

Fairness and Bias in Algorithmic Decision-Making

Daniel B. Neill, Ph.D.

Carnegie Mellon University (Heinz College)
and NYU (Center for Urban Science & Progress)

E-mail: neill@cs.cmu.edu / daniel.neill@nyu.edu

Joint work with Zhe Zhang (CMU Heinz College)

Carnegie Mellon University

EPD Lab

EVENT AND PATTERN DETECTION LABORATORY

Why should we care about fairness?

Online algorithms can exacerbate demographic and socioeconomic disparities, e.g., through price discrimination or targeted advertising.

Sensitive decisions at the individual level: school admissions, job applications, loan/credit approval, insurance premiums...

Patient care: access to quality care, medications; prevention; health insurance coverage; appropriately targeted treatments → **outcomes**.

Public health: policy choices can reduce or exacerbate disparities. Food deserts, poverty, environmental risks, pollution, fresh water...

Policing and criminal justice: geographic and demographic biases in targeted patrolling; biases in sentencing/parole/probation decisions.

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and

ASHKAN SOLTANI

December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was \$15.79, while the price on Trude Frizzell's screen, just a few miles away, was \$14.29.

A key difference: where Staples seemed to think they were located.

<http://www.wsj.com/articles/SB10001424127887323777204578189391813881534>

What happened: lower store density in poor & ethnic minority neighborhoods → higher prices → racially disparate impact.

Q: Is this happening in health care as well?
Pharmacies/medications? Insurance companies?



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. Josh Ritchie for ProPublica

*Source:
Julia Angwin,
Jeff Larson,
Surya Mattu and
Lauren Kirchner, ProPublica*

Machine Bias

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

Humans have biases too– what’s your baseline for comparison?

An Analysis of the New York City Police Department’s “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias

Andrew GELMAN, Jeffrey FAGAN, and Alex KISS

Recent studies by police departments and researchers confirm that police stop persons of racial and ethnic minority groups more often than whites relative to their proportions in the population. However, it has been argued that stop rates more accurately reflect rates of crimes committed by each ethnic group, or that stop rates reflect elevated rates in specific social areas, such as neighborhoods or precincts. Most of the research on stop rates and police–citizen interactions has focused on traffic stops, and analyses of pedestrian stops are rare. In this article we analyze data from 125,000 pedestrian stops by the New York Police Department over a 15-month period. We disaggregate stops by police precinct and compare stop rates by racial and ethnic group, controlling for previous race-specific arrest rates. We use hierarchical multilevel models to adjust for precinct-level variability, thus directly addressing the question of geographic heterogeneity that arises in the analysis of pedestrian stops. **We find that persons of African and Hispanic descent were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation.**

KEY WORDS: Criminology; Hierarchical model; Multilevel model; Overdispersed Poisson regression; Police stops; Racial bias.

How can we use machine learning to identify and reduce biases?

How can we avoid introducing new biases, or exacerbating existing biases, when we perform data-driven analyses?

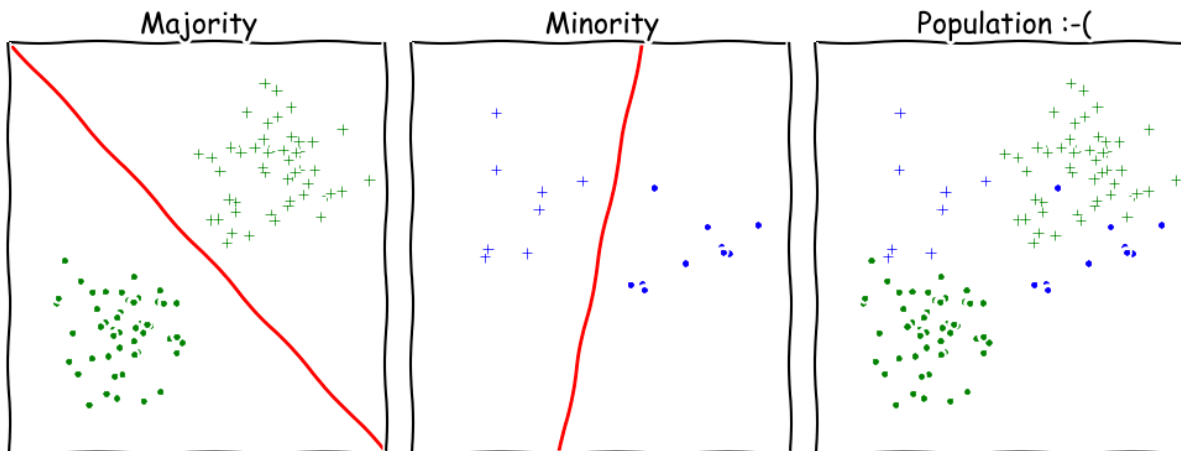
How algorithms (or humans) can discriminate

1. Problem Specification

- Predicting past outcomes can reinforce undesirable status quo in hiring, treatment decisions, etc.
(would we have _____ vs. should we have _____)
- Biased proxies for the target variable of interest.
Predicting re-arrest rather than re-offending; failing to account for differences in medication adherence.

How algorithms (or humans) can discriminate

2. Problems w/ **training data** used to learn predictive models:
- The data generating process itself was inherently discriminatory (e.g., counting cell phones; 311 calls)
 - Imbalanced data → poor models of minority patients; failure to account for heterogeneity in treatment effects.



Note: Biases can be introduced in many other ways as well, for example, model misspecification.

Our work in algorithmic fairness

- Discovering heterogeneous treatment effects, in both experimental and observational data.
 - Re-analysis of Tennessee STAR experiment (class size and educational outcomes): teacher's aide may help for inner city schools when teachers sufficiently experienced.
 - Analyzing Highmark claims data: glucocorticoids lead to poor outcomes in hypertensive, overweight males.
- **Bias scan**: a general approach for **auditing** (and correcting) black-box algorithms for fairness.
 - Case study in criminal justice: is the COMPAS algorithm for predicting re-offending risk **fair**, or is it **biased** against some subpopulation defined by observed characteristics (race, gender, age, etc.)?

Broward County data

- Source: ProPublica's data on criminal defendants in Broward County, FL, in 2013-2014
- Outcome: re-arrests (!) assessed through April 2016.
- Score: **COMPAS** score from 1 (low risk) to 10 (high risk)

Background	Black ($n = 3696$)		White ($n = 2454$)
Age	32.7 (10.9)	<	37.7 (12.8)
Male (%)	82.4	>	76.9
Number of Priors	4.44 (5.58)	>	2.59 (3.8)
Any priors? (%)	76.4	>	65.9
Felony (%)	68.9	>	60.3
COMPAS Score	5.37 (2.83)	>	3.74 (2.6)

What does it mean to be “fair”?

There are at least three possibilities (and probably more):

1) Group Fairness: The same proportion of each group should be classified as “high risk”.

- Makes sense for analyzing discrimination in employment: about the same proportion of each group should be hired.
- Doesn't seem reasonable for COMPAS: observed reoffending rates are not constant across groups. For Broward County, 51% of black defendants and 39% of white defendants reoffended.

2) Disparate Impacts: Comparing false positive and false negative rates across groups.

- Impacts depend on how predictions are used (particularly if the prediction is a probability). Can we separate **fairness of prediction** from **fair decisions** using these predictions?

What does it mean to be “fair”?

There are at least three different ways to think about fairness (see):

1) **Group**
should

3) We focus on **unbiasedness** of probability estimates.

Individual risk probabilities should be predicted accurately, **without systematic biases** based on any observed attributes or combinations of attributes.

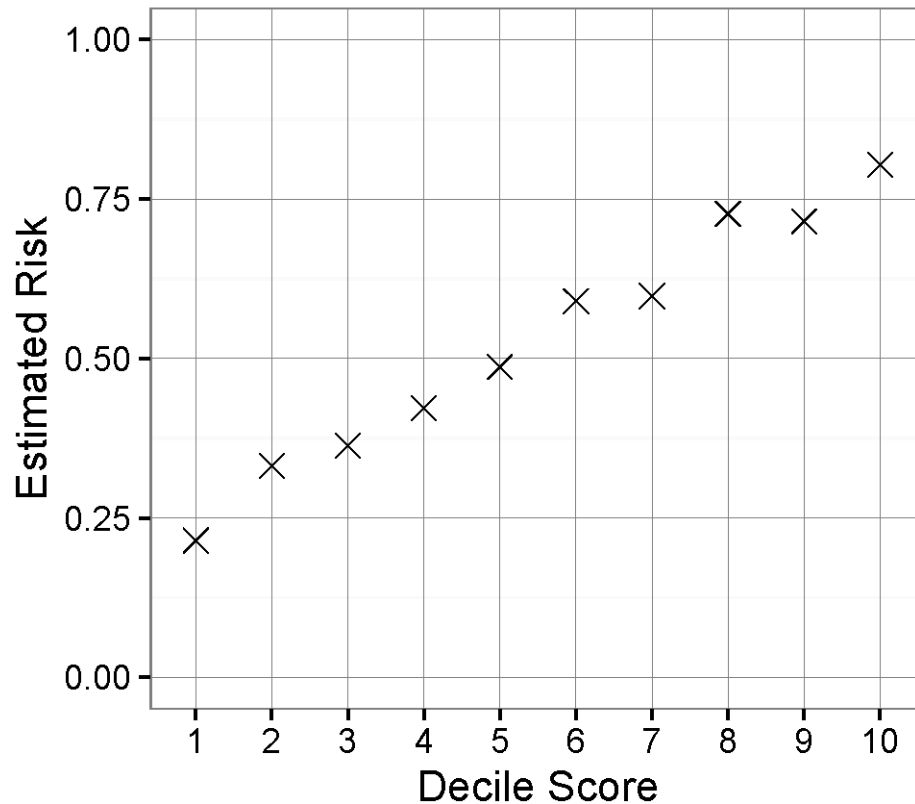
→ Are there any **statistically significant** biases?

→ Can we automatically **correct** these systematic biases, in order to improve fairness of prediction?

2) **Individual** **false**
negative

- Impacts depend on how **predictions** are used (particularly if the prediction is a probability). Can we separate **fairness of prediction** from **fair decisions** using these predictions?

Results of bias scan on COMPAS

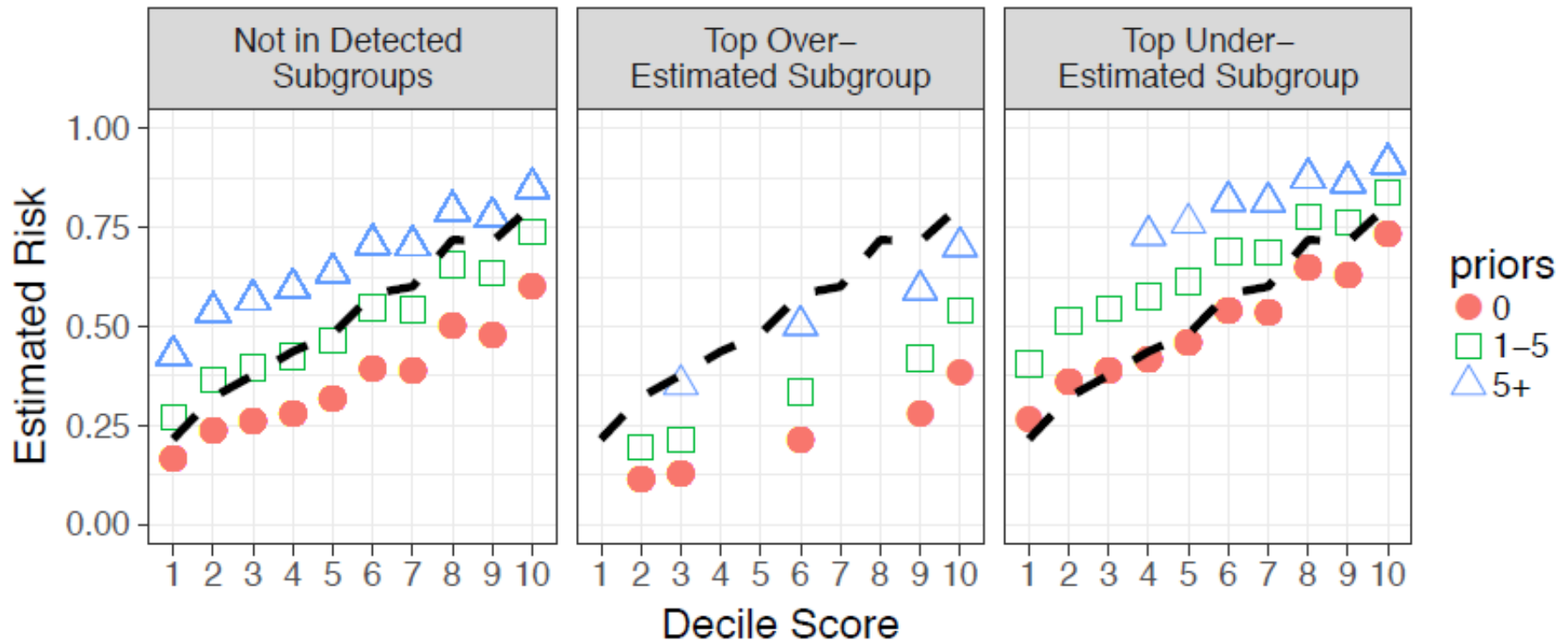


Start with maximum likelihood risk estimates for each COMPAS decile score.

Detection result 1: COMPAS underestimates the importance of prior offenses, overestimating risk for 0 priors, and underestimating risk for 5 or more priors.

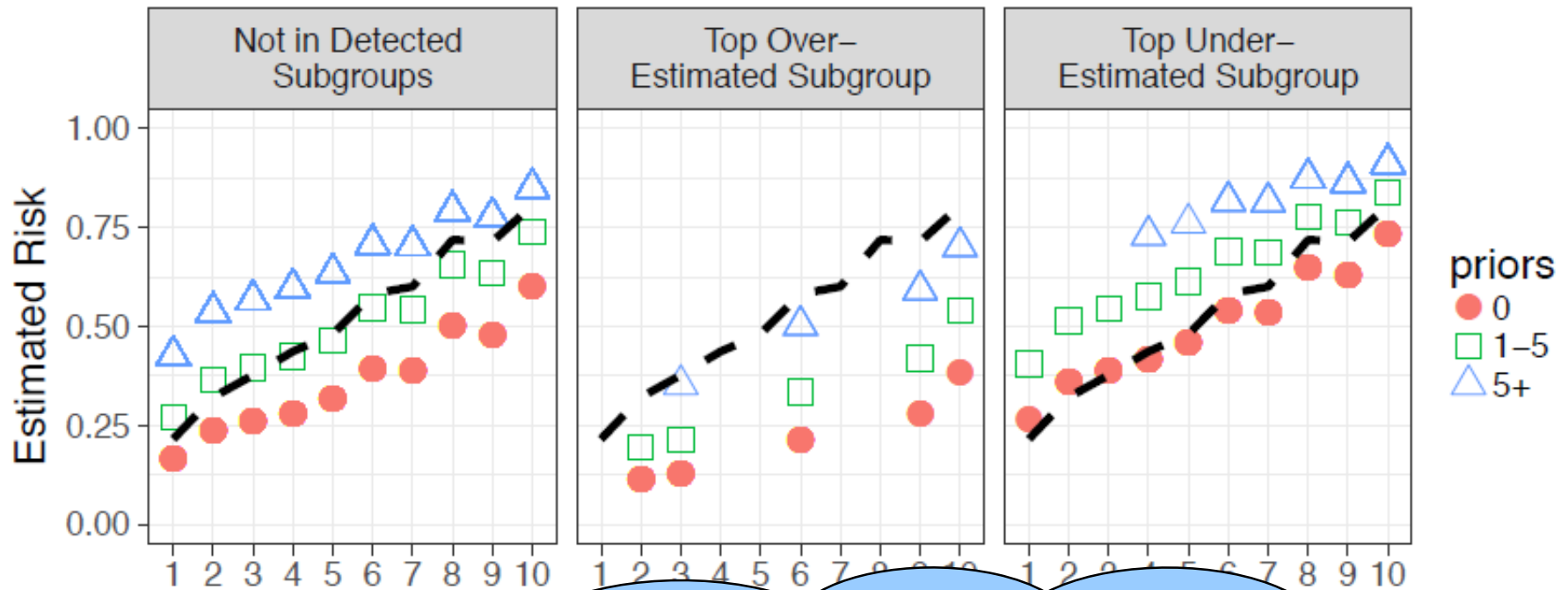
Detection result 2: Even controlling for prior offenses, COMPAS still underestimates risk for males under 25, and overestimates risk for females who committed misdemeanors.

Results of bias scan on COMPAS



After controlling for number of prior offenses and for membership in the two detected subgroups, there are no significant systematic biases in prediction.

Results of bias scan on COMPAS



The resulting probabilistic classifier has greater interpretability (though still based partially on a black box) and is less biased than the original COMPAS predictions... but does this mean it is “fair”?

Discussion: predictive fairness in context


- The method does not account for **target variable bias**: we predict re-offending risk but the gold standard is based on re-arrests not re-offenses.
 - Big problem with drug possession, weapon possession charges. Leads to feedback loops.
- How to avoid **disparate impacts** when making decisions based on even unbiased predictions?
 - Integration with other data sources? Probability matching?
 - Different cutoff thresholds for different groups.
 - Optimize predictive models subject to a constraint on balance (non-discrimination) between majority and minority classes; typically tradeoffs between balance and predictive accuracy.

The bigger picture

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016



PredPol
Predict Crime in **Real Time**

PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.



Chronicle Of Social Change

Chronicle Webpage



California Bets on Big Data to Predict Child Abuse

CADE METZ BUSINESS 07.11.16 7:00 AM

ARTIFICIAL INTELLIGENCE IS SETTING UP THE INTERNET FOR A HUGE CLASH WITH EUROPE



References

- Resources for fairness, accountability, and transparency in ML: <http://www.fatml.org/resources.html>
- A. Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*, 5(2): 153-163, 2017. <http://online.liebertpub.com/doi/full/10.1089/big.2016.0047>
- Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. <https://arxiv.org/pdf/1611.08292.pdf>. In *NIPS Workshop on Interpretable Machine Learning*, 2016.
- A. Romei & S. Ruggieri. A multidisciplinary survey on discrimination analysis. <http://www.di.unipi.it/~ruggieri/Papers/ker.pdf>
- S. Barocas and A.D. Selbst. Big Data's Disparate Impact. In *104 California Law Review* 671, 2016.
- Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4): 1060-1089, 2017. <http://www.zliobaite.com/publications>
- M. Bilal Zafar, et al. Learning Fair Classifiers. Tech. report, 2016. <http://arxiv.org/pdf/1507.05259v3.pdf>
- S. Feldman et al.. Certifying and removing disparate impact. In *Proc. KDD 2015*, http://sorelle.friedler.net/papers/kdd_disparate_impact.pdf