Office of Health Equity

Lauren Korshak, DHealth(c), MS, RCEP
Translation Lead
Office of Health Equity
Lauren.Korshak@va.gov

OFFICE OF HEALTH EQUITY

Created in 2012

Vision: To ensure that VHA provides appropriate individualized health care to each Veteran in a way that-

- Eliminates disparate health outcomes and
- Assures health equity

OFFICE OF HEALTH EQUITY TEAM

https://www.va.gov/healthequity

OUR PRESENTERS

Suzanne Tamang, PhD, MS, is a Research Associate with the Veterans Health Administration, VA Palo Alto. She is also an Instructor in the Department of Biomedical Data Science, at Stanford University School of Medicine; a Research Economist, with the National Bureau of Economic Review; an Intramural Investigator at the National Institutes of Health and the Assistant Faculty Director of Data Science, Stanford Center for Population Health Sciences.

OUR PRESENTERS

Amol Navathe, MD, PhD, is a staff physician and core investigator at the Center for Health Equity and Research Promotion at the Corporal Michael J. Crescenz VA Medical Center in Philadelphia. He is an Assistant Professor of Health Policy and Medicine and a Senior Fellow at the Leonard Davis Institute for Health Economics at the University of Pennsylvania. Dr. Navathe is also the Co-Director of the Health Transformation Institute and the Director of the Payment Insights Team at the University of Pennsylvania.

Ravi B. Parikh, MD, MPP, is a Staff Physician at the Corporal Michael J. Crescenz VA Medical Center and an Assistant Professor in the Department of Medical Ethics and Health Policy and Medicine at the University of Pennsylvania. Dr. Parikh is a practicing oncologist with expertise in delivery system reform and informatics.

Racial and Ethnic Bias in Real World Risk Prediction Models

Suzanne Tamang, PHD

Program Evaluation Resource Center, Office of Mental Health and Suicide Prevention, Department of Veterans Affairs

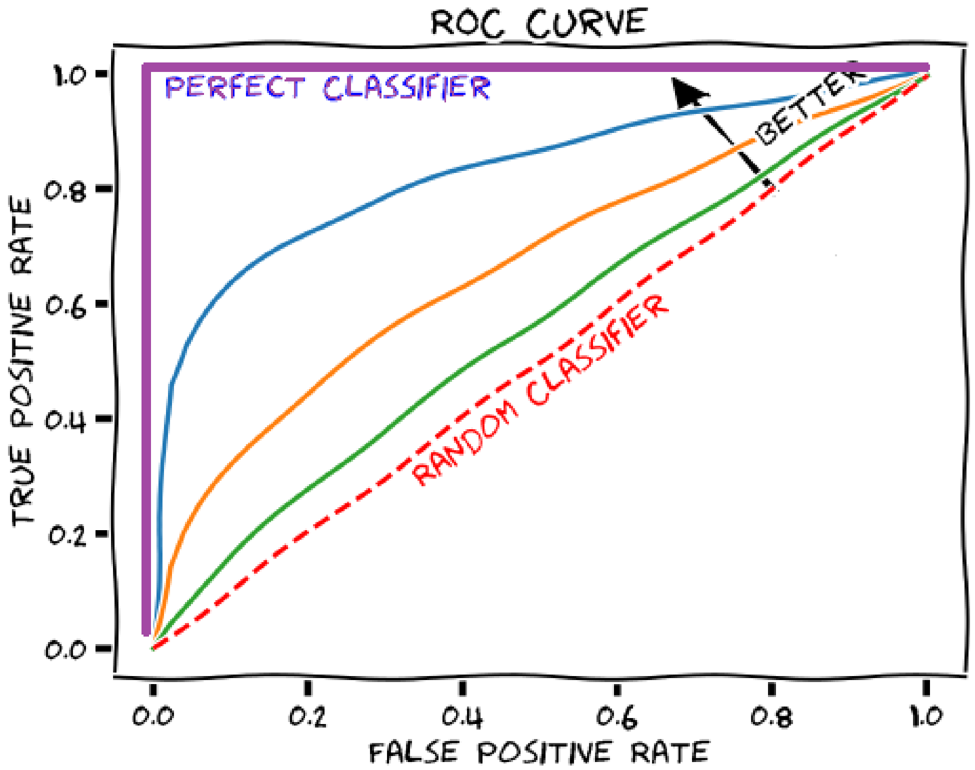
Instructor, Department of Biomedical Data Science, Stanford University

Assistant Faculty Director, Data Science, Stanford Center for Population Health Science

Algorithmic Bias: *Should we be concerned?*

M1: Receiver Operating Characteristic Curve

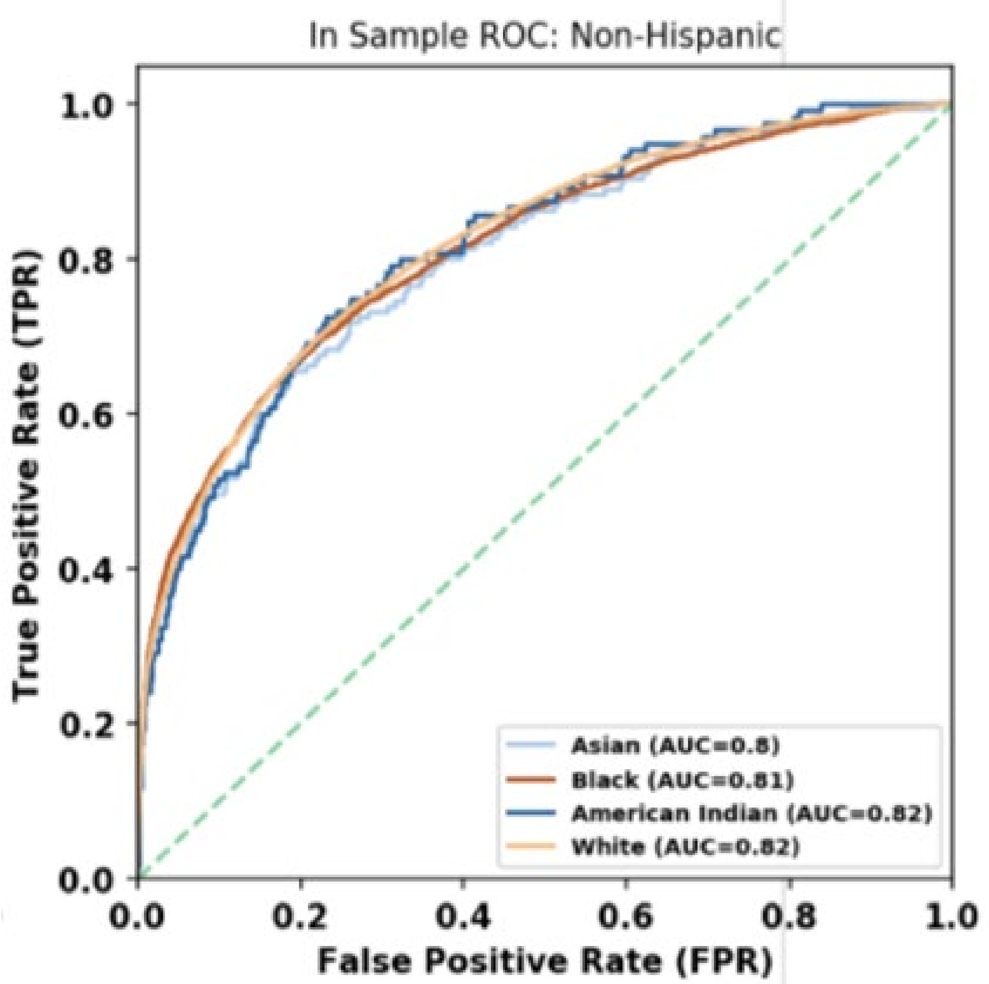
The ROC curve is created by plotting the **true positive rate** (**TPR**) against the **false positive rate** (**FPR**) at various threshold settings.



- TPR, aka **sensitivity, recall** or ***probability of detection*** in machine learning.
- FPR, aka ***probability of false alarm***, can be calculated as (1 – specificity).
- The ROC curve is the sensitivity or recall as a function of fall-out.

Example ROC Curve & AU-ROC: Race x Ethnicity

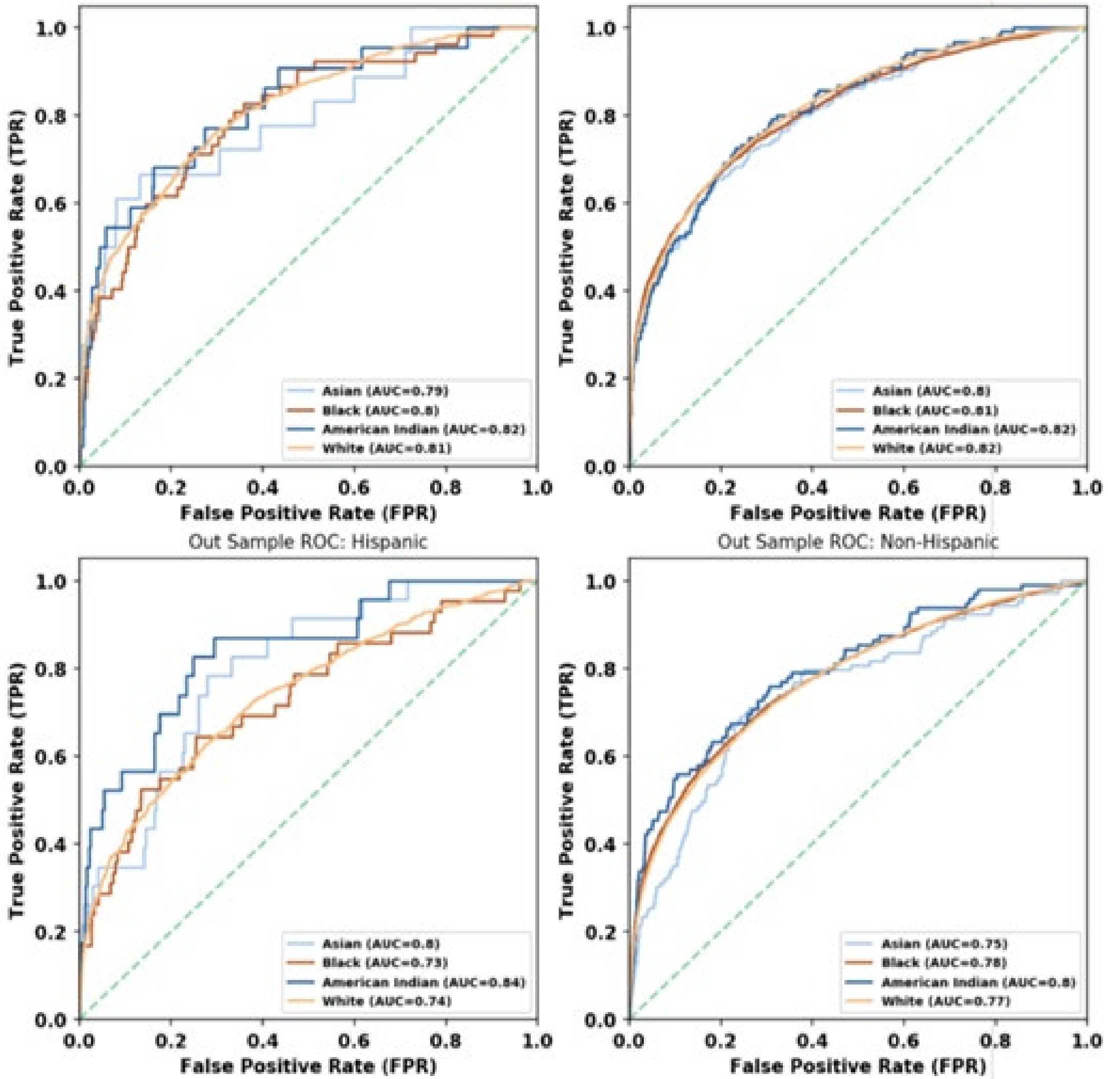
2016-2017



Hispanic | Non-Hispanic

In Sample

Out Sample



M2: Precision Recall Curve

Precision Recall curves are created by plotting the **<u>Precision</u>**, also known as the ***positive predictive value,*** and **<u>Recall</u>**, TPR. Recall is more commonly called ***sensitivity*** in medicine and HSR and is the probability the model will predict all positive cases for the outcome.

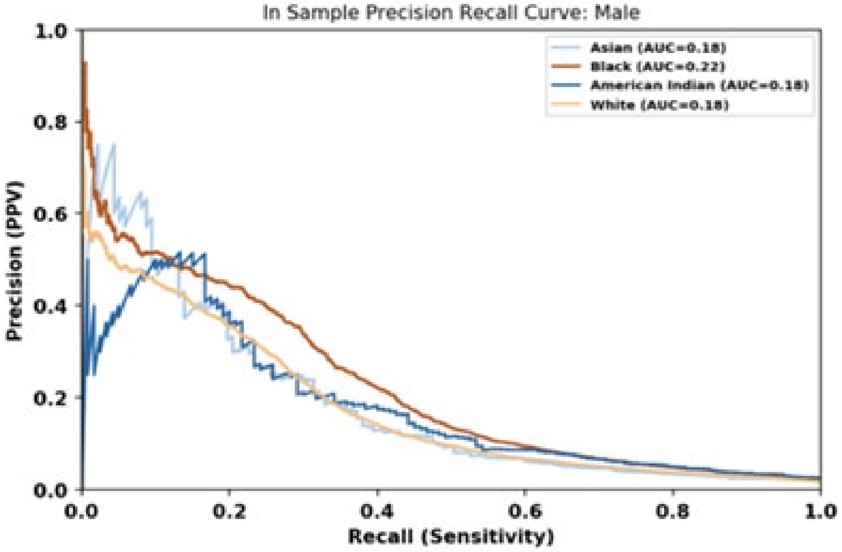
In contrast to the ROC curves and ROC-AUC statistics, the Precision-Recall Curve and the PR-AUC performance metric provide *more information on prediction scenarios that involve rare binary events*.

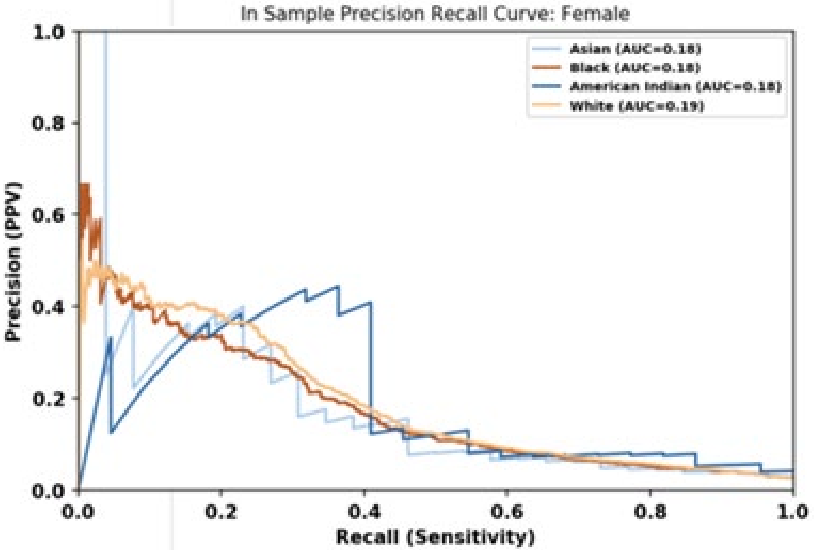
Precision Recall Curve: Race x Sex

Male

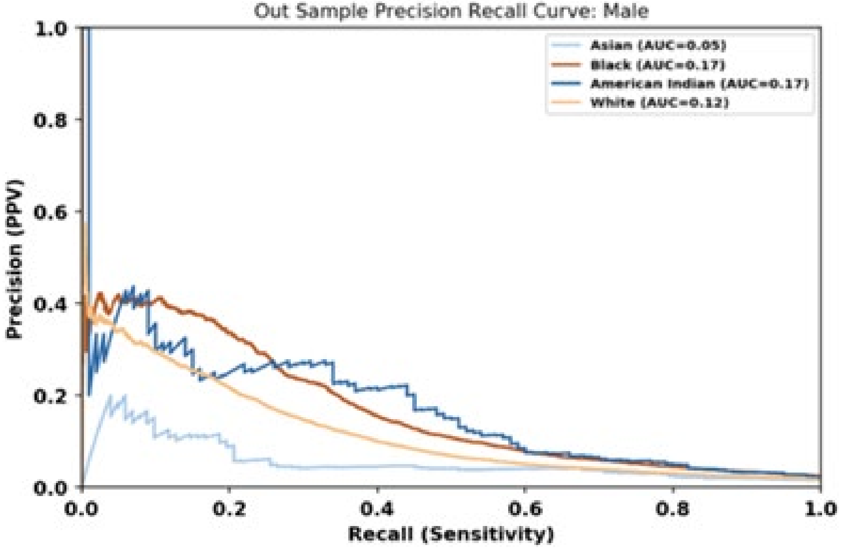
Female

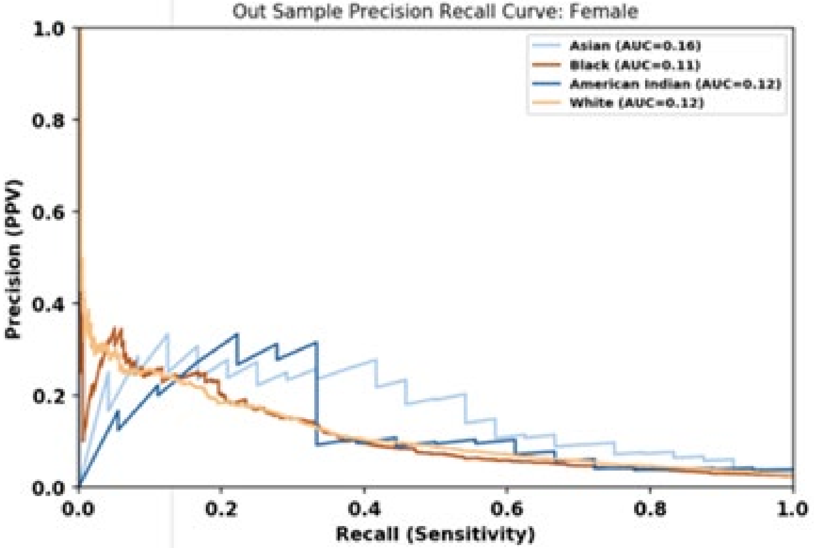
In Sample





Out Sample



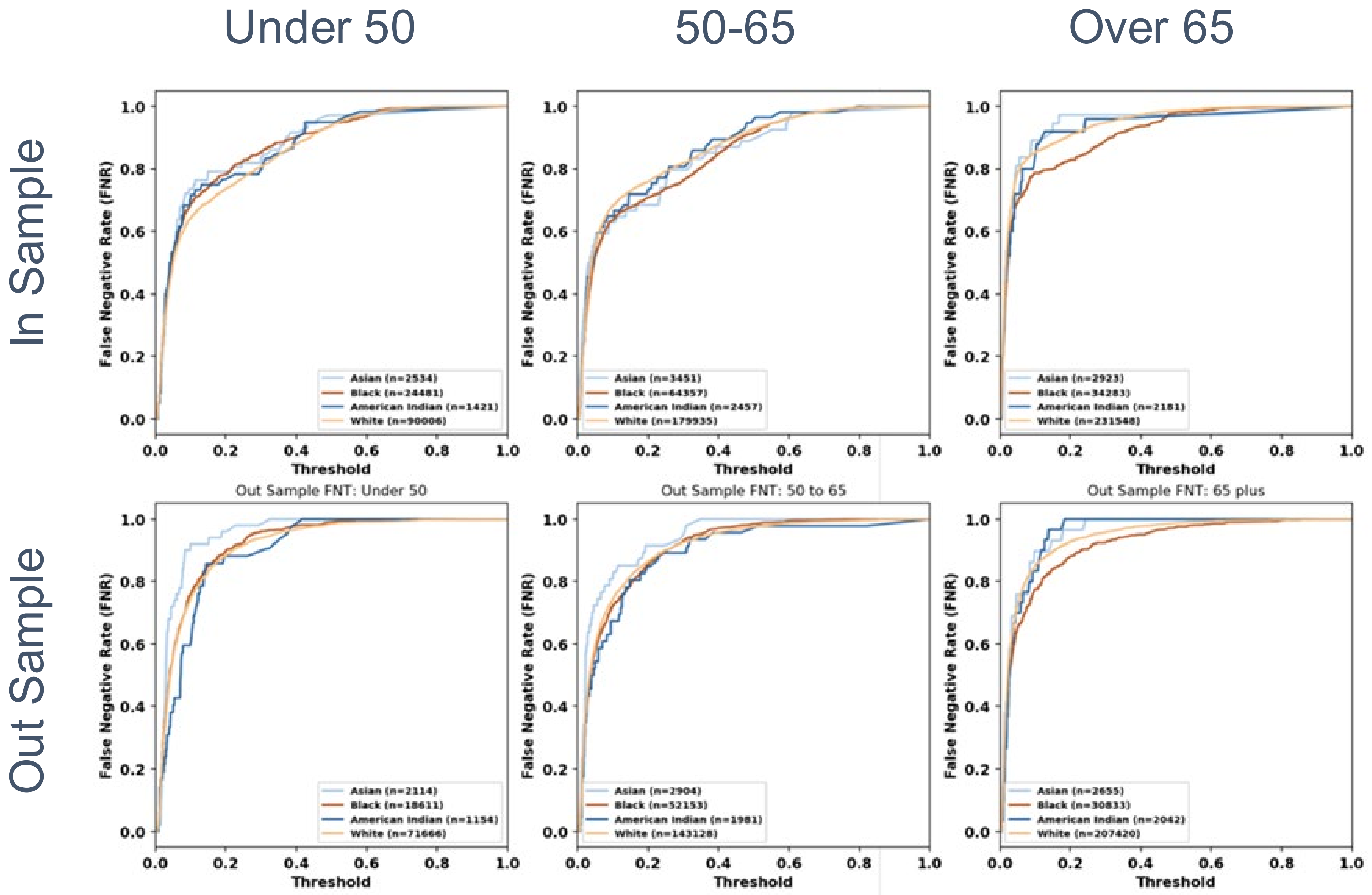


M3: False Negative Parity

The ***false-negative rate*** represents the percentage of true positives missed by the prediction model.

False-negative parity describes the closeness of the **FPR** (false positives/true positives) across different subgroups of interest. It is a commonly reported in algorithmic bias analyses.

False Negative Parity: Race x Age



M4: Calibration

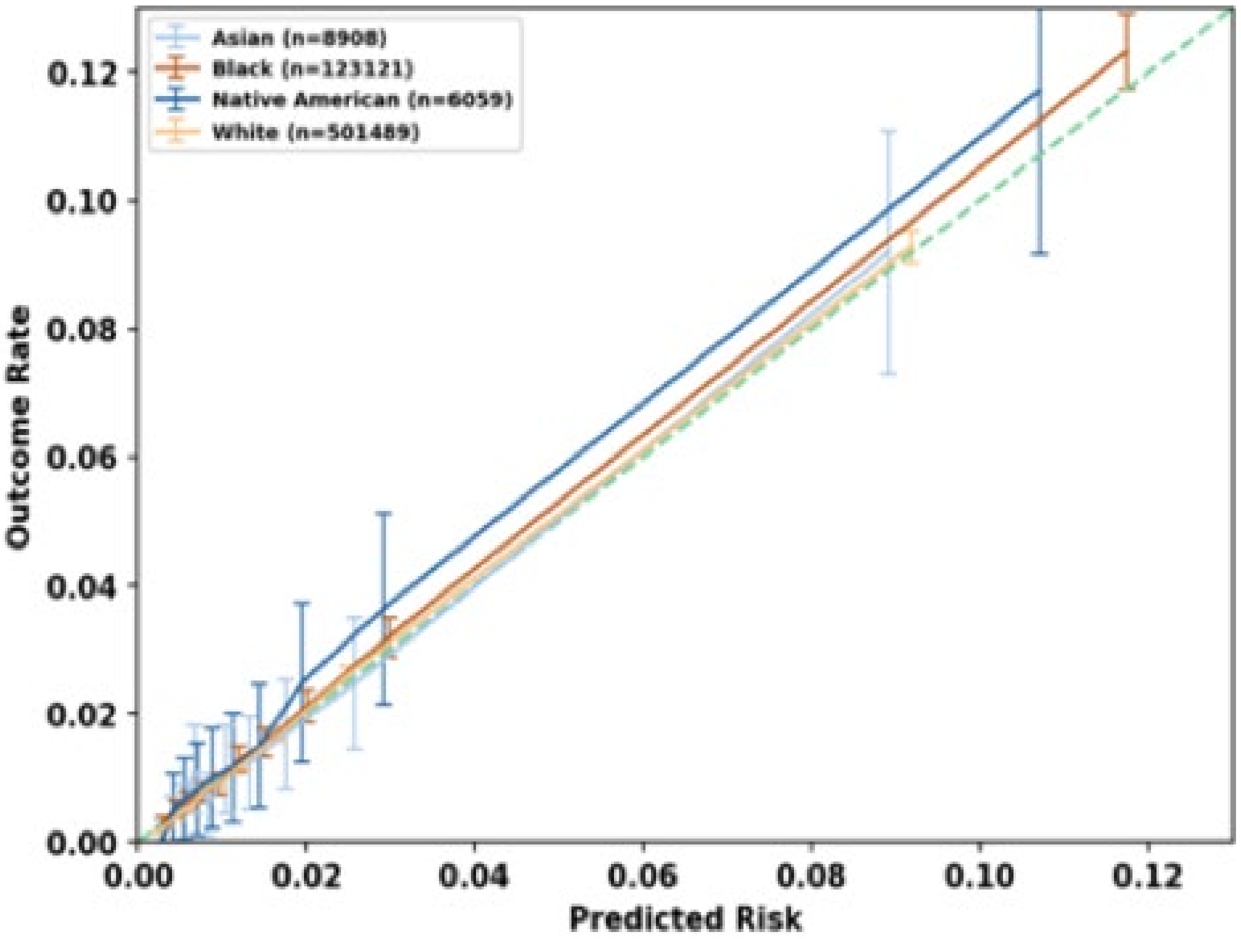
Calibration is defined as the following property:

> *“ If we assign some group a risk of **x**, the actual outcome incidence rate should also be **x** ”*

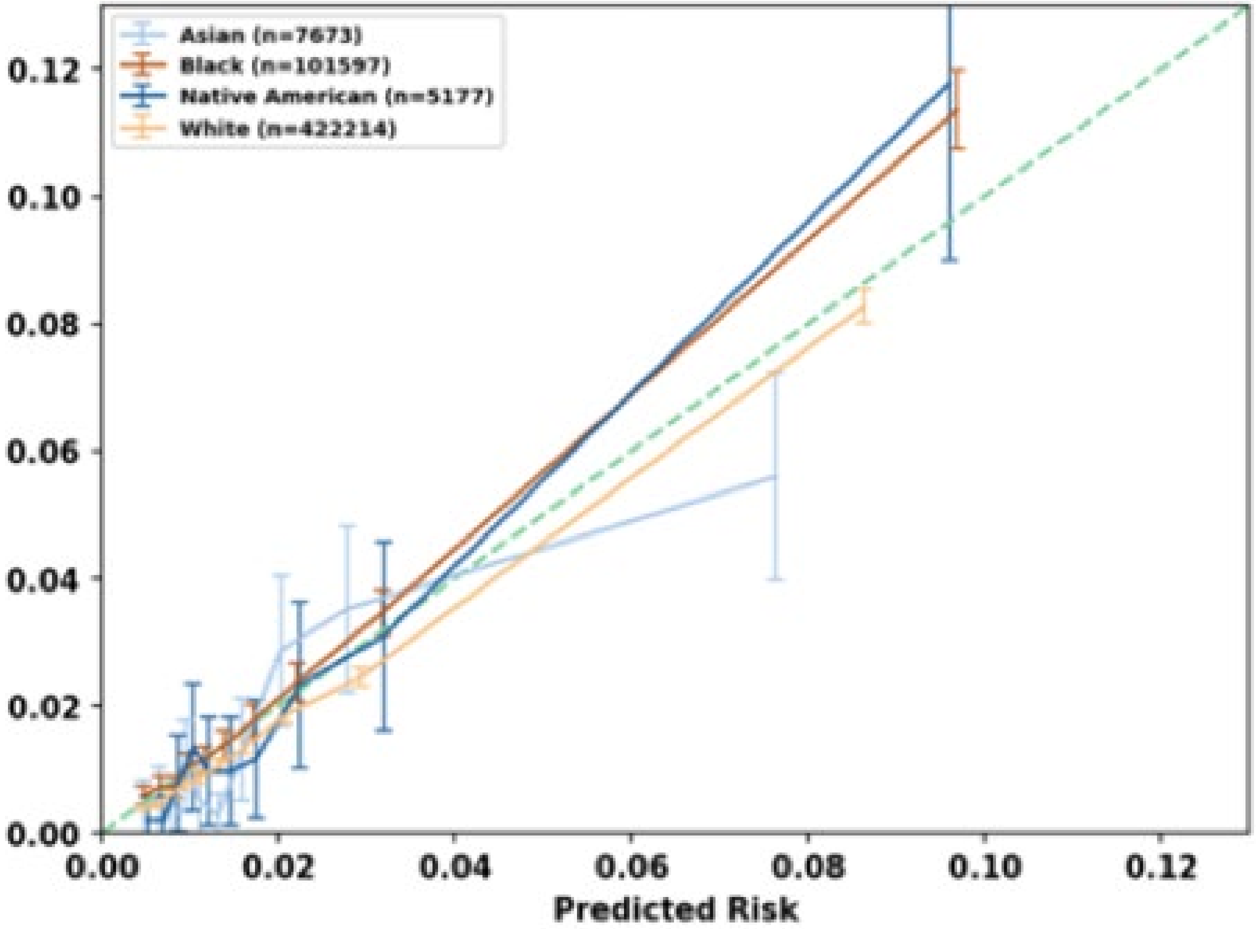
For example, if we assign a group of people a risk of 10%, the actual overdose/suicide-related incidence rate should also be 10%.

Calibration: Race

In Sample



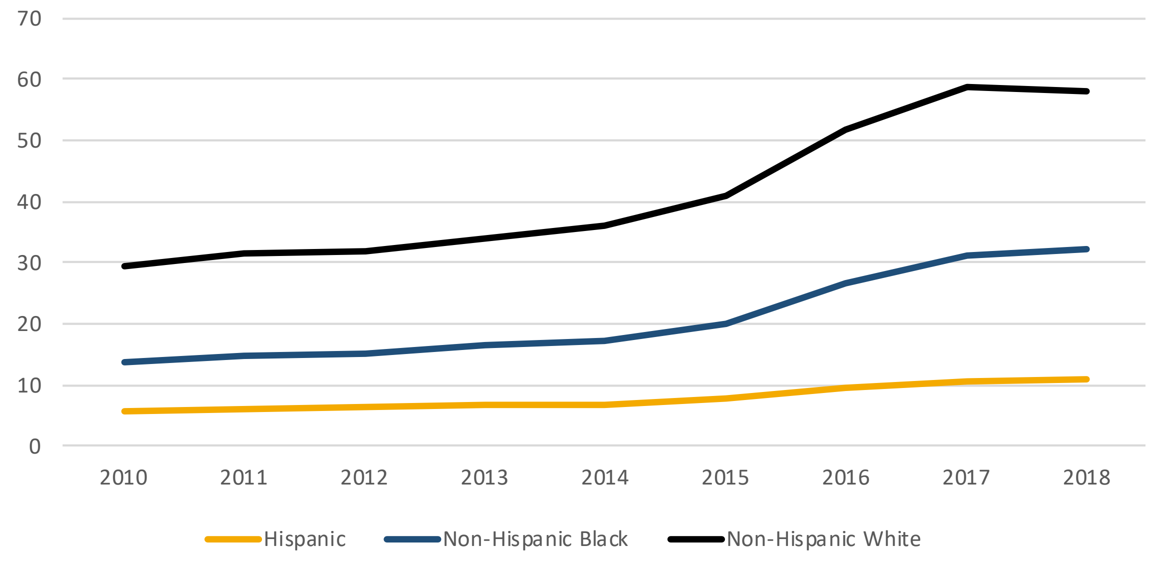
Out Sample



SAE Trends X Race during modeling period

Age-adjusted drug poisoning rates

from: https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/



Sharp jump in drug poisoning rates between 2015 and 2018

Increase varied by race/ethnicity

Large relative increase in drug poisoning rates in Black population:

Year	Black	White
2015	12.2	21.1
2017	20.6	27.5

Emphasizes the need for on-going calibration of predictive models, particularly when population risk is evolving rapidly.

Conclusions

- Algorithmic bias related to race is observed in the STORM algorithm and likely associated with other stratification tools for opioid risk mitigation
- Sets of measures that provide model summary statistics provide key context
- Visualization techniques that provide model diagnostics can convey important information to SMEs
- Due to their role in MH operations, similar analyses should be conducted on REACHVET models (STORM 2, RV) to inform strategies for addressing bias

Improving Algorithmic Fairness at the VA

Ravi B. Parikh, MD, MPP
Kristin A. Linn, PhD
Jiali (Helen) Yan, MS
Kevin A. Jenkins, PhD
Matthew Maciejewski, PhD
Sumedha Chhatre, PhD
Amol S. Navathe, MD, PhD

Penn Medical Ethics & Health Policy
CHERP Center for Health Equity Research and Promotion
VA HSR&D Center of Innovation

Funding and Disclosures

Funding: VA HSR&D I01HX003371-01

Disclosures

Ravi Parikh: Grants from Humana; grants from National Palliative Care Research Center; grants from Prostate Cancer Foundation; grants from Conquer Cancer Foundation; personal fees and equity from GNS Healthcare, Inc. and Onc.AI; personal fees from Cancer Study Group and Nanology; contributor at Medscape; board member at the Coalition to Transform Advanced Care

Amol Navathe: Grants from Hawaii Medical Service Association, grants from Anthem Public Policy Institute, grants from Commonwealth Fund, grants from Oscar Health, grants from Cigna Corporation, grants from Robert Wood Johnson Foundation, grants from Donaghue Foundation, grants from Pennsylvania Department of Health, grants from Ochsner Health System, grants from United Healthcare, grants from Blue Cross Blue Shield of NC, grants from Blue Shield of CA, personal fees from Navvis Healthcare, personal fees from Agathos, Inc., personal fees and equity from Navahealth, personal fees from YNHHSC/CORE, personal fees from Maine Health Accountable Care Organization, personal fees from Maine Department of Health and Human Services, personal fees from National University Health System - Singapore, personal fees from Ministry of Health - Singapore, personal fees from Elsevier Press, personal fees from Medicare Payment Advisory Commission, personal fees from Cleveland Clinic, personal fees from Analysis Group, personal fees from VBID Health, and equity from Embedded Healthcare, and other from Integrated Services, Inc.

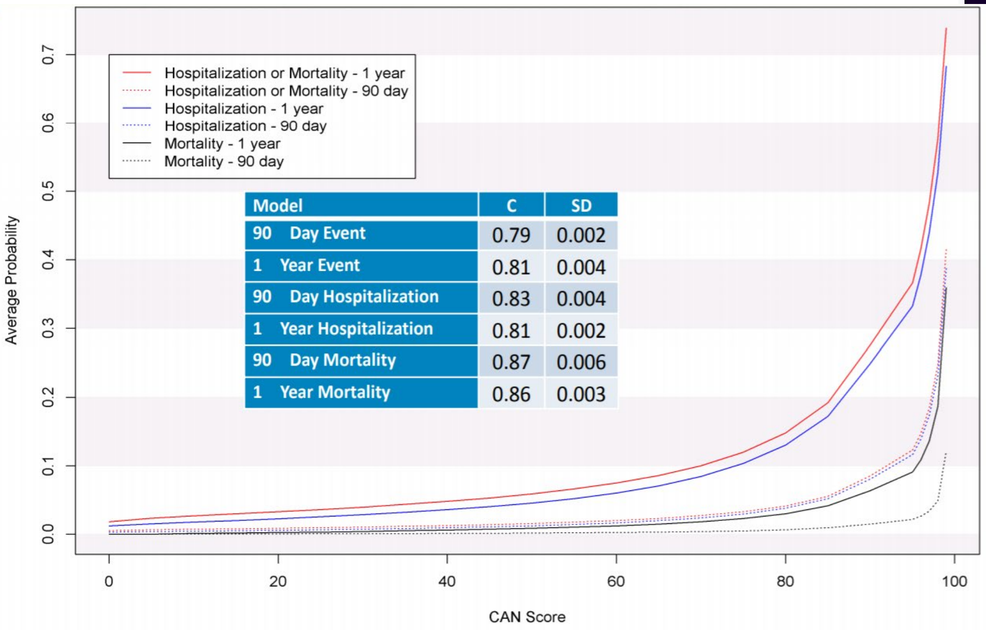
Agenda

- What is algorithmic unfairness?
- Detecting unfairness in the VA CAN score
- What contributes to unfairness in the CAN score?
- How can we reduce unfairness in the CAN score?

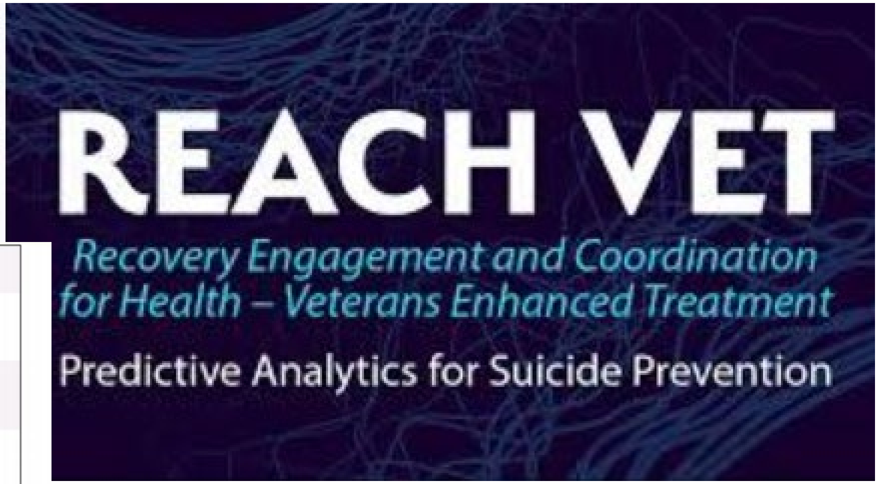
VA is at the leading edge of clinical predictive analytics

By Stephan D. Fihn, Joseph Francis, Carolyn Clancy, Christopher Nielson, Karin Nelson, John Rumsfeld, Theresa Cullen, Jack Bates, and Gail L. Graham

Insights From Advanced Analytics At The Veterans Health Administration



Model	C	SD
90 Day Event	0.79	0.002
1 Year Event	0.81	0.004
90 Day Hospitalization	0.83	0.004
1 Year Hospitalization	0.81	0.002
90 Day Mortality	0.87	0.006
1 Year Mortality	0.86	0.003



Fihn et al, Health Affairs, 2013

The Care Assessment Needs (CAN) Score

- Predicts risk of **hospitalization and/or death** for VHA's entire primary care population
- Accessed **4000 times** by **1200 VA clinicians and health workers each month**
- Used to
  - Create individualized care plans
  - Make care management referrals
  - Determine geographic sites of new health care services
- Standardized to a percentile risk
- C-stat for one year mortality or hospitalization: **0.79**

Wang et al., Med Care, 2013

The Care Assessment Needs (CAN) Score

- Relies on six data domains
  - Demographics
  - Diagnoses (inpatient and outpatient)
  - Vital signs
  - Medications
  - Laboratory results
  - Prior use of health services
- **Updated weekly** at the patient-level

Wang et al., Med Care, 2013

Project goals

- Our goal is to improve equity in health care resource allocation for Veterans through a more fair Care Assessment Needs (CAN) score
  - Ensure that the CAN score is promoting equity for racial and ethnic minorities
  - To generate an algorithmically fair CAN score with respect to African-American race that will serve as an example for VA predictive algorithms.
    - Will yield a generalizable methodology to address unfairness in the current CAN score

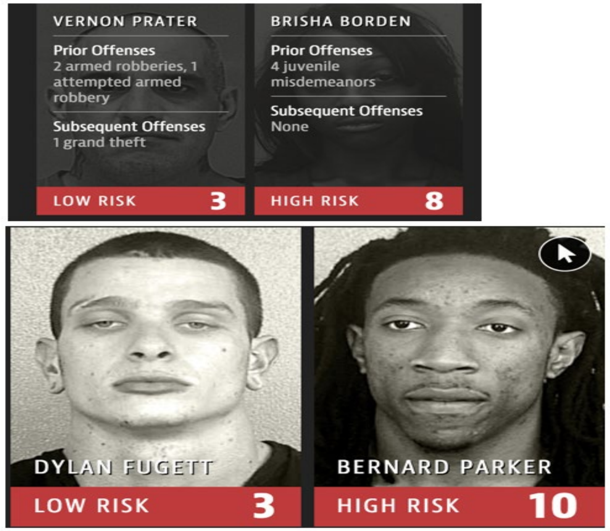
What is algorithmic unfairness?

Algorithm unfairness is a major concern

Algorithm Unfairness

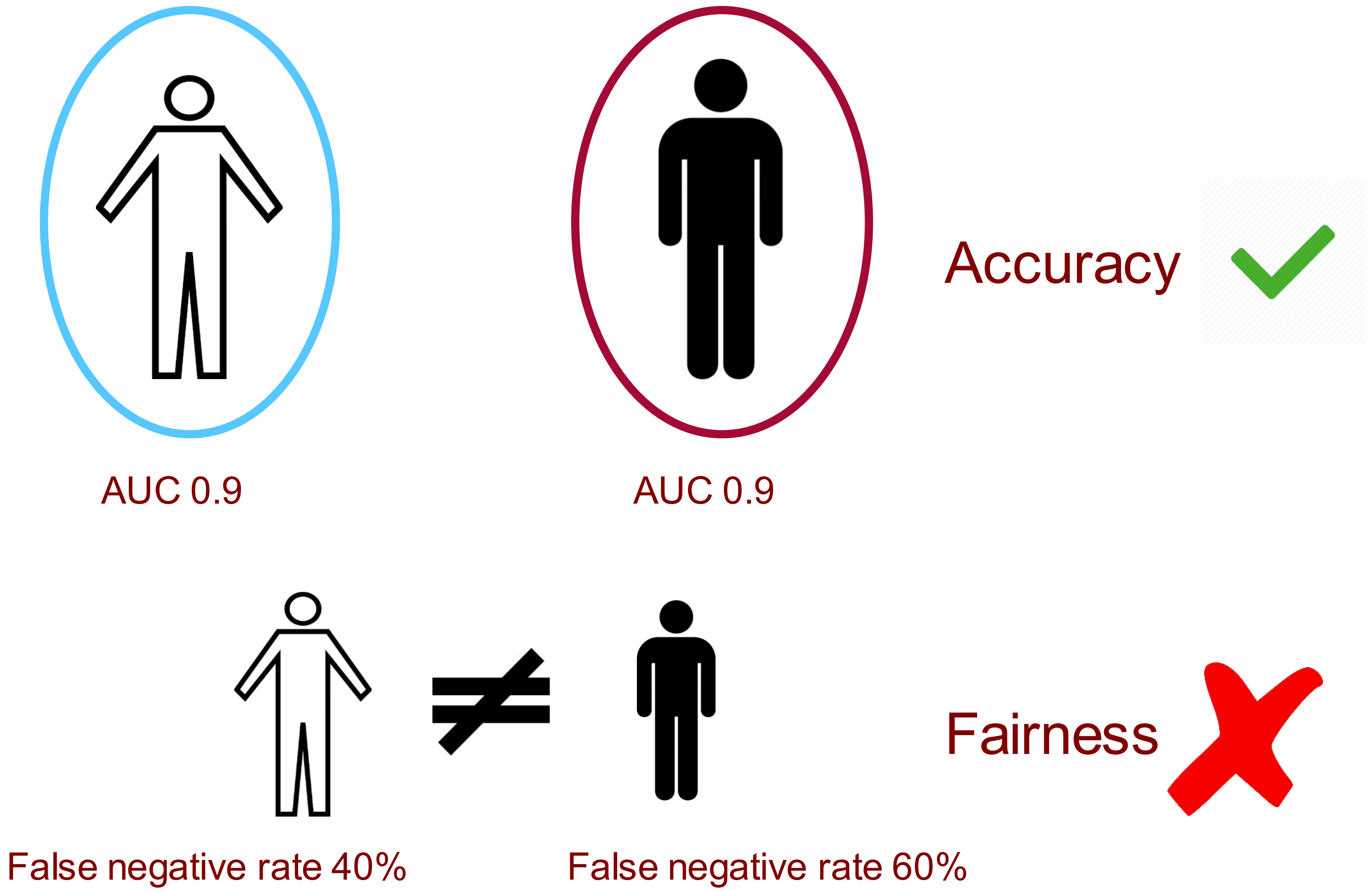
Does an algorithm systematically mischaracterize risk for a certain subgroup of individuals?

- Algorithm unfairness has been well-characterized in several non-clinical fields
  - Recidivism in crime
  - Banking loans
  - Facial recognition



African-Americans appear to be particularly disadvantaged by algorithm unfairness

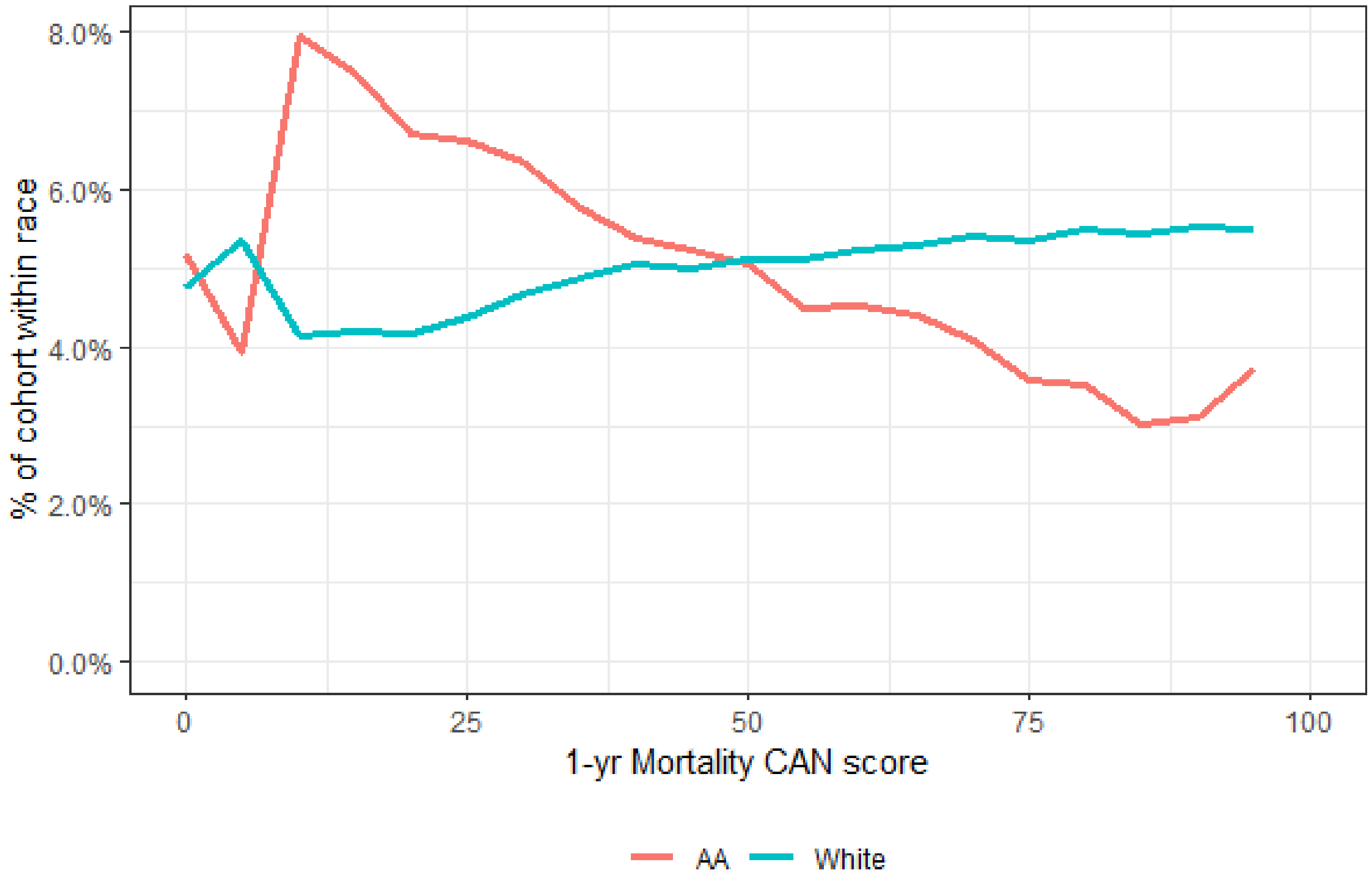
Algorithmic fairness ≠ accuracy



Could the CAN score be more fair?

Case Study: CAN 2.0 score for one-year mortality, based on 2014 data (#AA=859,598, #White=4,014,927)

African-Americans have lower CAN scores than Whites



Algorithmic unfairness

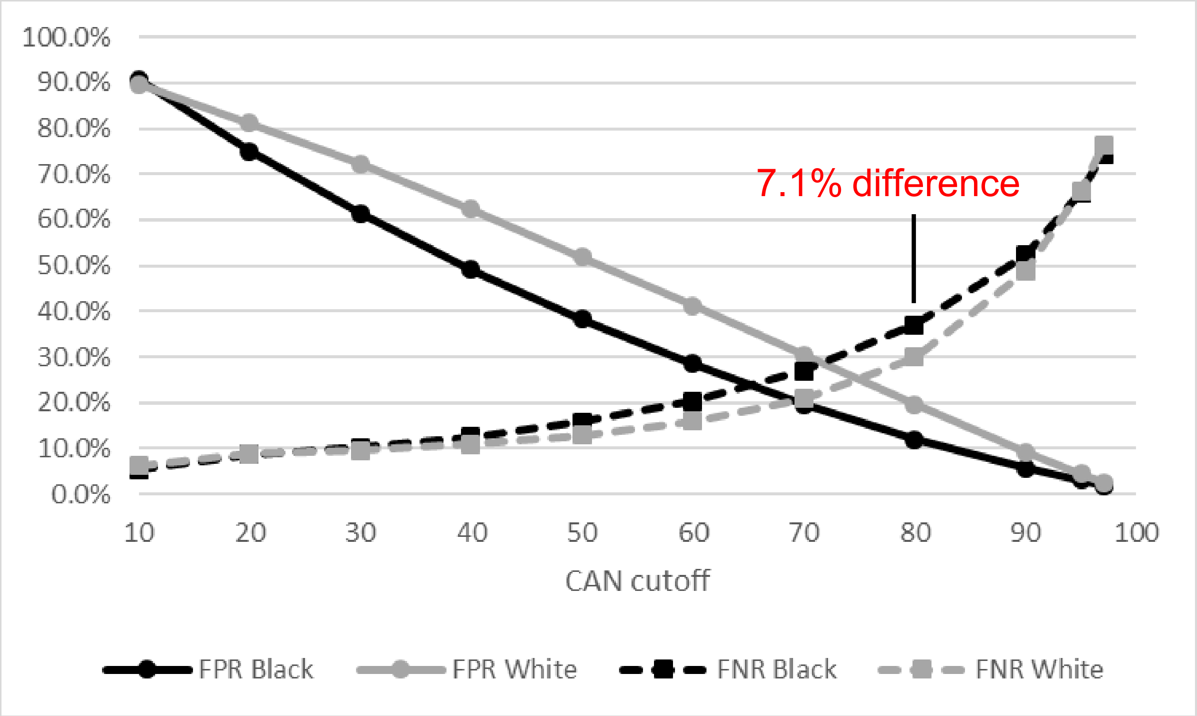
Our preliminary data suggests that the VA CAN score may be *algorithmically unfair* towards African-Americans using common definitions of fairness:

- *Equality of opportunity*
- *Individual fairness*

Lower scores for African-Americans may impact referrals for and receipt of specific VA services

AA Veterans may be falsely classified as low-risk

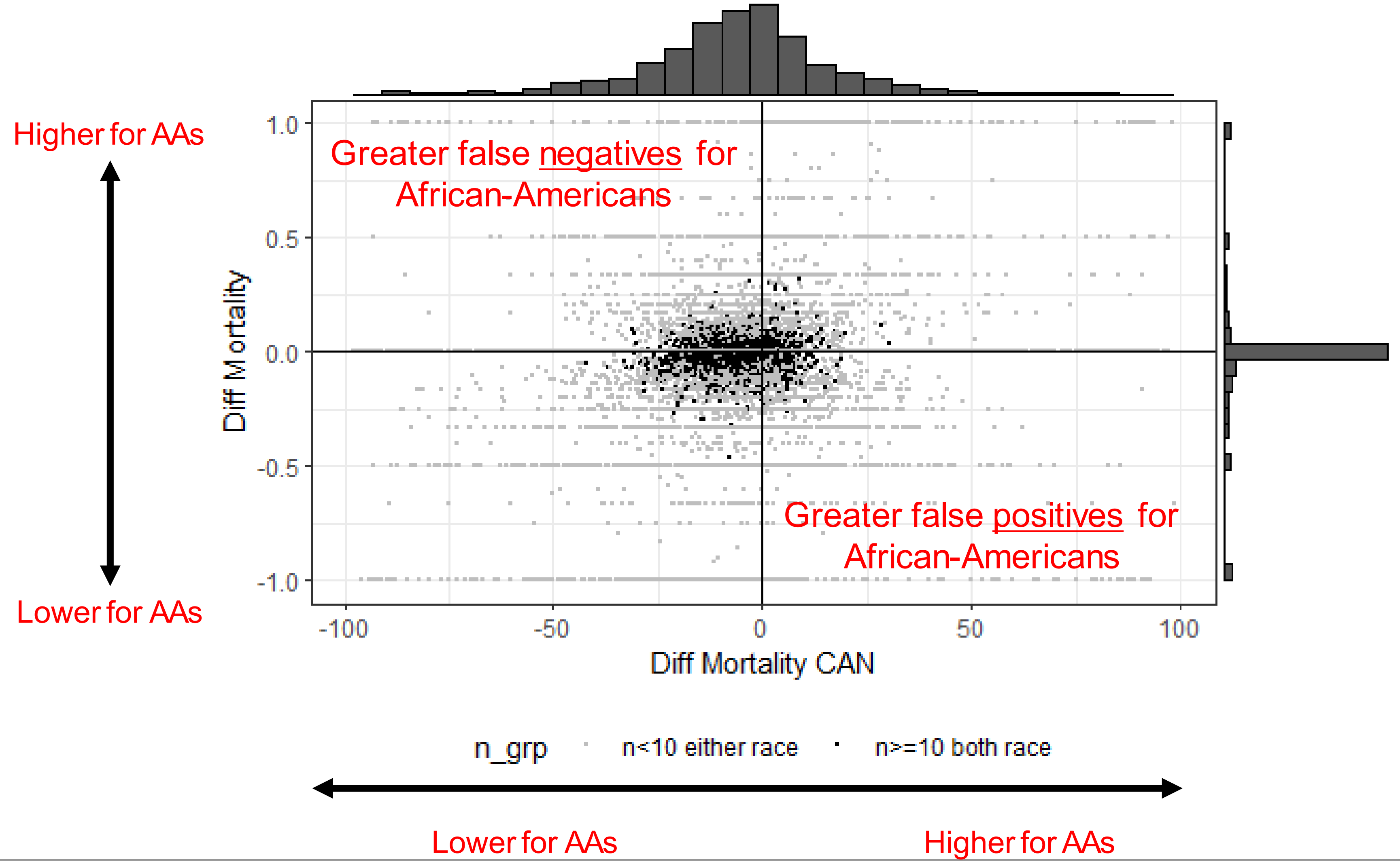
Equality of Opportunity: Do AA and White Veterans with the same CAN score die at the same rate in the following year?



Across all CAN scores, African-American Veterans are slightly more likely to be falsely classified as low-risk

White Veterans with similar comorbidities have greater CAN scores than AA Veterans

Individual Fairness: Do AA and White Veterans who are similar on non-race variables have similar CAN scores?



What may contribute to algorithmic unfairness?

- **Internal data issues**
  - **<u>Class imbalance</u>**
  - Measurement error
  - “Labels problem”: Selection of biased outcomes
  - Heterogeneity of covariate relationships with outcomes
- **External data issues**
  - **<u>Omitted variables</u>**
  - **<u>Unmeasured mediators</u>**
  - Rare events

Class imbalance between races may contribute to unfairness

Class Imbalance:

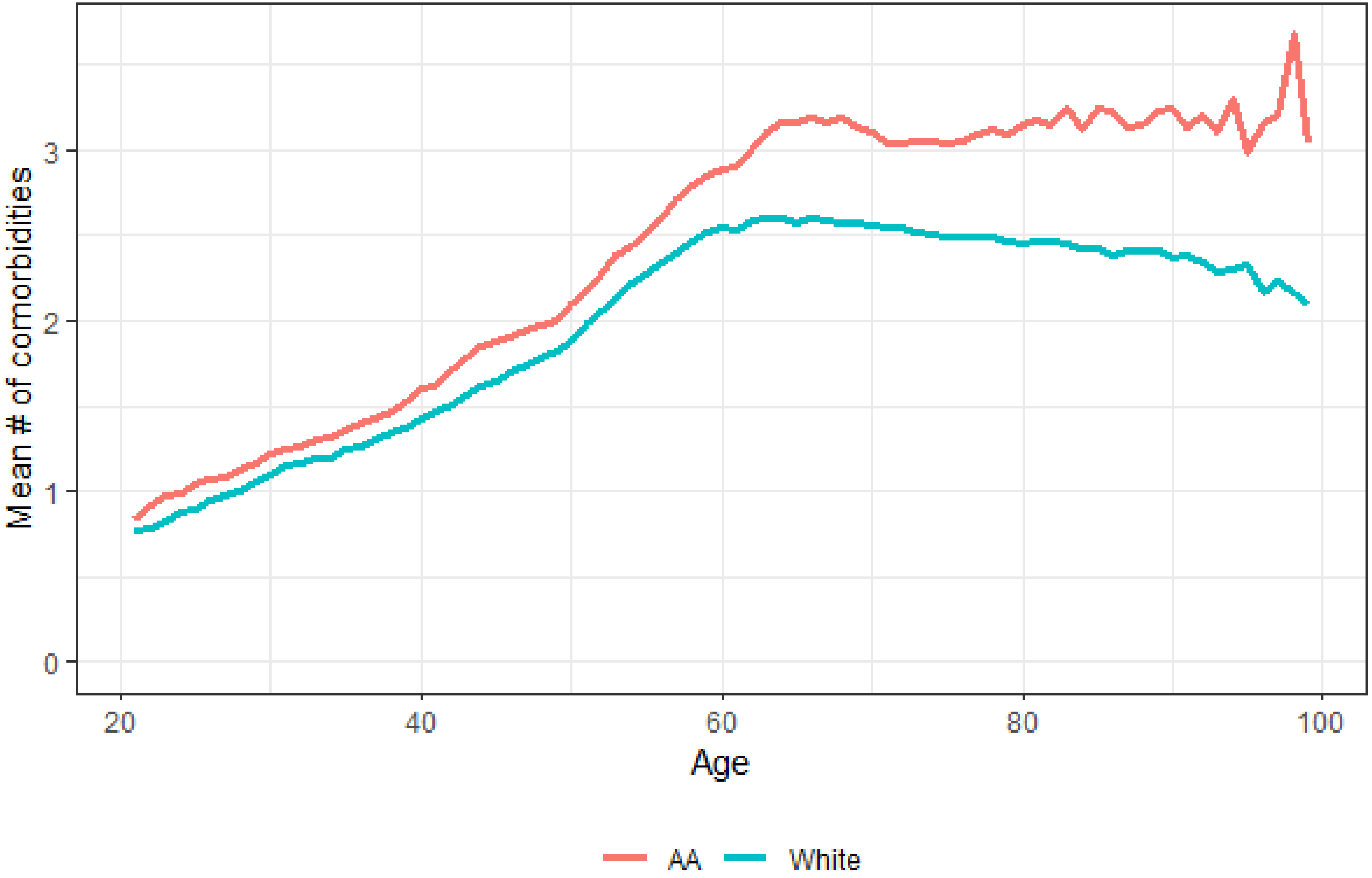
Distribution of a particular class (e.g. White) in an algorithm training set is not equal to another class (e.g. African-American)

Variable	Both (n=4,874,525)	AAs (n=859,598)	Whites (n=4,014,927)
Female, n (%)	343,077 (7.0%)	110,711 (12.9%)	232,366 (5.8%)
Age, median (IQR)	64.0 (53.0, 72.0)	57.0 (48.0, 65.0)	65.0 (55.0, 74.0)
Elixhauser groups, mean (SD)	2.3 (2.0)	2.5 (2.1)	2.3 (2.0)

Class imbalance between races may be a mechanism of unfairness in the CAN

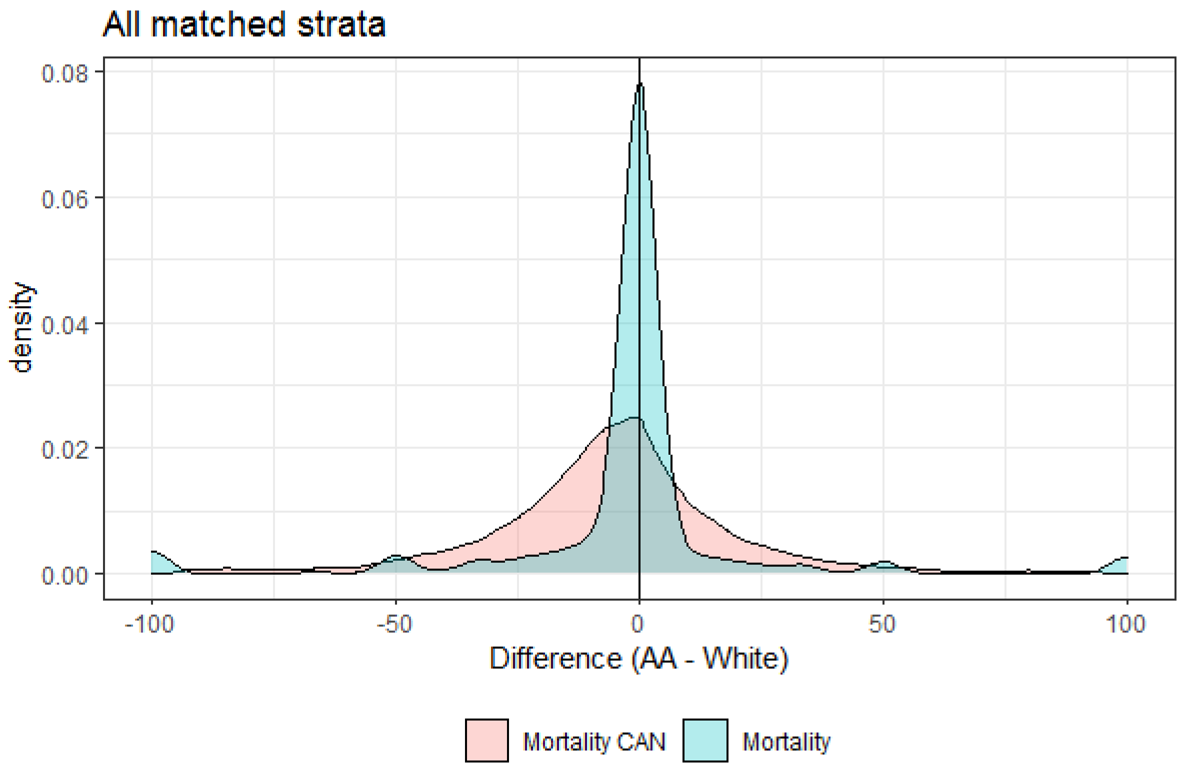
Systematic racial differences in age/comorbidities could contribute to unfairness

African-American Veterans have high comorbidity burden



What if we match on comorbidities?

CAN score is fair if AA and White Veterans who are similar on non-race variables have similar CAN scores

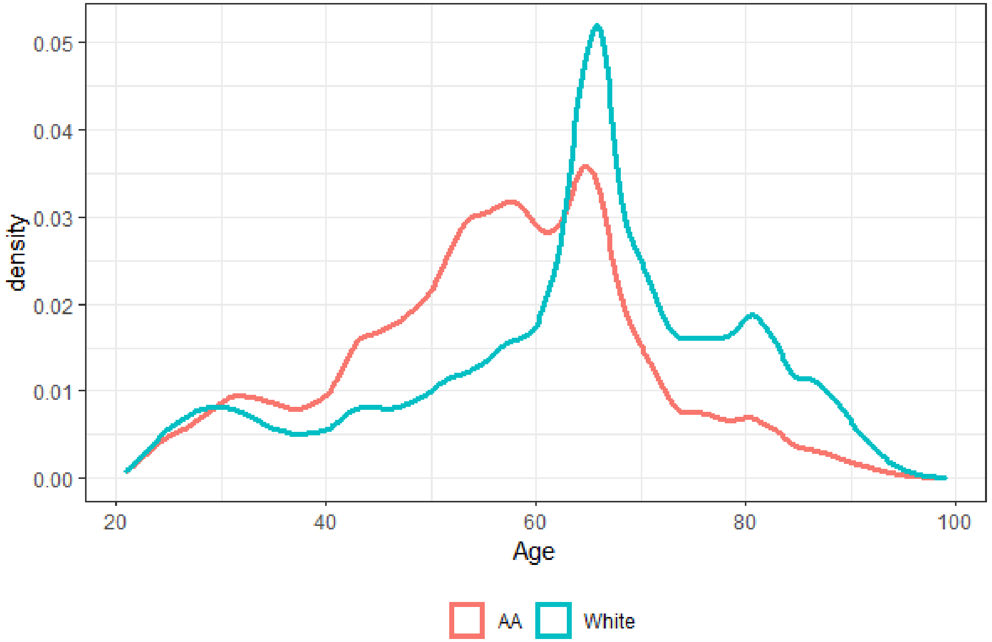


Negative values: Higher for Whites

Positive values: Higher for AAs

White Veterans who have similar comorbidity burden as AAs tend to have greater CAN scores than AAs.

Age differences may drive algorithmic unfairness



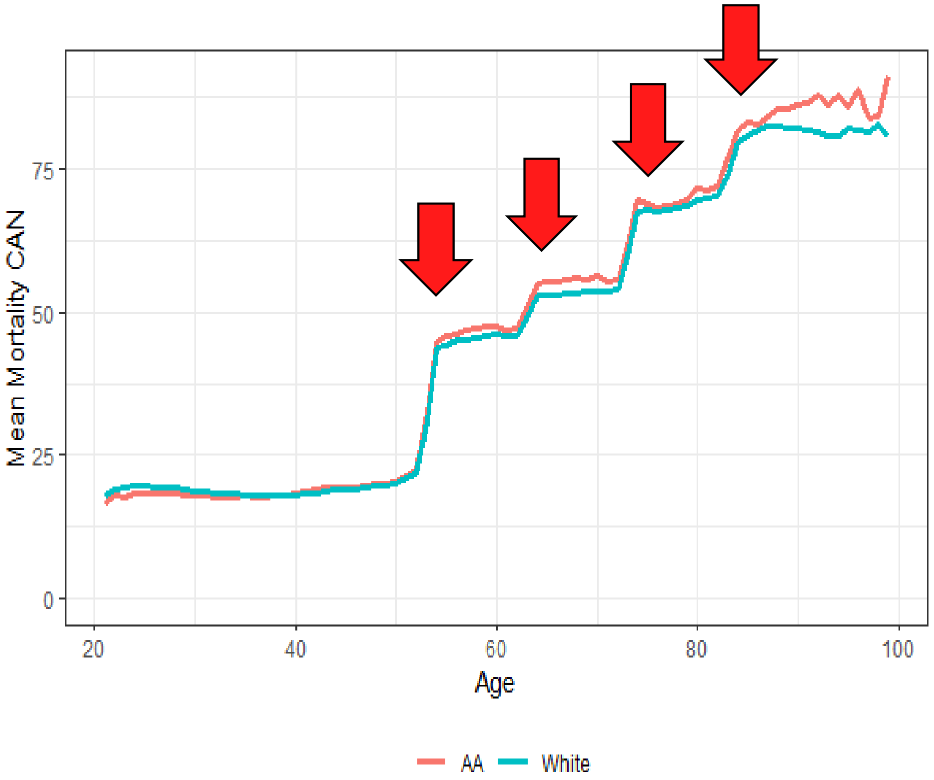
Predictors	Reference	Model Coefficients Estimate
Age < 55	Age ≥85	-2.795
Age 55-64		-1.752
Age 65-74		-1.407
Age 75-84		-0.702
Others		~-0.5 – 0.5

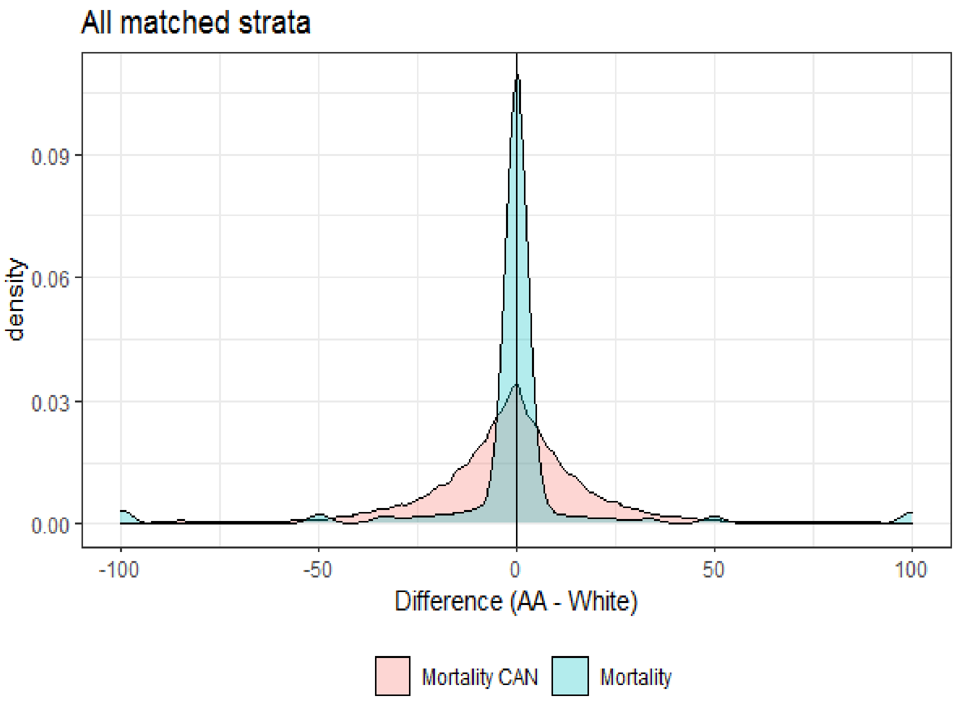
<u>Hypothesis:</u>

Lower age of African-Americans is a large contributor to algorithmic unfairness in the CAN score

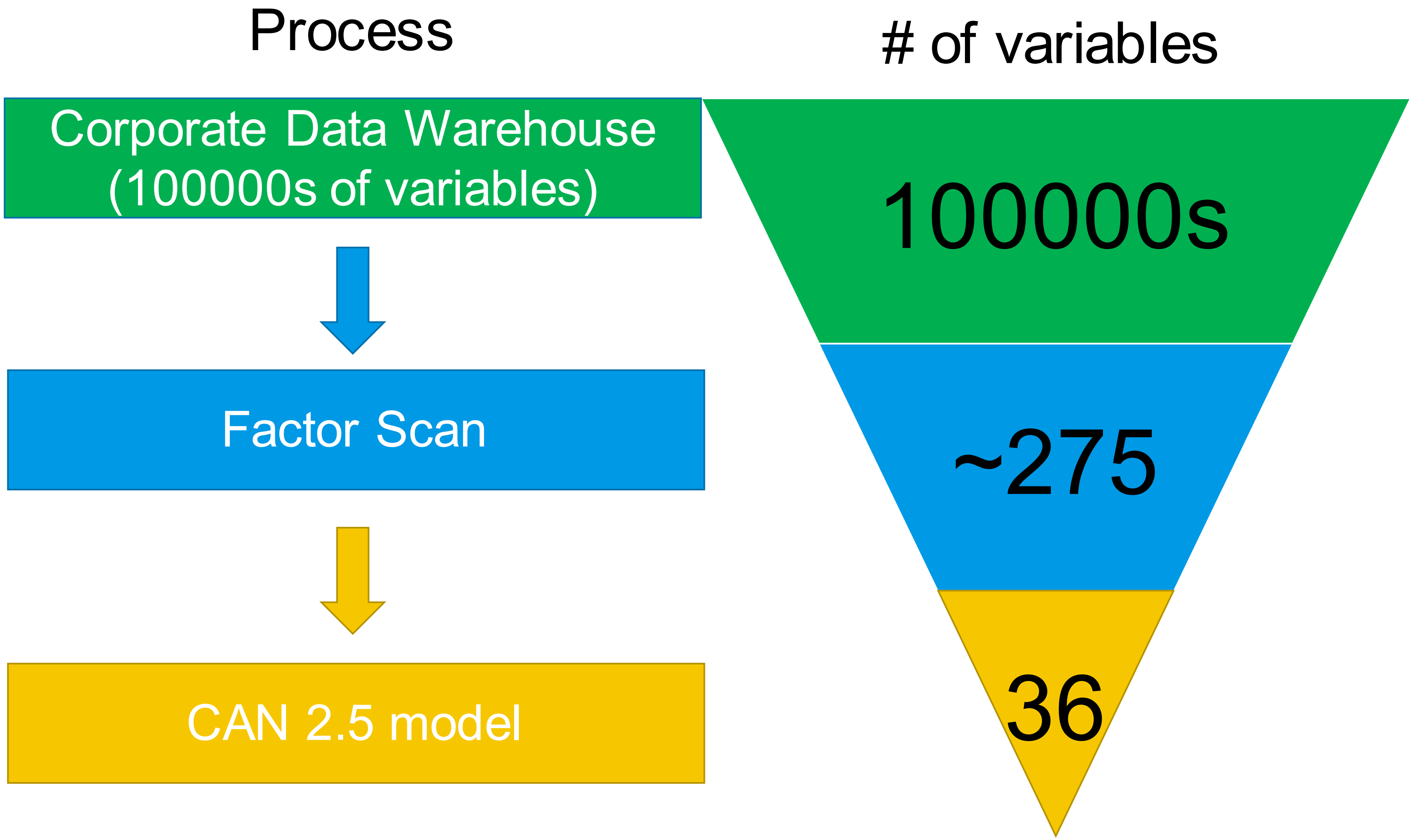
Accounting for age may mitigate unfairness

Exact matching based on age (5yr bin) and Elixhauser groups





Could the data generating process be unfair?



Variable importance by race

Pooled model	AA-specific model	White-specific model
Statin	**Priority level 0**	Variation for weight
Respiration vital	**Albumin**	Statin
Variation for weight	Variation for weight	Respiration vital 18-20
Beta Blocker med fill	**Metastatic Cancer**	Beta-blocker
TIU teleph notes	**Dementia**	TIU teleph notes
Age 75-84	**Albumin variation**	Age 75-84
Pulse 60-90	Statin	Age 65-74
SBP 110-140	**Phone 21-30m**	No office visit prior 90d

<u>Future Direction:</u>

Re-train CAN models with race-specific variable selection

What are strategies to ameliorate unfairness?

Preliminary data using CAN 2.5 (a recalibration of the CAN 2.0 model weights)

False negative rate is our metric of interest

- False negative rate is a correlate of equality of opportunity
- FNR = % of low-risk Veterans who die in the next 12 months
  - Set at 80^{th} percentile threshold

A fair algorithm has no difference in FNR for African-Americans and Whites

Statistical techniques to mitigate false negative rates

Technique	Overall	AAs	Whites	Difference
CAN 2.5	27.6	35.3	26.5	**8.8**
Weighted by race	26.5	34.1	25.5	**8.6**
Weighted by event rate	26.0	33.7	25.0	**8.7**
Weighted by both	26.1	33.7	25.0	**8.7**
Race*age group interaction	26.3	34.9	25.1	**9.8**
Continuous age and race*age interaction	26.0	35.5	24.7	**10.8**
Separate models	-	34.8	25.0	**9.8**
Gradient boosting - pooled	26.2	34.0	25.1	**8.9**
Gradient boosting - separate models	-	25.7	27.7	**-2.0**
Random forest - pooled	35.7	34.7	35.8	**-1.1**
Random forest- separate models	-	29.2	37.8	**-8.6**

Future Directions: Social determinants of health

Adverse social determinants of health may be disproportionate contributors to risk for AAs and thus contribute to unfairness

	White	African-American
n	4,014,927	859,598
Age (mean [SD])	63.13 (16.03)	56.33 (13.90)
Enrollment priority, n (%)		
1-2	1363201 (34.0)	371224 (43.2)
3-8	2644007 (65.9)	486234 (56.6)
Missing	7719 (0.2)	2140 (0.2)
Location, n (%)		
Highly rural	65568 (1.6)	2136 (0.2)
Rural	1530487 (38.1)	143751 (16.7)
Urban	2306618 (57.5)	697908 (81.2)
Marital Status, n (%)		
Married	2355242 (58.7)	368337 (42.8)
Single	411361 (10.2)	157598 (18.3)
Other	1233998 (30.7)	330424 (38.4)
Disability (%)	497934 (12.4)	137477 (16.0)

Identifiable SDoH from the CDW

Metric	Data source
Violence/military sexual trauma	ICD codes; Stop codes; Health factors
Housing instability	ICD codes; Stop codes; Health factors
Financial and employment problems	ICD codes; Enrollment priority; Stop codes; Health factors
Legal problems	ICD codes; Stop codes
Family/social problems (e.g. problem related to upbringing)	Health factors
Inadequate transportation	Health factors; Rural/urban indicator
Non-specific psychosocial needs	ICD codes; Health factors

Blosnich et al, 2019; Blosnich et al, 2017[68]

Thank you!

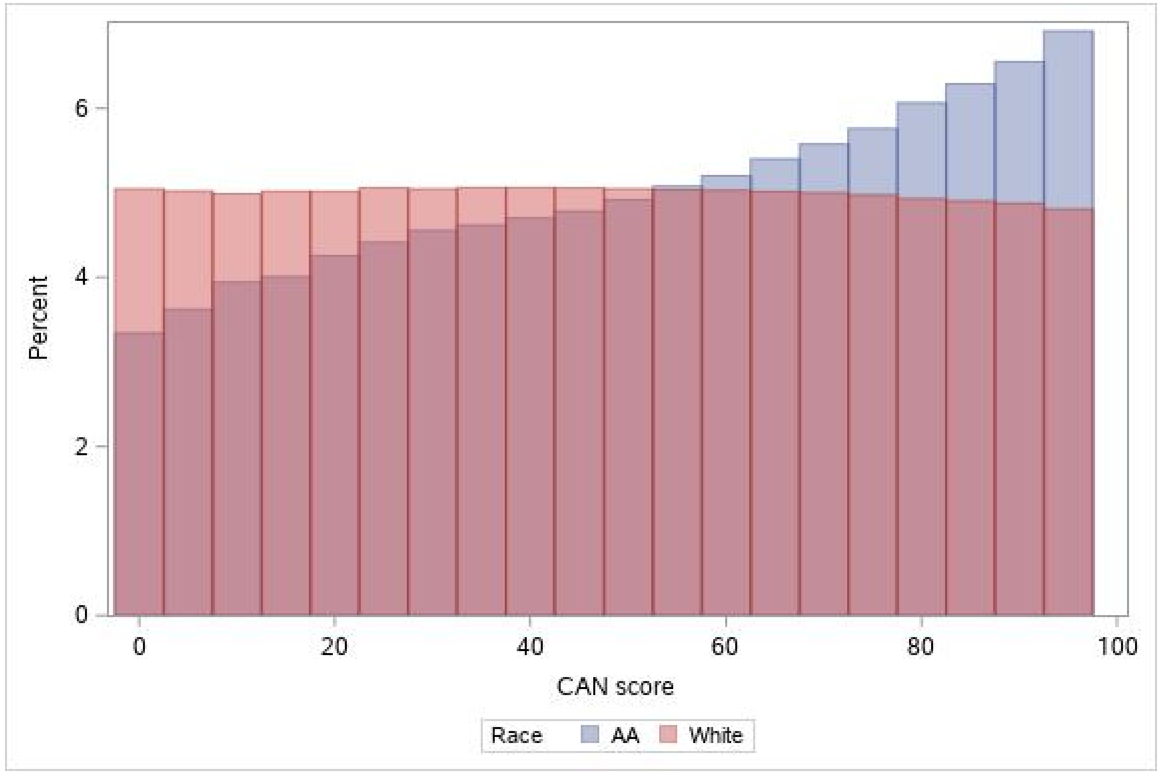
Ravi.parikh@va.gov
Ravi.parikh@pennmedicine.upenn.edu
@ravi_b_parikh

Amol.navathe2@va.gov
Amol@pennmedicine.upenn.edu
@amolnavathe

APPENDIX

Demographic parity

Is the distribution of CAN scores similar between African-American and White Veterans?



<u>However</u>, there may be legitimate reasons for observed differences in CAN scores by race

Social determinants of health

Adverse social determinants of health may be disproportionate contributors to risk for AAs

	White	African-American
n	4,014,927	859,598
Age (mean [SD])	63.13 (16.03)	56.33 (13.90)
Enrollment priority, n (%)		
1-2	1363201 (34.0)	371224 (43.2)
3-8	2644007 (65.9)	486234 (56.6)
Missing	7719 (0.2)	2140 (0.2)
Location, n (%)		
Highly rural	65568 (1.6)	2136 (0.2)
Rural	1530487 (38.1)	143751 (16.7)
Urban	2306618 (57.5)	697908 (81.2)
Marital Status, n (%)		
Married	2355242 (58.7)	368337 (42.8)
Single	411361 (10.2)	157598 (18.3)
Other	1233998 (30.7)	330424 (38.4)
Disability (%)	497934 (12.4)	137477 (16.0)

- Over the past decade, the VA has increasingly implemented routine screening for certain SDoH in clinical settings
- Question: *What SDoH indicators are available in structured data and operationally important?*
- Question: *Are there means to identify adverse SDoH outside of structured fields?*

Identifiable SDoH from the CDW

Table 3. Validated SDoH metrics in VA literature	
Metric	Data source
Violence/military sexual trauma	ICD codes; Stop codes; Health factors
Housing instability	ICD codes; Stop codes; Health factors
Financial and employment problems	ICD codes; Enrollment priority; Stop codes; Health factors
Legal problems	ICD codes; Stop codes
Family/social problems (e.g. problem related to upbringing)	Health factors
Inadequate transportation	Health factors; Rural/urban indicator
Non-specific psychosocial needs	ICD codes; Health factors

- Question: *Are these the right sources of information to find adverse SDoH?*
- Question: *What level of underreporting do we expect?*
- Question: *Are there other indicators that you would add or replace?*

Blosnich et al, 2019; Blosnich et al, 2017[68]

Questions?

Thank you!