FP)</i>
Will not commit a crime	<i>False Negatives (FN)</i>	True Negatives (TN)

- Accuracy = Number of correct predictions/Total predictions
- Accuracy = $TP + TN / (TP + TN + FP + FN)$
- Precision = $TP / (TP + FP)$ *TP / Predicted P*
- Recall = $TP / (TP + FN)$ *TP / Actual P*
- Sensitivity = TP rate = $TP / (TP + FN)$
- Specificity = TN rate = $TN / (TN + FP)$
- False positive rate = $FP / (TN + FP) = 1 - \text{Specificity}$





COMPAS performance

- Accuracy
 - 20% accuracy for violent crimes
 - 61% when all crimes are considered
- Asymmetry of mistakes: who bears the costs?
 - False negatives
 - False positives
- Racial disparity: different types of mistakes for whites and blacks



Racial disparities in COMPAS risk scores

Significant racial disparities: different types of mistakes for whites and blacks

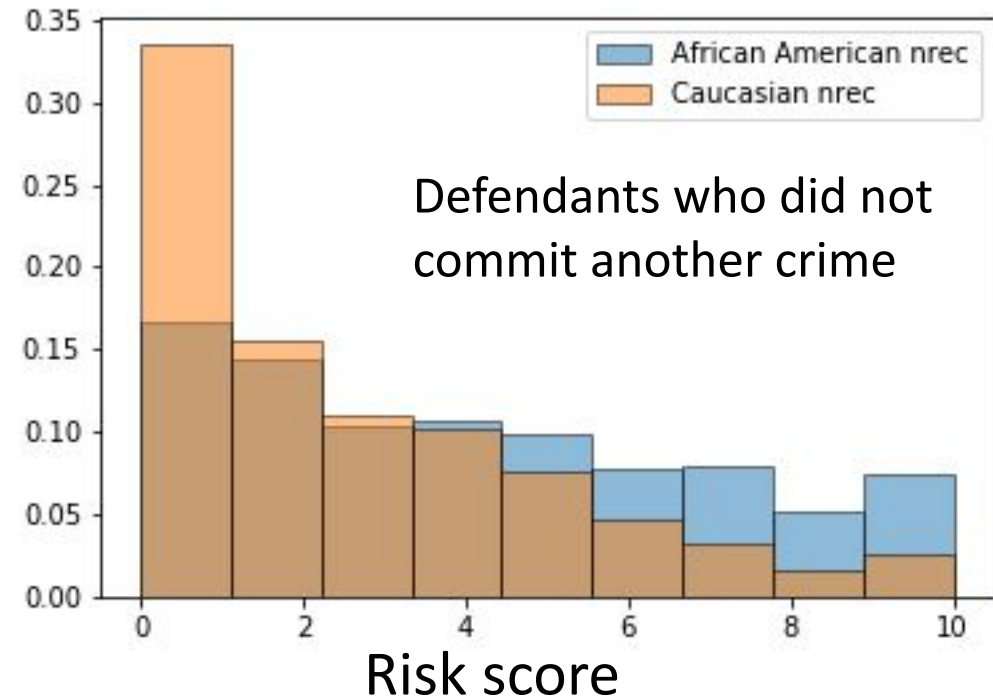
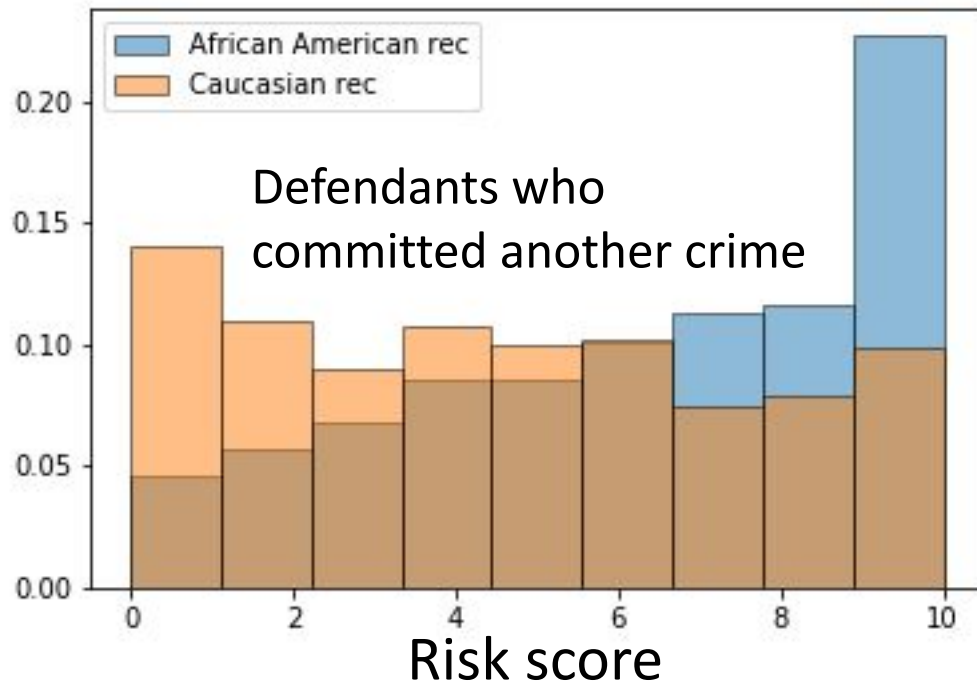
False Positive rate high for blacks:

Algorithm falsely flagged black defendants as future criminals at 2X the rate for white defendants

False Negative rate high for whites

White defendants were mislabeled as low risk more often than black defendants.

This disparity cannot be explained by prior crimes, the type of crimes, their age and gender





Northpointe response

Response to ProPublica: Demonstrating accuracy equity and predictive parity

- ProPublica focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites.
- When the correct classification statistics are used, the data does not substantiate the ProPublica claim of racial bias towards blacks.
- The interpretation of the results in the samples used by ProPublica demonstrate that the General Recidivism Risk Scale (GRRS) and Violent Recidivism Risk Scale (VRRS) are equally accurate for blacks and whites.



Northpointe response

- COMPAS has **predictive parity** and **accuracy equity**
- Classifier has *predictive parity* if it has similar predictive values for two different groups (blacks and whites),
 - I.e., Probability of recidivism given a high risk score is similar for blacks and whites.
- A classifier has *accuracy equity* if it can discriminate recidivists and nonrecidivists equally well for two different groups (blacks and whites).
 - Use AUC to quantify discriminative ability. The ROC curve is a plot of Sensitivity (tpr) and Specificity (tnr) for all possible cut points of a risk scale. The AUC is a summary measure of discriminative across all the thresholds. The AUC is interpreted as the probability that a randomly selected recidivist will have a higher risk score than a randomly selected non-recidivist.

Inherent Trade-Offs in the Fair Determination of Risk Scores



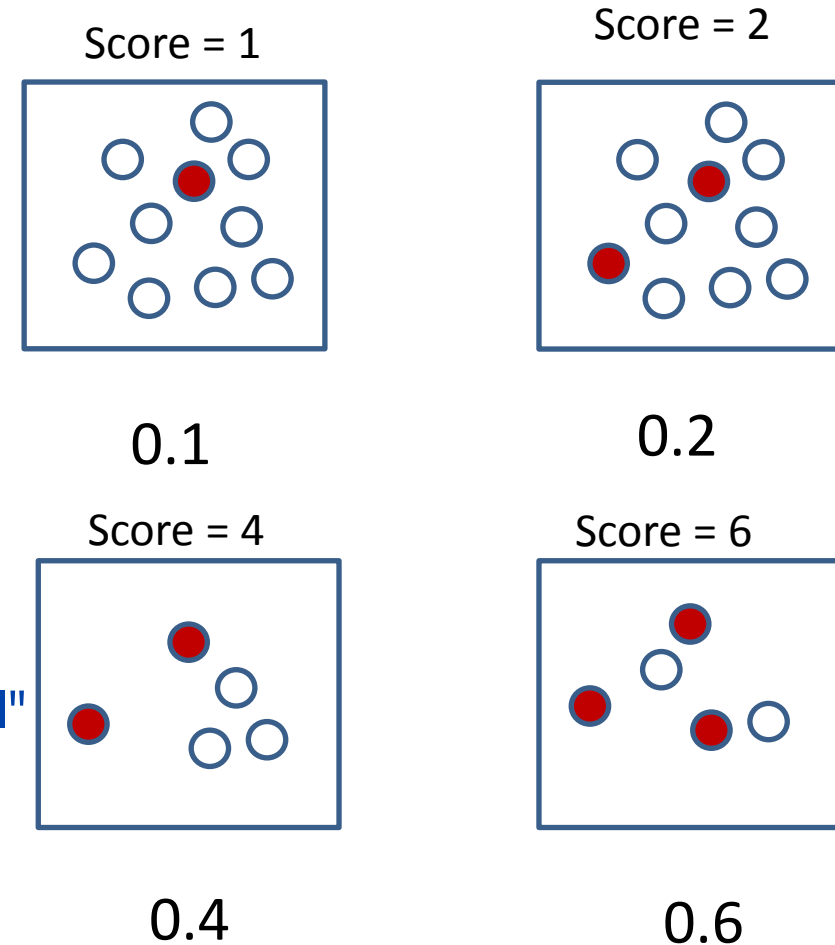
- Calibration, the balance conditions for the positive and negative classes intuitively all seem to be variants of the same general goal — algorithmic predictions should have the same effectiveness regardless of group membership
- But, they are incompatible with each other and can only be simultaneously satisfied in a few special cases.



Fairness properties of risk assessments: Calibration

- **Goal:** predictions should be *well-calibrated*
- If the algorithm identifies a set of people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances
- this condition should hold when applied separately in each group
- Societal benefits to calibration

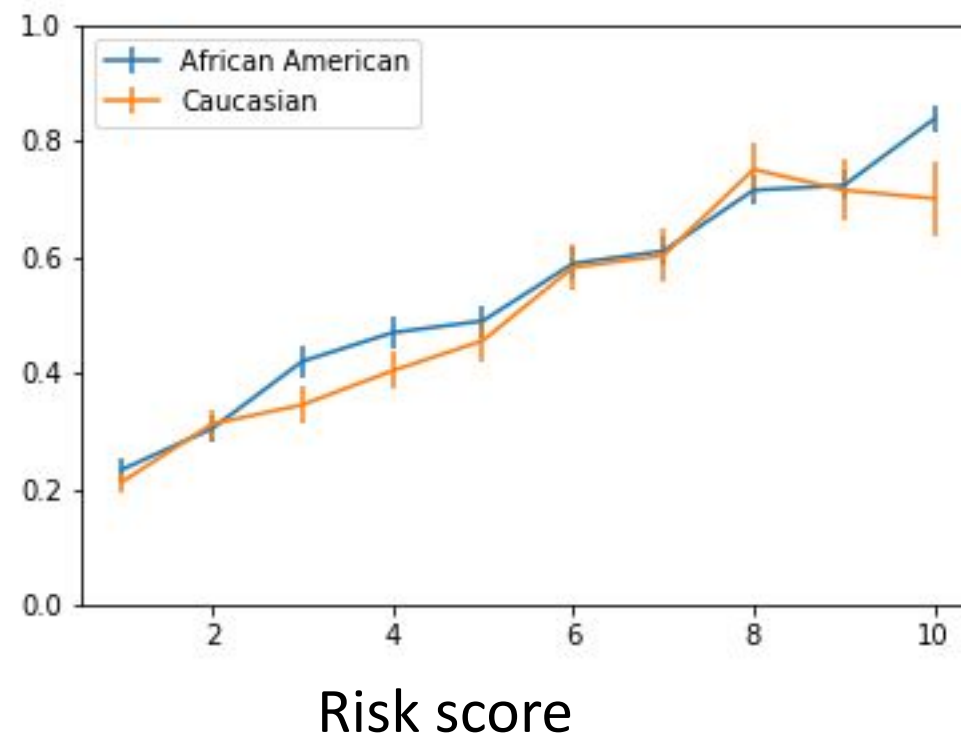
"Societal" refers to large-scale social issues, while "social" refers to the everyday interactions between people



COMPAS is well calibrated



Probability of recidivism





Balance of Predictions

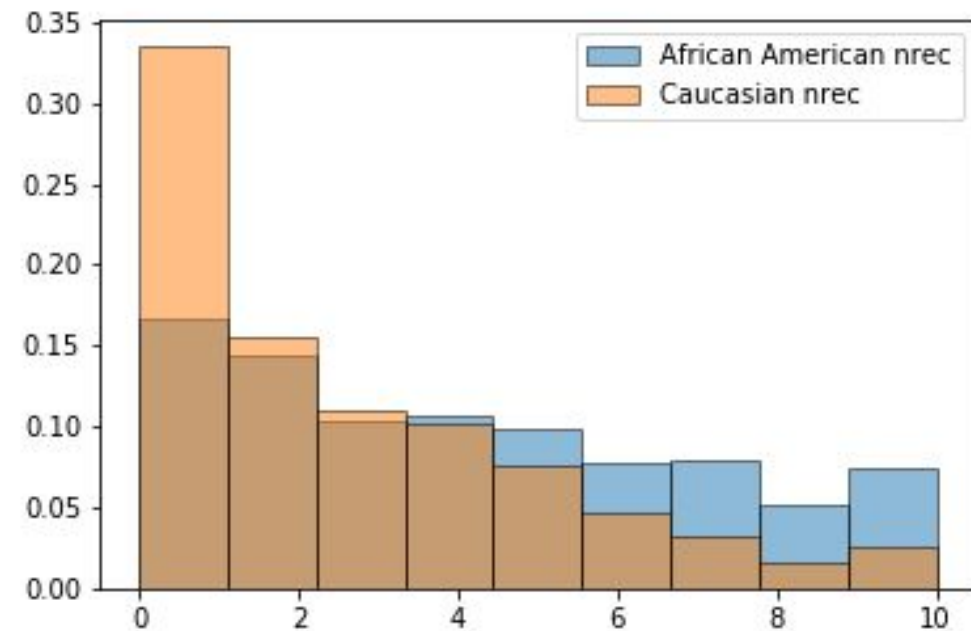
- **Goal:** Predictions should be well *balanced*
 - False negative for group A should be the same as the false negative rate for group B
 - Groups should have equal false positive rates.



Fairness properties of risk assessments: **Balance**

- Predictions should be well *balanced*
 - False positive rate should be the same for all groups
- *Balance for the negative class*: the average scores of negative members in group A (do not re-offend) should be the same as the average score of negative members in group B
 - i.e., risk scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

Negative instances: Defendants who do not commit another crime



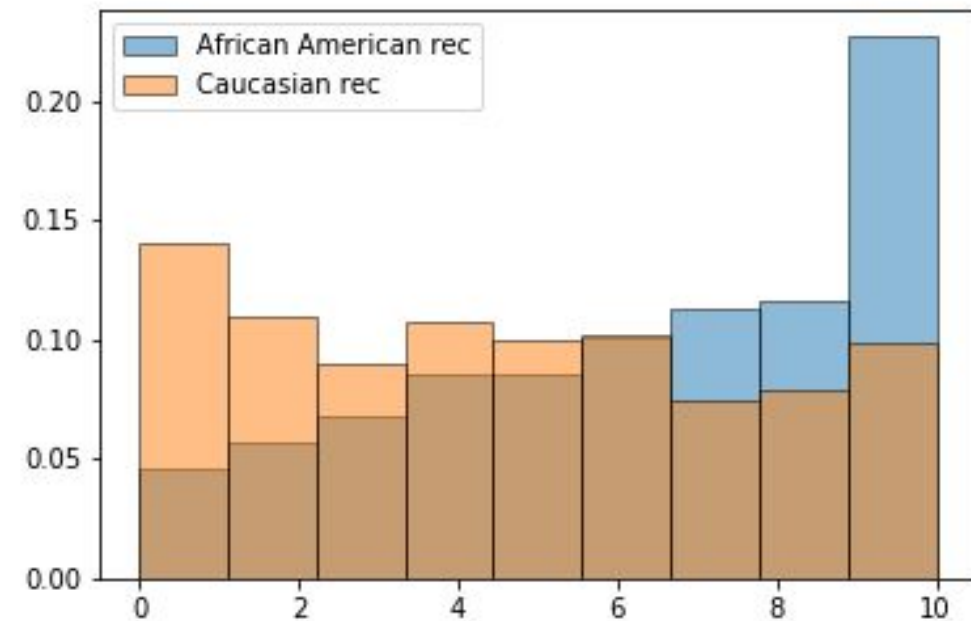
Risk score



Fairness properties of risk assessments: Balance

- *Balance for the positive class* symmetrically requires that average score of defendants who do commit another crime should be the same in each group
- both groups should have equal false positive and false negative rates

Positive instances: Defendants who commit another crime



Risk score



Dueling notions of fairness

- Calibration requires that we treat people with the same score similarly to one other, regardless of their group (race):
 - In other words, **conditioned on the predicted risk score**, the likelihood that the individual is a member of the positive class (i.e., recidivist) is independent of the group to which the individual belongs. They should be given similar bails, and similar sentences (Northpointe's fairness claims)
- Balance requires that if two people in different groups have the same future behavior (recidivate or not), they should be treated the same way
 - In other words, **conditioned on actual behaviors (outcomes)**, members of one group (never commit another crime) should not have consistently higher predicted scores than members of the other group (ProPublica's fairness claims)



Fundamental limits of fairness

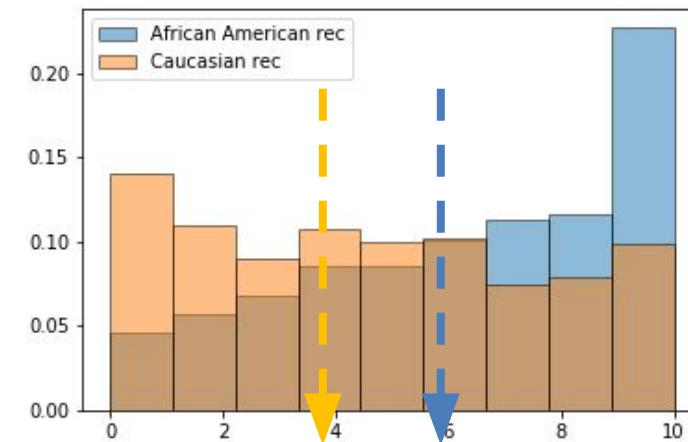
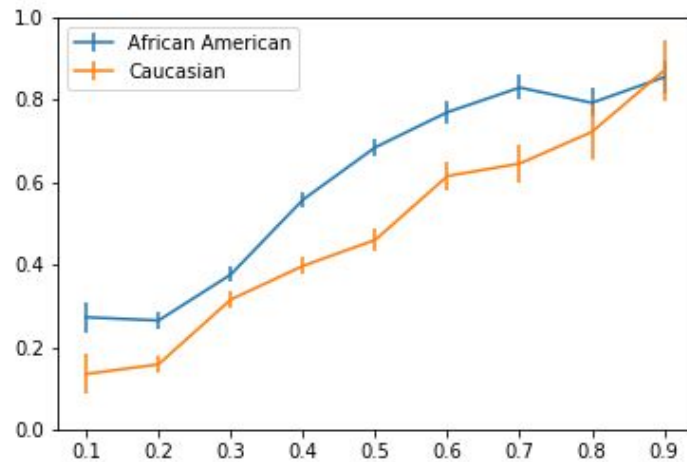
- There are only two cases in which a risk assignment can achieve all three fairness guarantees simultaneously.
three???
 - Perfect prediction
 - Equal base rates
- ?
- In all other cases, you must sacrifice at least one fairness condition

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.



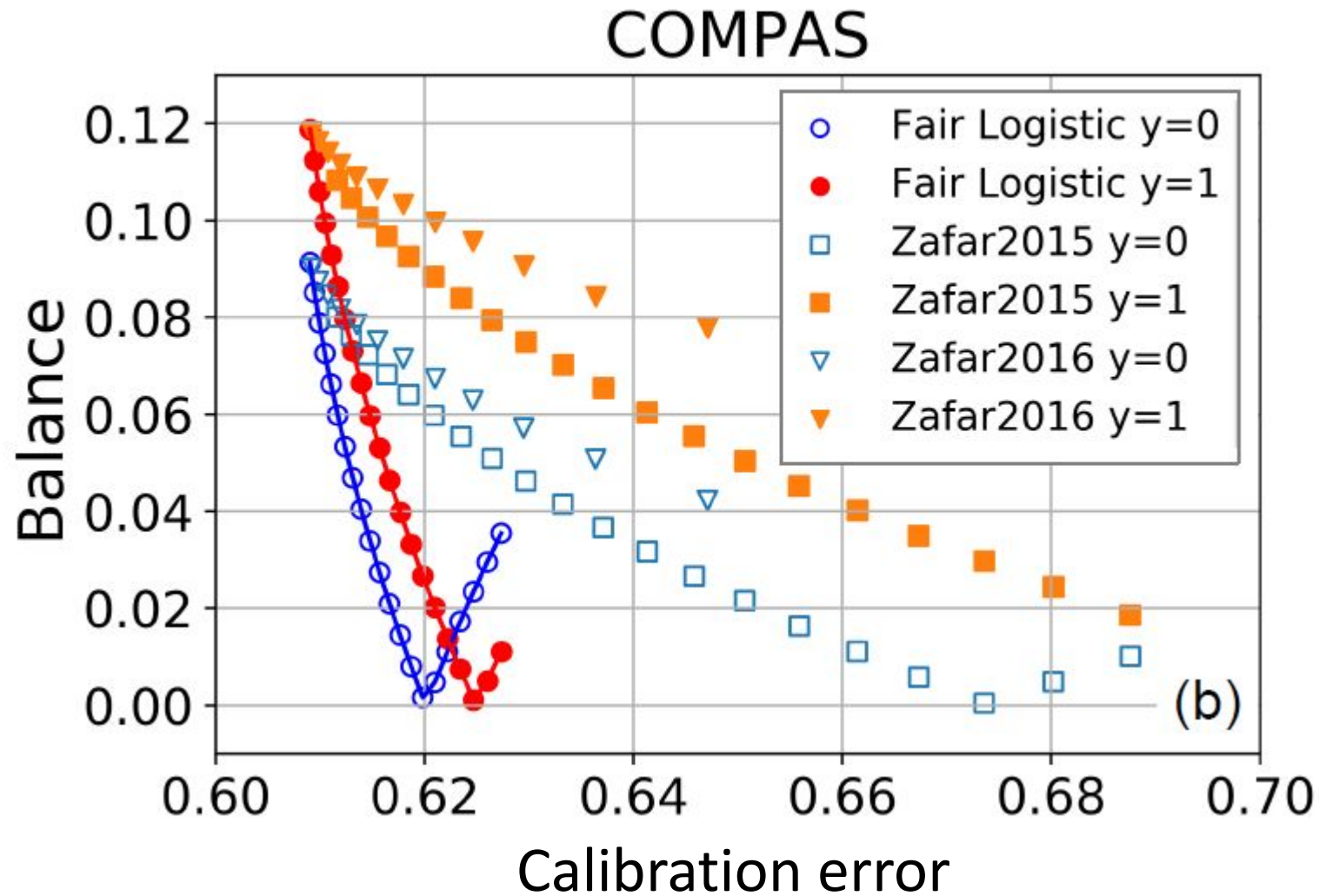
Balance vs Calibration error tradeoffs

- Calibration error: area between two curves
- Balance: Difference between the means for the two groups





Balance vs Calibration error tradeoffs



Difference
between the
means for the
two groups



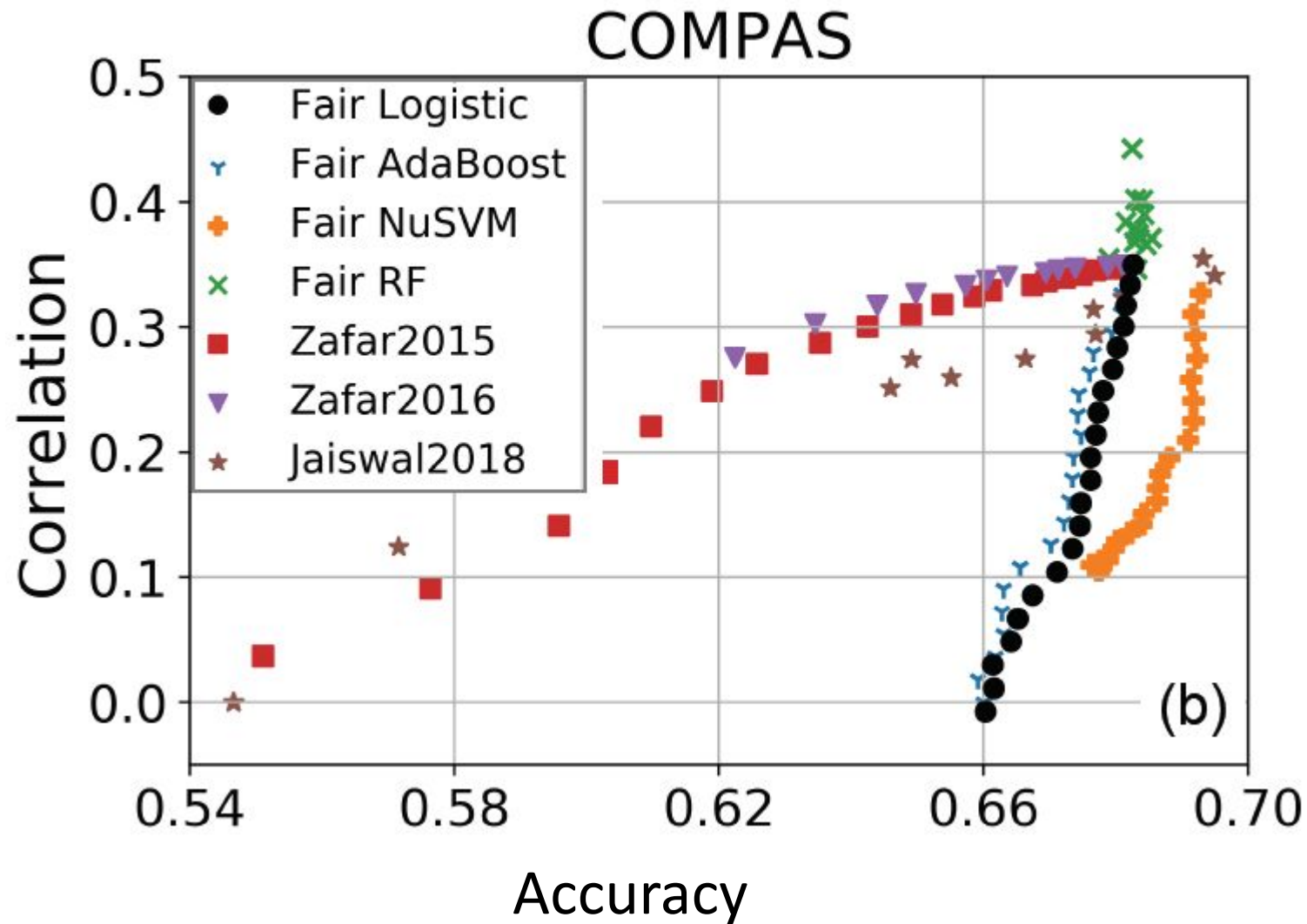
Alternate ways to measure fairness



- Do predictions reveal information about the protected feature?
 - Imagine an evil actor who reverse engineers values of the protected feature from the predictions
 - Measuring dependency
 1. Pearson correlation between the predictions and the protected feature
 2. Mutual information between the predictions and the protected feature
 3. Accuracy of predicting the protected features from predictions



Fairness vs Accuracy tradeoffs





How does bias arise?



Mechanism of bias in COMPAS

- Where does the COMPAS data come from?
 - Level of Service Inventory – Revised (LSI) *is a tool used to assess the risk of an offender reoffending*
- A lengthy questionnaire for the prisoner to fill out, with questions like
 - “How many prior arrests have you had?”
 - “What part did alcohol and drugs play?”
 - “The first time you were involved with the police?”
- The questions themselves treat minority classes differentially and contribute to the feedback loop amplifying bias (masking)
- Also, differences in policing norms across neighborhoods lead to different arrest rates for different groups

Lessons learned: How algorithms can discriminate



- Transparency
 - How is the data collected? Is the bias built into data collection?
 - Choices of what data to collect, what variables to model reflect ideology and blind spots
 - Are participants aware of being modeled? Do they know how their data is used?
 - Does the model rely on proxies? Are they valid?
- Asymmetry
 - Who is harmed by the mistakes? Is the damage born by all groups equally?
 - Is the harm of False Positives comparable to the harms of False Negatives?
- Feedback
 - Does the model learn from mistakes? Is it updated when ground truth changes?
 - Does it create a feedback loop that exacerbates the damage?



RISK PREDICTION IN HEALTH CARE

OBERMEYER, Z., POWERS, B., VOGELI, C., & MULLAINATHAN, S. (2019). DISSECTING RACIAL BIAS IN AN ALGORITHM USED TO MANAGE THE HEALTH OF POPULATIONS. *SCIENCE*, 366(6464), 447-453.



Main findings

- Health systems use algorithms to identify and help patients with complex health needs,
- Algorithms are racially biased: Conditioned on risk score, Black patients are considerably sicker than White patients
- Removing the disparity would increase the percentage of Black patients receiving care from 17.7 to 46.5%.
- The bias arises because health care costs are used as a proxy of health. However, unequal access to health care means that less money is spent on caring for Black patients than White patients.
- The choice of convenient, seemingly effective proxies for ground truth can be a source of algorithmic bias.



The problem

- Health systems rely on algorithms to identify patients who will benefit from costly health interventions
- Identifying patients who will benefit from interventions is a challenging causal inference problem. Instead, health systems use a key assumption: Patients with the greatest care needs will benefit the most from the program.
- This is a prediction problem: use past data to predict future health care needs.
- This study: Access to data, specifically inputs, outputs, and eventual outcomes, allows to quantify racial disparities in algorithms and isolate the mechanisms by which they arise.



The study

- Patients
 - 6,079 Black; 43,539 White
- Predict $R_{i,t}$ risk score for patient i in year t ,
 - Predicted on the basis of claims data $X_{i,(t-1)}$ from the prior year.
 - Patients above the 97th percentile are automatically identified for enrollment in the program.
- Assess performance and fairness
 - Compare the risk score $R_{i,t}$ to patient's health outcomes $H_{i,t}$
 - Which fairness metrics?: Focus on calibration bias [comparing Blacks B and Whites W]:
 $E[Y|R;W]=E[Y|R; B]$ indicates the absence of bias
 - Assess how well the algorithmic risk score is calibrated across race for health outcomes $H_{i,t}$
 - How well the algorithm is calibrated for costs

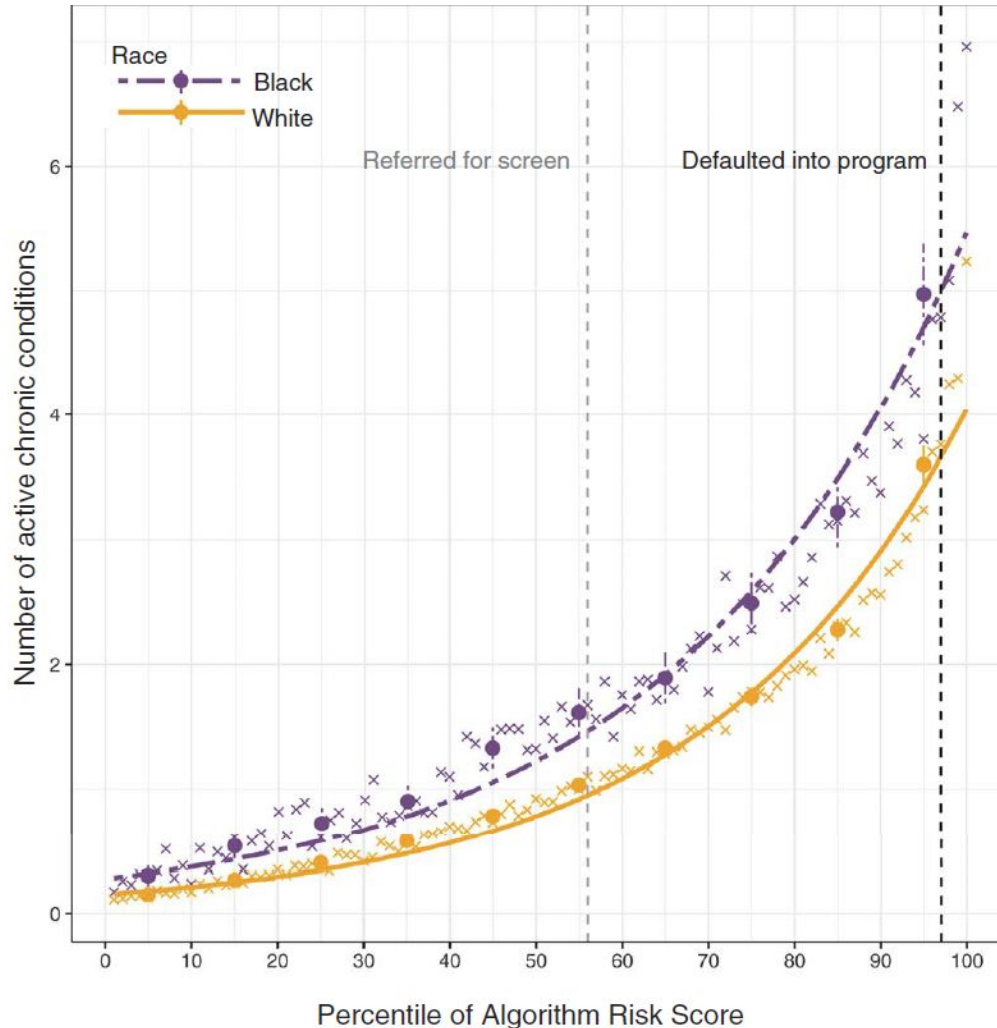
Racial differences in health

	White	Black
<i>n</i> (patient-years)	88,080	11,929
<i>n</i> (patients)	43,539	6079
<i>Demographics</i>		
Age	51.3	48.6
Female (%)	62	69
<i>Care management program</i>		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
<i>Care utilization</i>		
Actual cost	\$7540	\$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
<i>Mean biomarker values</i>		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
<i>Active chronic illnesses (comorbidities)</i>		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08





Health disparities conditioned on risk scores

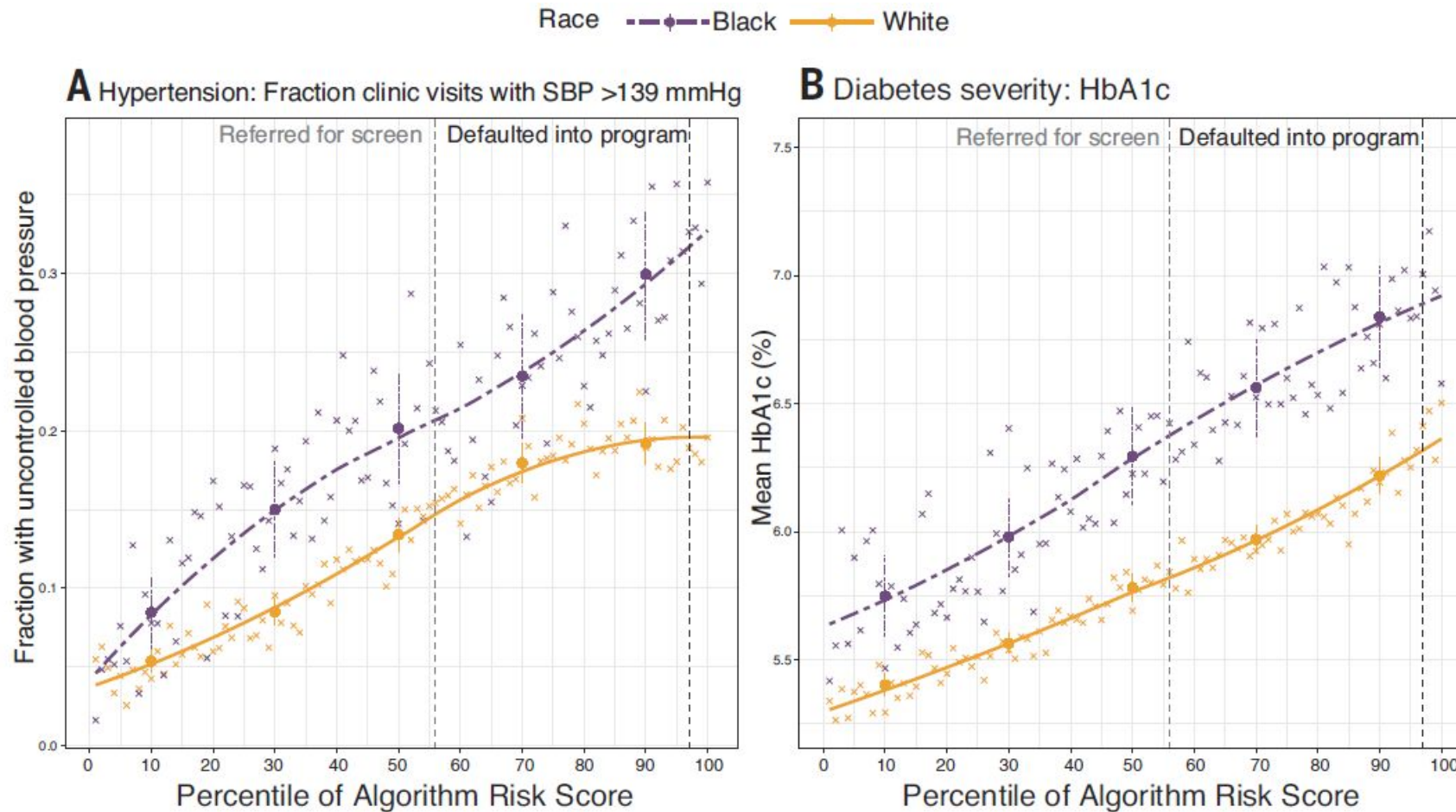


- Calculate the overall measure of **health** (number of active chronic conditions) by **race**, conditional on algorithmic **risk** score.
- At the same level of risk, Blacks have significantly more comorbidities than Whites.
 - E.g., at 97% risk score, Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$)

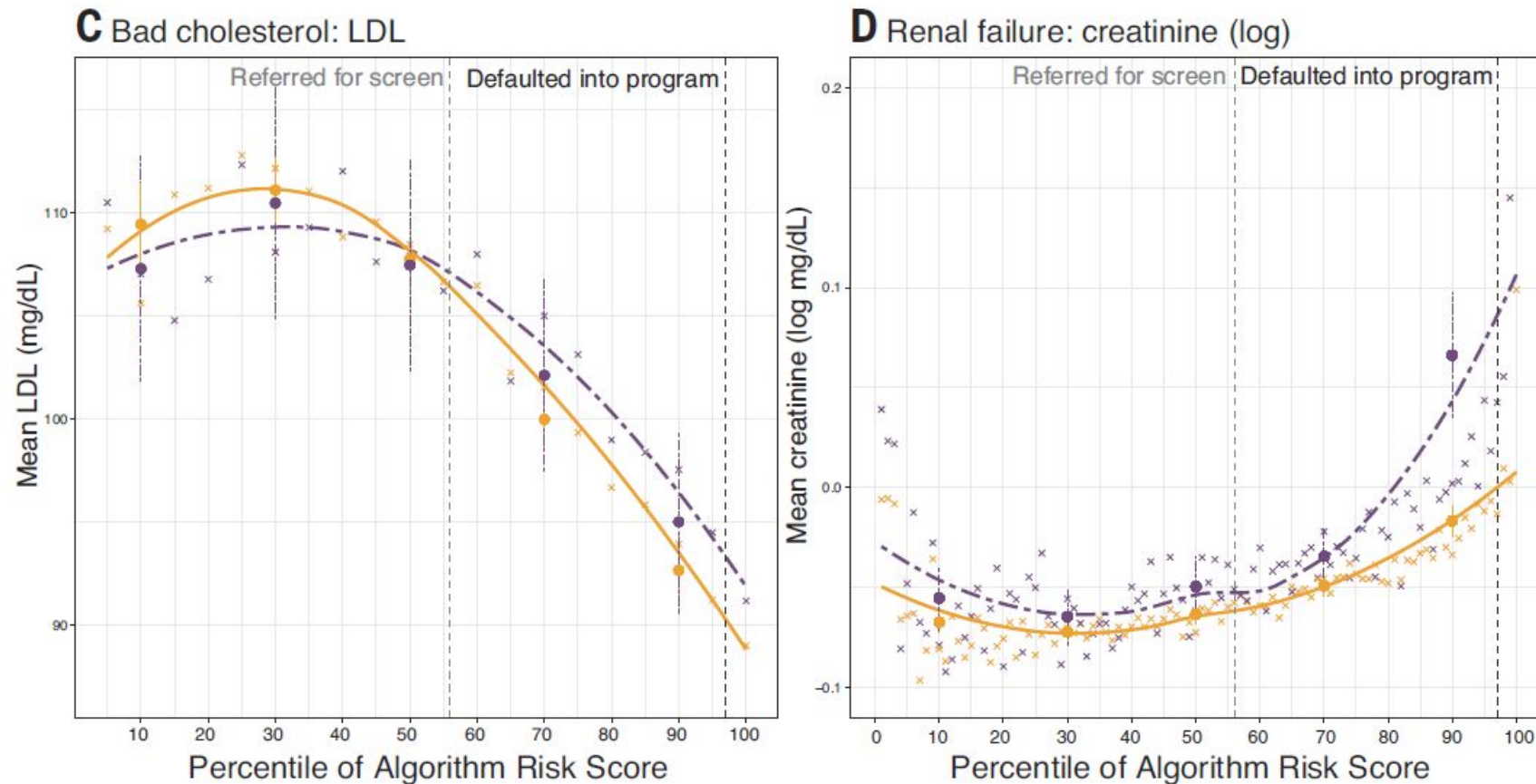
Racial disparities persist across all markers of health



Blacks have more-severe hypertension and diabetes: up to 30% differences in mortality



Racial disparities persist across all markers of health





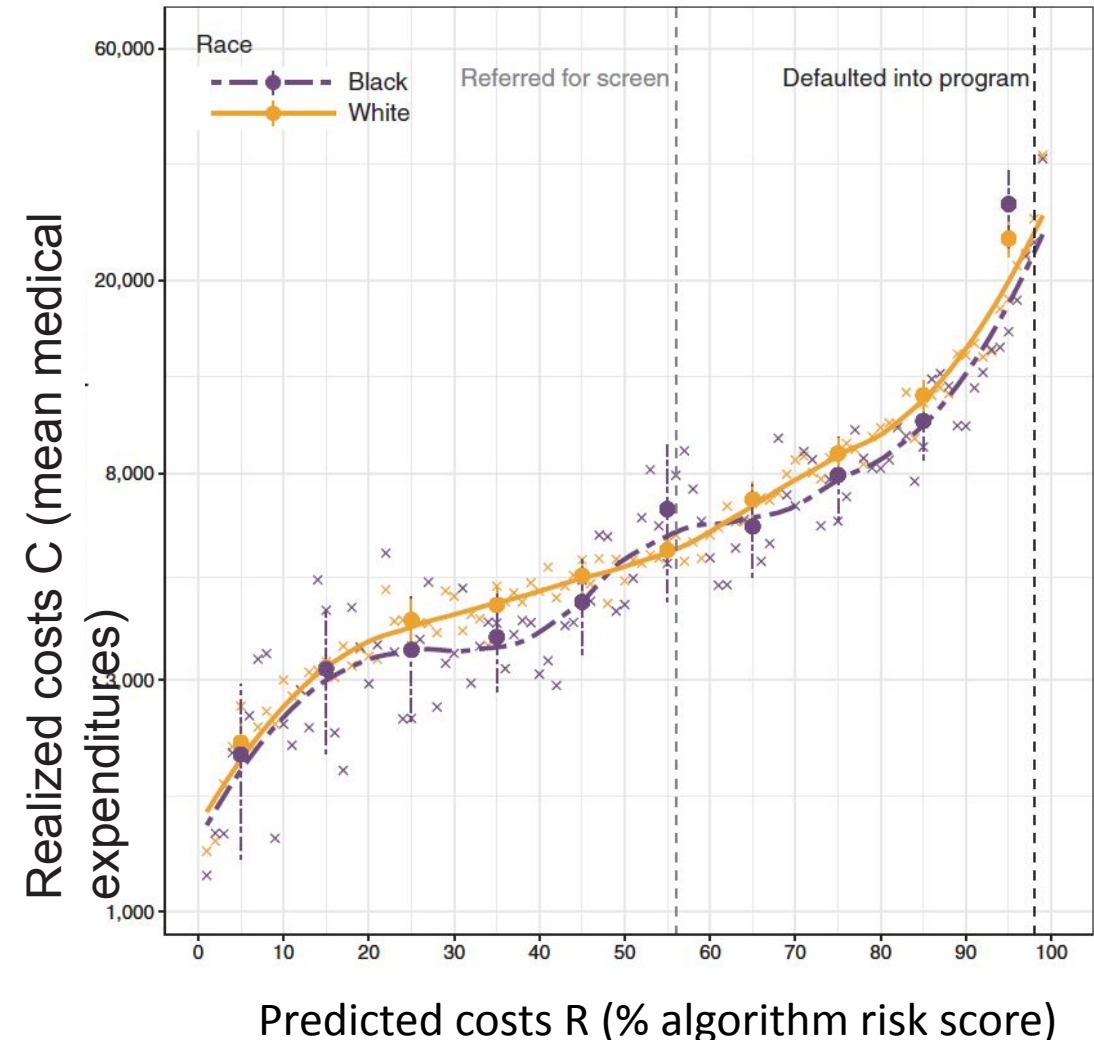
Mechanism of bias

- Study has all the data to investigate mechanisms giving rise to algorithmic health disparities
- Algorithm takes as **input**
 - set of raw insurance claims $X_{i,t-1}$ (features) over the previous year,
 - demographics (e.g., age, sex), **but not race**
 - insurance type, diagnosis and procedure codes,
 - medications, detailed costs.
- The algorithm uses these data to predict **outcome/label** $Y_{i,t}$ (health) in year t .
 - Design decision: use total medical costs C_t in year t as a **proxy of outcome**.
 - Health and costs are highly correlated: sicker patients receive more care, on average
 - Thus, **instead of predicting health needs, it predicts health costs**



Algorithm is well calibrated across races

- Algorithm is well calibrated: at every level of predicted risk, Blacks and Whites have (roughly) the same costs the following year.
 - Conditioned on risk, predictions do not favor Blacks or Whites
 - Eg, at the median risk score, Black patients had costs of \$5147 vs \$4995 for Whites;
 - in the top 5% of algorithm-predicted risk, costs were \$35,541 for Blacks versus \$34,059 for Whites.





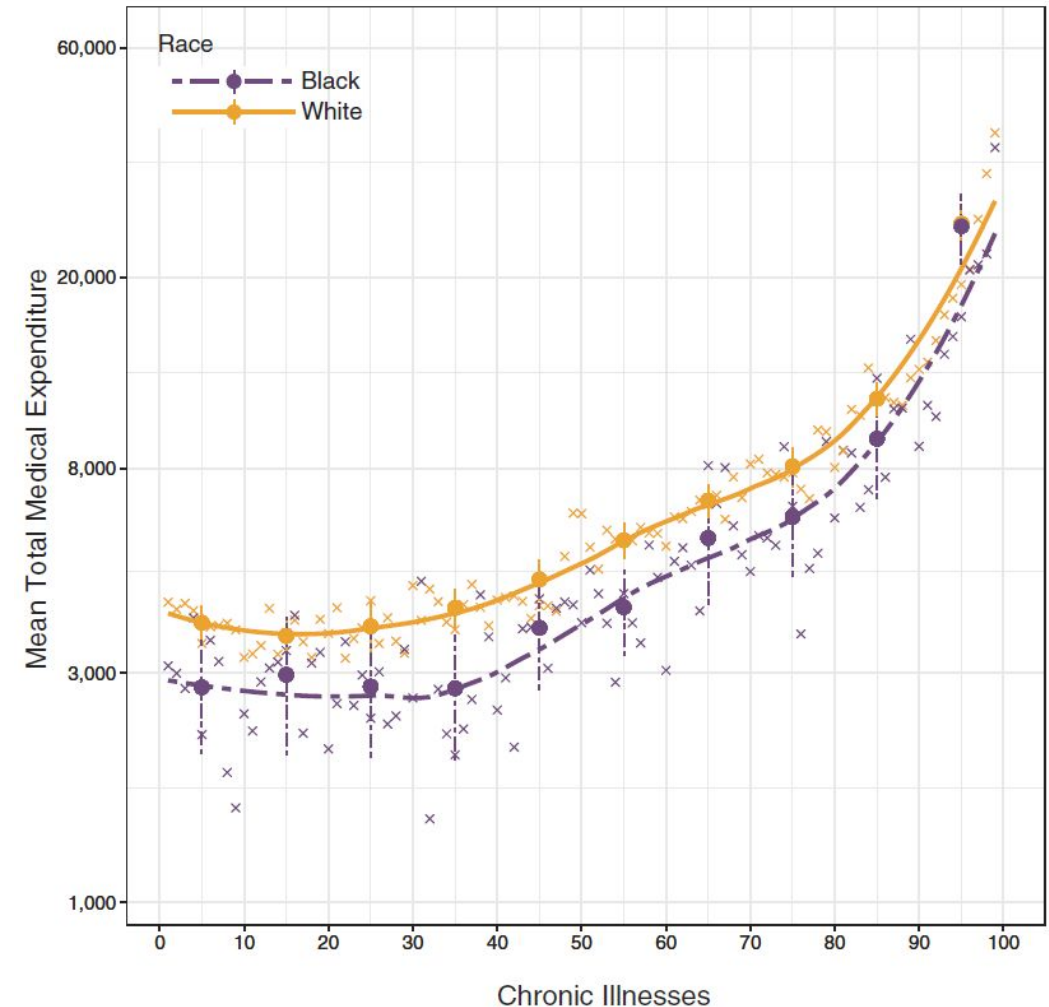
Why is this surprising?

- Findings:
 - There are substantial disparities in health conditional on risk,
 - But little disparity in costs.
- Source of disparity: the gap between needing health care and receiving health care is correlated with race.
 - i.e., sick Blacks do not receive adequate care to meet their needs.



Racial gap in health expenditures

- At a given level of health (measured by number of chronic illnesses), Blacks have lower costs than Whites—on average, \$1801 less per year, holding constant the number of chronic illnesses.
- Origin of bias: Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities.
- As a result, accurate prediction of costs necessarily means being racially biased on health.



Origins of racial disparities in health costs



Socioeconomic factors

- Poor patients (even when insured) face barriers to accessing health care. Poverty can lead to disparities in use of health care: geography and differential access to transportation, childcare and job demands.
- To the extent that race and socioeconomic status are correlated, these factors will differentially affect Blacks.

Prejudice (communication, trust or bias)

- Discrimination in doctor–patient relationship, etc. Physicians' perceptions of Black patients, in terms of intelligence, affiliation, or pain tolerance
- Black patients have reduced trust in the health care system due to adverse experiences.

The collective effect is to lower health spending for Black patients, conditional on need

Choosing a label on which to train the algorithm



- On the one hand, the choice to predict future costs is reasonable: The program's goal is to reduce costs, and it stands to reason that patients with the greatest future costs could have the greatest benefit from the program.
- On the other hand, future cost is by no means the only reasonable choice. For example, care management programs do not reduce costs, but prevent catastrophic health care utilization. Thus, avoidable future costs, i.e., emergency visits and hospitalizations, could be a useful label to predict
- Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions.
- Health care costs, though easily quantifiable and readily available in insurance claims, are also the result of a complex aggregation process with a number of distortions due to structural inequality, incentives, and inefficiency.



Alternate labels/outcomes

- How does label choice affect predictive performance and racial bias?
- Algorithms trained to predict the following outcomes:
 - total cost in year t (this tailors cost predictions to our own dataset rather than the national training set),
 - avoidable cost in year t (due to emergency visits and hospitalizations),
 - health in year t (measured by the number of chronic conditions that flare up in that year).
- 3-fold cross validation of all models
 - trained on a random $\frac{2}{3}$ training set and show all results only from the $\frac{1}{3}$ holdout set.
- Exclude race from the feature set



Performance of predictors trained on alternate labels

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	

- Performance: All algorithms perform well for predicting the outcome on which they were trained and other outcomes: The concentration of realized outcomes in those at or above the 97th percentile is notably similar for all algorithms across all outcomes.
- Bias: the racial composition of highest-risk group varies across algorithms: fraction of Black patients at or above risk levels ranges from 14.1% to 26.7%.
- Although there are many reasonable choices of label—all predictions are highly correlated, and any could be justified as a measure of patients' likely benefit from the program—they have markedly different implications in terms of bias, with nearly 2x variation in composition of Black patients in the highest-risk groups.



Label choice is a big source of algorithmic bias

- Labels meant to capture ineffable factors (“health”, “good worker”) are often measured with errors that reflect structural inequalities.
 - Health: Using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-White populations
 - Credit-scoring: algorithms that predict outcomes related to income incorporate disparities in salary.
 - Policing: algorithms predict crime, also reflect increased scrutiny to marginalized communities
 - Hiring: algorithms predict employment decisions or worker ratings are affected by race and gender biases.
 - Retail: algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to higher prices as a result.



Reducing bias is possible

Labels are the key determinant of both predictive performance and predictive bias;

- Careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their harms.
- Producing new labels, however, requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment.



Lessons learned: Why models fail

- Transparency
 - Is the model transparent or opaque? Or even invisible?
- Are participants aware of being modeled?
 - Do people know model's conclusions?
 - Are the mistakes the model makes symmetric?
- Does the model learn from mistakes?
 - Is it updated dynamically when ground truth changes?
 - Does it create a feedback loop that exacerbates the damage
- Does the model use proxies?
 - Models using actual data about the behavior, rather than proxies, are more likely to be valid



How algorithms can go wrong

- Blackbox algorithms are unfair
 - They cannot explain their decisions, and you cannot appeal their decisions
- Algorithms are unfair when their mistakes are asymmetric
 - Damage of False Positives vs damage done by False Negatives
- Algorithms are unfair when they do not learn from their mistakes
- Algorithms may be unfair when they rely on proxies of the behavior or outcome they are measuring
- Algorithms are unfair when they create pernicious feedback loops
- Algorithms may be unfair by design
 - Designer decides what data to collect, what variables to model
 - Choices reflect designer's ideology and blind spots