

Machine Learning for Equitable Healthcare

Irene Y. Chen

PhD Student, Electrical Engineering and Computer Science



Joint work with David Sontag, Marzyeh Ghassemi,
Fredrik D. Johansson, Rahul G. Krishnan, Sherri Rose, Emma Pierson, Shalmali Joshi,
Kadija Ferryman, Bharti Khurana, Emily Alsentzer, Hyesun Park, Richard
Thomas, Babina Gosangi, Rahul Gujrathi

MIT Clinical ML
www.clinicalml.org

LOST MOTHERS

How Hospitals Are Failing Black Mothers

A ProPublica analysis shows that women who deliver at hospitals that disproportionately serve black mothers are at a higher risk of harm.

by Annie Waldman, Dec. 27, 2017, 8 a.m. EST

JACC: HEART FAILURE
© 2018 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

VOL. 6, NO. 5, 2018

EDITORIAL COMMENT

Narrowing the Disparities in Heart Failure

Treat the Event or Try to Prevent?*



Hena Patel, MD, Kim Allan Williams, Sr., MD

The Secret to Keeping Black Men Healthy? Maybe Black Doctors

In an intriguing study, black patients were far more likely to agree to certain health tests if they discussed them with a black male doctor.

The New York Times

The Upshot

THE NEW HEALTH CARE

A Sense of Alarm as Rural Hospitals Keep Closing

The potential health and economic consequences of a trend associated with states that have turned down Medicaid expansion.



The Growing Gap in Life Expectancy by Income: Recent Evidence and Implications for the Social Security Retirement Age

Katelin P. Isaacs

Analyst in Income Security

Sharmila Choudhury

Section Research Manager

May 12, 2017

Disparities filter into observational data

Need and Goldstein, *Cell* 2009; U.S. Food and Drug Administration,
National Cancer Institute, Riley Wong for *Propublica*, 2018.

Disparities filter into observational data

Table 1. Ethnicity of participants in genome-wide association studies^a

Race/ethnicity	Number of studies	Total participants ^d
European only ^b	320	1 581 776
Asian only	26	52 841
Hispanic only	3	1019
Native American only	2	1102
Jewish only	2	3479
Gambian only	1	2340
Micronesian only	1	2346
Mixed ^c	11	European ^{b,e} African-American Asian Papua-New Guinean Other ^f



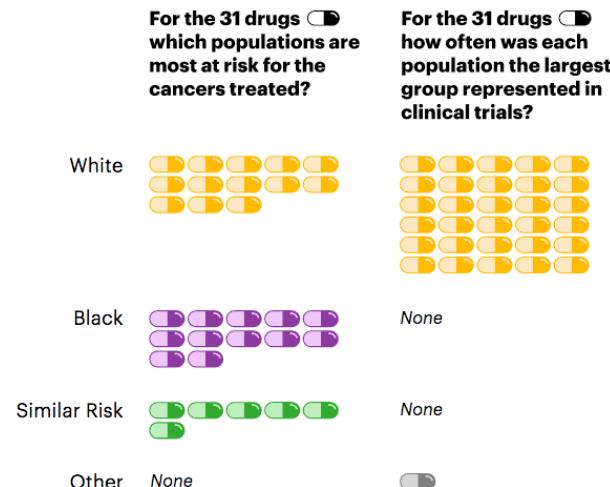
96% of participants in **GWAS studies**
were of European descent

Need and Goldstein, *Cell* 2009; U.S. Food and Drug Administration,
National Cancer Institute, Riley Wong for *Propublica*, 2018.

Disparities filter into observational data

Table 1. Ethnicity of participants in genome-wide association studies^a

Race/ethnicity	Number of studies	Total participants ^d
European only ^b	320	1 581 776
Asian only	26	52 841
Hispanic only	3	1019
Native American only	2	1102
Jewish only	2	3479
Gambian only	1	2340
Micronesian only	1	2346
Mixed ^c	11	European ^{b,e} African-American Asian Papua-New Guinean Other ^f
		92 437 7500 33 276 269



96% of participants in GWAS studies were of European descent

Cancer clinical drug trials do not match the populations most at risk.

- ▶ **Potentially biased** observational data
- ▶ **Opaque** and **hard to certify** as “bias-free”

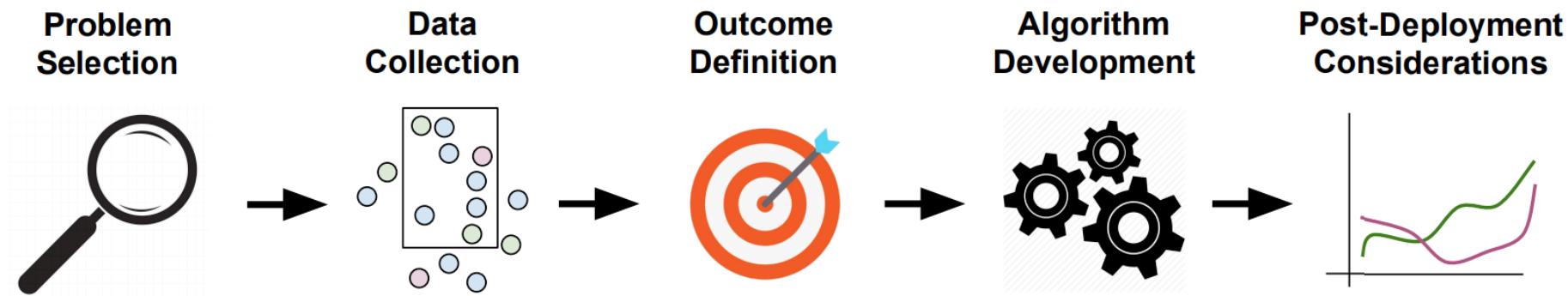


- ▶ **Potentially biased** observational data
 - ▶ **Opaque** and **hard to certify** as “bias-free”
-



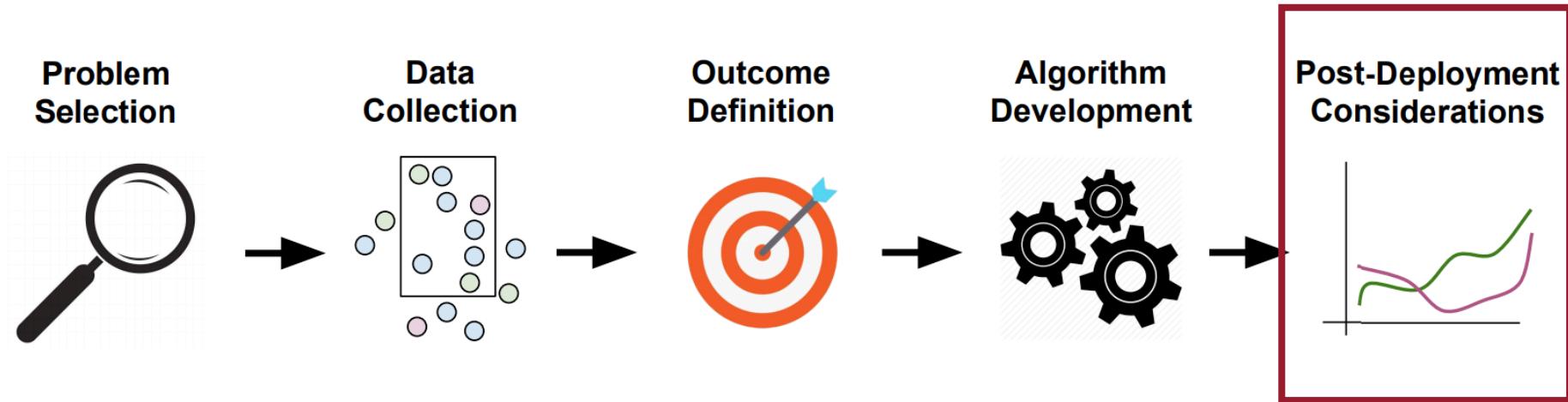
- ▶ **Potentially biased** observational data
 - ▶ **Opaque** and **hard to certify** as “bias-free”
-
- ▶ Incorporate **massive datasets**
 - ▶ Find latent patterns in **underserved populations**
 - ▶ **Scale** quickly and widely

Machine Learning for Equitable Healthcare



Chen et al, "Ethical Machine Learning for Health Care," *Annual Reviews for Biomedical Data Science* 2021.

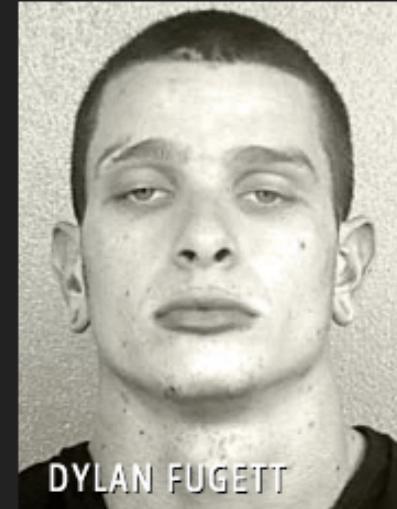
Machine Learning for Equitable Healthcare



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Two Drug Possession Arrests



DYLAN FUGETT

LOW RISK

3



BERNARD PARKER

HIGH RISK

10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

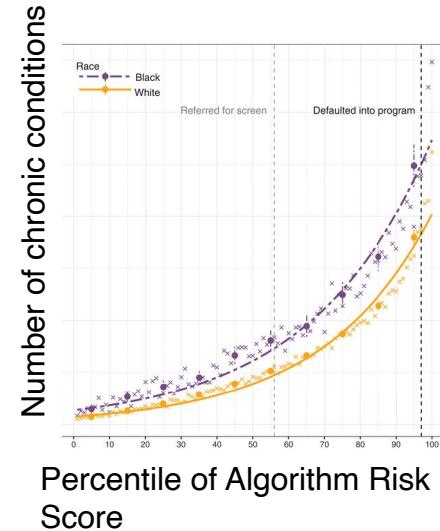
<http://gendershades.org/overview.html>

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

We are finding evidence of bias through audits

New images	Output
	95% malignant 5% benign
	20% malignant 80% benign

Dermatology algorithms are trained primarily on data from fair-skinned patients



Care management algorithms show racial bias due to training on the “wrong” outcome

- [1] Adamson and Smith, “Machine Learning and Health Care Disparities in Dermatology,” *JAMA Dermatology* 2018.
[2] Obermeyer et al, “Dissecting racial bias in algorithm used to manage the health of populations”, *Science* 2019.

How do we define “bias”?

- ▶ Fairness through unawareness
- ▶ Group fairness
- ▶ Calibration
- ▶ Error rate balance
- ▶ Representational fairness
- ▶ Counterfactual fairness
- ▶ Individual fairness

How do we define “bias”?



Arvind Narayanan

@random_walker

- ▶ Fairness thresholds

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!

Here it is (with minor edits):

docs.google.com/document/d/1bn...

See you on Feb 23/24.

- ▶ Error rate thresholds

- ▶ Representativeness

- ▶ Counterfactuals

- ▶ Individual fairness



Arvind Narayanan

@random_walker · Nov 6, 2017

When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread]
[twitter.com/random_walker/...](https://twitter.com/random_walker/)

[Show this thread](#)

4:24 PM · Jan 8, 2018 · Twitter Web Client

60 Retweets 208 Likes

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg *

Sendhil Mullainathan †

Manish Raghavan ‡

Abstract

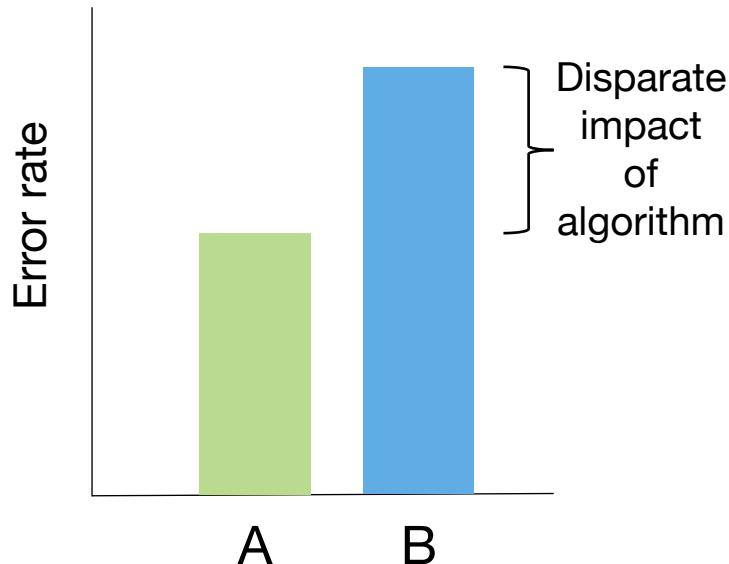
Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

Inherent Trade-Offs in the Fair Determination of Risk Scores

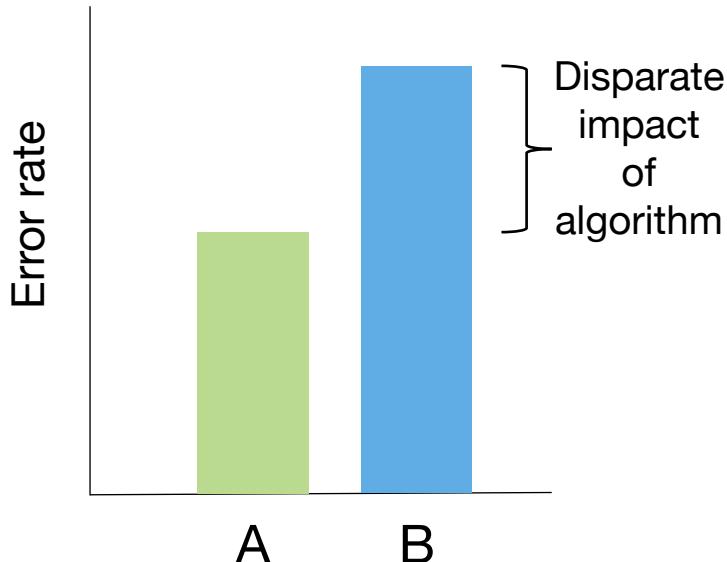
“We prove that except in highly constrained special cases, **there is no method** that satisfies these three [fairness] conditions simultaneously.”

version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

Why might my algorithm be unfair?

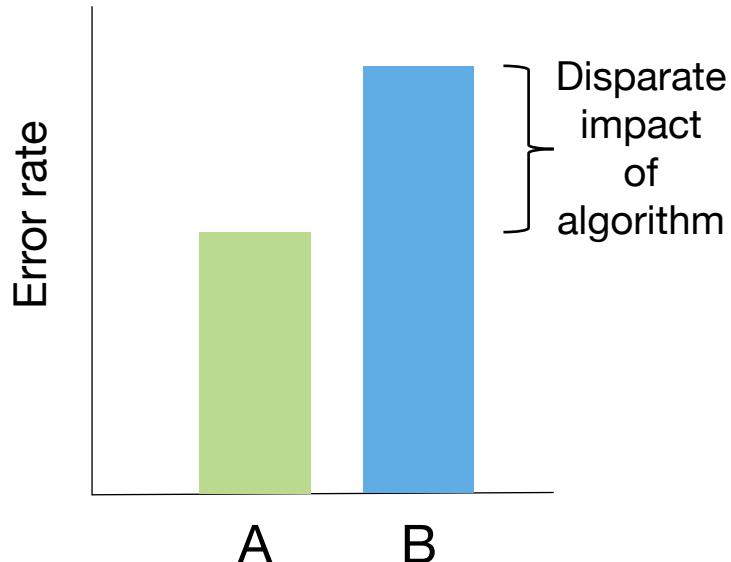


Why might my algorithm be unfair?



1. Group B is much smaller than Group A.
2. Group B has patterns in the data require more complex computational tools.
3. Measurements from Group B are less reliable.

Why might my algorithm be unfair?



1. Group B is much smaller than Group A. **VARIANCE**
2. Group B has patterns in the data require more complex computational tools. **BIAS**
3. Measurements from Group B are less reliable. **NOISE**

Bias, variance, and noise

	Description	How to fix
Bias	How well model fits data	Change model class
Variance	How much sample size affects accuracy	Increase training data size
Noise	Error independent of model class and sample size	Increase number of features

Chen et al, “Why is My Classifier Discriminatory?” NeurIPS 2018.

Sources of unfairness

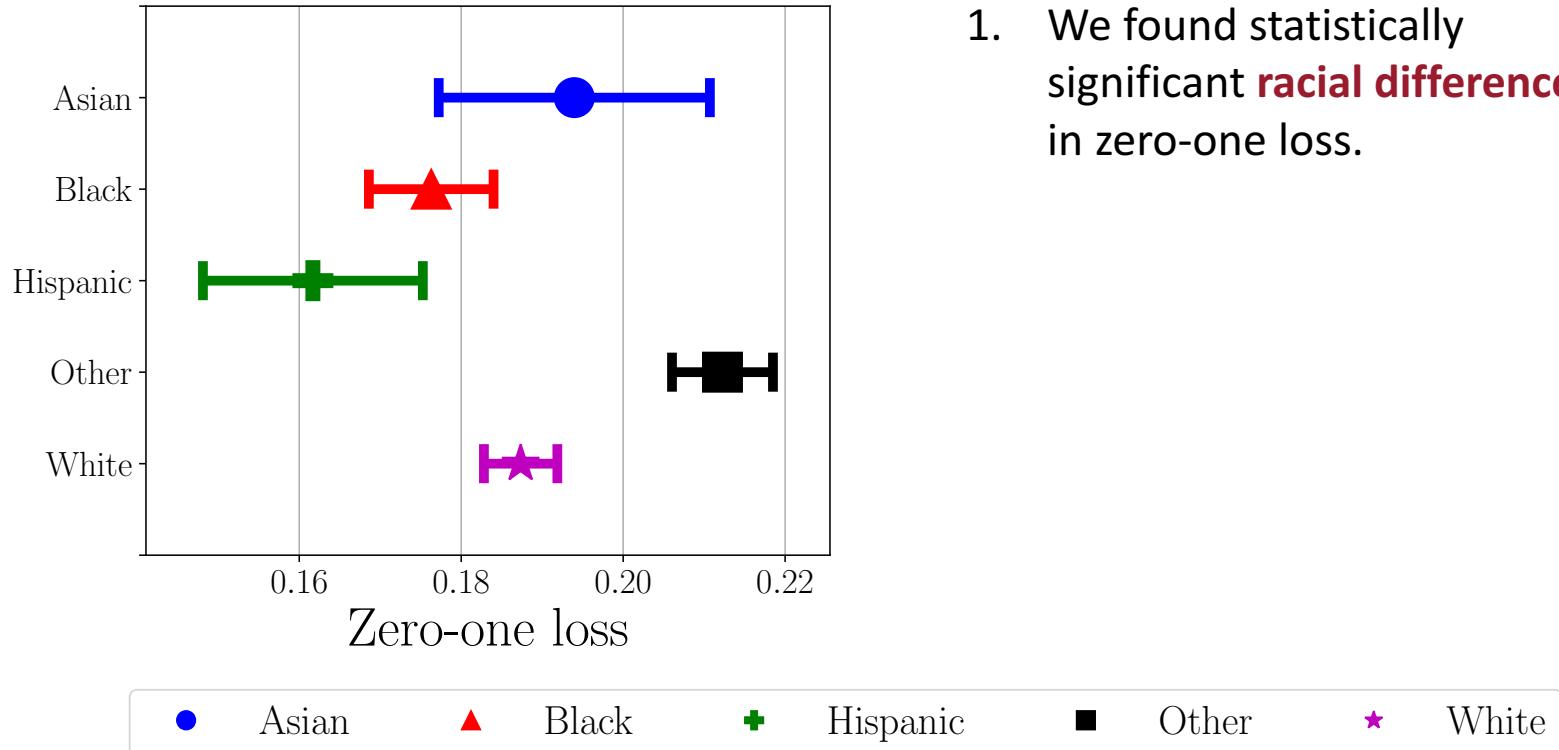
“unfairness”

$$\bar{\Gamma} = |(\bar{B}_1 - \bar{B}_0) + (\bar{V}_1 - \bar{V}_0) + (\bar{N}_1 - \bar{N}_0)|$$

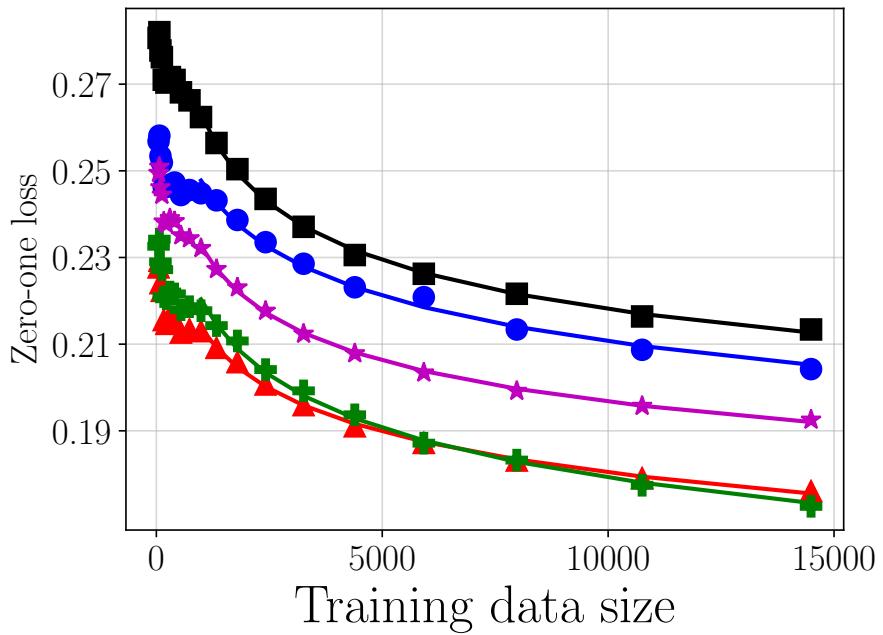
difference in bias **difference in variances** **difference in noise**

- ▶ How can we realistically estimate \bar{B}_a , \bar{V}_a , and \bar{N}_a ?
- ▶ What happens if $\bar{N}_0 \neq \bar{N}_1$?

Mortality prediction from MIMIC-III clinical notes



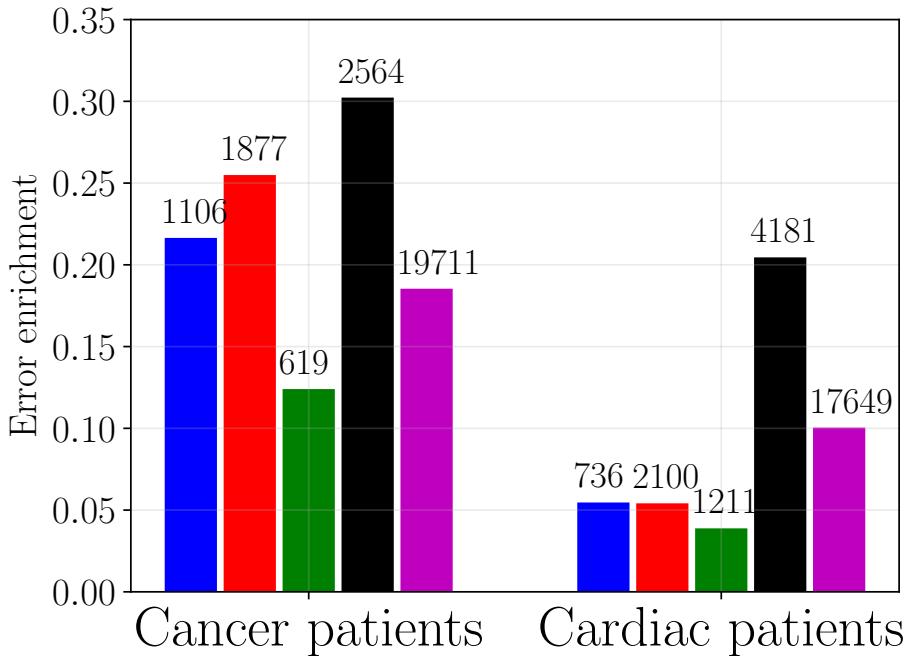
Mortality prediction from MIMIC-III clinical notes



1. We found statistically significant **racial differences** in zero-one loss.
2. By subsampling data, we fit inverse power laws to estimate the benefit of **more data** and reducing variance.

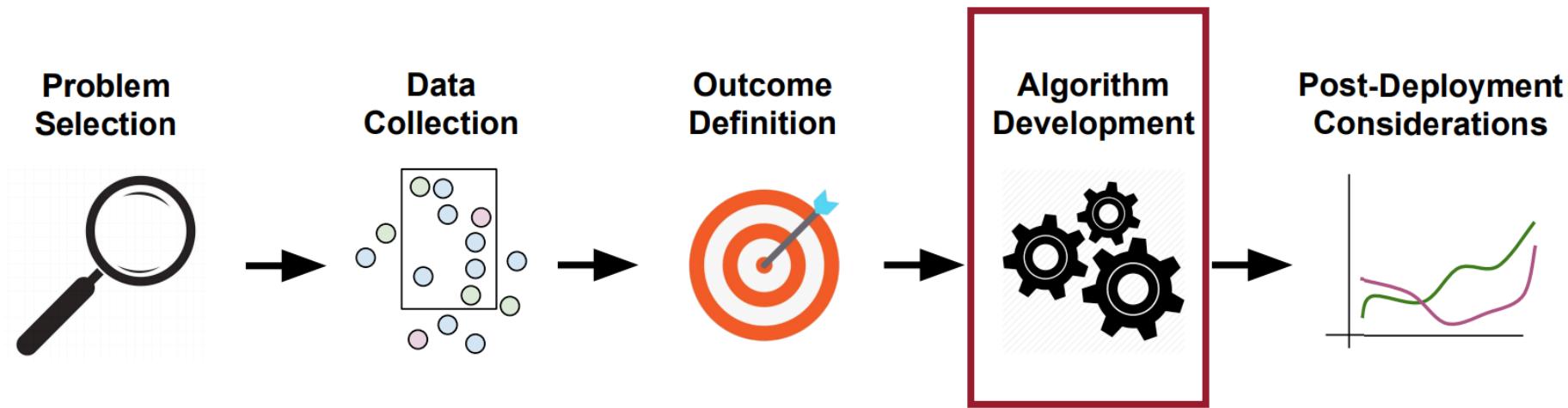
● Asian ▲ Black + Hispanic ■ Other ★ White

Mortality prediction from MIMIC-III clinical notes



1. We found statistically significant **racial differences** in zero-one loss.
2. By subsampling data, we fit inverse power laws to estimate the benefit of **more data** and reducing variance.
3. Using topic modeling, we identified **subpopulations** to gather more features to reduce noise.

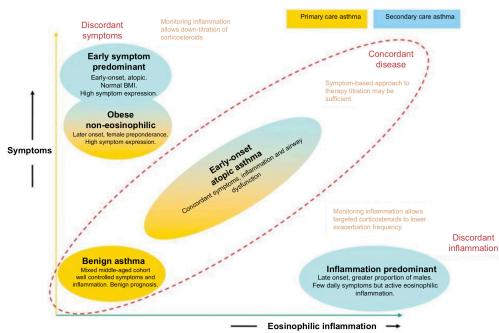
Machine Learning for Equitable Healthcare



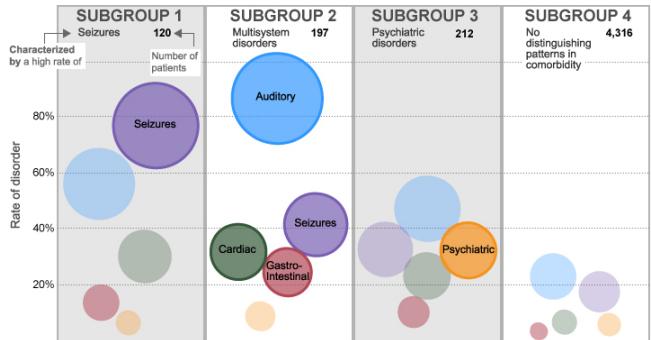
Systemic health disparities

- ▶ **Disparities in access to care**
 - ▶ Rural hospitals closing, insurance coverage, trust in healthcare system, medical adherence
- ▶ **Disparities in treatment**
 - ▶ Different treatments for same conditions, same treatments for different physiological systems
- ▶ **Disparities in outcomes**
 - ▶ Life expectancy by socioeconomic status, maternal morbidity/mortality by race

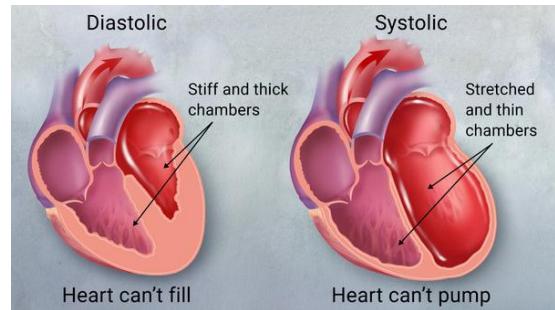
Many diseases are biologically heterogeneous despite a common diagnosis



Asthma



Autism

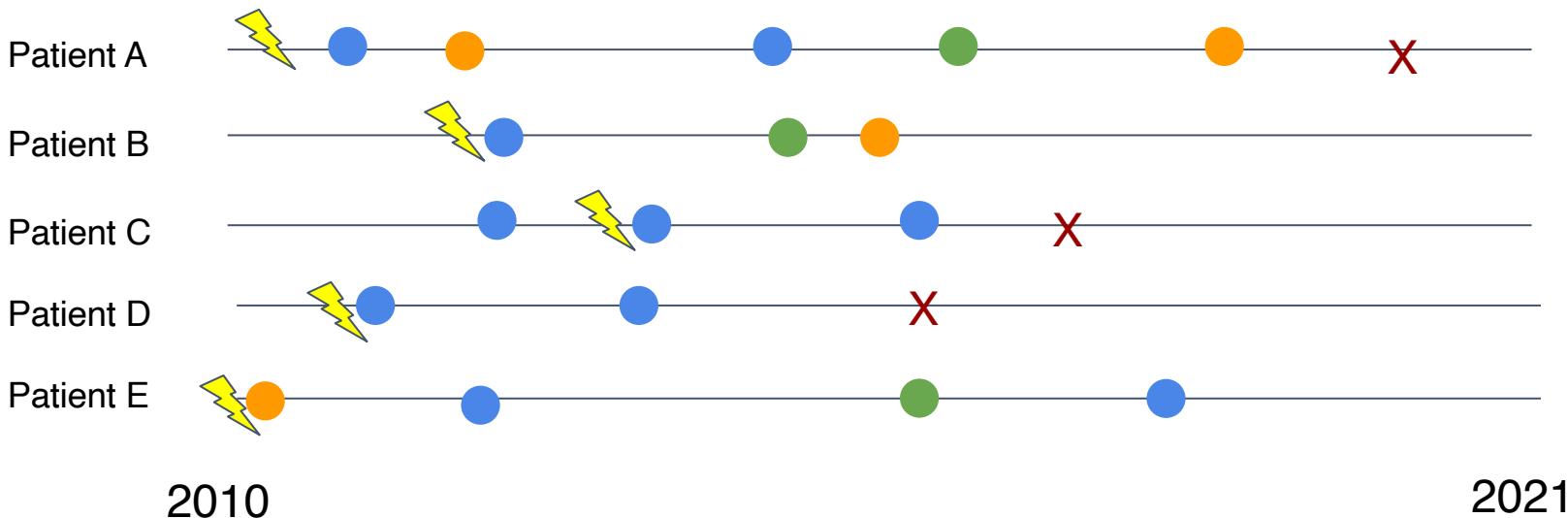


Heart Failure

- [1] Nissen et al, *Journal of Asthma and Allergy* 2018.
- [2] Kohane et al, *PLoS One*, 2012.
- [3] Mayo Clinic

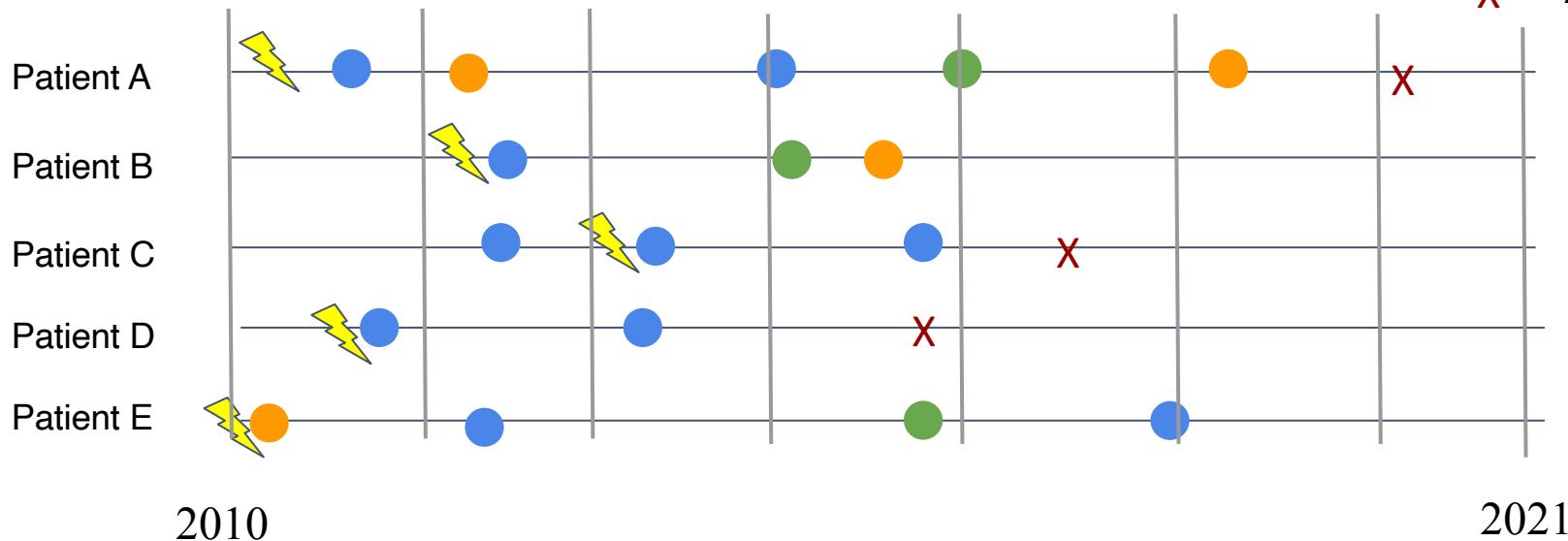
Clinical data can be sparse, multivariate, and irregularly spaced

- = Diagnosis
- = Biomarker 1
- = Biomarker 2
- = Biomarker 3
- = Adverse Event



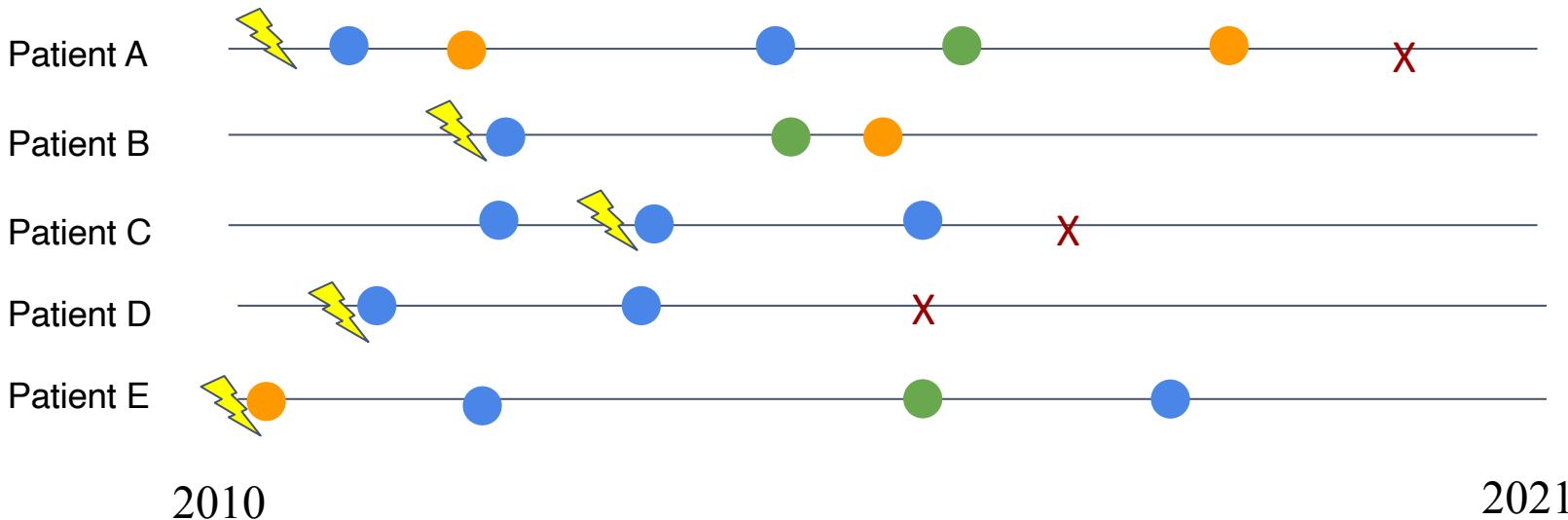
We can perform clinical prediction of adverse events.

- = Diagnosis
- = Biomarker 1
- = Biomarker 2
- = Biomarker 3
- = Adverse Event



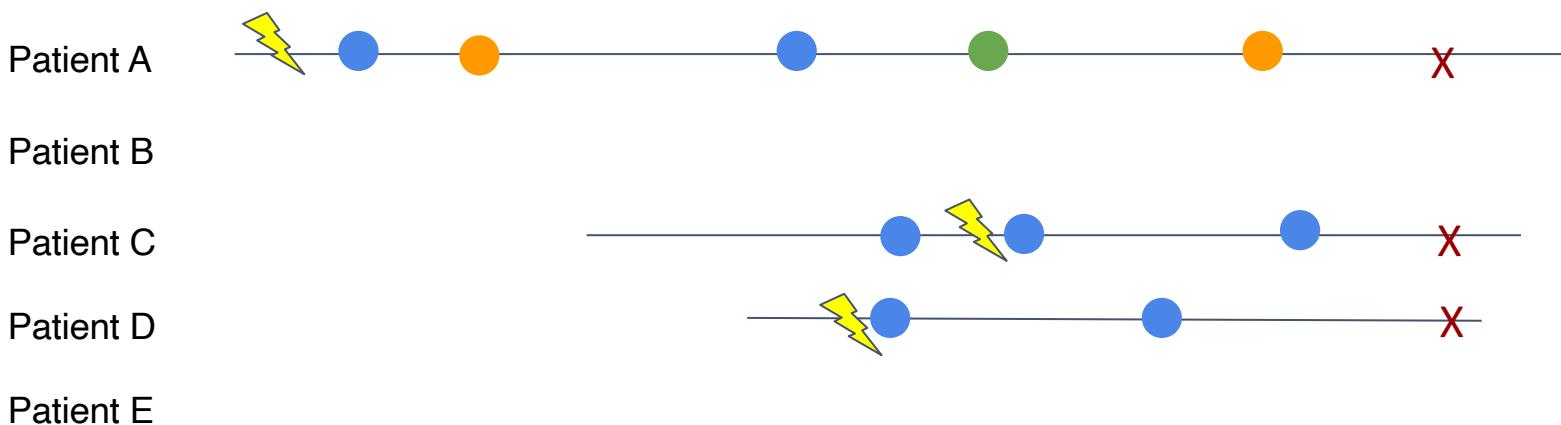
What is we wanted to learn about general disease progression?

- = Diagnosis
- = Biomarker 1
- = Biomarker 2
- = Biomarker 3
- = Adverse Event



We could align by adverse event, but this limits our dataset.

- = Diagnosis
- = Biomarker 1
- = Biomarker 2
- = Biomarker 3
- = Adverse Event



Adverse Event
Moment

Learning disease progression usually requires aligning by diagnosis.

- = Diagnosis
- = Biomarker 1
- = Biomarker 2
- = Biomarker 3
- = Adverse Event

Patient A



Patient B



Patient C



Patient D



Patient E



Time 0

Interval-censoring can introduce bias



Access to health insurance

Feb 24, 2020, 02:13pm EST | 11,106 views

1 In 4 Rural Hospitals Are At Risk Of Closure And The Problem Is Getting Worse



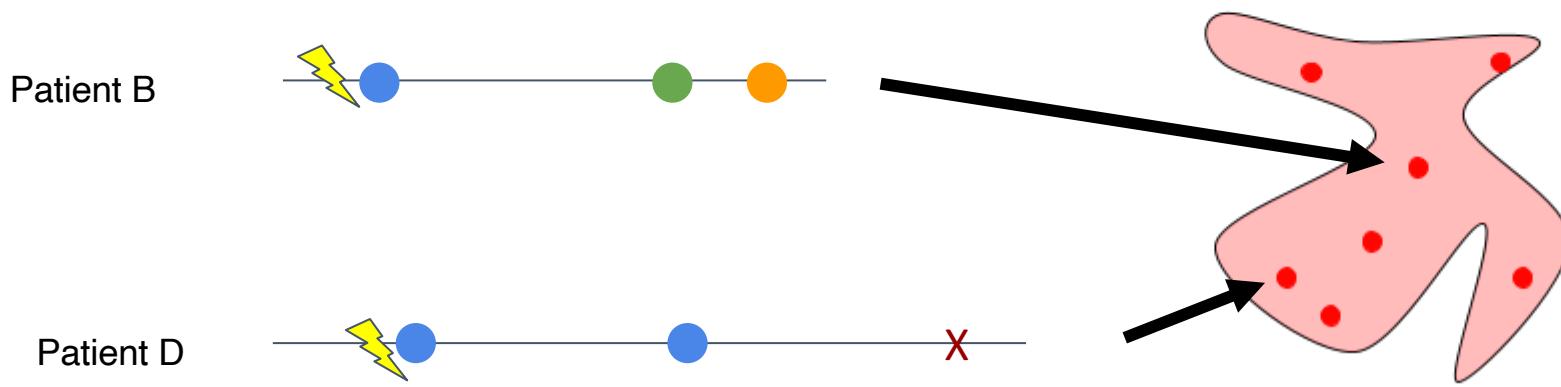
Clary Estes Former Contributor +
Healthcare

Geographic proximity to hospitals



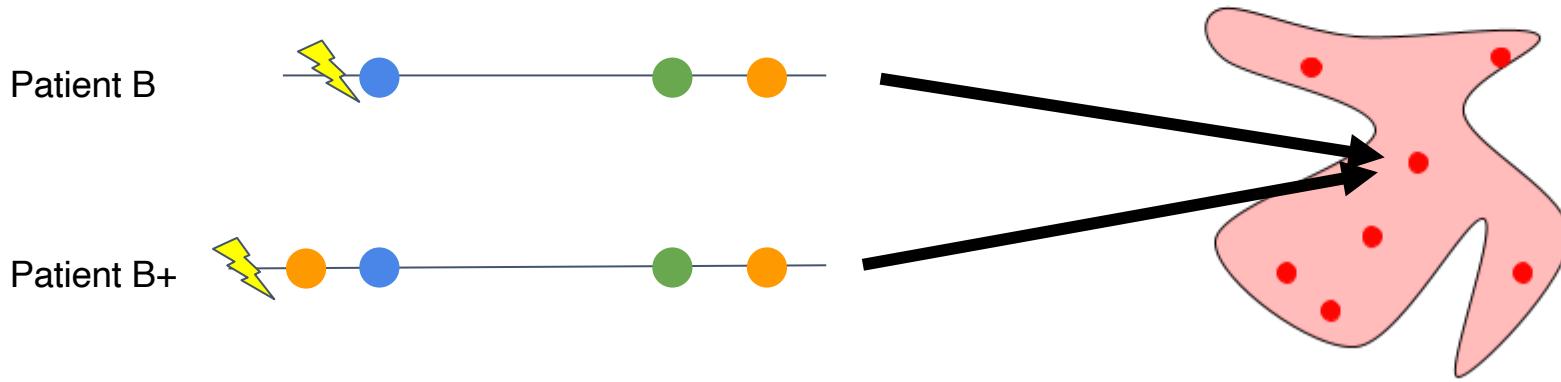
Medical mistrust

A deep generative model maps patients to a low-dimensional latent space



Patients close together are more similar.

A deep generative model maps patients to a low-dimensional latent space



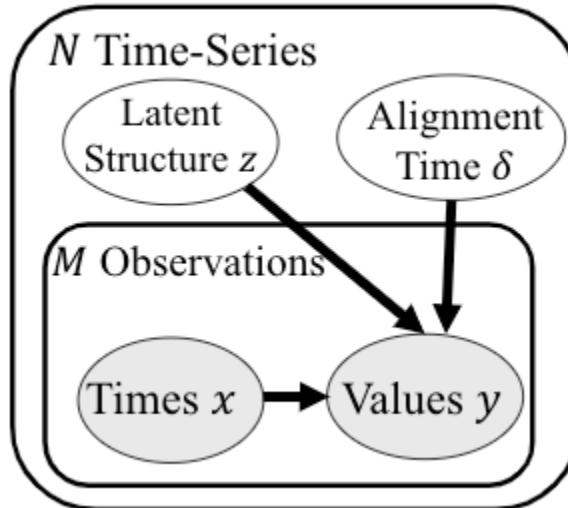
Similar patients with different left-censorship
should still be close together.

SubLign is a deep generative model to learn subtype and alignment

$$P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$$

for all $\theta_1, \theta_2 \in \Theta$.

Identifiability results
show sufficient
conditions

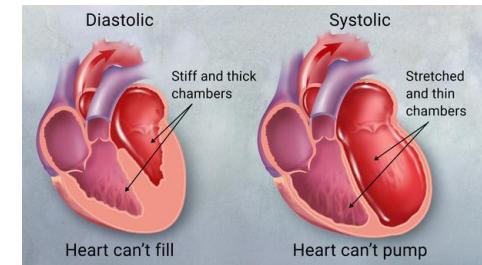


Variational inference to
approximate likelihood



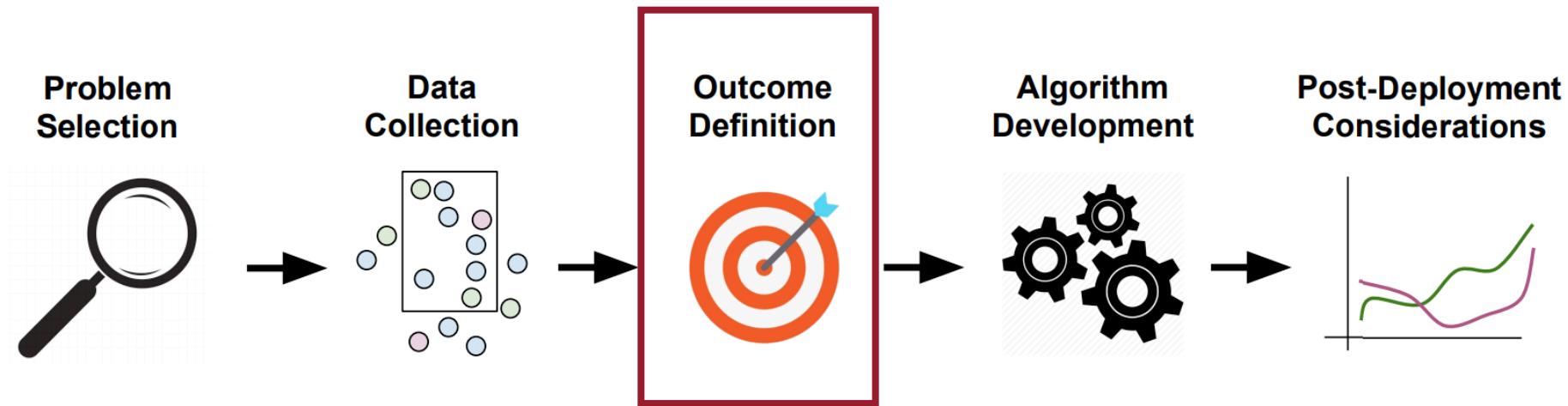
PARKINSON'S
PROGRESSION
MARKERS
INITIATIVE

Play a Part in Parkinson's Research



Experiment results
recover known clinical
findings

Machine Learning for Equitable Healthcare



Racial bias in predictive healthcare algorithm

Science

Contents ▾ News ▾ Careers ▾ Journals ▾

SHARE

RESEARCH ARTICLE



Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer^{1,2,*}, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan^{5,*†}

* See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Article

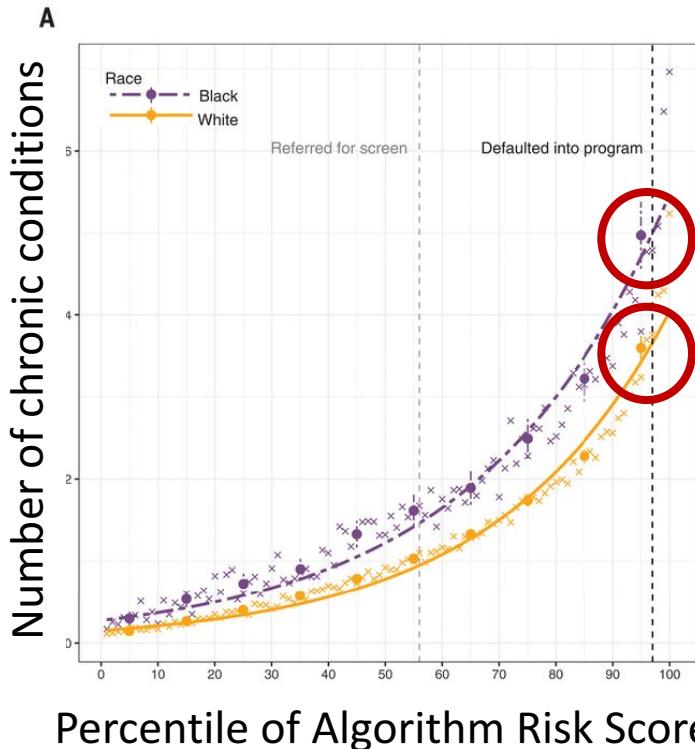
Figures & Data

Info & Metrics

eLetters

PDF

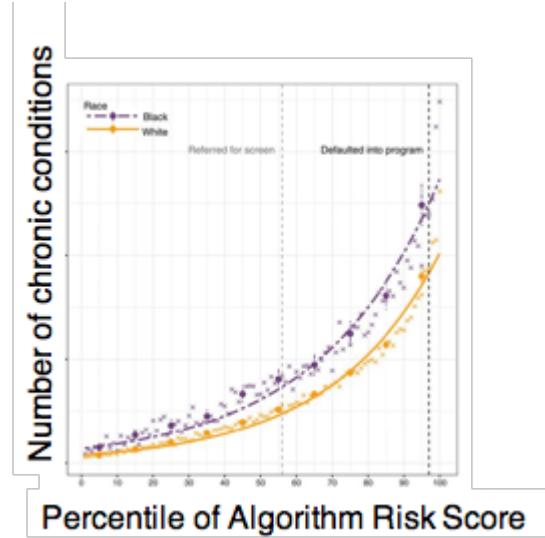
Racial bias in predictive healthcare algorithms



1. Health insurance companies identify high-risk patients for **care management**
2. **Predictive algorithms** are trained on how much patients would cost the healthcare system in the future
3. For the same percentile of algorithm risk score, white patients have **fewer chronic conditions**
4. Proposed solution is to train on **health need** instead of cost

Dissecting racial bias in an algorithm used to manage the health of populations

- ▶ **Available Data:** Risk scores and clinical data for patients in electronic health records
- ▶ **Risk scores:** Model output of prediction of whether a patient will be “high risk” in future year
- ▶ **Features:** Number of chronic conditions, measures of disease severity including hypertension and diabetes
 - ▶ Note that the actual features used for risk scores are unknown



Dissecting racial bias: Results

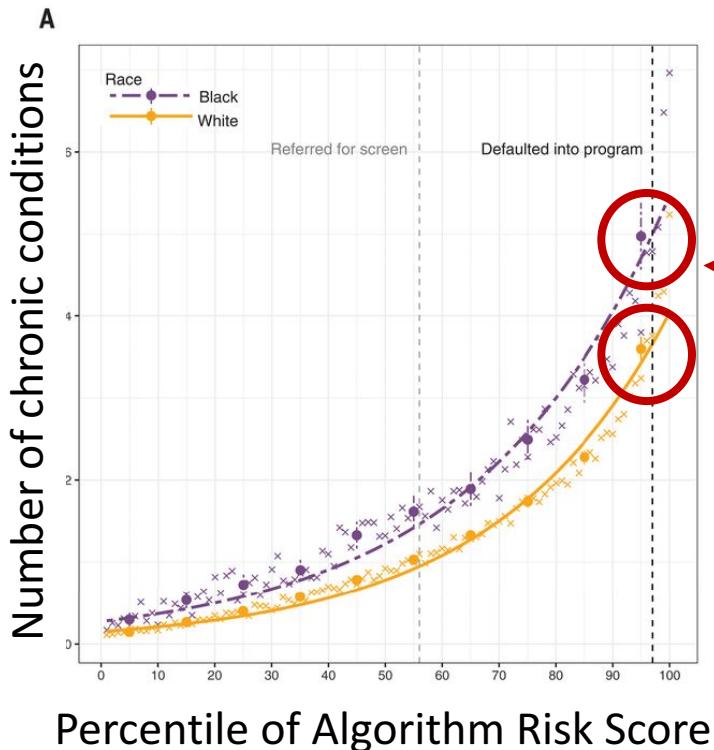


Figure 1.A.: Mean number of chronic illnesses versus algorithm-predicted risk, by race.

A person in this decile has <4 (White) or 5 (Black) chronic conditions and a risk score in the 99-percentile

Dissecting racial bias: Results

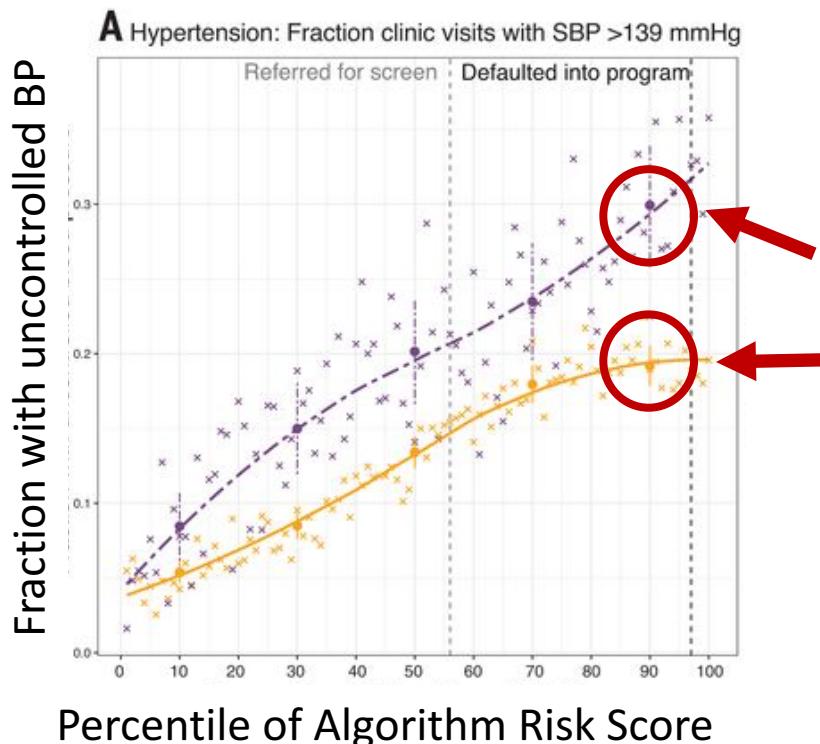


Figure 2.A.: Fraction of clinic visits with uncontrolled blood pressure.

A person in this decile has a 30% chance (Black) or <20% (White) chance of having hypertension for the same risk score.

Similar analysis conducted for diabetes, renal failure, anemia, and cholesterol based from extracted values in electronic health records.

Dissecting racial bias: Proposed Fix

Increase the fraction of Black patients in highest risk group from 14% to 26%

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		
1) Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141 (0.003)
2) Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210 (0.003)
3) 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

Table 2: Results from L1-regularized logistic regression for three different labels.

2019 Paper Aftermath

- ▶ **Press:** The paper was covered widely across news outlets
- ▶ **Policy:** Senators Ron Wyden and Cory Booker addressed letters to CMS and FTC asking for information
- ▶ **Industry vigilance:** Significantly more collaboration and interest from insurance companies on algorithmic fairness

United States Senate
WASHINGTON, DC 20510
December 3, 2019

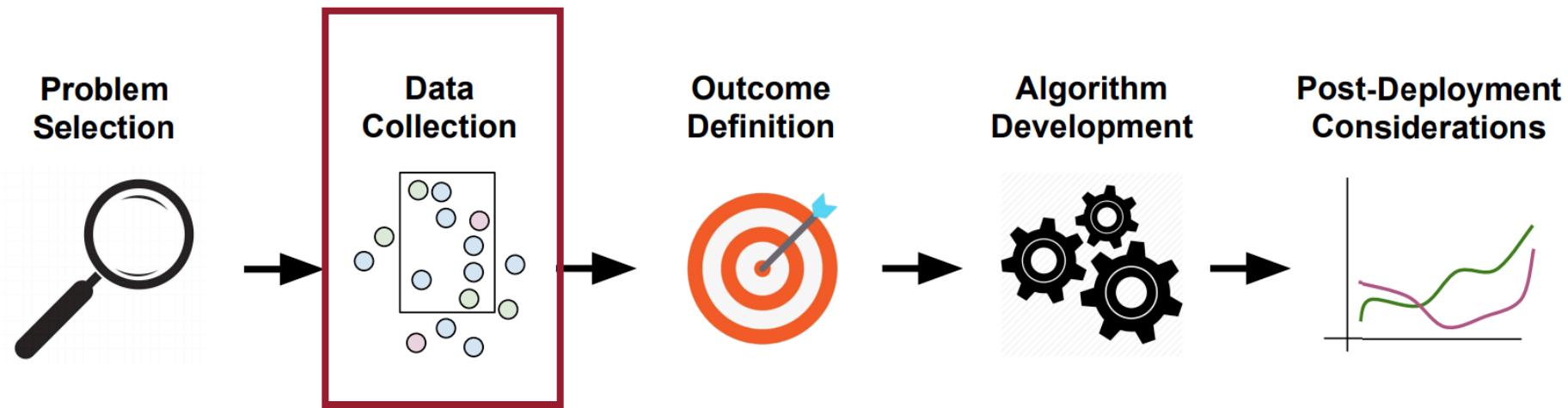
The Honorable Seema Verma
Administrator
The Centers for Medicare & Medicaid Services
Department of Health & Human Services
Room 445-G, Hubert H. Humphrey Building
200 Independence Ave., S.W.
Washington, DC 20201

Dear Administrator Verma:

We write today to request information regarding any actions that the Centers for Medicare & Medicaid Services (CMS) is taking or plans to take to assess the potential for algorithms used throughout the health care system to perpetuate biases.

Algorithms are increasingly embedded into every aspect of modern society, including the health care system. Organizations use automated decision systems, driven by technologies ranging from advanced analytics to artificial intelligence (AI), to organize and optimize the complex choices they need to make on daily basis. CMS and commercial health insurers have begun to explore ways to incorporate algorithms that automate decisions like predicting health care needs and outcomes, targeting resources, improving quality of care, and detecting waste, fraud, and abuse.

Machine Learning for Equitable Healthcare



How can data collection be biased?

- ▶ Group membership can be absent
 - ▶ Canada and France do not record race and ethnicity in nationalized health databases (Leonard, *Humanity and Society* 2014)
- ▶ Data can be imbalanced
 - ▶ Acute kidney injury model trained on 6.4% female dataset (Tomasev et al, *Nature* 2019)

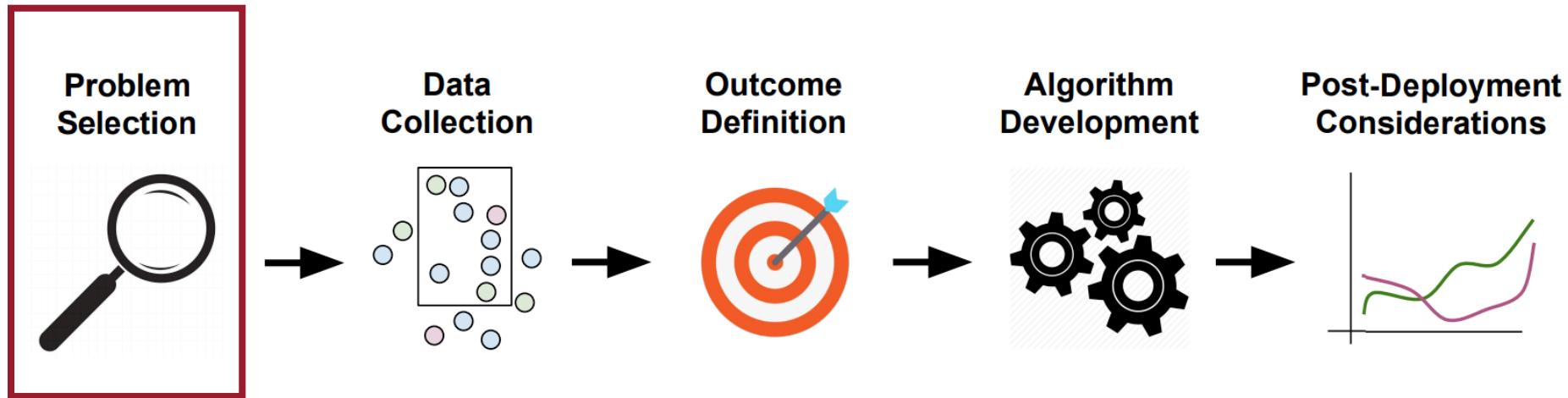
Heterogeneous Data Losses

- ▶ Randomized Controlled Trials
 - ▶ In 24 of 31 most recent cancer drugs, fewer than 5% of study participants were black (Wong, *Propublica* 2019)
 - ▶ 94% of adult asthmatics would not be eligible for trials (Travers et al, *Thorax* 2007)
- ▶ Electronic Health Records
 - ▶ MIMIC dataset has 71% White patients, 9% Black, 3% Hispanic, and 2% Asian

Population-specific Data Losses

- ▶ Low- and Middle- Income Nationals
 - ▶ 9 of 46 member states in Sub-Saharan Africa had death statistics about burden of disease (Jamison et al, *World Bank Publication 2006*)
- ▶ Transgender and Gender Non-conforming Individuals
- ▶ Undocumented Immigrants
- ▶ Pregnant Women

Machine Learning for Equitable Healthcare



How can we detect IPV victims early?



Half of all women killed globally are killed by intimate partners or family.¹



IPV victims report higher rates of clinical visits.²

nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva , Brett Kuprel , Roberto A. Novoa , Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 

Nature 542, 115–118 (02 February 2017) | Download Citation 

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney , Marcin Sieniek, [...] Shravya Shetty 

Nature 577, 89–94(2020) | Cite this article

53k Accesses | 164 Citations | 3524 Altmetric | Metrics

Algorithms can screen patients with performance that exceeds humans.

1. U. N. O. on Drugs and Crime, Global Study on Homicide: Gender-related Killing of Women and Girls (UNODC, United Nations Office on Drugs and Crime, 2018).

2. C. Wisner, T. Gilmer, L. Saltzman and T. Zink, Intimate partner violence against women, *Journal of family practice* 48, 439 (1999).

How do we get accurate IPV labels?

- Biggest barrier to early intervention is **underreporting** by the patient because of shame, economic dependency, or lack of trust in healthcare providers
- IPV victims use healthcare services like the **emergency department** or **imaging studies** at higher rates than other patients
- We examine 1,479 victims and control patients at Brigham and Women's Hospital (BWH) in Boston

Chen et al, "Intimate Partner Violence and Injury Prediction from Radiology Reports," *PSB* 2021.

What kind of labels could we use?

1. **ICD codes**: Based on clinical staff assessment
2. **Patient self-report**: Based on patient enrollment in violence prevention program
3. **Radiologist labeling**: Based on injuries in radiology reports

1) Self-report labels

► Inclusion Criteria

- ▶ IPV victims: Identified as entering a violence prevention program at BWH, for IPV, with at least one radiology study at BWH
- ▶ Control cohort: Age- and sex-matched patients in the BWH patient population with at least one radiology study at BWH

► Features

- ▶ Radiology report text, extracted from template

► Label

- ▶ Was this person a **self-report to the BWH violence prevention program?**

Passageway – Domestic Abuse Intervention and Prevention

CCHHE's Passageway program works to improve the health, wellbeing, and safety of those experiencing abuse from an intimate partner. We offer the following support services to hospital and health center patients, employees, and community members:

- Free and confidential advocacy services*
- Safety planning
- Individual counseling and support
- A safe place to talk
- Information about the health effects of domestic violence
- Support groups
- Medical advocacy
- Legal and court advocacy
- Referrals to community resources (health care, housing, shelter, lawyers, and others)



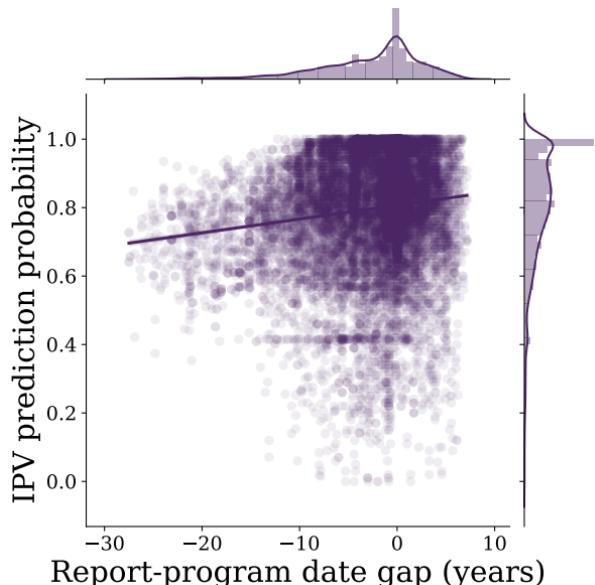
2) Radiology injury label

- ▶ **Inclusion Criteria**
 - ▶ Data from BWH
- ▶ **Features**
 - ▶ Radiology report text, extracted from template
 - ▶ Each report text treated as separate
- ▶ **Label**
 - ▶ **Fellowship-trained emergency radiologists** provided injury labels

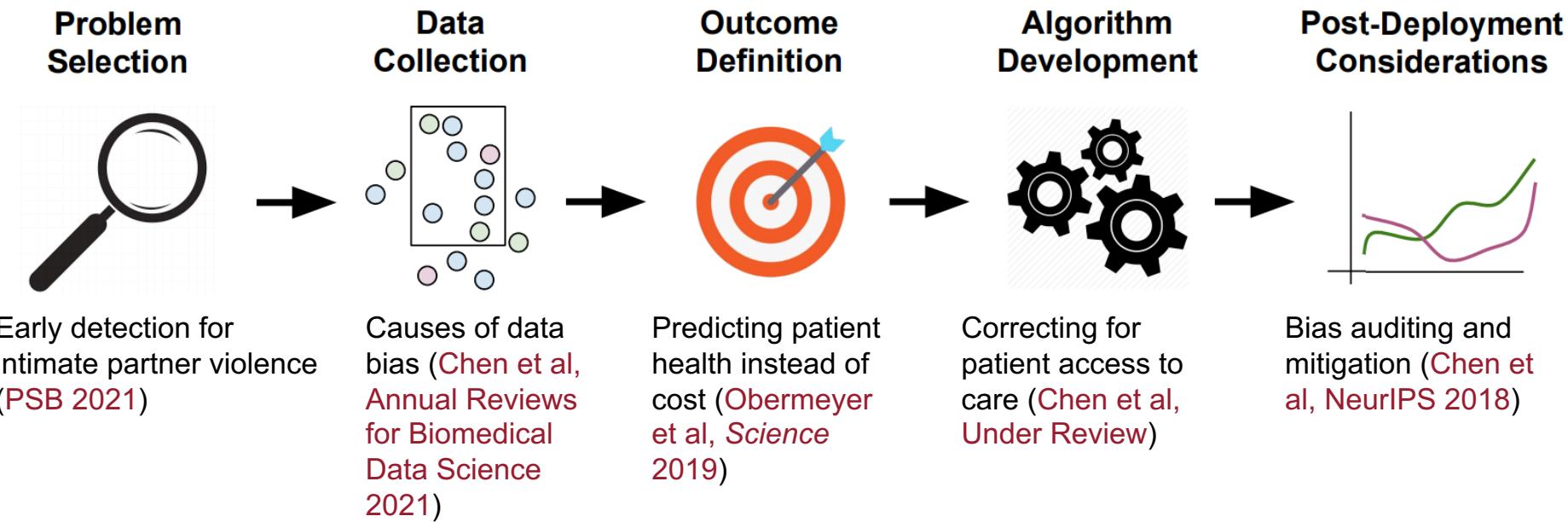


How do predictions differ on the two label sets?

- ▶ Models performance for both labels are **comparable**
 - ▶ Self-report label: 0.84 ± 0.03
 - ▶ Radiologist label: 0.87 ± 0.01
- ▶ We can use **self-report labels**, which are much less time intensive than radiologist labels.
- ▶ We can detect IPV a **median of 3.08 years** before program entry (sensitivity 64%, specificity 95%)



Machine Learning for Equitable Healthcare



Challenge to the community

1. **Expand beyond mathematical definitions.** Consider historical and systemic causes to define and fix health disparities.



Challenge to the community

1. **Expand beyond mathematical definitions.** Consider historical and systemic causes to define and fix health disparities.
2. **Seek and promote different perspectives.** Interdisciplinary work and a more diverse research community bring more people to the table.



Challenge to the community

1. **Expand beyond mathematical definitions.** Consider historical and systemic causes to define and fix health disparities.
2. **Seek and promote different perspectives.** Interdisciplinary work and a more diverse research community bring more people to the table.
3. **Aim for higher fruit.** Short-term clinical prediction is only the first step in improving the healthcare system.

