Fairness

Material from Berkeley's CS 294: Fairness in Machine Learning (https://fairmlclass.github.io)
and
N[eur]IPS2017 tutorial
(https://vimeo.com/248490141)
by Solon Barocas (Cornell)
and Moritz Hardt (Berkeley)

Massachusetts Institute of Technology

NASEM Committee on Science, Technology, and Law

March, 2018

Co-Chairs:
David Baltimore, Caltech
David S. Tatel, U.S. Court of Appeals for the District of Columbia Circuit

- Blockchain and Distributed Trust
- **Artificial Intelligence and Decision-Making**
  - **Hank Greely, Stanford**
  - **Cherise Fanno Burdee, Pretrial Justice Institute**
  - **Matthew Lungren, Stanford**
  - **Peter Szolovits, MIT**
  - **Suresh Venkatasubramanian, U. Utah**
- Privacy and Informed Consent in an Era of Big Data
- Science Curriculum for Law School
- Emerging Issues in Science, Technology, and Law
- Using Litigation to Target Scientists
- Communicating Advances in the Life Sciences to a Skeptical Public

Algorithms and Justice

- Government use of decision automation for
    - determining eligibility for services
    - evaluating where to deploy health inspectors and law enforcement personnel
    - defining boundaries around voting districts
- In the law
    - “To the extent they inject clarity and precision into bail, parole, and sentencing decisions, algorithmic technologies may minimize harms that are the products of human judgment.”
    - “Conversely, the use of technology to determine whose liberty is deprived and on what terms raises significant concerns about transparency and interpretability.”

Critique of Bail Algorithms

Instead of leaving cash as collateral for freedom before a trial in court, those accused of crimes in California will be graded by an algorithm, starting in October 2019. A county official will then take that grade and use it to recommend whether the accused should be released or remain in jail.

- "… the machine learning systems used to calculate these risk scores throughout the criminal justice system, have been shown to hold severe racial biases, scoring people of color more likely to commit future crimes."
- "Furthermore, since private companies have been typically contracted to offer these services, the formulas derived by machine learning algorithms to calculate these scores are generally withheld as intellectually property that would tip competitors to the company's technology."
- "… you have data collection that's flawed with a lot of the same biases as the criminal justice system."

https://qz.com/1375820/california-just-replaced-cash-bail-with-algorithms/

COURTS ARE USING AI TO SENTENCE CRIMINALS. THAT MUST STOP NOW.
-- Wired.

- In the case of *Wisconsin v. Loomis*, defendant Eric Loomis was found guilty for his role in a drive-by shooting.
- During intake, Loomis answered a series of questions that were then entered into Compas, a risk-assessment tool developed by a privately held company and used by the Wisconsin Department of Corrections.
- The trial judge gave Loomis a long sentence partially because of the "high risk" score the defendant received from this black box risk-assessment tool.
- Loomis challenged his sentence, because he was not allowed to assess the algorithm.
- Last summer, the state supreme court ruled against Loomis, reasoning that **knowledge of the algorithm's output was a sufficient level of transparency**.

 https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/

HOW ALGORITHMS COULD HELP KEEP PEOPLE OUT OF JAIL. -- WIRED

- "… trying to use data to keep low-level offenders out of jail, figure out who needs psychiatric help, and even set bail and parole. In the same way that law enforcement uses data to deploy resources—so-called predictive policing—cities are using techniques borrowed from public health and machine learning to figure out what to do with people after they get arrested"

https://www.wired.com/2016/11/law-enforcement-mental-health-algorithms/

ScienceFriday

Why Machines Discriminate—and How to Fix Them

- Some believers in big data have claimed that, in big data sets, “the numbers speak for themselves.” Or in other words, the more data available to them, the closer machines can get to achieving objectivity in their decision-making. But data researcher Kate Crawford says that’s not always the case, because big data sets can perpetuate the same biases present in our culture, teaching machines to discriminate when scanning resumes or approving loans, for example.
- And when algorithms do discriminate, computer scientist Suresh Venkatasubramanian says he tends to hear expressions of disbelief, such as, “Algorithms are just code—they only do what you tell them.” But the decisions that machine-learning algorithms spit out are a lot more complicated and opaque than people think, he says, which makes tracking down an offending line of code a near impossibility.

What is Fairness?

- *your ideas…*

Bias, Technically

- {Selection, Sampling, Reporting} bias
- Case of Hypertrophic Cardiomyopathy
  - … risk stratification for hypertrophic cardiomyopathy has been enhanced by targeted genetic testing
  - Multiple patients, all of whom were of African or unspecified ancestry, received positive reports, with variants misclassified as pathogenic on the basis of the understanding at the time of testing.
  - Subsequently, all reported variants were re-categorized as benign.
  - The mutations that were most common in the general population were significantly more common among black Americans than among white Americans ($P<0.001$).
  - Simulations showed that the inclusion of even small numbers of black Americans in control cohorts probably would have prevented these misclassifications.

Bias, Technically

- {Selection, Sampling, Reporting} bias
- Bias of an Estimator
  - Generally, we have bias, variance, and noise
  - O = optimal possible model over all possible learners (model family)
  - L = best model learnable by this learner
  - A = actual model learned
  - Bias = O - L (limitation of learning method or target model)
  - Variance = L - A (error due to sampling of training cases)
    - Estimate significance by comparing against learning from randomly permuted data

- Inductive Bias — assumptions made by the learning algorithm about regularities that allow prediction on unseen cases

adapted from Solon Barocas

Aliferis, C. F., Statnikov, A., & Tsamardinos, I. (2006). Challenges in the analysis of mass-throughput data: a technical commentary from the statistical machine learning perspective. Cancer Informatics, 2, 133–162.

Isn't Discrimination the Very Point of Machine Learning?

- *Unjustified* basis for differentiation
- Practical irrelevance
- Moral irrelevance

- Fairness focuses on ethical concerns

- 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

adapted from Solon Barocas

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* (in some jurisdictions)

adapted from Solon Barocas

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?

adapted from Solon Barocas

Goals of (Anti-)Discrimination Law

- Disparate Treatment
  - Procedural fairness
  - Equality of opportunity
- Disparate Impact
  - Distributive justice
  - Minimize inequality of outcome

Conflict

E.g., affirmative action

- 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

adapted from Solon Barocas

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
- …

adapted from Solon Barocas

https://hrdag.org/2016/10/10/predictive-policing-reinforces-police-bias/

Ongoing Problems

- Limited features
  - Features may be less informative or less reliably collected for certain parts of the population
  - A feature set that supports accurate predictions for the majority group may not for a minority group
  - Different models with the same reported accuracy can have a very different distribution of error across population
- Sample size disparity
- Leakage
  - With rich data, protected class membership will be unavoidably encoded across other features
  - No self-evident way to determine when a relevant attribute is too correlated with proscribed features

adapted from Solon Barocas

Formalizing Fairness Discussion

- Hardt's example: advertising for a software engineer, question of gender bias
- Notation: $\mathbb{P}_a\{E\} = \mathbb{P}\{E \mid A = a\}$

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

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**
  - $Y \perp A \mid R$

adapted from Moritz Hardt

Independence
R ⊥ 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
  - $|P\{R=1 \mid A=a\} - P\{R=1 \mid A=b\}| > \epsilon$
  - $\left|\frac{P\{R=1 \mid A=a\}}{P\{R=1 \mid A=b\}} - 1\right| \geq \epsilon$
  - ε = 0.2 relates to 4/5 rule

adapted from Moritz Hardt

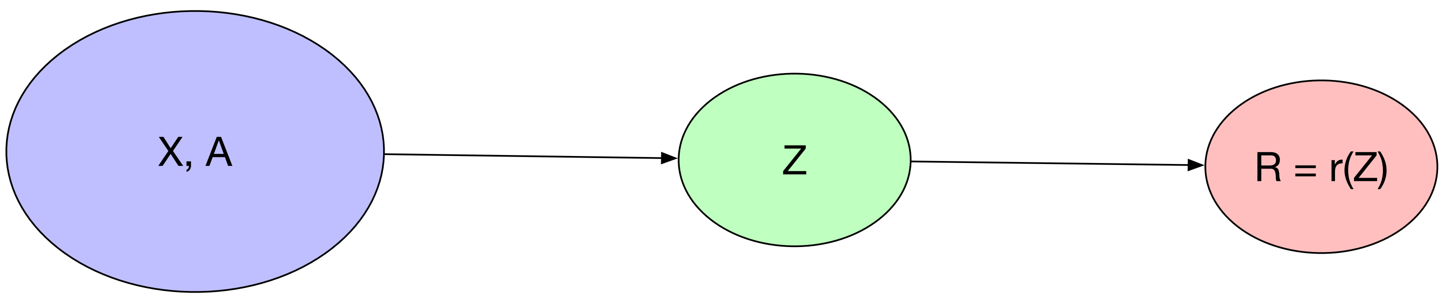
Problems with Independence

- 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?

adapted from Moritz Hardt

Potential Fixes to Achieve Independence

- Pre-processing:
  - Adjust the feature space to be uncorrelated with the sensitive attribute
    - Domain-specific
  - Representation learning



max I(X; Z)
min I(A; Z)

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

- Impose independence constraints at training time (for a given data set)
  E.g., include dependence in the loss function, differential sampling, …

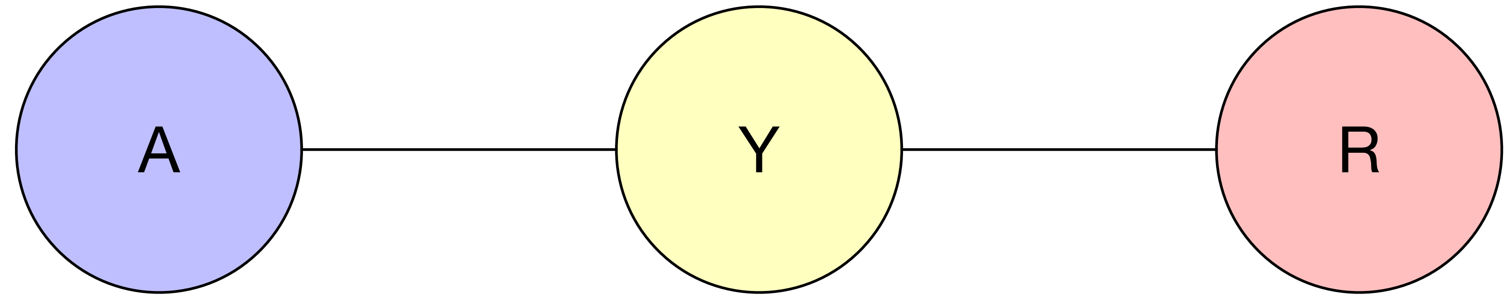
Calders, T., Kamiran, F., & Pechenizkiy, M. (2010). Building Classifiers with Independency Constraints (pp. 13–18). Presented at the 2009 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE. http://doi.org/10.1109/ICDMW.2009.83

- Post-processing
  - Create a new classifier F, $\hat{Y} = F(R, A)$
  - minimize cost of misclassification, perhaps more strongly for protected A

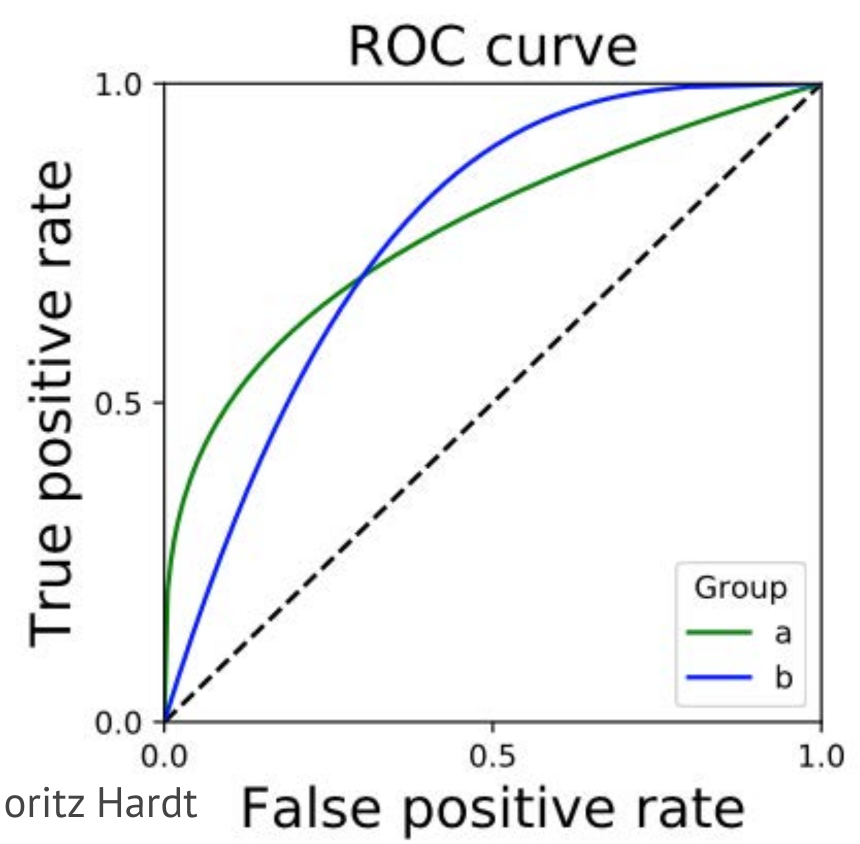
adapted from Moritz Hardt

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. KDD : Proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining, 259–268. http://doi.org/10.1145/2783258.2783311

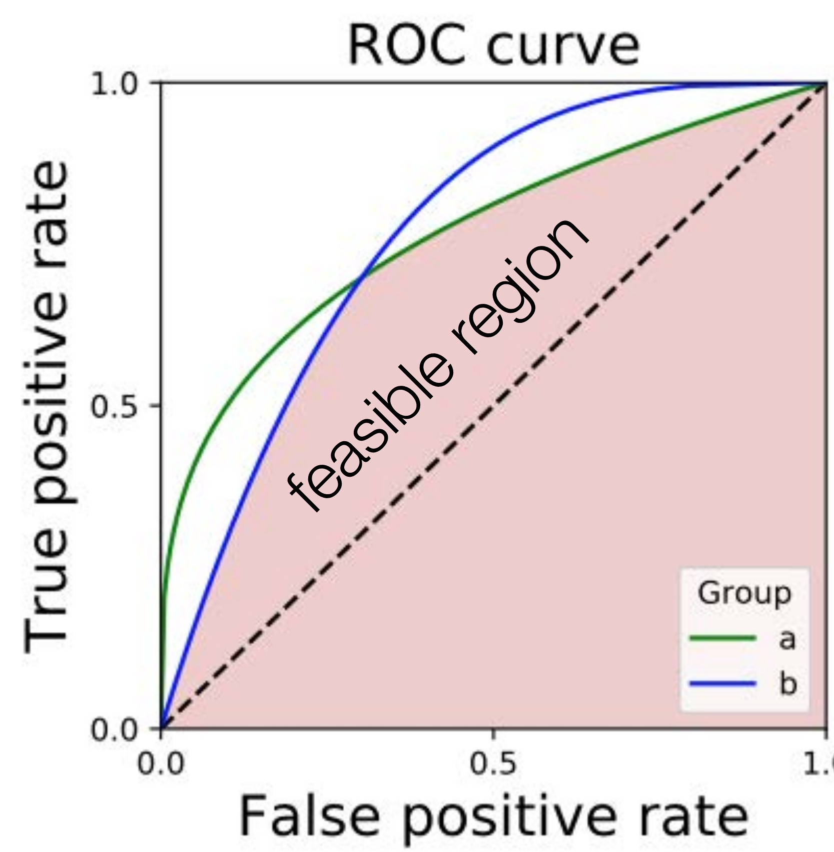
Separation
$R \perp A \mid Y$



- Recognizes that A may be correlated with the target variable
  - E.g., different success rates in a drug trial for different ethnic populations
- $$\begin{aligned} P\{R=1 \mid Y=1, A=a\} &= P\{R=1 \mid Y=1, A=b\} \\ P\{R=1 \mid Y=0, A=a\} &= P\{R=1 \mid Y=0, A=b\} \end{aligned}$$
  - 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



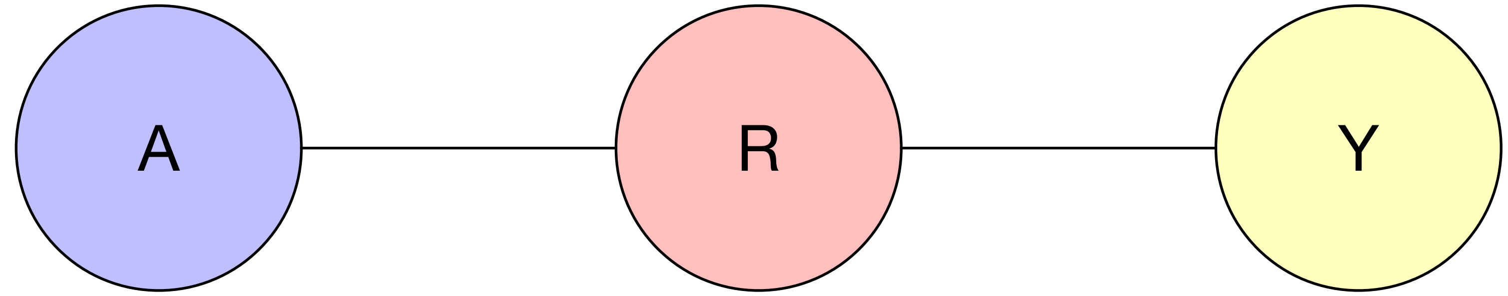
adapted from Moritz Hardt



Advantages of Separation over Independence

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

Sufficiency
$Y \perp A \mid R$

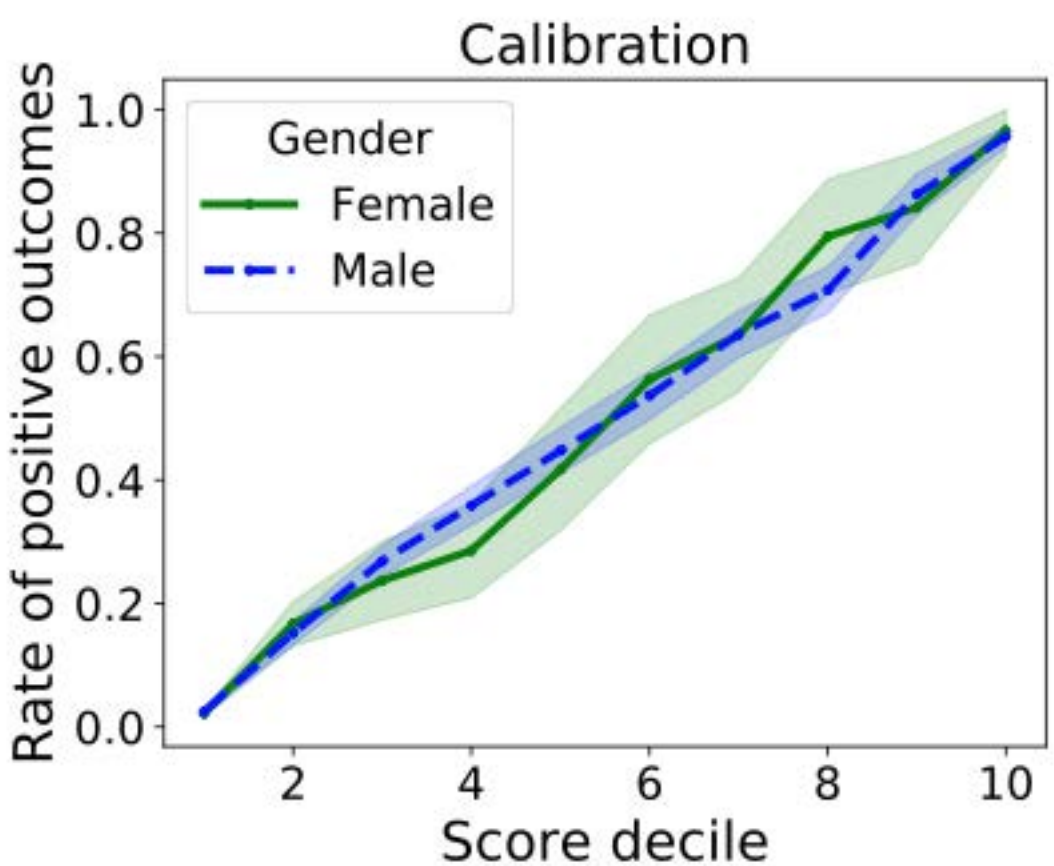


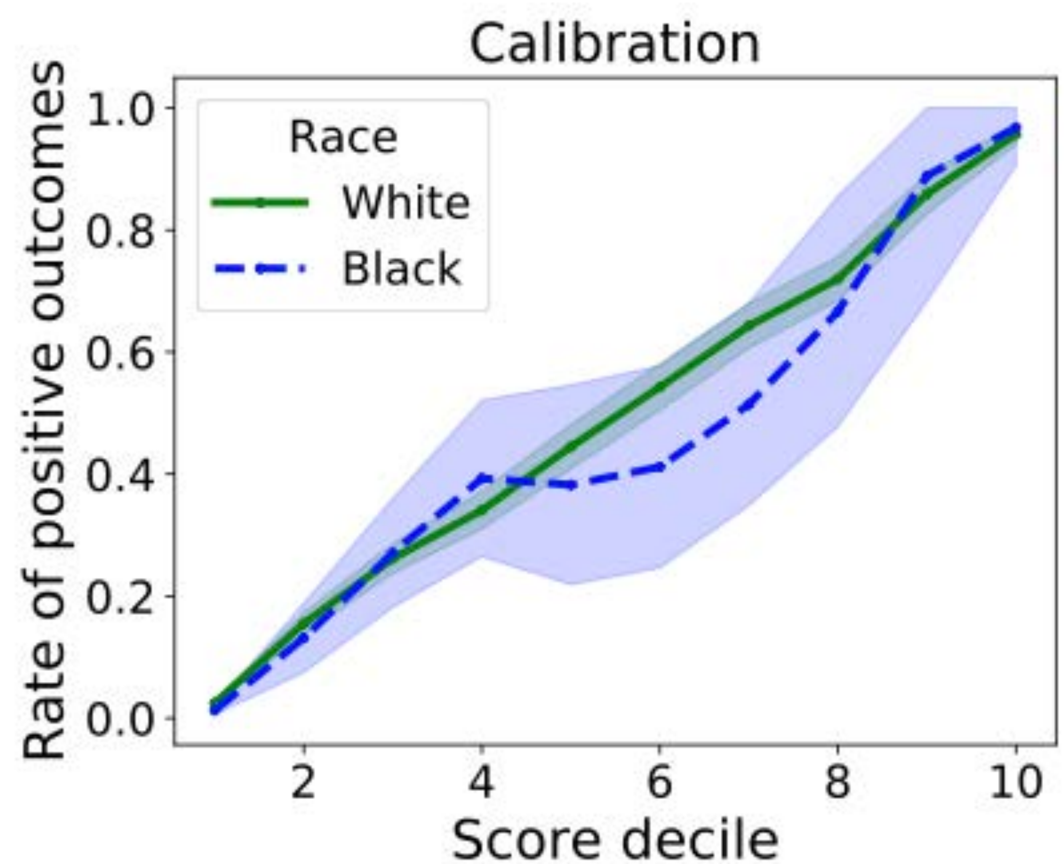
- $P\{Y = 1 \mid R = r, A = a\} = P\{Y = 1 \mid R = r, A = b\}$
- Requires parity of positive and negative predictive values across groups
- R is *calibrated* if $P\{Y = 1 \mid R = r, A = a\} = r$
  - I.e., if the scoring function is a probability of outcome, or
  - “the set of all instances assigned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

adapted from Moritz Hardt

Calibration Can be Good Without Even Trying

- E.g., UCI census data set, predicting income > $50,000/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, …





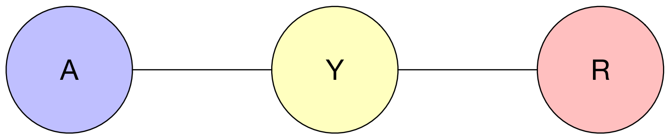
https://fairmlbook.org/introduction.html#demographic-disparities
https://archive.ics.uci.edu/ml/datasets/adult

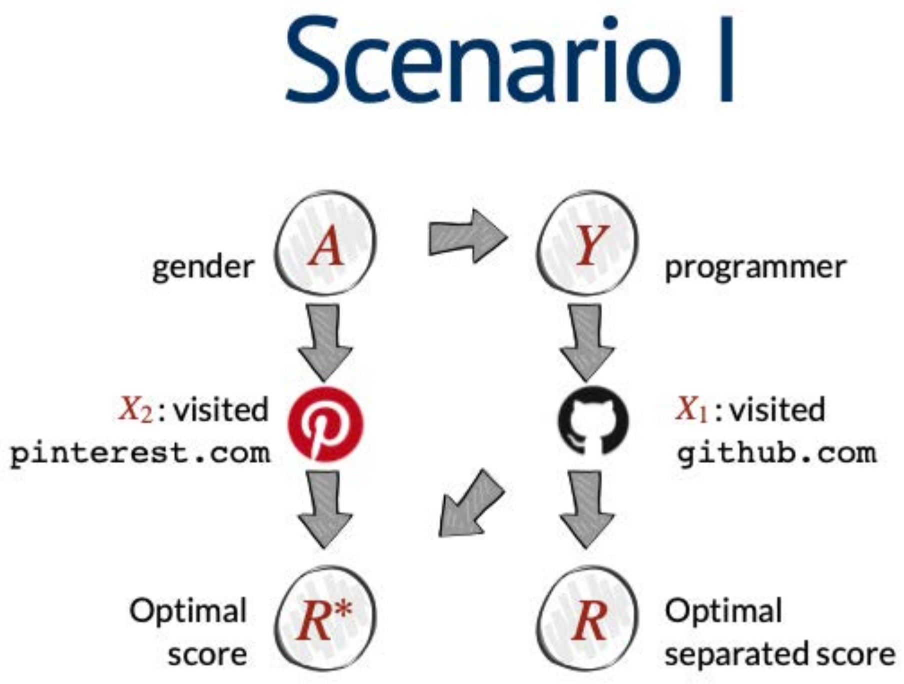
Bad News!

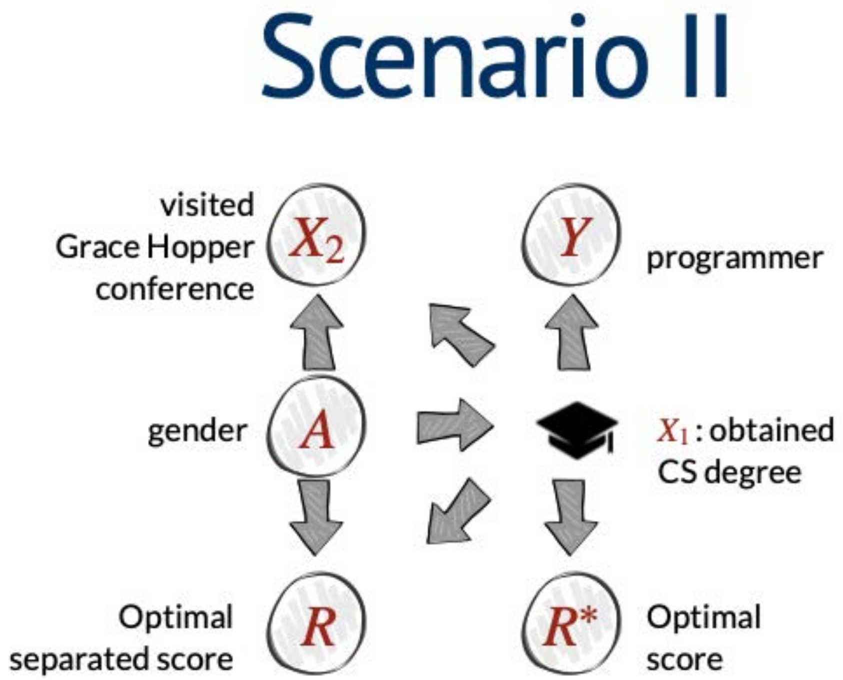
- It is not possible to jointly achieve any pair of these conditions
  - Independence *xor* Separation
  - Independence *xor* Sufficiency
  - Separation *xor* Sufficiency

- Nice illustration at
  - https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Different Scenarios Can Lead to Same Observed Distributions







- 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

Examined Error Rates in Two Data Sets

- Data: de-identified unstructured notes
  - MIMIC-III, predict ICU mortality
  - Psych inpatient data, predict 30-day psych readmission
- Is there bias, based on race, gender, insurance type (as proxy for socio-economic status)?
- Topic modeling on notes: 50 topics

Interpreting Notes by Topic Modeling

Topic Name[a]	Characteristic Words
Cancer	Mass, cancer, metastatic
Heart flow	Afib, atrial, Coumadin®, fibrillation
Kidney	Renal, dialysis, ESRD, line
Orthopedic	Liver, cirrhosis, hepatic, ascites
Pulmonary	COPD, home, BiPAP, chronic
Substance abuse	EtOH, abuse, CIWA, withdrawal
Anxiety	Anxiety, depression, disorder
Bipolar disorder	Bipolar, lithium, manic, episode
Chronic pain	Pain, chronic, mg
Depression	Depression, suicidal, depressive
Psychosis	Psychotic, psychosis, paranoia
Substance abuse	Use, substance, abuse, cocaine

Abbreviations: afib, atrial fibrillation; BiPAP, bilevel positive airway pressure; CIWA, Clinical Institute Withdrawal Assessment; COPD, chronic obstructive pulmonary disease; ESRD, end-stage renal disease; EtOH, ethanol.

[a] Topic name was inferred based on algorithmically found top words.

Psychiatry Results

- Race:
  - White patients had higher topic enrichment values for the **anxiety** and **chronic pain** topics
  - Black, Hispanic, and Asian patients had higher topic enrichment values for the **psychosis** topic
- Gender:
  - Male patients had higher topic enrichment values for **substance abuse** (0.024 v 0.015)
  - Female patients had higher topic enrichment values for **general depression** (0.021 v 0.019) and **treatment resistant depression** (0.025 v 0.015)
- Insurance:
  - private insurance patients have higher topic enrichment values than public insurance patients for **anxiety** (0.029 v 0.0156) and **general depression** (0.026 v 0.017)
  - public insurance patients have higher topic enrichment values for **substance abuse** (0.022 v 0.016)

ICU Results

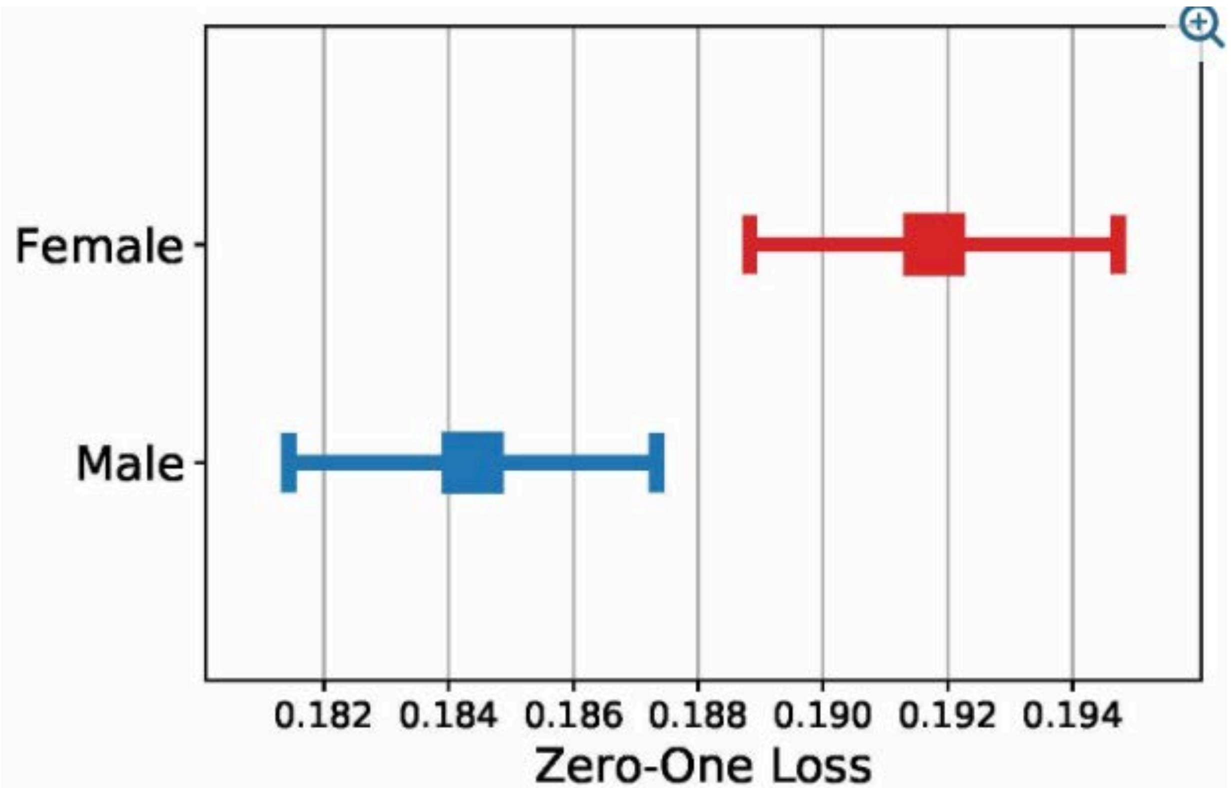
- Gender:
    - male patients have higher topic enrichment values for **substance use** (0.027 v 0.011)
    - female patients have higher topic enrichment values for **pulmonary disease** (0.026 v 0.016), potentially reflecting known underdiagnosis of chronic obstructive pulmonary disease in women
- Race:
    - Asian patients have the highest topic enrichment values for **cancer** (0.036), followed by white patients (0.021), other patients (0.016), and black and Hispanic patients (0.015)
    - Black patients have the highest topic enrichment values for **kidney problems** (0.061), followed by Hispanic patients (0.027), Asian patients (0.022), white patients (0.015), and other patients (0.014)
    - Hispanic patients have the highest topic enrichment values for **liver concerns** (0.034), followed by other patients (0.024), Asian patients (0.023), white patients (0.019), and black patients (0.014)
    - White patients have the highest topic enrichment values for **atrial fibrillation** (0.022), followed by other patients (0.017), Asian patients (0.015), black patients (0.013), and Hispanic patients (0.011)

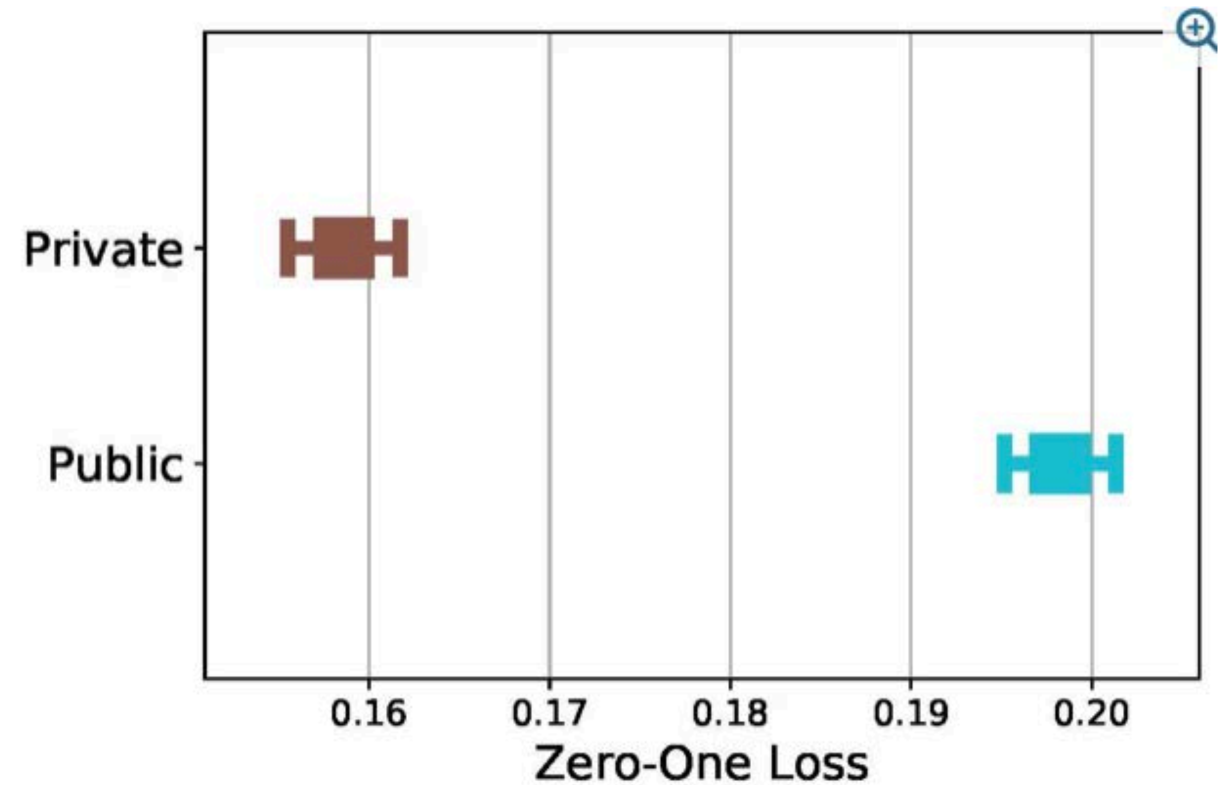
ICU Results, continued

- Insurance:
  - Those with public insurance often have **multiple chronic conditions** that require regular care
  - Public insurance patients have higher topic enrichment values for **atrial fibrillation** (0.24 v 0.013), **pacemakers** (0.023 v 0.014), and **dialysis** (0.023 v 0.013)
  - private insurance patients have higher topic enrichment values for **fractures** (0.035 v 0.012), **lymphoma** (0.030 v 0.015), and **aneurysms** (0.028 v 0.016)

- These results are consistent with known disparities from literature

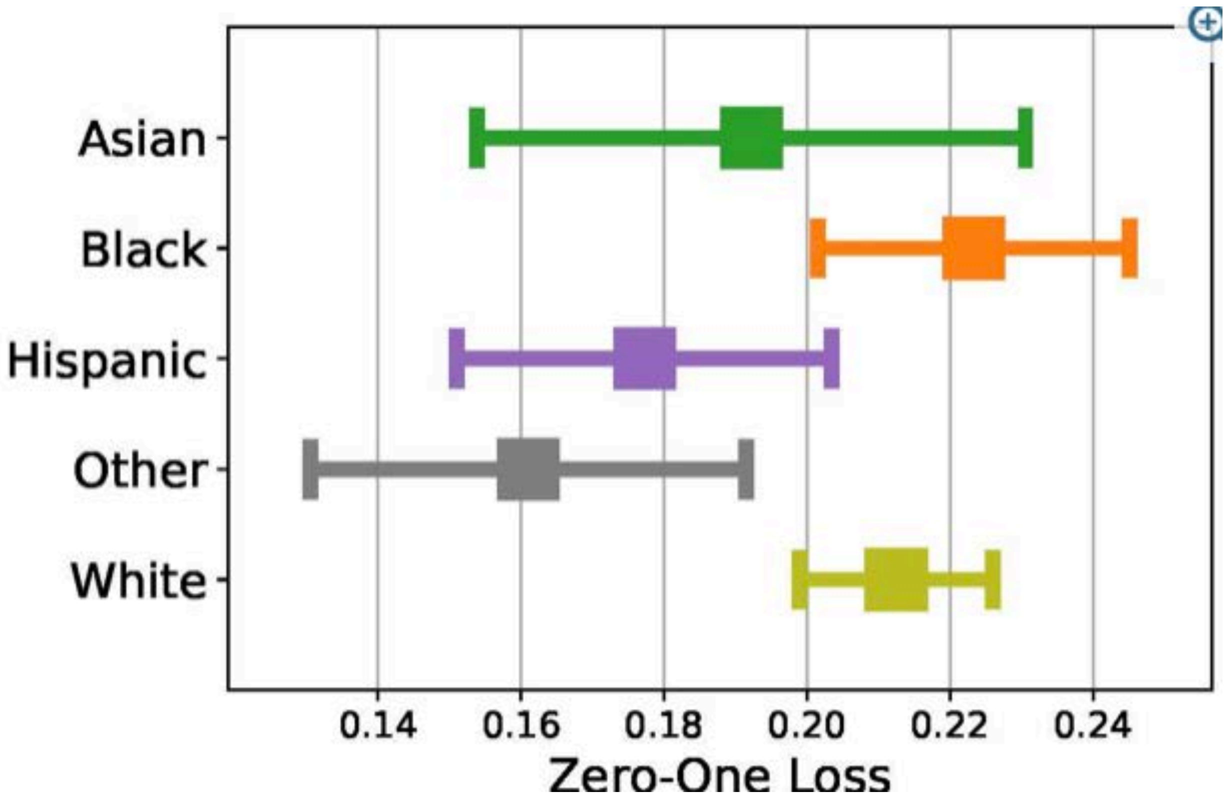
Prediction Errors in ICU (violation of Separation)

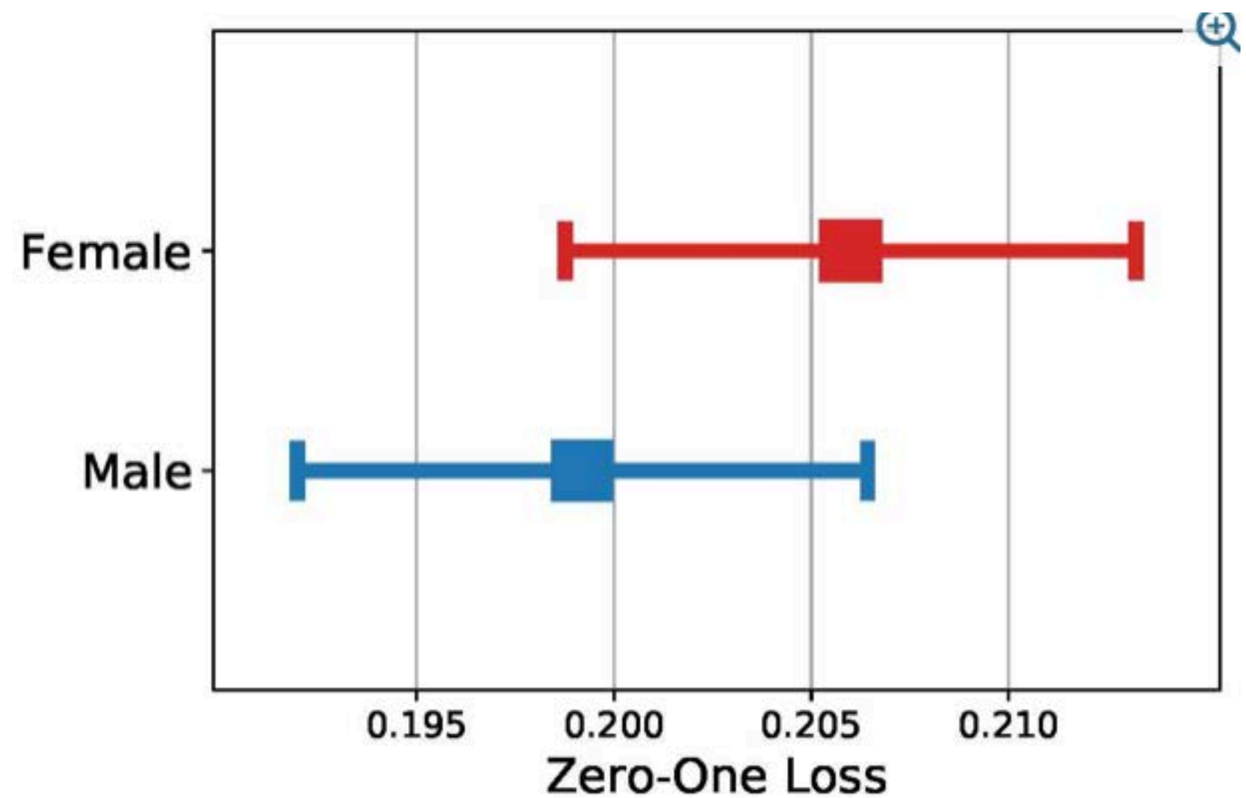




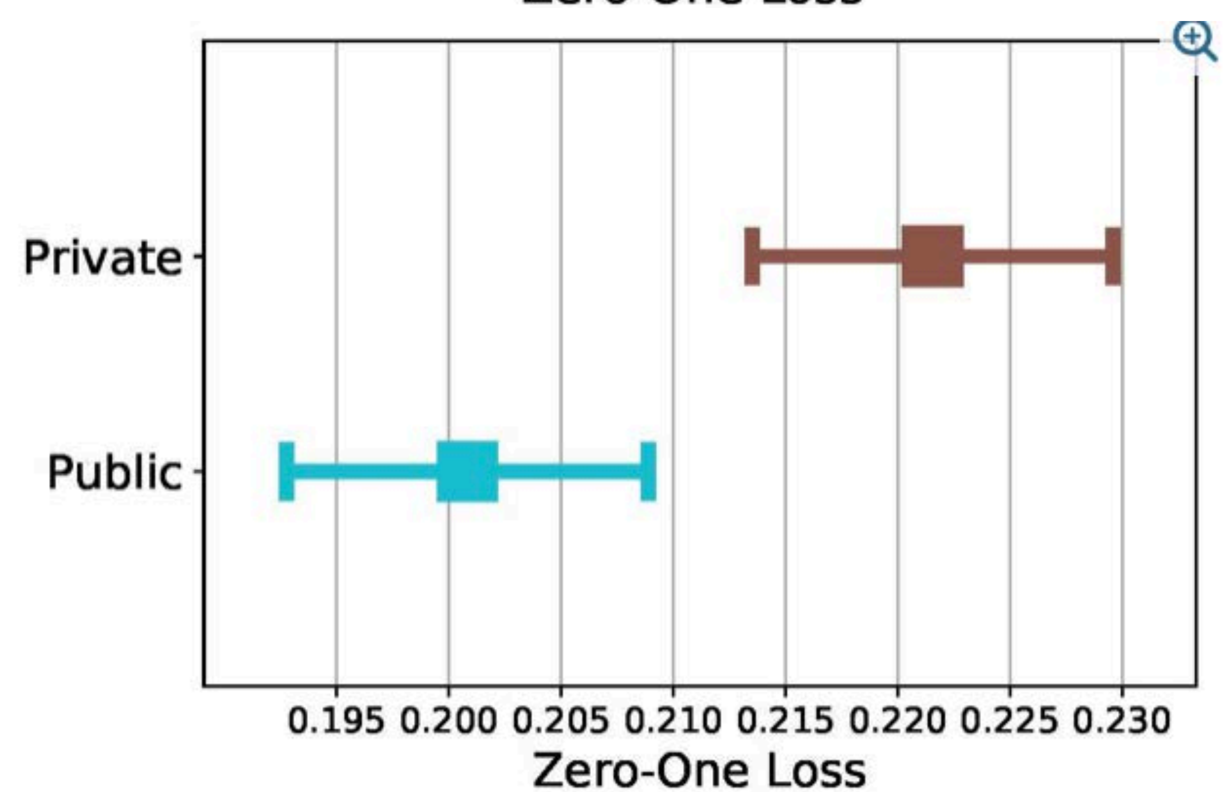
- 95% confidence intervals for zero-one loss differences across gender and insurance type

Prediction Errors in Psychiatry (violation of Separation)





- 95% confidence intervals for zero-one loss differences across race, gender and insurance type



Modeling Mistrust in End-of-Life Care

Willie Boag[1] Harini Suresh[1] Leo Anthony Celi[1] Peter Szolovits[1] Marzyeh Ghassemi[1 2 3 4]

Based on Boag, W. (2018, June). Quantifying Racial Disparities in End-of-Life Care. Master's Thesis, MIT EECS. Cambridge, MA.

- Replicate in MIMIC Racial Disparities expectation from previous studies
- Model Mistrust Algorithmically
- Compare Racial and Mistrust Disparities

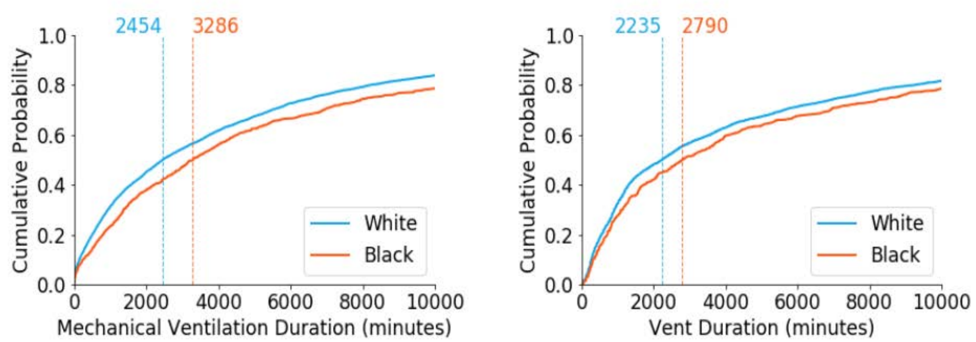
slide from Willie Boag

,
W Boag, H Suresh, L Celi, P Szolovits, and M Ghassemi (2018). Racial Disparities and Mistrust in End-of-Life Care, In Proceedings of the 3rd Machine Learning for Healthcare Conference, PMLR, Palo Alto, California, 85, pages 587-602, 17--18 Aug, 2018. URL: http://proceedings.mlr.press/v85/boag18a.html.

Racial Disparities in End-of-Life Care

African American patients receive longer durations of aggressive treatment during end-of-life care

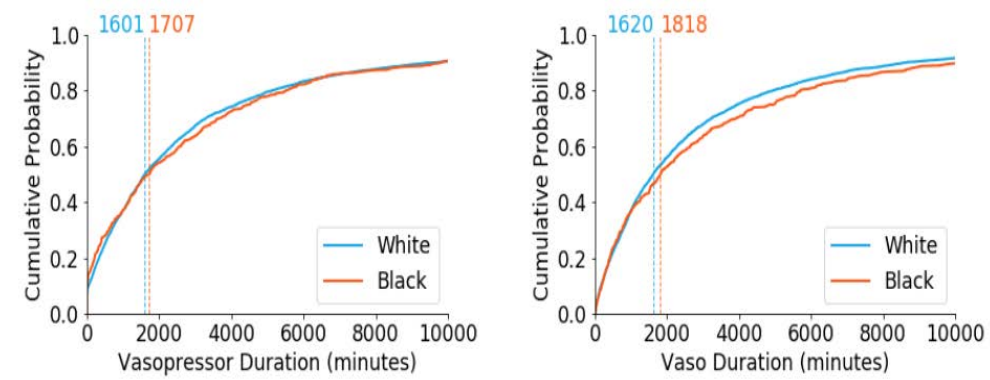
Figure 3-1: **Mechanical Ventilation:** CDF of ventilation duration by race, where dotted lines represent the median duration treatment for a population. In multiple datasets, the median black patient receives statistically significant longer ventilation durations than the median white patient.



(a) *MIMIC Mechanical Ventilation*
White: 4810 patients
Black: 510 patients
p=0.005

(b) *eICU Mechanical Ventilation*
White: 4911 patients
Black: 655 patients
$p < 0.001$

Figure 3-2: **Vasopressors:** In both datasets, the median black patient receives a longer duration of vasopressors than the median white patient. This trend is not statistically significant in either dataset..



(a) *MIMIC Vasopressors*
White: 4458 patients
Black: 453 patients
p=0.122

(b) *eICU Vasopressors*
White: 3479 patients
Black: 464 patients
p=0.422

Could this be the result of mistrust?

(e.g. If your doctor recommends hospice, do you accept their advice?)

slide from Willie Boag

Clues of Mistrust

Noncompliance in Clinical Notes

Social: Pt refused to sign ICU consent and expressed wishes to be DNR/DNI, seemingly very frustrated and mistrusting of healthcare system in relation to [redacted]. Also, w/ hx of poor medication compliance and follow-up

Autopsy Rates

Table 4.3: Autopsy rates by race in MIMIC III.

population	consent	decline	% consent
Asian	2	23	8.0%
White	161	505	24.2%
Other	56	102	32.9%
Black	32	51	38.6%
Hispanic	9	11	45.0%
ALL	260	692	27.3%

Problem: Not every patient has an “obvious” label.

Can we use the obvious examples as labels and train a model to interpolate every patient’s “mistrust” score onto the scale?

slide from Willie Boag

Chart Events Give Clues About Patient State Relevant to True

Table 4.1: Coded interpersonal feature types from chartevents.

1:1 sitter present?	baseline pain level (0 to 10)	received bath?	bedside observer
behavioral intervent	currently experiencing pain	disease state	consults
education barrier	education learner	education method	feamily meeting?
education readiness	harm by partner?	education topic	judgement
follows commands?	family communication method	gcs - verbal response	informed?
hair washed?	goal richmond-ras scale	headache?	health care proxy?
pain management	non-violent restraints?	orientation	pain (0 to 10)
pain assess method	understand & agree with plan?	pain level acceptable?	reason for restraint
restraint device	richmond-ras scale (-5 to +4)	rsbi deferred	riker-sas scale
safety measures	violent restraints ordered?	security	security guard
side rails	status and comfort	sitter	skin care?
spiritual support	behavior during application	support systems	stress
verbal response	teaching directed toward	wrist restraints?	social work consult?

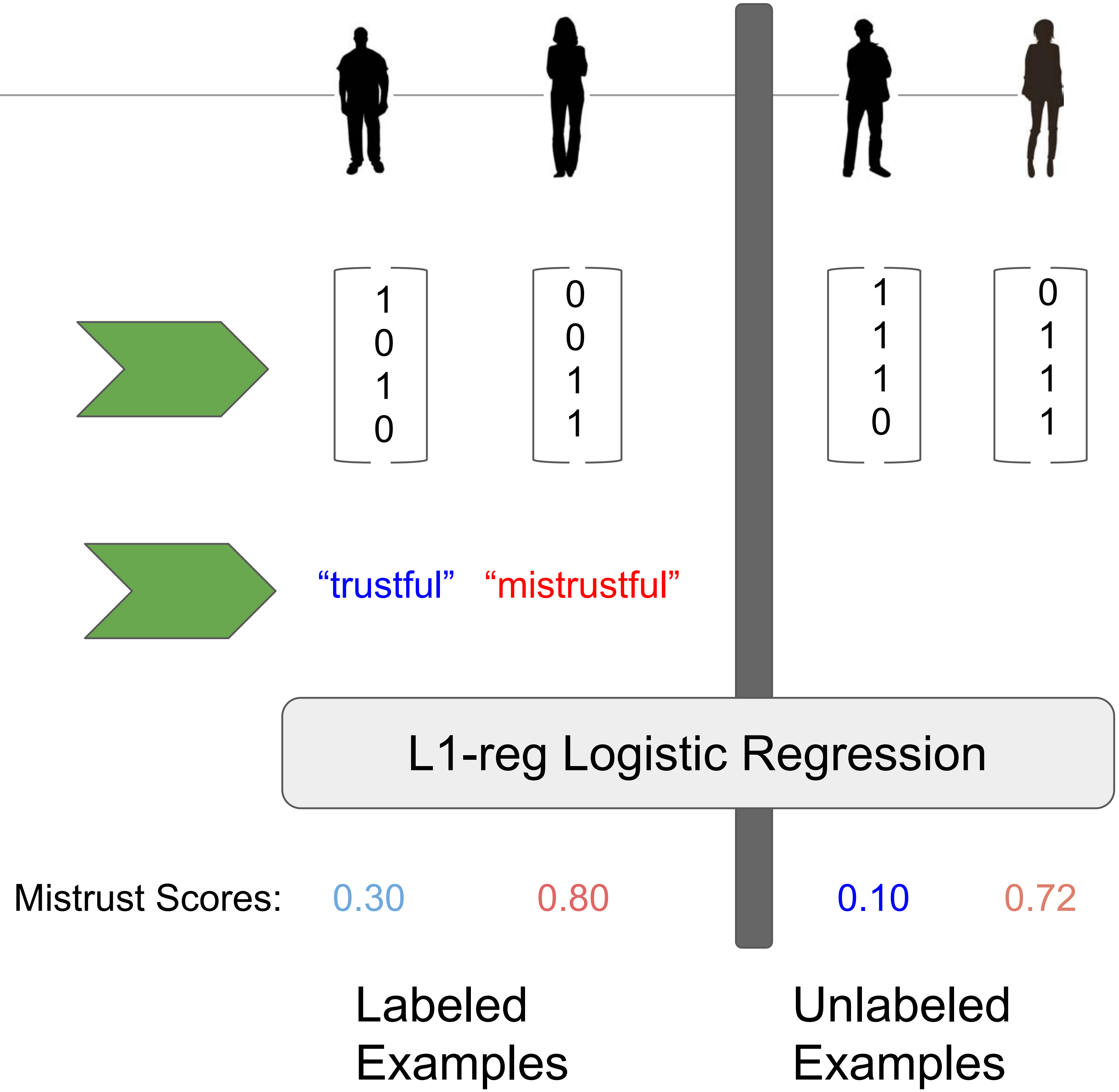
Structured data in the EHR documenting interpersonal variables, including:

- Is the patient's comfort being taken seriously?
- Is the patient being treated as a threat?
- Is the patient's pain being managed?
- Are there good communication between staff and the family?

slide from Willie Boag

Modeling Mistrust

620 binary indicators of trust
indication of family meetings
patient education engagement
patient needed to be restrained
pain is being monitored and treate
healthcare literacy
has a healthcare proxy
has a support system (such as family, social workers, and religion)
agitation scales (Riker-SAS and Richmond-RAS)



Inspecting the Mistrust Metrics

Mistrustful patients: Agitated & in pain
Trustful patients: No pain & calm

Feature	Weight
state: alert	-1.0156
riker-sas scale: agitated	0.7013
pain: none	-0.5427
richmond-ras scale: 0 alert and calm	-0.3598
education readiness: no	0.2540
pain level: 7-mod to severe	0.2168

(a) Noncompliance-derived Mistrust

Mistrustful patients: Restrained
Trustful patients: No pain & healthcare literacy

Feature	Weight
pain present: no	-0.2689
spokesperson is healthcare proxy	-0.2271
family communication: talked to m.d.	-0.1184
reapplied restraints	0.1153
restraint type: soft limb	0.0980
orientation: oriented 3x	0.0363

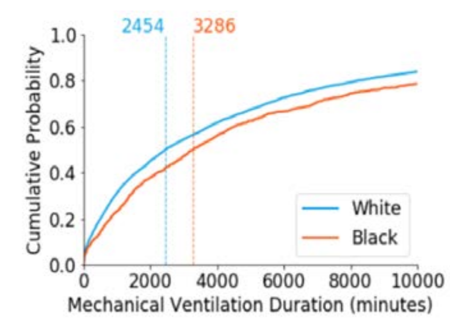
(b) Autopsy-derived Mistrust

slide from Willie Boag

Treatment Disparities are much larger across trust cohorts than race.

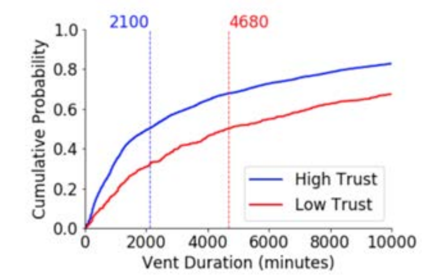
	Race-based Disparity	Trust-based Disparity
Mechanical ventilation	see below	see below
Vasopressors	see below	see below

Mechanical ventilation — Race-based Disparity



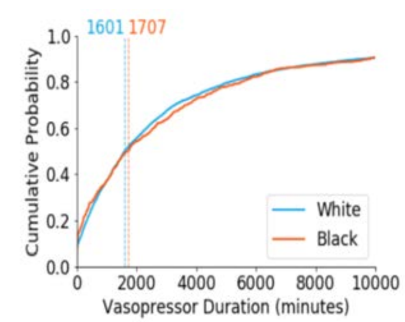
(a) *MIMIC Mechanical Ventilation*
White: 4810 patients
Black: 510 patients
p=0.005

Mechanical ventilation — Trust-based Disparity



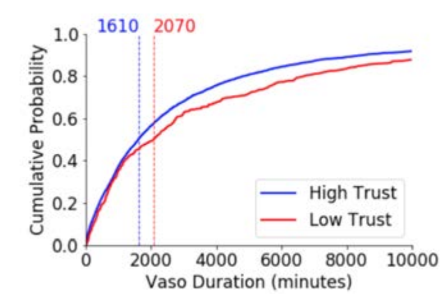
(a) **Mechanical Ventilation**
High Trust: 4810 patients
Low Trust: 510 patients
$p < 0.001$

Vasopressors — Race-based Disparity



(a) *MIMIC Vasopressors*
White: 4458 patients
Black: 453 patients
p=0.122

Vasopressors — Trust-based Disparity



(b) **Vasopressors**
High Trust: 4456 patients
Low Trust: 453 patients
p=0.001

slide from Willie Boag

Population Mistrust

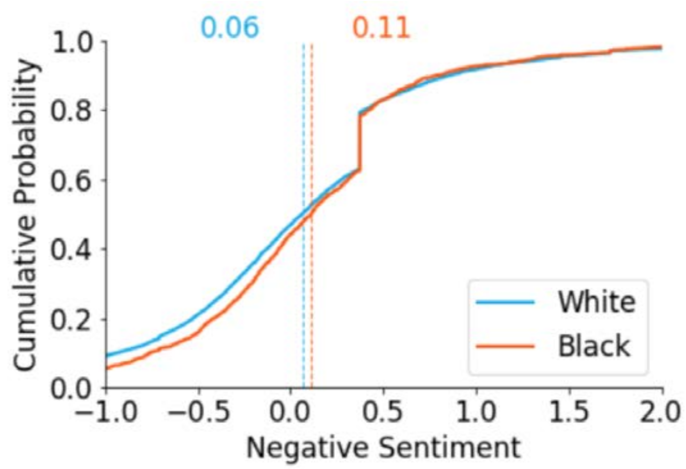


Figure 4-4: Racial disparity in (negative) sentiment.
White: 9669 patients
Black: 1173 patients
p=0.007

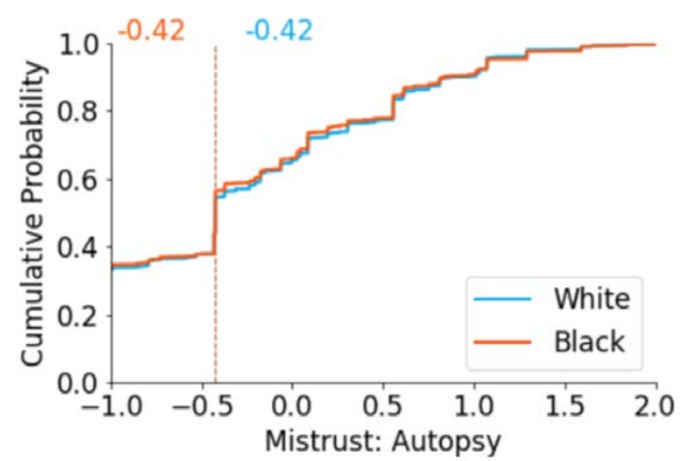


Figure 4-3: Racial disparity in autopsy-derived mistrust metric.
White: 9923 patients
Black: 1202 patients
p=0.126

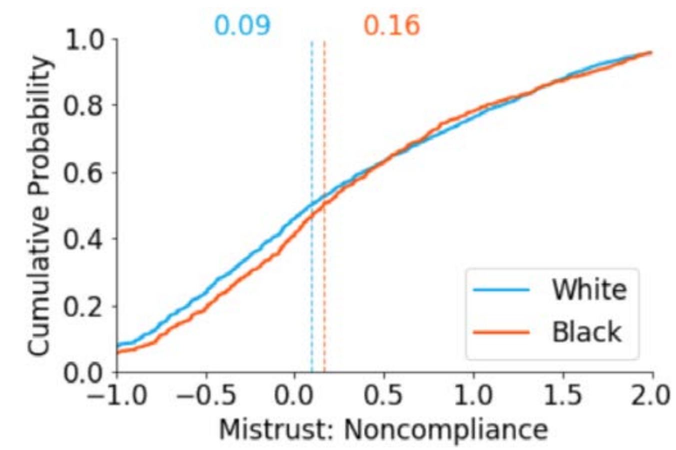


Figure 4-2: Racial disparity in noncompliance-derived mistrust metric.
White: 9923 patients
Black: 1202 patients
$p < 0.001$

For 2/3 metrics, the median black patient has a statistically significantly higher mistrust score than the median white patient.

slide from Willie Boag

Much Work and Education to be Done

- Conferences and Workshops
  - Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop
  - ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)
  - Machine Learning for Healthcare Conference (MLHC)
  - ACM CHI Conference on Human Factors in Computing Systems (CHI)
- Popular Press
- Classes
  - Berkeley CS 294: Fairness in Machine Learning
  - U. Penn CIS 399 The Science of Data Ethics