PS/HS Bias: Bias Types

Katja Markert

Uni Heidelberg

WS 2020

Overview

Case Study: Obermeyer et al 2019

Disparate treatment/impact

Causes of bias

Overview

Case Study: Obermeyer et al 2019

Disparate treatment/impact

Causes of bias

Paper

Obermeyer et al: *Dissecting racial bias in an algorithm used to manage the health of populations.* Science (2019)

- Area: Health care
- Algorithm Aim: Predict patients at high risk (with complex health needs) in order to provide additional ressources for them
- Bias/Problem: Gives black patients lower risk scores although they are equally or more at risk $\longrightarrow$ they receive additional ressources less often than whites at equal or lesser health

Political Background

- Affordable Care Act (Obamacare)
- Flat annual fee per patient or end-of-year monetary adjustments relative to negotiated cost targets (fee-per-patient instead of fee-per-service)
- Led to high-risk care management programs: additional ressources before health deteriorates. Need precise targeting of patients that are likely to benefit and/or that otherwise cost more.

Algorithms

- Key Assumption: Patients with high risk (= greatest care needs) will benefit the most from additional treatment
- Commercial algorithms predict which patients have the most complex health care needs and get entry into program.
- Mostly done via cost prediction

Algorithm in Paper

- Commercial, used for 200m Americans
- Aim: *flag individuals for intervention before their health becomes catastrophic*
- If risk score above 97th percentile $\rightarrow$ automatic enrollment into health care program
- If above 55th percentile $\rightarrow$ referred to primary care physician

Algorithm

1. $R_{i,t}$: algorithmic risk score for patient i in year t. Label to be predicted.
2. $R_{i,t}$ is predicted via insurance data from prior year $X_{i,(t-1)}$.
3. What they really try to predict are the actual, realised costs $C_{i,t}$ in the next year.

Chain of Assumptions

treatment value $\approx$ *health* $H_{i,t} \approx$ *pred. risk score* $R_{i,t} \approx$ *realised costs* $C_{i,t}$

Algorithm in detail

Label to predict: $C_{i,t}$

Features from previous year $X_{i,(t-1)}$

- Demographics (e.g., age and sex, but specifically excluding race),
- Insurance type,
- ICD-9 diagnosis and procedure codes,
- Prescribed medications,
- Encounters, categorized by type of service (e.g., surgical, radiology, etc.),
- Billed amounts, categorized by type (e.g., outpatient specialists, dialysis, etc.).

Inner workings of algorithm unknown

Data

- 6079 self-identified black and 45 539 self-identified white patients
- 2013-2015 in one hospital

How to measure bias

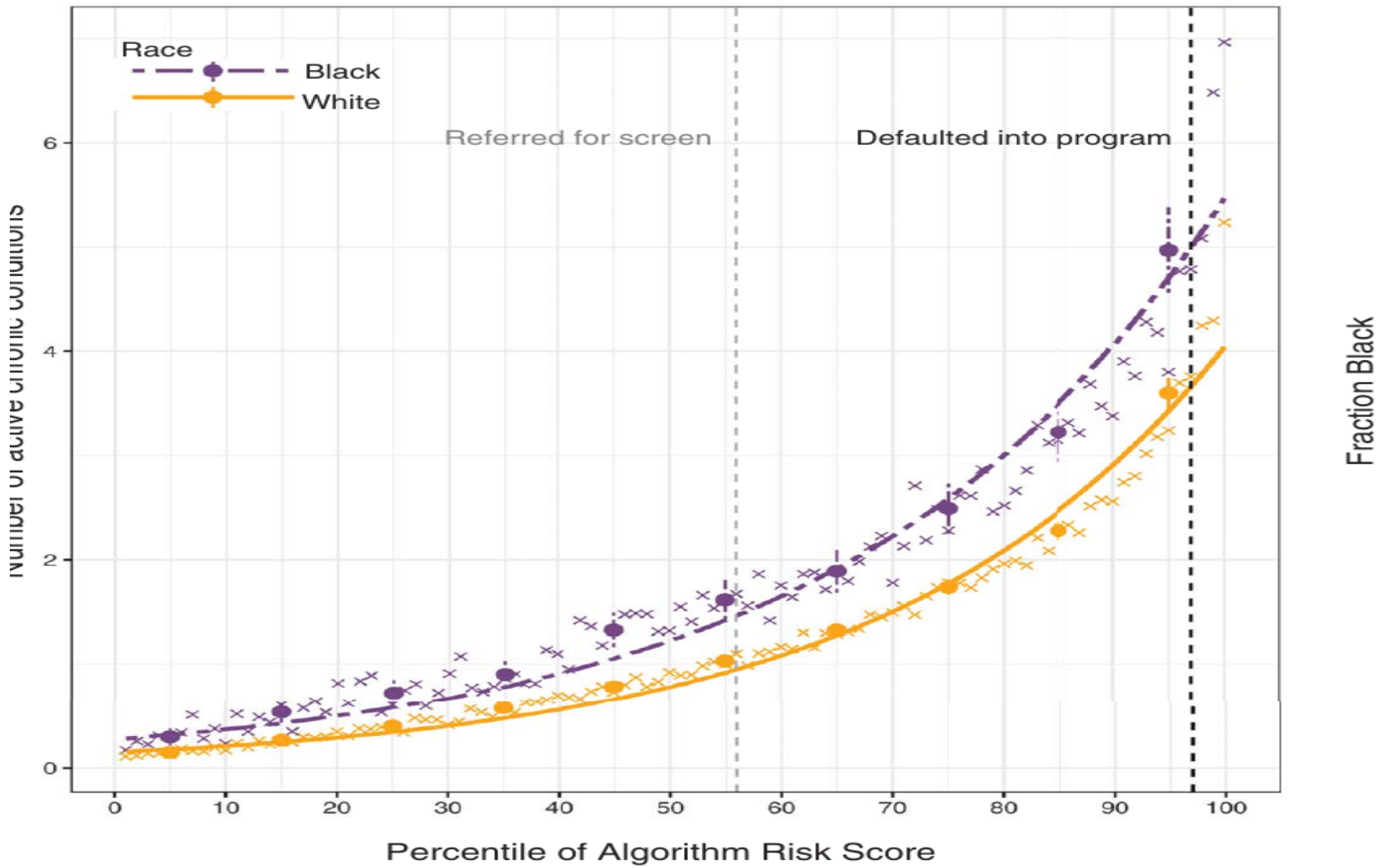
- Measure real health $H_{i,t}$: use electronic health records for diagnoses as well as lab measurements
- Compare $R_{i,t}$ to $H_{i,t}$
- Calibration for Health: If algorithm is unbiased, what should hold is

$$E[H|R, W] = E[H|R, B]$$

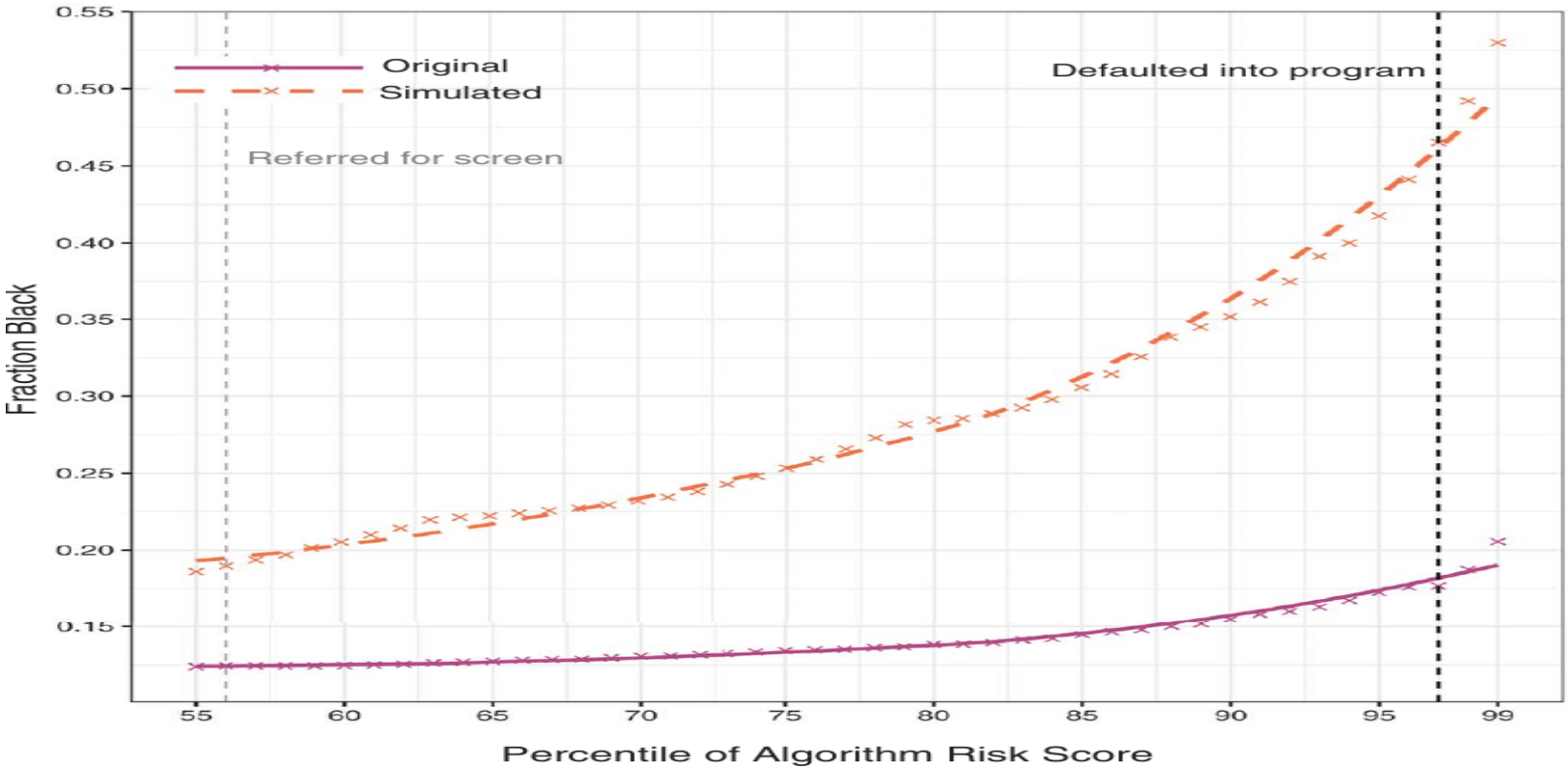
- Also looked at calibration for costs $C_{i,t}$ measured via insurance claims

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

Number of chronic illnesses active per year per risk score



Simulation if no such bias existed



Fraction of blacks at the 97th percentile rises from 17.7% to 46.5%, i.e almost half of the referred patients should be black.

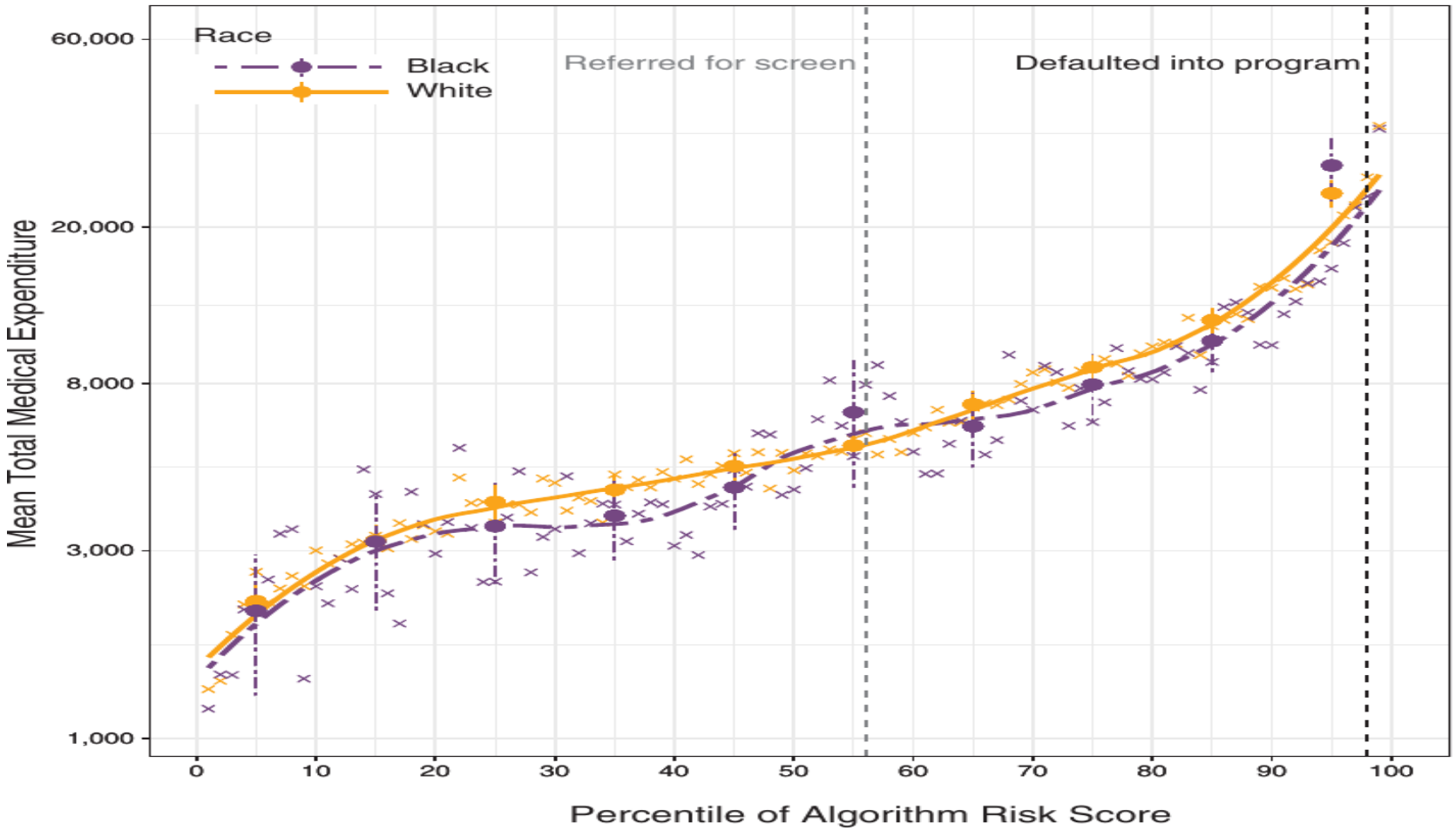
Could it be that this comes from actual enrolment in program?

No:

- Same effect when using H for years before program enrollment
- Same effect for enrolled and unenrolled patients

Calibration for costs

First idea: maybe algorithm risk score does not match costs $C_{i,t}$ very well.



So what does that mean

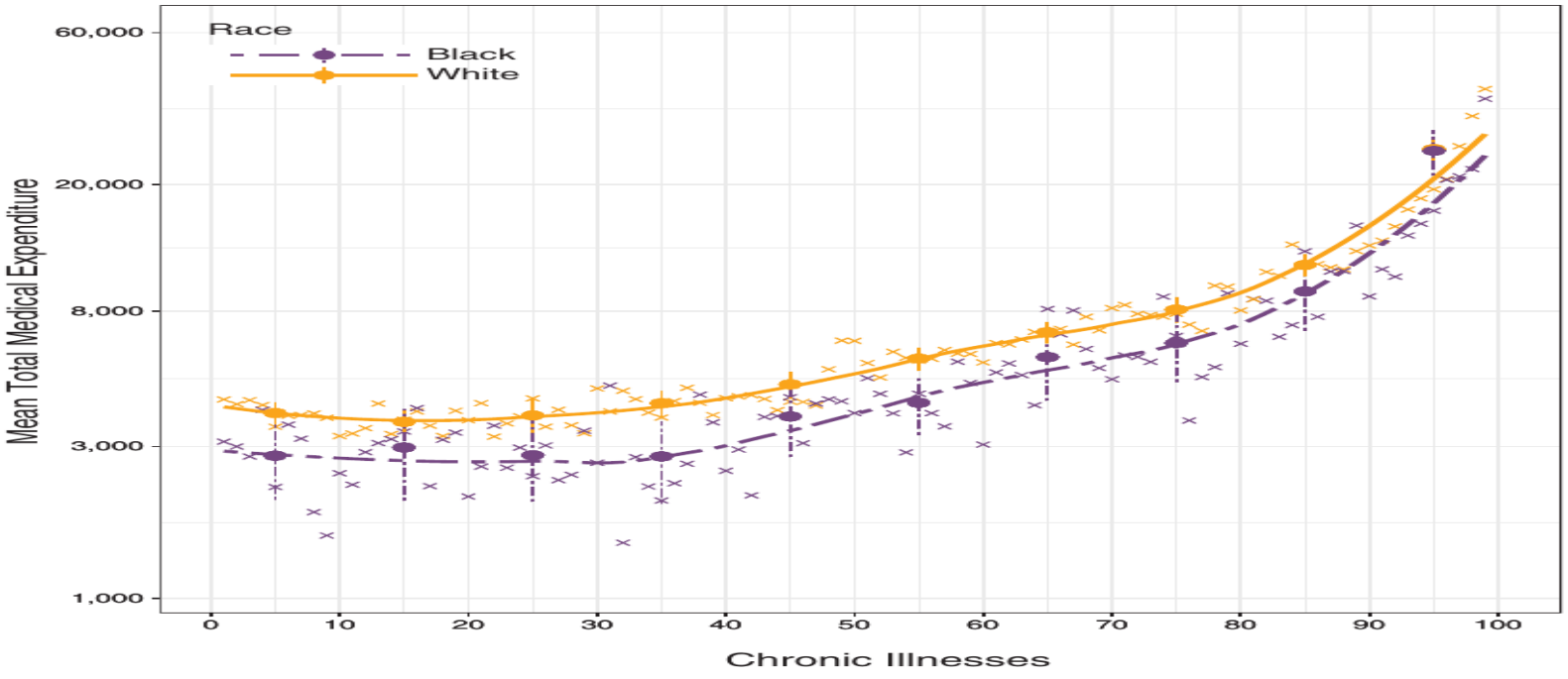
- Well calibrated for cost across races, not well calibrated for health across races
- Substantial differences in health at same risk scores but none on costs

This must mean that the equivalency health $\approx$ costs breaks down when it comes to race.

Assumptions

$$\textit{treatment value} \approx \textit{health } H_{i,t} \approx \textit{pred. risk score } R_{i,t} \approx \textit{realised costs } C_{i,t}$$

Evidence



- at a given level of health, blacks generate lower costs than whites (and different costs)
- accurate predictions of costs is necessarily racially biased on health
- costs not a good proxy label for health $\rightarrow$ label bias

Why might blacks generate lower costs?

- Insurance
- Poverty: access to transport, more demands from jobs, single parent families
- Trust relationship to doctors and health systems: not many doctors black, Tuskegee Study of Untreated Syphilis in the Negro Male, different treatment by doctors

Label bias

Label Bias Definition

- Choose a proxy label to predict
- This choice leeds to bias

Proxy labels are often used:

- Convenience of access (for example 1-year survival rate in pharmacological trials)
- Hard to measure "real label"
- Proxy label might be in interest of manufacturer or provider

Experiments on label choice

Other labels with own algorithm *Predictor*

1. total cost in year t (tailored to dataset, not overall national training set)
2. avoidable cost in year t (emergency and hospitalisations)
3. health in year t (= number of chronic conditions flaring up in t

Results turn out to be well correlated but:

Label	% of black in highest risk group
total cost	14.1%
avoidable cost	21.0%
health	26.7%

How does *Predictor* work?

- L1-regularised regression
- 149 features (without race)
- 2/3rd training, 1/3rd test
- regularization penalty tuned via ten-fold cross-validation n training

Other examples for potential label bias

Label	Proxy
good employee	supervisory ratings
good hospital	readmissions/mortality
likely to reoffend	criminal rates of groups/family/friends

Discussion Questions

1. What kind of bias does the paper describe (intentional/unintentional; to which protected or minority group; stereotyping or algorithmic decisions)?
2. What is the cause of the bias exhibited?
3. Research the terms *disparate treatment* vs *disparate impact*: are these two terms relevant to the bias exhibited?
4. Has the *Spiegel* article presented the research appropriately?
5. What additional points of the forum participants under the *Spiegel* article are (ir)relevant and which exhibit what kind of misunderstandings? (See for example posts 12,15,28 but please also look at others).
6. How could you avoid label bias in the examples on the previous slide?

Overview

Case Study: Obermeyer et al 2019

Disparate treatment/impact

Causes of bias

Disparate treatment

Disparate treatment (= unmittelbare Diskriminierung)

Discriminatory outcomes due to choices made explicitly based on membership in a protected class.

Sometimes also called **blatant explicit discrimination**

When the protected class is used directly in the model, called **direct discrimination**.

Questions

- Is it still disparate treatment if majority members are rejected due to majority class status? (see also Ricci vs. DeStefano case in US and Dwork et al (2012))
- Is it still disparate treatment if the membership of the protected class is taken into account but leads to more positive outcomes for protected class members?

Operationalisation of disparate treatment

Fairness through unawareness

An algorithm is fair as long as any protected attributes A are not explicitly used in decision-making process.

Grgic/Hlaca et al: *The case for process fairness in learning: feature selection for fair decision making* in NIPS Symposium on Machine Learning and the Law 2016.

Disparate Impact Definition

Disparate Impact (= mittelbare Diskriminierung)

outcomes should not be different based on individuals' protected class membership, even if process does not explicitly base decision on protected attributes

Disparate Impact

- predominant theory in US (and Germany)
- covers unintended and some forms of intended discrimination
- no rigid formulae in law: case-based decision
- US Equal Employment Opportunity Commission (EEOC): 80% rule on ratio of hiring rates:

$$\frac{p(YES|minority)}{p(YES|majority)} \leq 0.8$$

- Disparate impact can be allowed if employer can show that necessary for safe and efficient performance of job! (Explainable Discrimination) `https://www.eeoc.gov/policy/docs/factemployment_procedures.html`

Protected Attributess

A	FHA	ECOA	EEOC	Allg. Gl.
Race	x	x	x	x
Color	x	x	x	
National origin	x	x	x	x
Religion	x	x	x	x
Sex	x	x	x	x
Disability	x			x
Marital status		x		
Recipient of public assistance		x		
Age		x	x	x

FHA: Fair Housing, ECOA: Equal Credit Opportunity, EEOC: Employment; Allg. Gl.: allgemeines Gleichbehandlungsgesetz

Example decisions I: Griggs vs. Duke Power 1971

- Paving the way for disparate impact
- Duke Powers' San River Steam Station
- 1950s: blacks can only be in one of four departments (the one with lowest pay)
- After civil rights act of 1964: high school diploma or employment tests to transfer to higher paying departments
- Blacks much less likely to fulfil requirements
- Tests and criteria ruled illegal as tests not directly related to job — employer needs to prove that test is necessary

Example Decisions II: Ricci vs. De Stefano (2009)

- 12 white and one Hispanic firefighter claimed discrimination after they were not promoted although they passed test
- Reason: no black firefighter passed the test and therefore test was considered invalid by employer who was scared of being sued
- Supreme court: decision to ignore test result wrong
- Supreme court: exams were fair and valid

Terminology

In the following slides:

- Y is a binary class to be predicted (recidivism yes/no, at-risk yes/no)
- $\tilde{Y}$ is a (binary) predictor
- A is a protected attribute
- L are legitimate attributes
- S is the predicted score of an algorithm

Disparate impact operationalisations I

Individual Fairness

An algorithm is fair if it gives similar predictions to similar individuals. Needs a similarity metric.

- Dwork et al (2012): *Fairness through awareness* In Proc of the 3rd Innovations in Theoretical Computer Science Conference.
- Zemel et al (2013): *Learning fair representations*. In ICML 2013.

Disparate impact operationalisations II

Group-based, related closely to 80% rule

Demographic/Statistical Parity (equal positive rates)

$$p(\tilde{Y} = 1|A = 0) = p(\tilde{Y} = 1|A = 1)$$

For binary outcomes, this means that the marginal distributions are the same. For non-binary outcomes, we can generalise to all marginal distributions.

Conditional Statistical Parity

$$p(\tilde{Y} = 1|L = 1, A = 0) = p(\tilde{Y} = 1|L = 1, A = 1)$$

- Zafar et al (2016): *Learning fair classifiers.*
- Feldman et al (2015): *Certifying and removing disparate impact.* In Proc. of the SIGKDD Conference on Knowledge Discovery and Data Mining.

Disparate Impact Operationalisations

Group-based:

Equal opportunity (equal true positive rates)

$$p(\tilde{Y} = 1|A = 0, Y = 1) = p(\tilde{Y} = 1|A = 1, Y = 1)$$

Equalized Odds (equal true positive and false positive rates)

$$p(\tilde{Y} = 1|A = 0, Y = y) = p(\tilde{Y} = 1|A = 1, Y = y)$$

Aligns with high accuracy, but wants high accuracy in all demographics

Hardt et al (2016): Equality of opportunity in supervised learning.

Other operationalisations exist

Examples:

- Overall accuracy equality: overall procedure accuracy is the same for each protected group category
- Test fairness (Calibration): A score S is well-calibrated of

$$p(Y = 1|S = s, A = 0) = p(Y = 1|S = s, A = 1)$$

These definitions...

can unfortunately be incompatible i.e. might not be able to be satisfiable in parallel

Problem

Except in trivial cases, it is impossible to maximize accuracy and fairness at the same time, and impossible simultaneously to satisfy all kinds of fairness. Berk et al (2017): Fairness in criminal Justice Risk Assessments: The state of the art.

One reason: different Base Rates

Overview

Case Study: Obermeyer et al 2019

Disparate treatment/impact

Causes of bias

Causes of bias: Label bias

Label Bias Definition

- Choose a proxy label to predict
- This choice leeds to bias
- Other attributes and data representation can be completely fair

See Obermeyer et al (2019).

Related to measurement bias.

Causes: Bias because of attribute/feature choice

- Direct use of protected attribute
- Encoding protected attribute A indirectly:
  - Non-protected attribute B strongly correlated to A: redlining
  - Redundant encoding
- Omitted variable bias: some features that are essential are omitted

Population and Sampling Biases

- Historical bias/negative legacy (algorithm "just depicts" reality)
- Representation and Sampling bias: We just sample from one part of the population more
  - More white faces in our face recognition database
  - ...
- Population bias: User population in dataset differs from original target population
- Subset targeting and definition of protected attribute: Maybe no bias against blacks but only against black women

Caution!

There are many different types of bias. They are also named differently by different researchers.
For one overview see `https://arxiv.org/pdf/1908.09635.pdf`