

Data 146: Foundations for CPH

Fairness, Bias, and Ethics

Irene Y. Chen


PUBLIC HEALTH
SERVICE -
- CLINIC HOURS
U.S. PHS STUDY
TUSKEGEE
STUDY
TUSKEGEE
STUDY

When is scientific progress unethical? How can we protect the rights of the underserved?

Outline

- **Problem Selection** (10 mins)
- **Data Collection** (10 mins)
- **Outcome Definition** (5 mins)
- **Algorithm Development** (10 mins)
- **Post-Deployment Considerations** (5 mins)



How can we make Data 146 better for you?

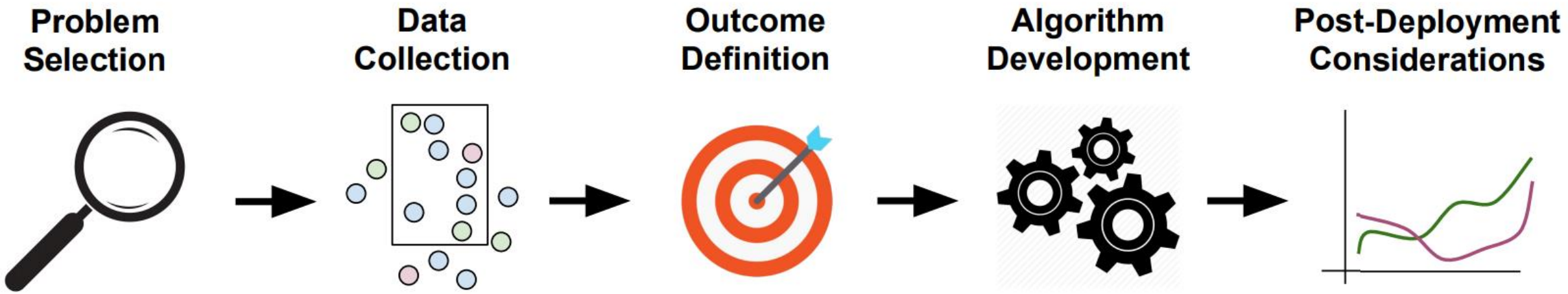
Learning Objective: Understand how questions of ethics enter stages of ML pipeline

Healthcare as-is is not equitable

- Black women are **3.2 times** as likely as White women to die from maternal causes
  - This is not explained by education!
- Life expectancy differs by almost **15 years** across racial groups (Asian vs American Indian or Alaska Native).
- Lower-income adults are nearly **5 times** as likely to be uninsured.

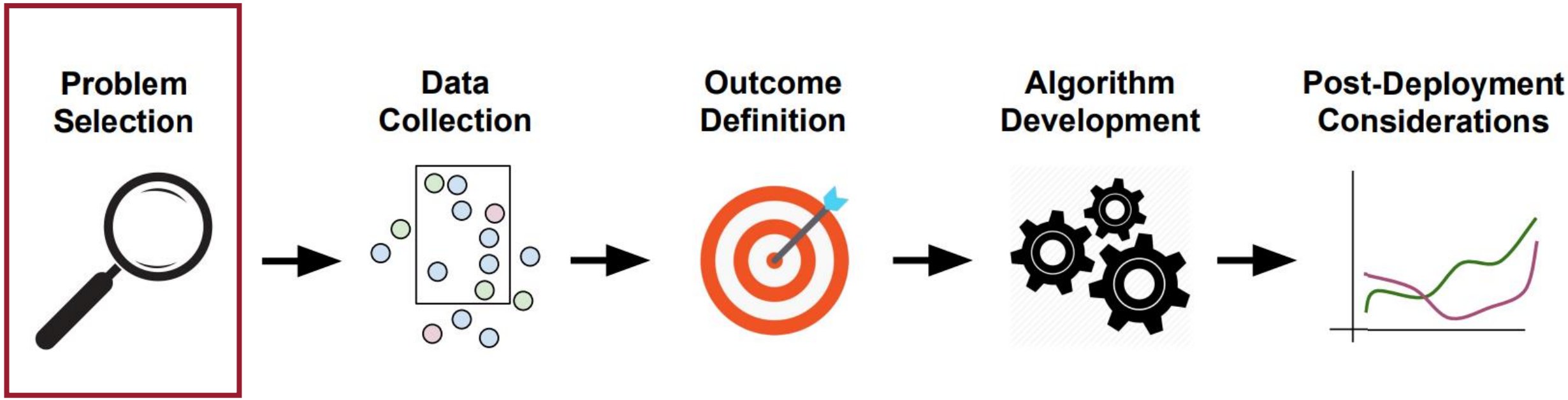
[1] National Center for Health Statistics. (2026). *Maternal Mortality Rates in the United States, 2024*. CDC, Health E-Stats No. 113.
[2] National Center for Health Statistics. (2026). *United States Life Tables, 2024*. National Vital Statistics Reports, 75(5).
[3] National Center for Health Statistics. (2026). *Health Insurance Coverage: Early Release of Estimates From the National Health Interview Survey, 2025*. CDC.

Machine Learning for Equitable Healthcare

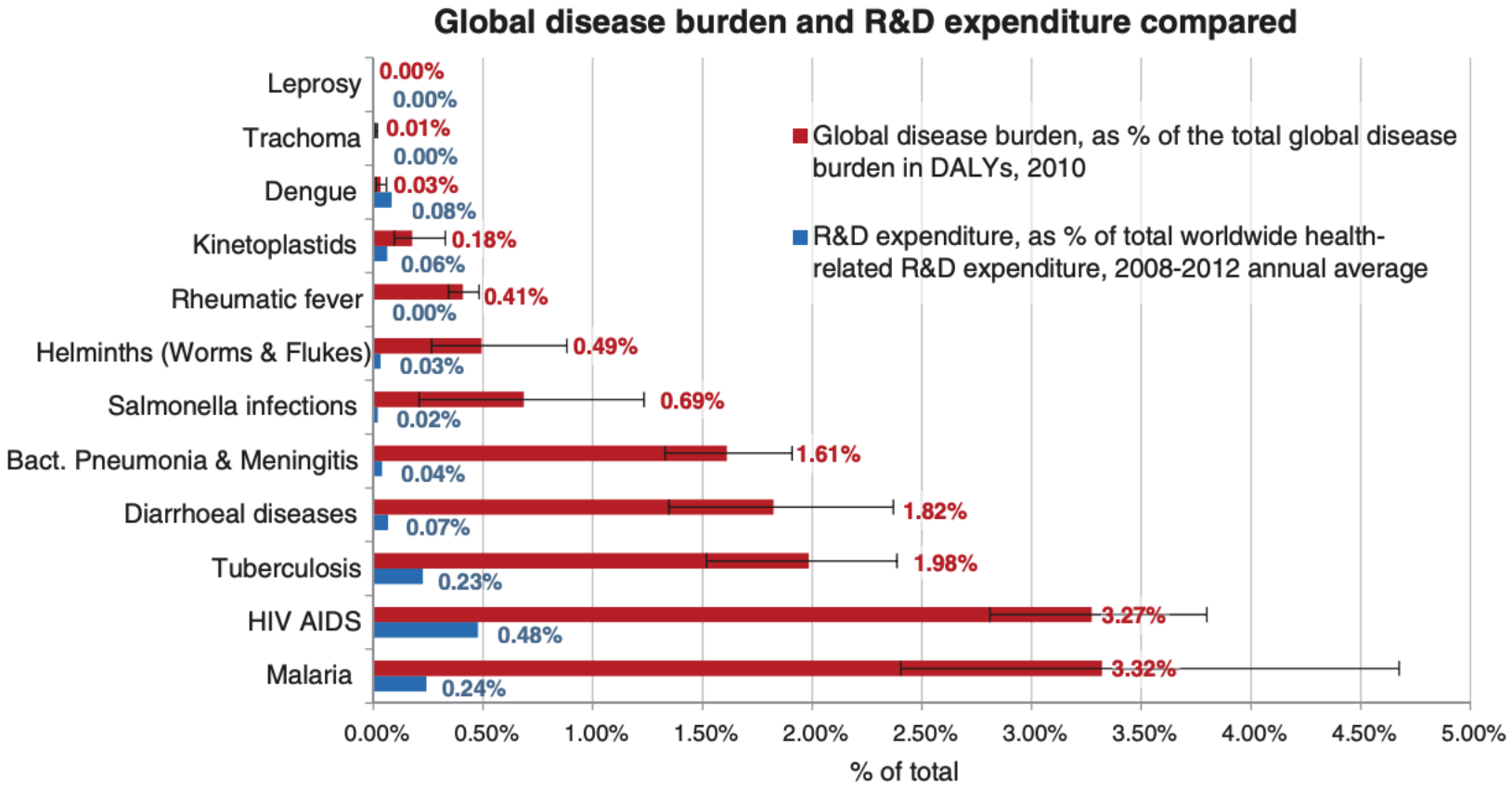


Chen et al, “Ethical Machine Learning for Health Care,” *Annual Reviews for Biomedical Data Science 2021.*

Machine Learning for Equitable Healthcare



Global Health Injustice



[Philipsborn et al, Global Health Action 2014.]

Racial Injustice

- Sickle cell disease (SCD) vs cystic fibrosis (CF)
  - Both are rare inherited blood disorders
  - SCD affects Black individuals at a higher rate (1 in 365)
  - CF affects White individuals at a higher rate (1 in 2500)
  - SCD affects 90k individuals; CF affects 30k individuals (2008-2018)
- Large funding and research differences
  - NIH funding for CF was 3.4x larger than SCD from 2008-2016
  - Private funding for CF was 971x larger for SCD than SCD
  - There are 1.8x more PubMed publications for CF compared to SCD

[Farooq et al, Health Policy 2020.]

Gender Injustice

- Prevalent health conditions affecting women are understudied
  - Endometriosis affects 10% of women
  - Menstrual cycle is not logged in studies
  - Until 2019, NIH did not have R01 (original grant mechanism) related to sex or gender on health and disease
- Male mental health is understudied
  - Emphasis on between-genders may be obscuring research on within-gender differences

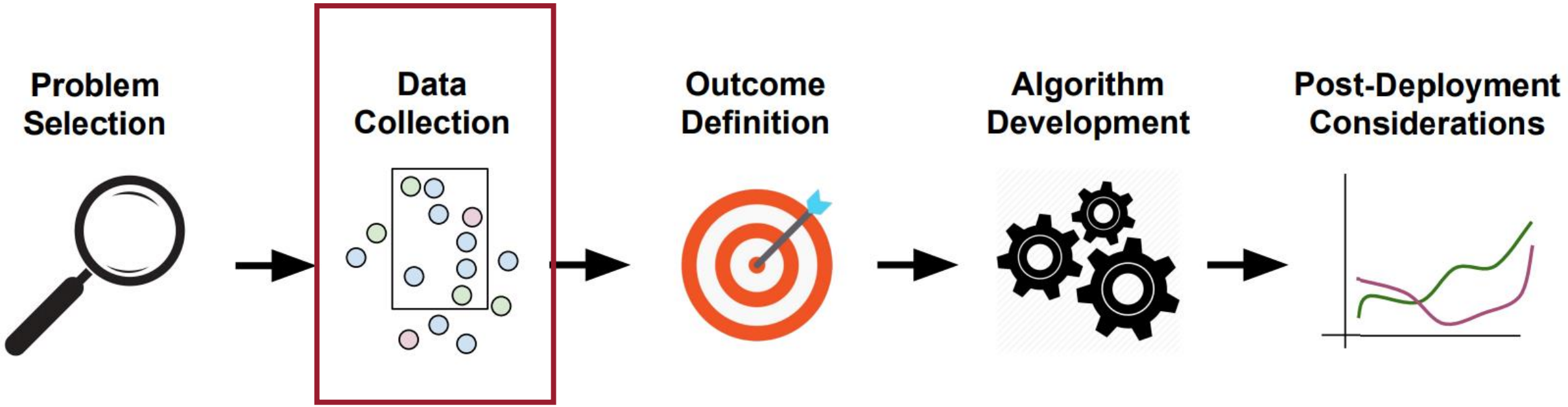
[Eisenberg et al, BJOG; NIH 2020; Smith et al, American Journal of Men's Health 2018.]

Diversity in the Workforce

- 1.2 million US doctoral recipients from 1977 to 2015
- Underrepresented groups produce higher rates of scientific novelty, likely due to focusing on understudied areas
- Novel contributions by gender and racial minorities are taken up by other scholars at lower rates than novel contributions by gender and racial majorities
- Equally impactful contributions of gender and racial minorities are less likely to result in successful scientific careers than for majority groups

[Hofstra et al, Diversity-Innovation Paradox in Science, PNAS 2020.]

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)

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

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

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

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

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

Using Search Queries to Understand Health Information Needs in Africa

- Why is health data important?
  - Patients use health data to determine when to seek treatment
  - Clinicians use data to make clinical decisions
  - Organizations use data to determine where resources should be allocated
- Problem: health data is not always readily available
  - ”Unless gaps are identified early and accurately, simply providing a raft of general interventions will not meet the real health needs of the people in the Region” – 2014 African Union regional report

Images Videos Maps News | MSN Office Online Outlook.com

Bing

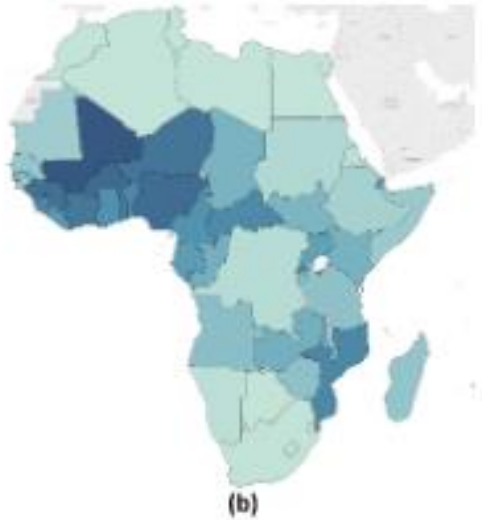
54 African nations; Jan 2016-June 2017

Containing “HIV” or “AIDS”



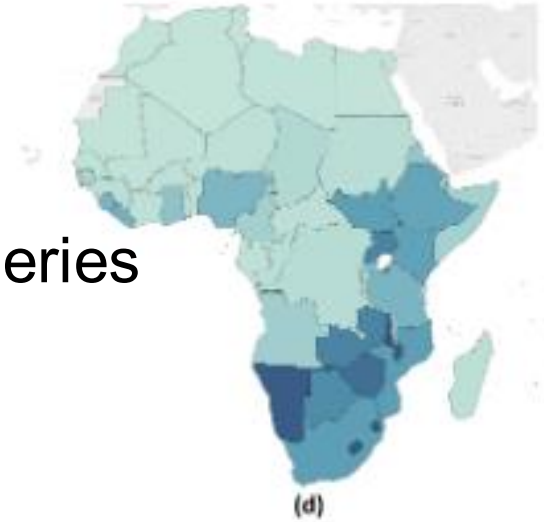
World Bank Data







Bing search queries







HIV/AIDS

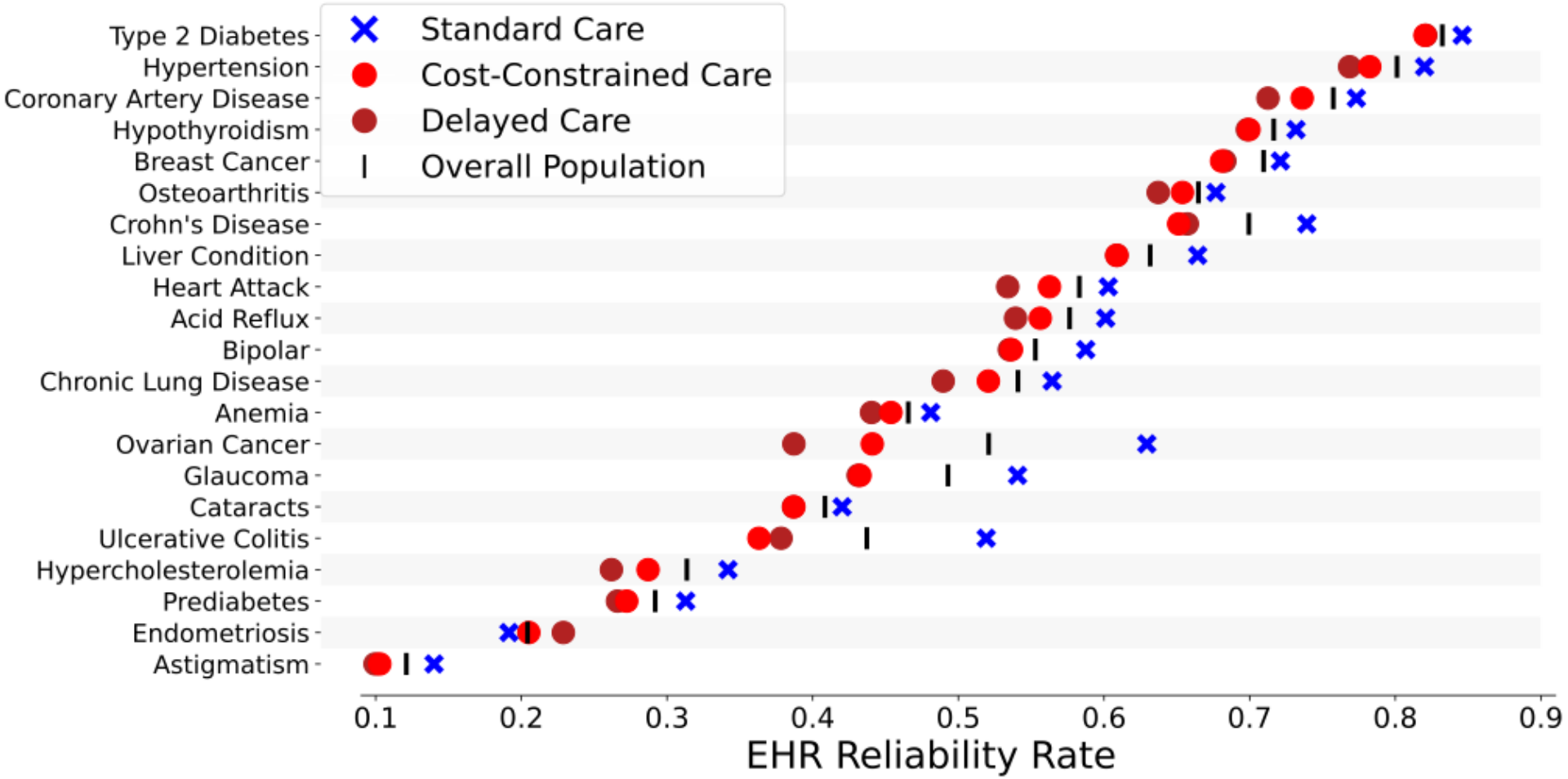
Malaria

Tuberculosis

Disease	Topic	20 Most Representative Words	Sample Queries from Top 100
HIV/ AIDS	*Symptoms* (2.28%)	pain, sign, lymph, swollen, nodes, sore, symptom, symptoms, throat, infection, body, back, positive, pains, stomach, fever, neck, headache, glands, patient	hiv painfull jaw hiv swollen lymph nodes hiv swollen gland throat
	Natural Cure (0.74%)	cure, oil, black, healing, heal, healed, seed, herbs, natural, cures, moringa, kill, cured, testimonials, coconut, traditional, god, garlic, lemon, aloe	prophet bushiri hiv miracles hiv garlic lemon honey coloidal silver hiv testimonials
	Epidemiology (0.59%)	statistics, report, 2015, global, unaids, 2016, united, epidemic, besigye, kizza, children, 2014, progress, 2010, response, nations, nigeria, prevalence, million, sa	unaids global aids report mia khalifa hiv hiv 2030
	Drugs (0.85%)	drug, treatment, patients, abuse, therapy, drugs, resistance, antiretroviral, substance, adherence, alcohol, art, failure, spread, leads, patient, relationship, transmission, effect	stanford hiv drug resistance hiv drug therapy resistance virological failure hiv
	Breastfeeding (0.66%)	positive, baby, mother, breastfeeding, breast, mothers, child, born, feeding, babies, birth, give, breastfeed, infant, infected, feed, milk, pregnant, safe, exposed	hiv exclusive fomular feeding exclusive breast feeding and hiv hiv mom can breast feed baby
	Stigma (0.46%)	stigma, issues, discrimination, related, ethical, legal, prevention, safety, pdf, workplace, relating, precaution, work, dies, surrounding, universal, reduce, address	hiv aids ethical dilema safty issues relating to hiv-aids aids stigma in garissa

Latent Dirichlet Allocation topics can find structure in unstructured data

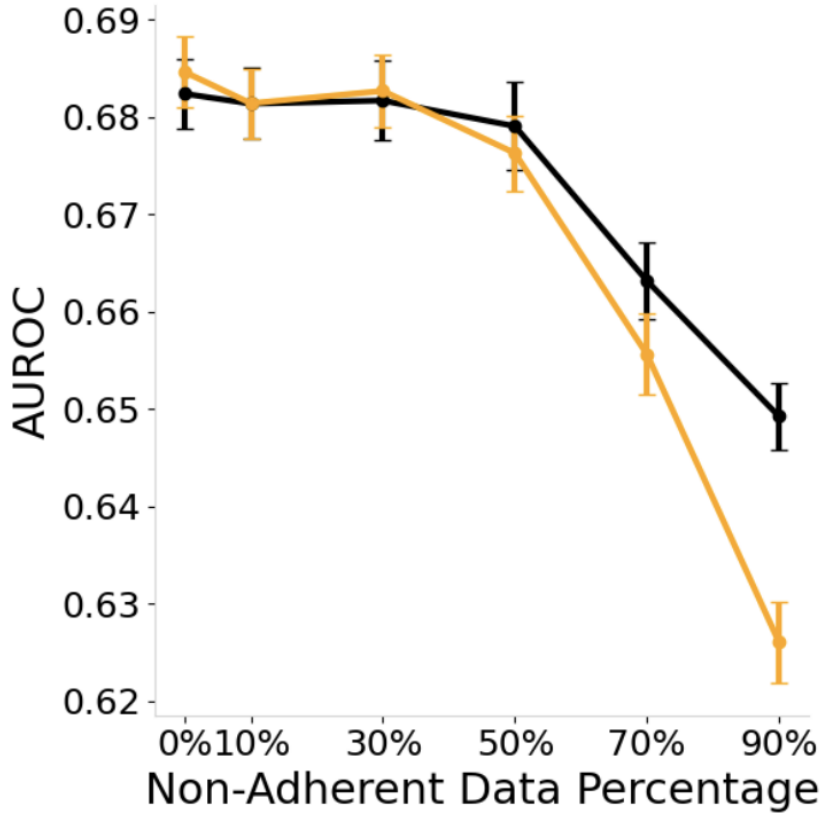
Access to healthcare affects data reliability



Zink et al, Nature Health 2025.

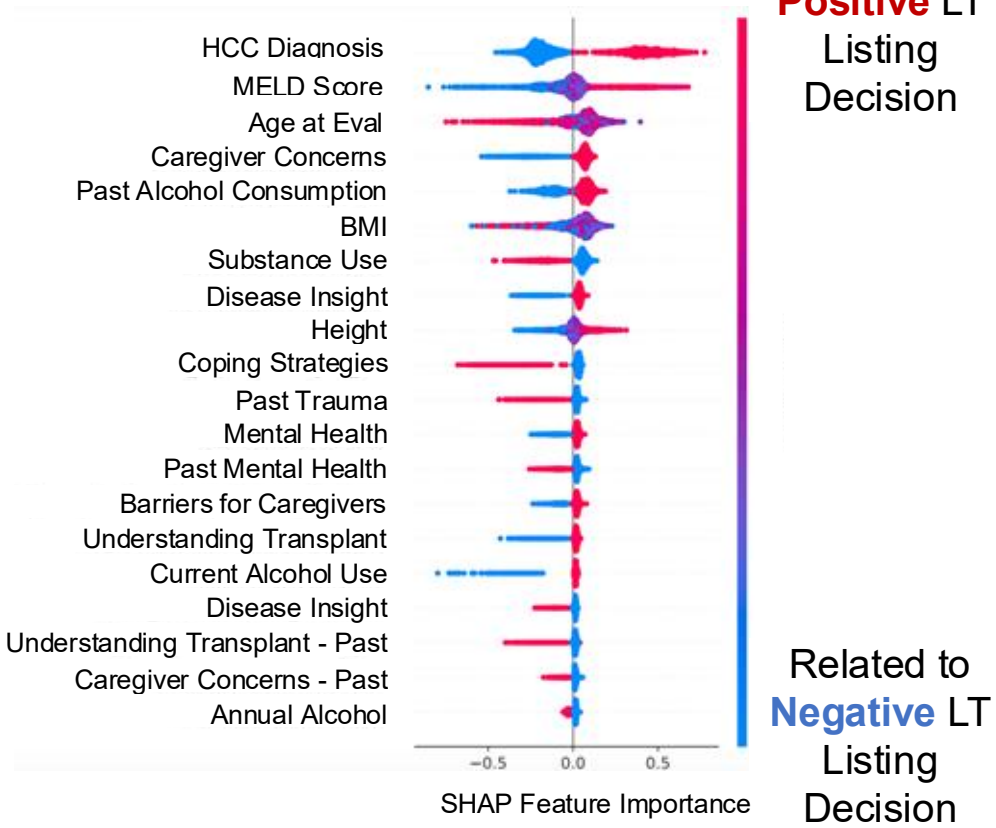
LLMs to Correct Non-Adherence Bias

- **40-50%** of patients fail to consistently take hypertension medication
- We correct patient records using LLMs to find mentions of **non-adherence** in clinical notes
- Our method improves prediction performance and **causal inference** results



Liver transplant listing decisions

- **54%** of UCSF patients referred for transplant eligibility are placed on transplant waiting lists
- Using **LLMs**, we extract features from clinical notes
- We model the **effect of SDOH** factors on listing decisions

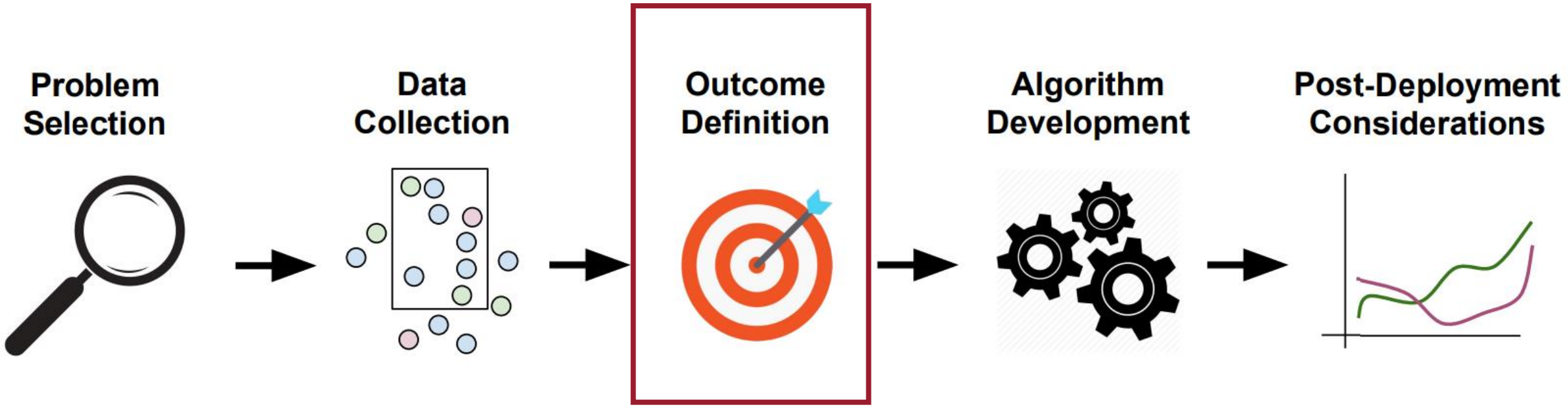


Robitshek et al, *npj Digital Medicine* 2025; **Leonidas H. Berry Health Equity Research Award**, ACG 2024; **Best Paper**, ML for Health Symposium 2024

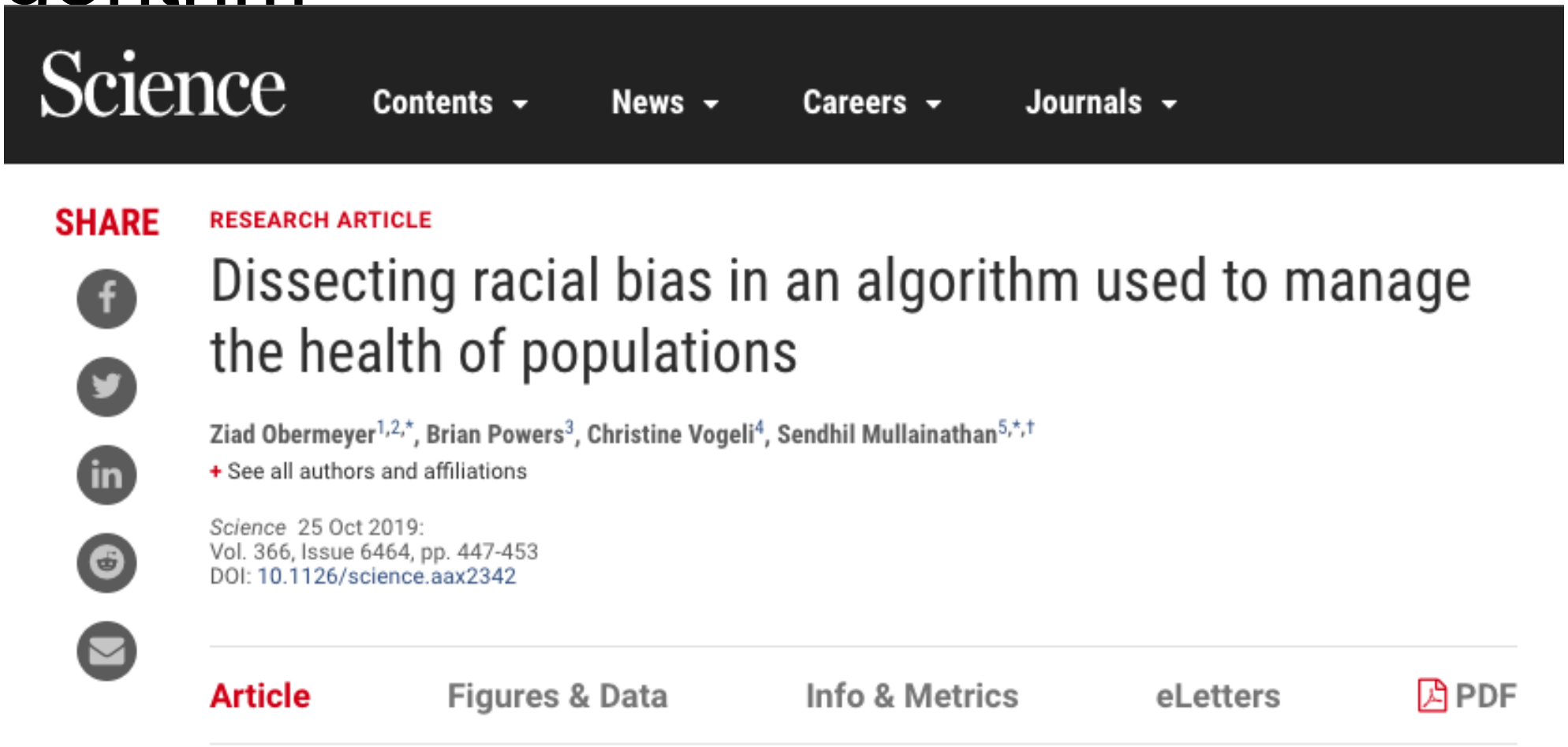
Data collection discussion questions

- More and more types of data are becoming available. How can we think about the privacy and ethics of using wider health data?
  - Wearable (e.g., Fitbit) data
  - Search query logs
  - Social media posts
  - Mobility data
- Part of data collection involves individual consent and also that individuals benefit from the data.

Machine Learning for Equitable Healthcare

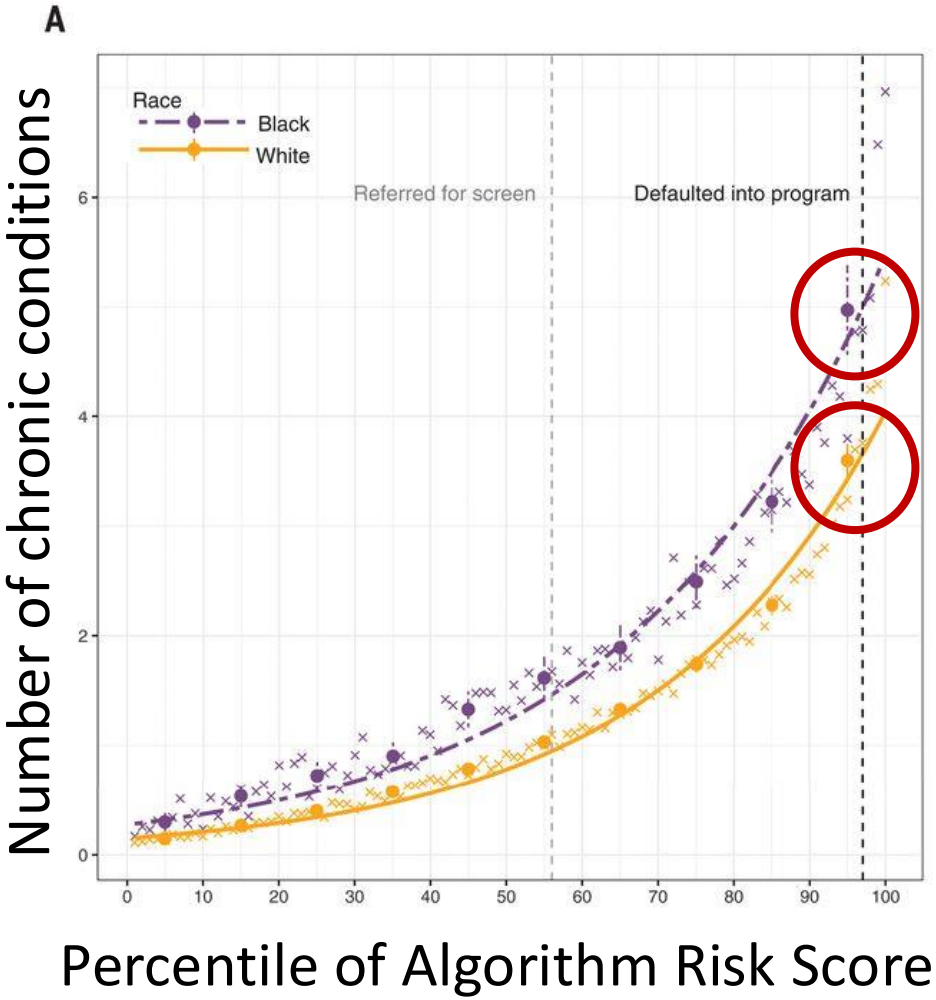


Racial bias in predictive healthcare algorithm



Obermeyer et al, 2019, “Dissecting racial bias in an algorithm …”, *Science*.

Racial bias in predictive healthcare algorithms

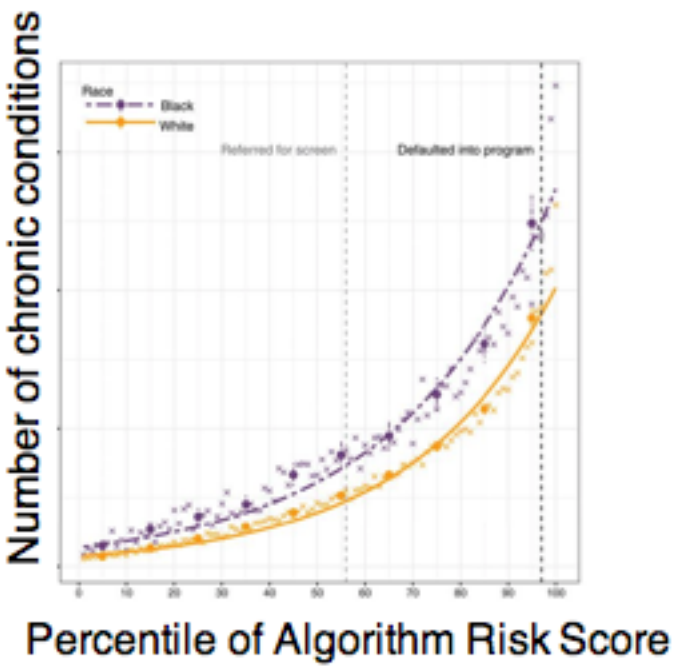


1. Health insurance companies identify high-risk patients for **care management**
2. **Predictive algorithms** are trained on how much patients would cost the healthcare system in the future
3. For the same percentile of algorithm risk score, white patients have **fewer chronic conditions**
4. Proposed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

Obermeyer et al, 2019, "Dissecting racial bias in an algorithm …", *Science*.

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

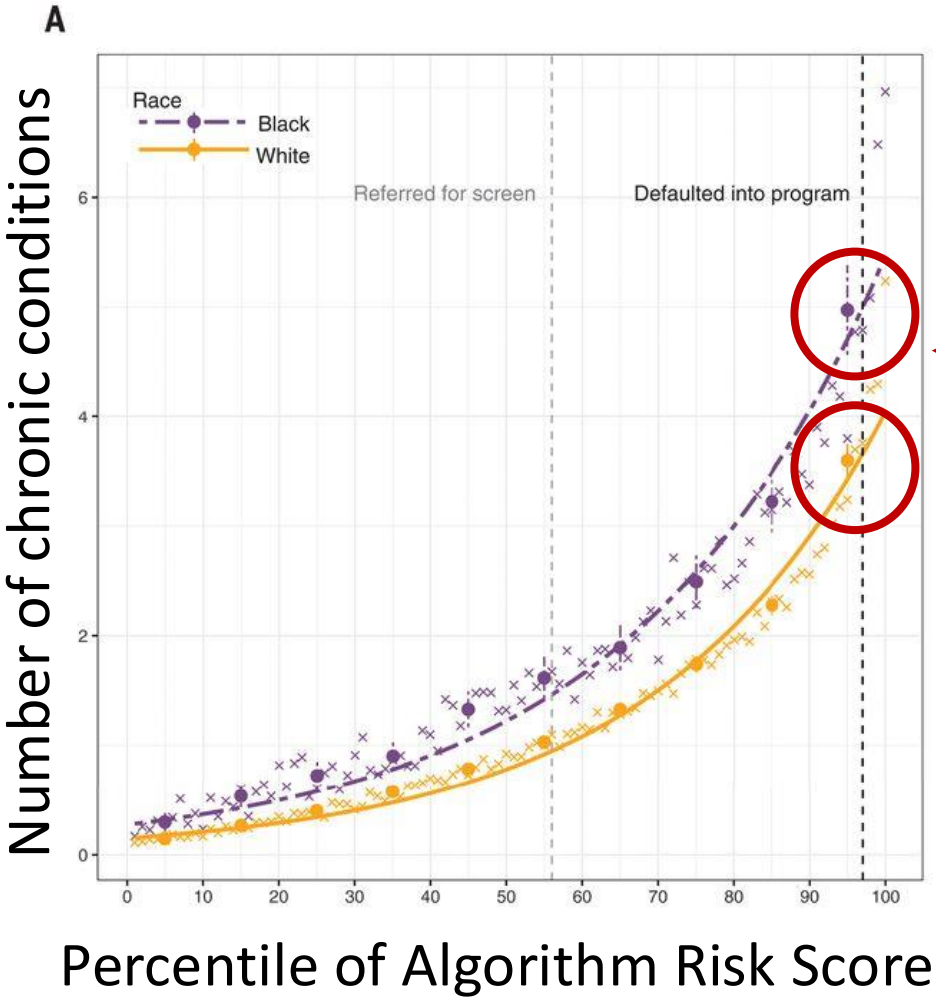


Figure 1.A.: Mean number of chronic illnesses versus algorithm-predicted risk, by race.

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

Slide: Stephanie Gervasi. **Figure:** Obermeyer et al, 2019. *Science*.

Dissecting racial bias: Results

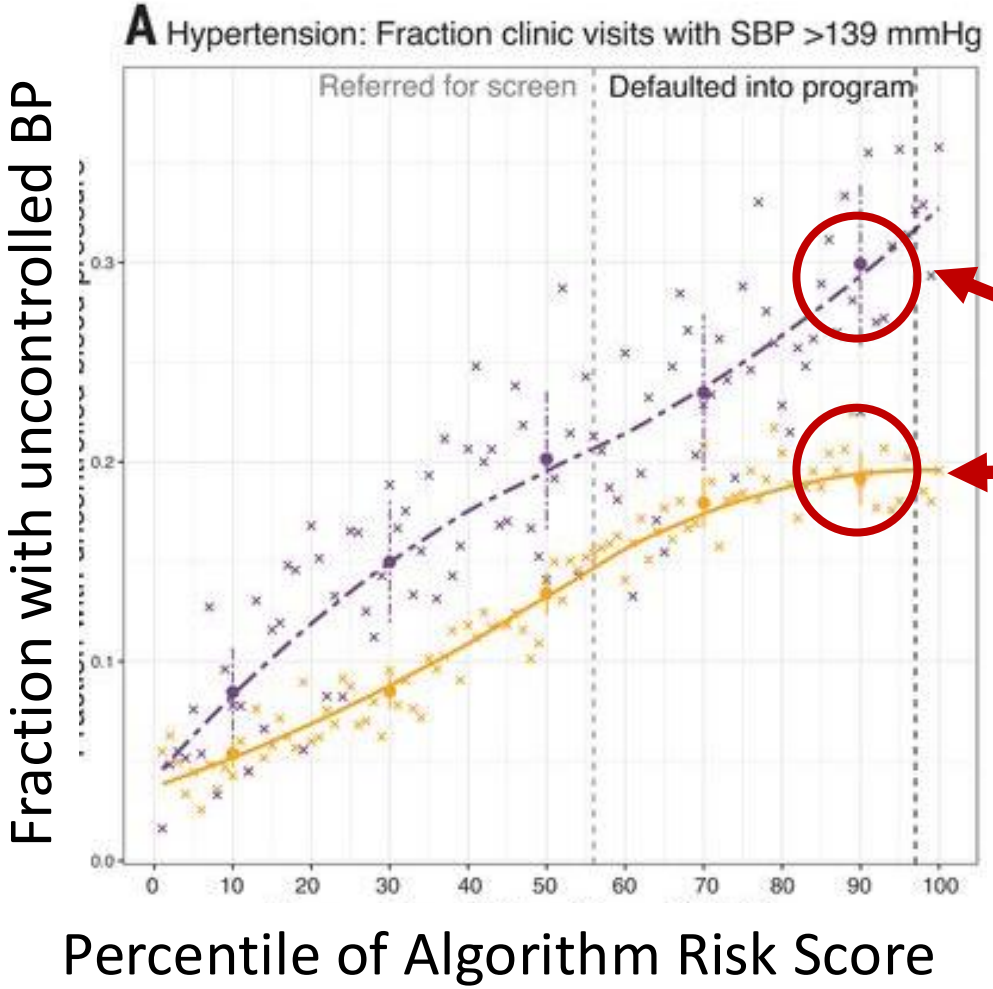


Figure 2.A.: Fraction of clinic visits with uncontrolled blood pressure.

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.

Obermeyer et al, 2019, "Dissecting racial bias in an algorithm …", *Science*.

Dissecting racial bias: Proposed Fix

Increase the fraction of Black patie
in highest risk group from 14% to 2

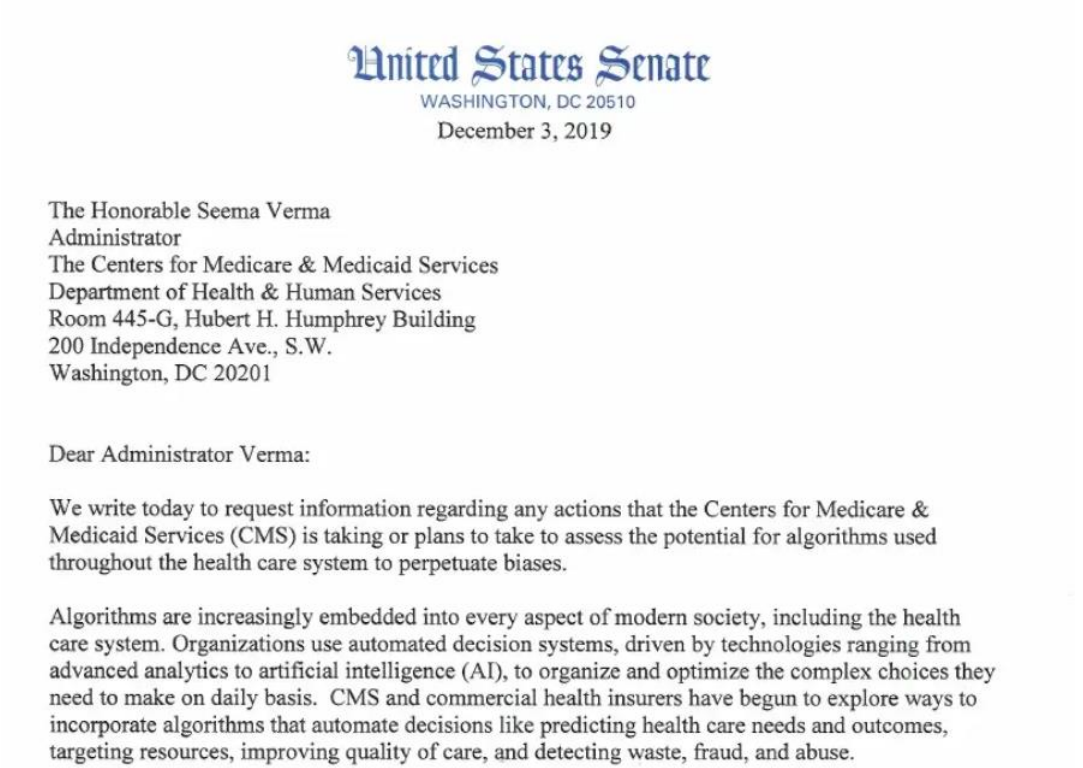
Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
1) Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
2) Avoidable costs	0.142	(0.003)	0.215	(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.

Obermeyer et al, 2019, “Dissecting racial bias in an algorithm …”, *Science*.

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:

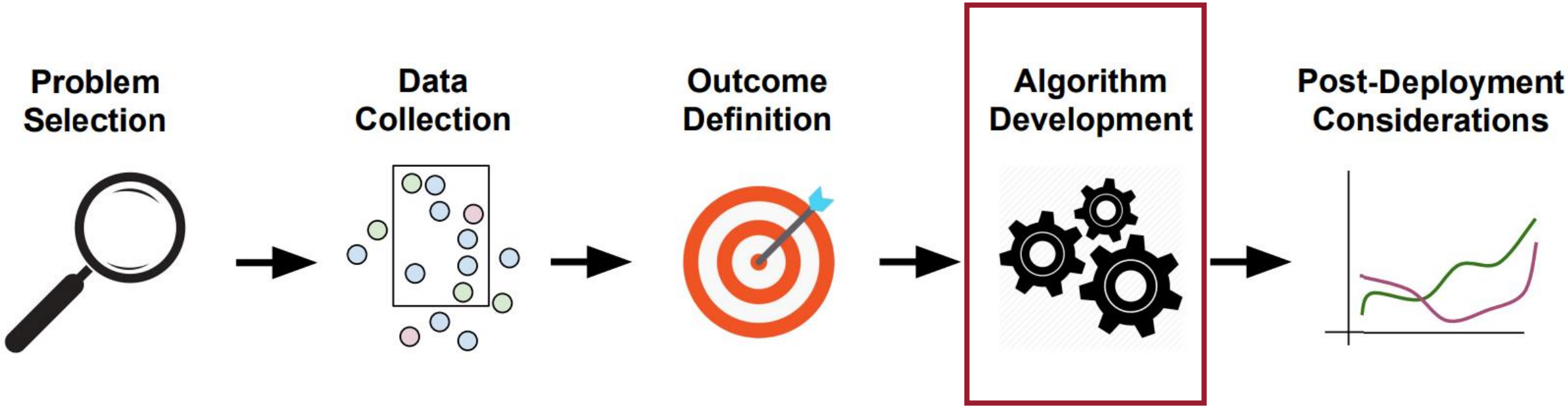
We write today to request information regarding any actions that the Centers for Medicare & Medicaid Services (CMS) is taking or plans to take to assess the potential for algorithms used throughout the health care system to perpetuate biases.

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.

Labels that can be biased or noisy

- Health need
- Clinical diagnosis
  - Over- and under-reporting based on financial incentives
  - Access to care affecting measurement
- Readmission
  - Seeking healthcare in other venues
- Prior history
  - Dependent on relationship with healthcare provider
- Mortality
  - Source of truth?

Machine Learning for Equitable Healthcare



How do we define “bias”?

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

Fairness through Blindness

- **Plan**: Remove any sensitive group from data
- **Example**: Predict diabetes risk *Y* from clinical features *X* and race *A* using $P(\hat{Y} = Y|X)$ instead of $P(\hat{Y} = Y\ |X, A)$
- **Problems:**
  - *A* might have predictive value. What if *Y* = *A*?
  - Other features of *X* might be correlated with *A*



Demographic parity

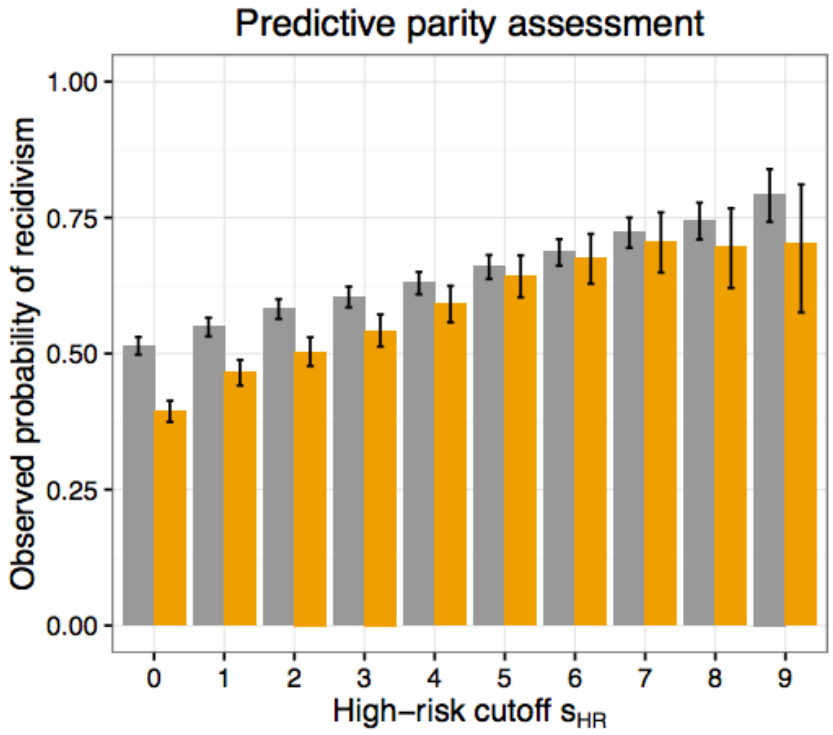
- **Plan**: Require same fraction of $\hat{Y} = 1$ for each group *A*
- **Example**: Predict diabetes risk *Y* from clinical features *X* and race *A* such that $P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$
- **Problems:**
  - What if true *Y* perfectly correlates with *A*?
  - Too strong: even perfect prediction $Y = \hat{Y}$ doesn't satisfy requirements
  - Too weak: doesn't control error rate, could be perfectly biased (wrong for all *A* = 1, correct for A = 0) and still have demographic parity

Calibration

- **Plan**: Same positive predictive value across groups
- **Example**: Predict diabetes risk *Y* from score *S* with threshold *T* from clinical features *X* and race *A* such that

$$P(Y = 1 | S > T, A = 0) = P(Y = 1 | S > T, A = 1)$$

- **Problems:**
  - Might be in conflict with error rate balance



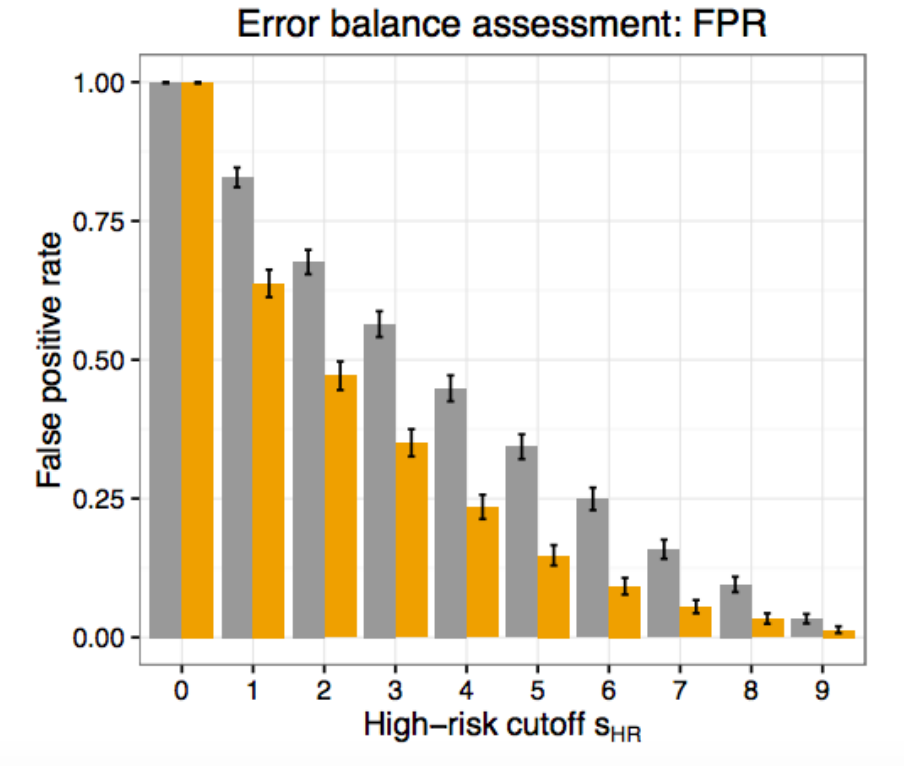
[Chouldechova, 2018]

Error rate balance

- **Plan**: Same positive predictive value across groups
- **Example**: Predict diabetes risk Y from score S with threshold T from clinical features X and race A such that

$$P(S > T|Y = 0, A = 0) = P(S > T|Y = 0, A = 1)$$

- **Problems:**
  - Might be in conflict with calibration

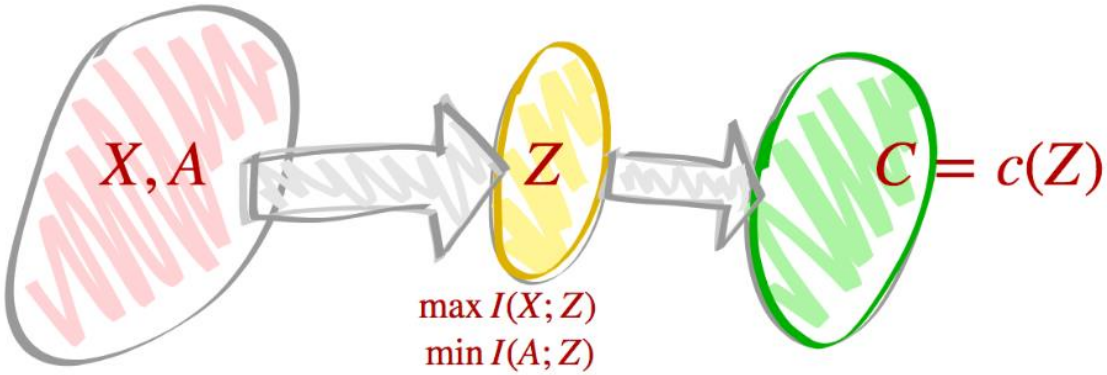


[Chouldechova, 2018]

Representation learning

- **Plan**: Learn latent representation to minimize group information
- **Example**: Predict diabetes risk *Y* from score *S* with threshold *T* from clinical features *X* and race *A* such that

  $\max I(X;Z)$ and $min\, I(A;Z)$
- **Problems:**
  - How to ensure you are not losing too much info and learning right representation?



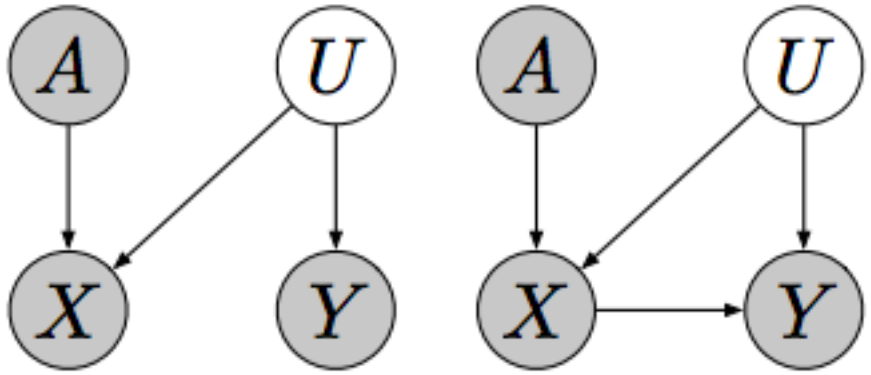
[Zemel et al, 2013]

Causal inference and fairness

- **Plan**: Group *A* should not be cause of prediction $\hat{Y}$
- **Example**: Predict diabetes risk *Y* from clinical features *X* and race *A* such that

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a)$$

- **Problems:** $= P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$
  - Creating a structural model encodes prior beliefs about world
  - Causal inference often requires ignorability assumptions



[Kusner et al, 2017]

How do we define “bias”?

- Fairness thr
- Group fairn
- Calibration
- Error rate b
- Representa
- Counterfact
- Individual fa



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.

Fairness algorithm development

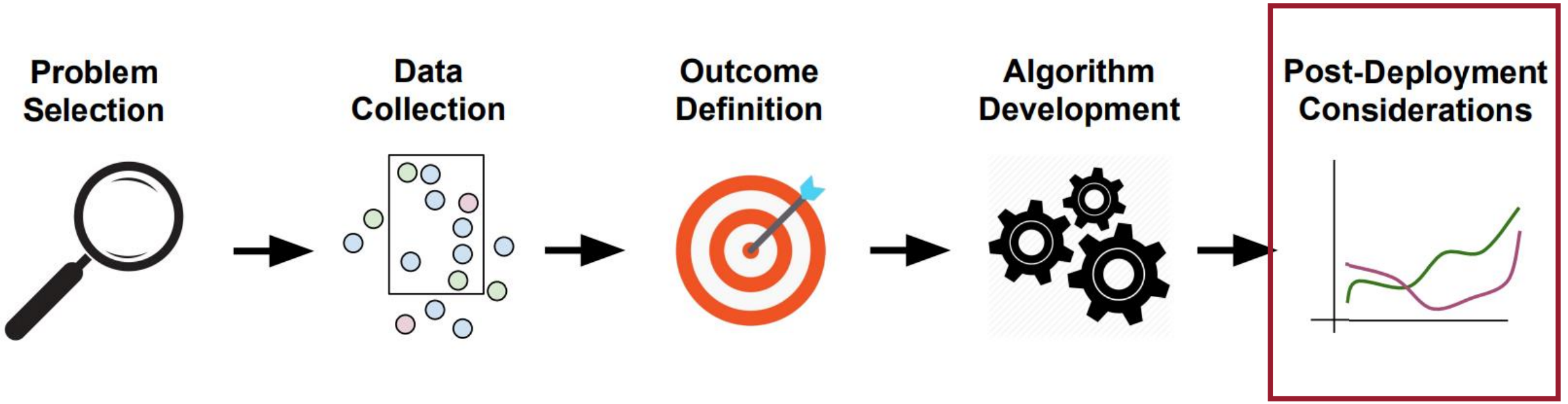
- In practice, practitioners will pick one fairness metric to optimize for
- Depending on the data constraints, the practitioner may create loss functions that try to optimize overall performance while minimizing ”unfairness”

Other algorithm considerations

- Understanding confounding
  - Factors can influence both outcome and covariates
- Feature engineering

- Tuning parameters
- Performance metrics

Machine Learning for Equitable Healthcare



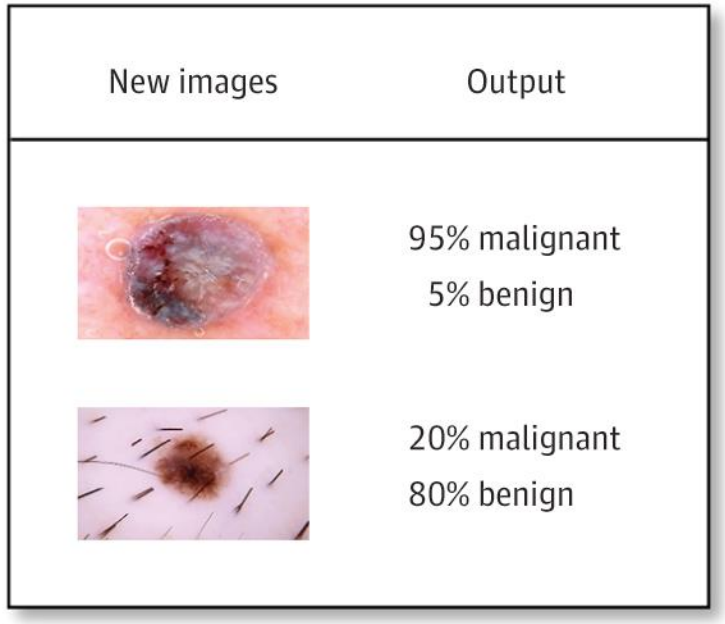
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

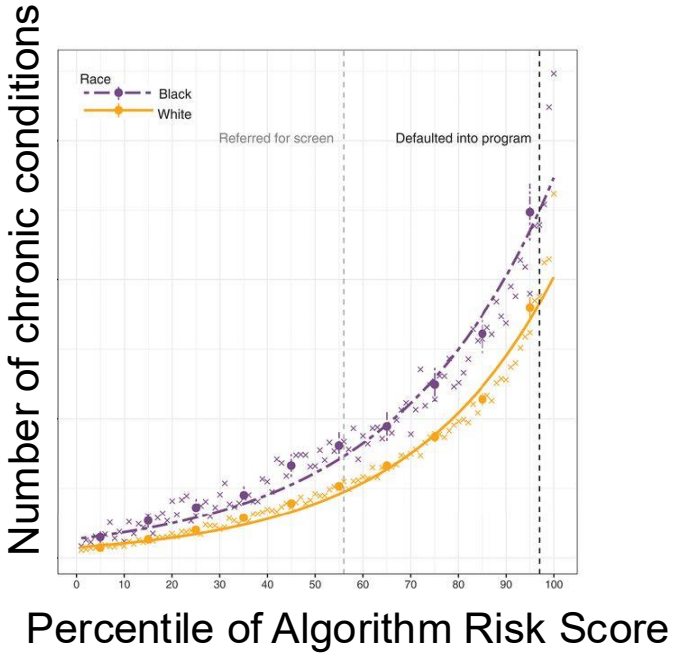




http://gendershades.org/overview.html
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

We are finding evidence of bias through audits





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

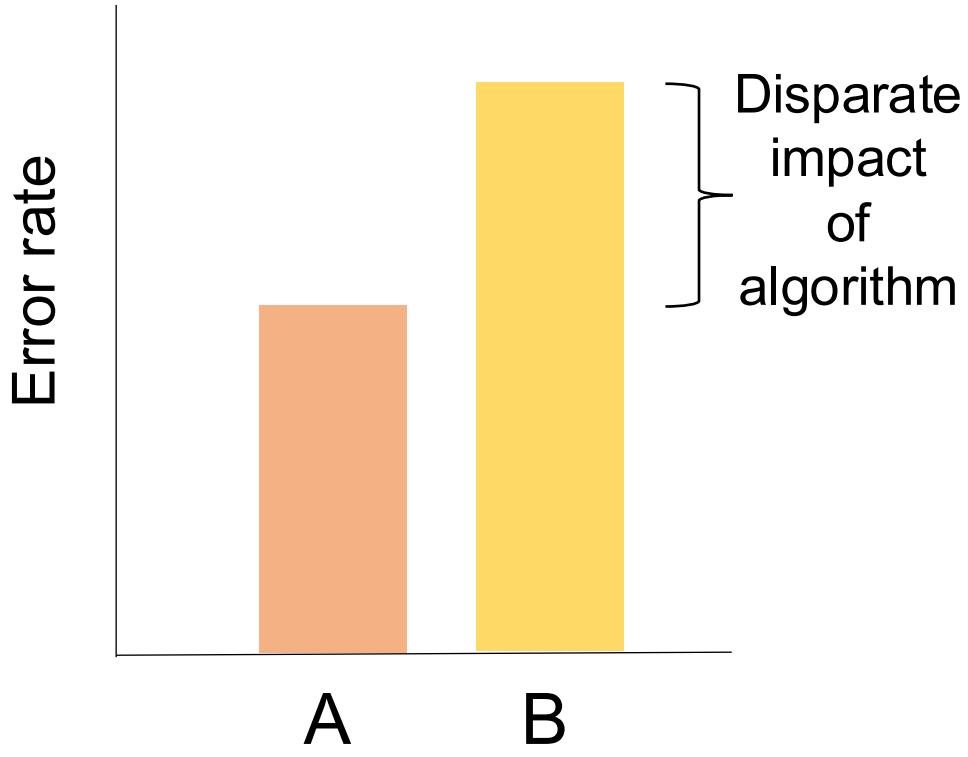
Care management algorithms show racial bias due to training on the "wrong" outcome

[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.

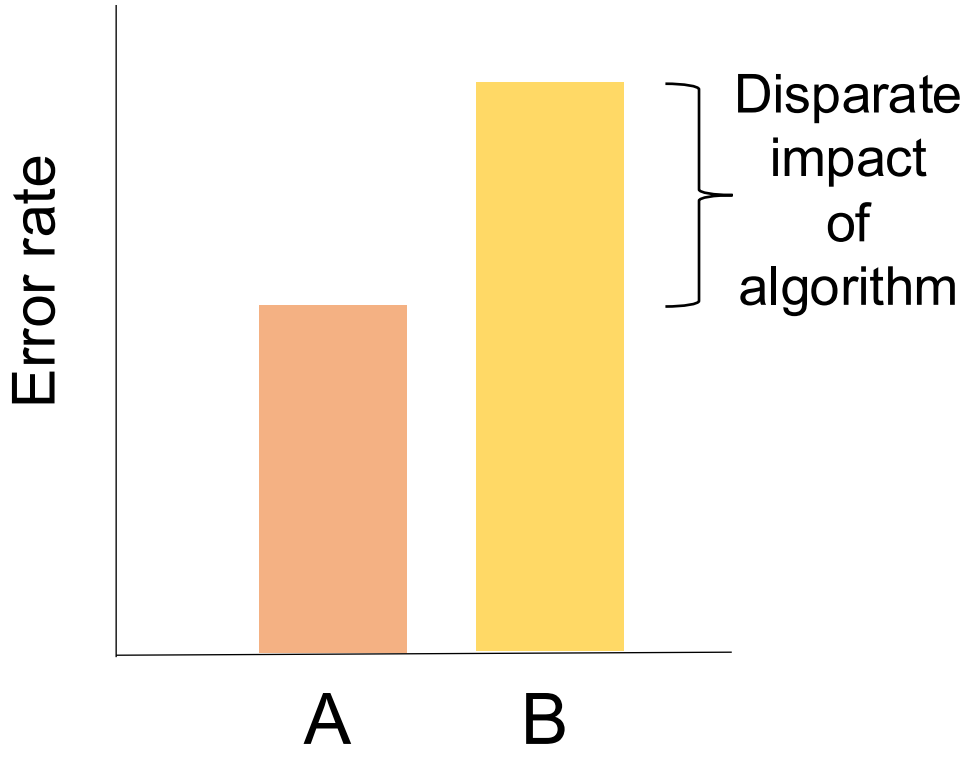
Post-Deployment Considerations

- Quantifying impact
  - High model performance does not translate to large health impact
  - Unintended side effects on clinical workflow, e.g., referring more people to radiologist
- Model generalizability
  - Years
  - Subgroups
- Model and data documentation
- Regulation

Why might my algorithm be unfair?

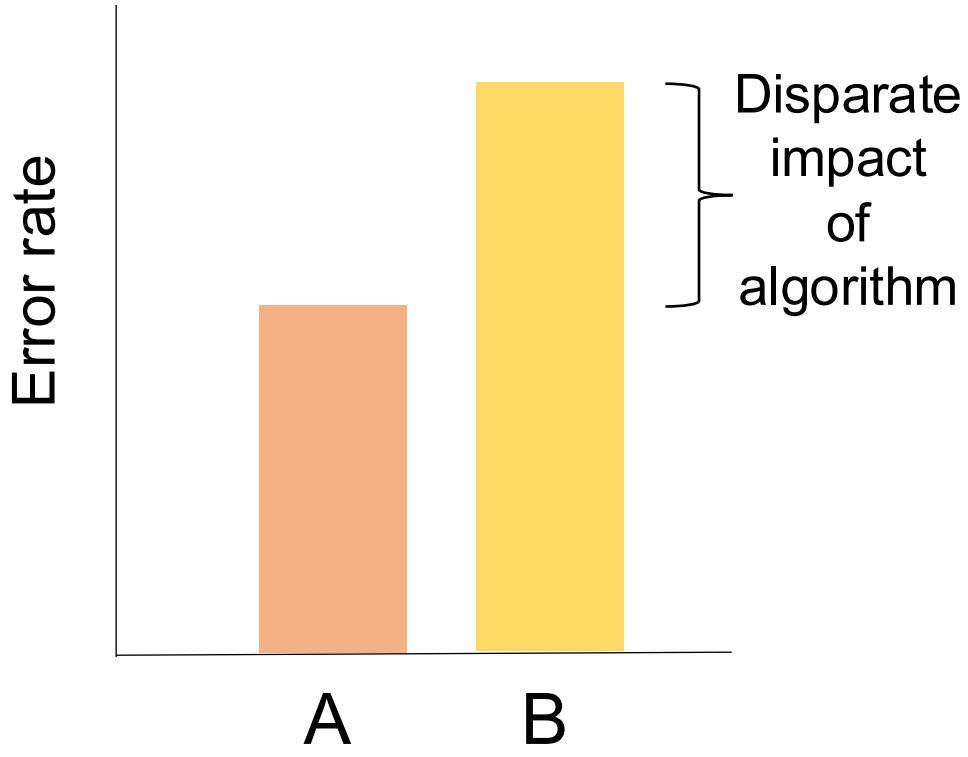


Why might my algorithm be unfair?



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”

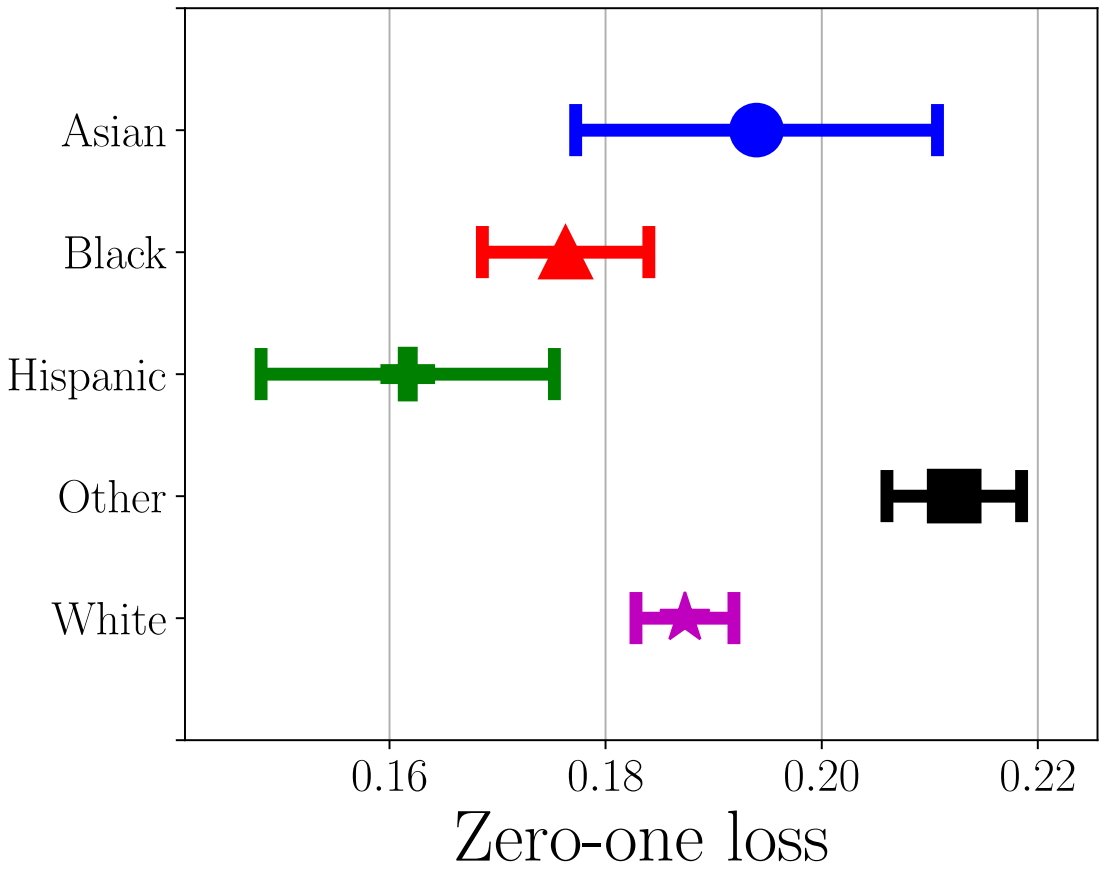
$$\bar{\Gamma} = |(\bar{B}_1 - \bar{B}_0) + (\bar{V}_1 - \bar{V}_0) + (\bar{N}_1 - \bar{N}_0)|$$

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

How can we realistically estimate $\bar{B}_a$, $\bar{V}_a$, and $\bar{N}_a$?

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

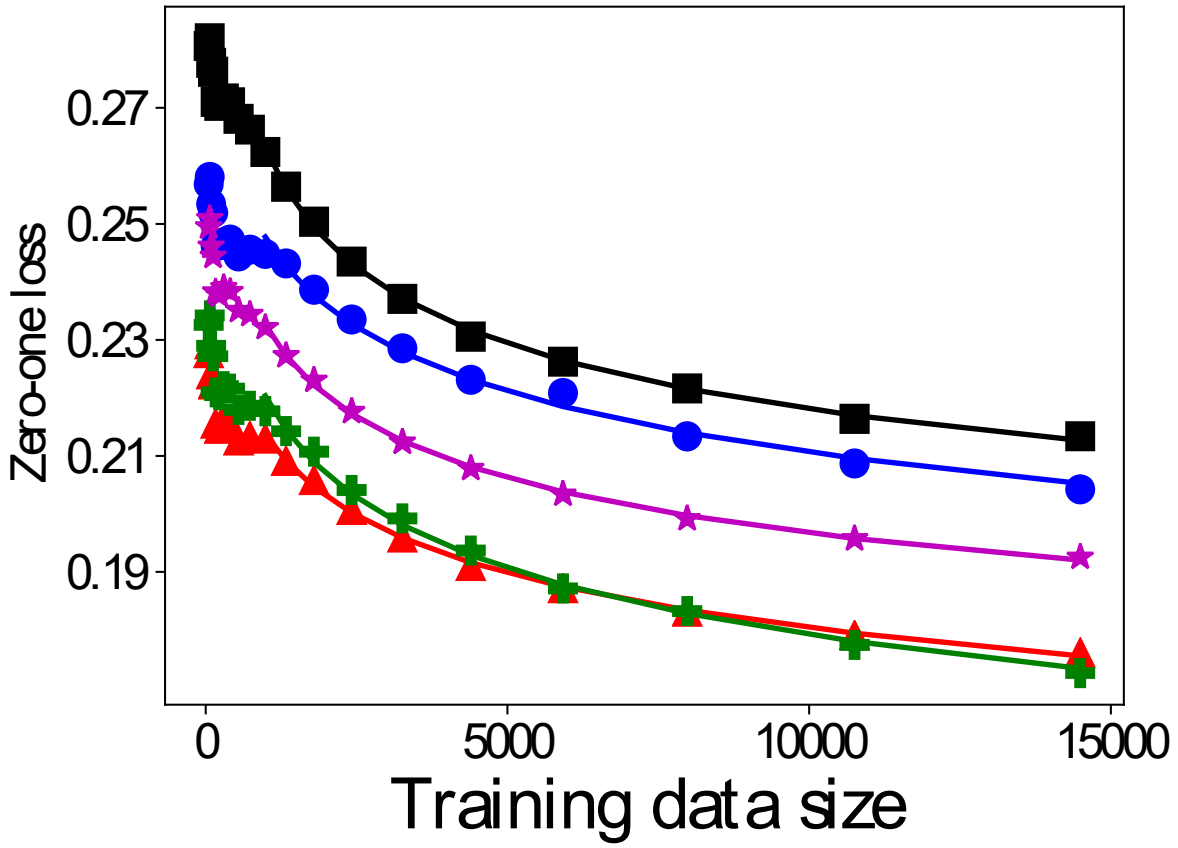
Mortality prediction from MIMIC-III clinical notes



1. We found statistically significant **racial differences** in zero-one loss.

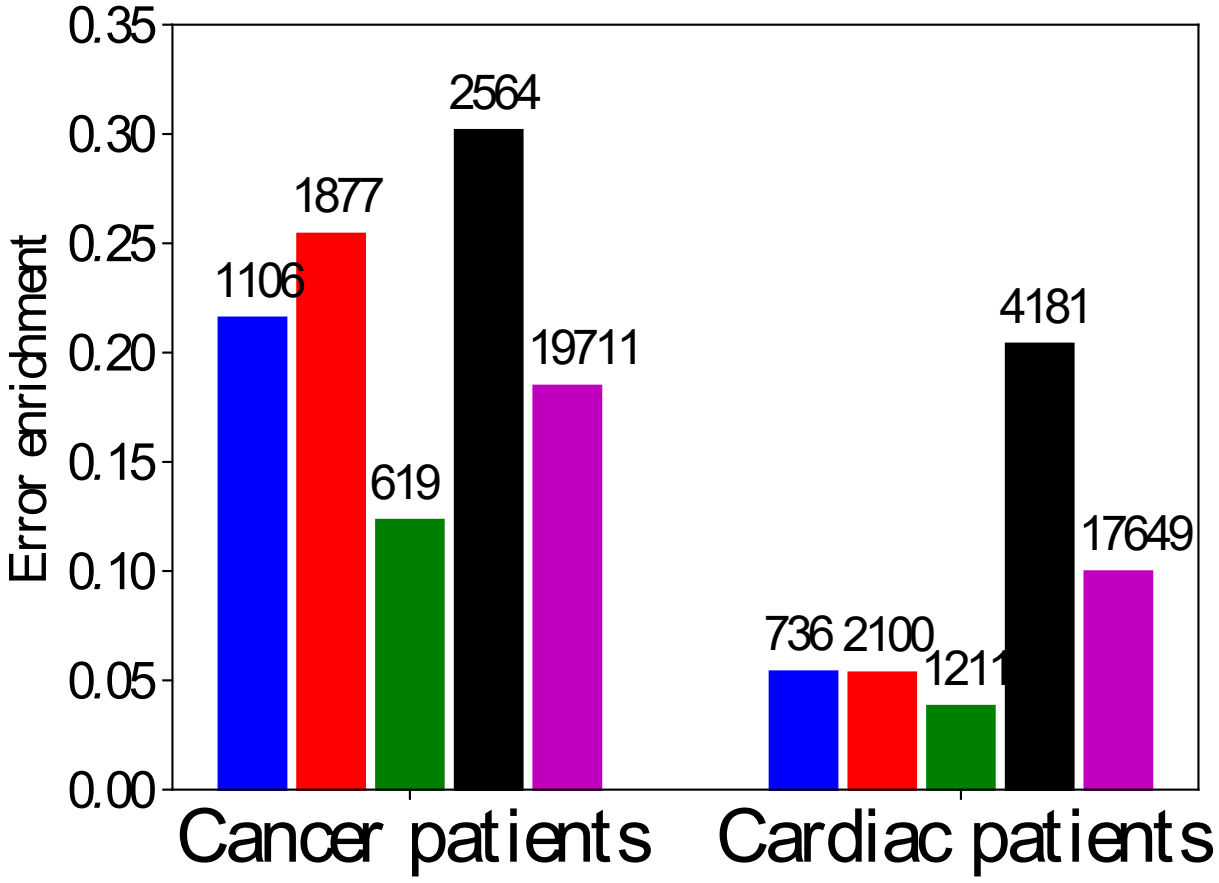
● Asian ▲ Black ✚ Hispanic ■ Other ★ White

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.

Mortality prediction from MIMIC-III clinical notes



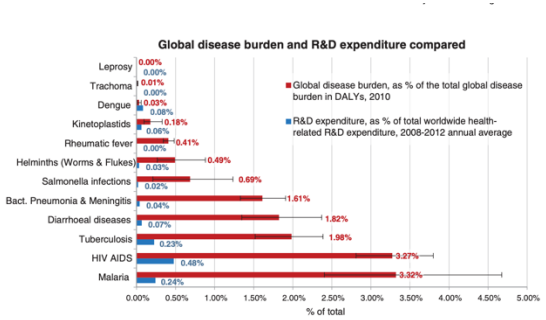
1. We found statistically significant **racial differences** in zero-one loss.
2. By subsampling data, we fit inverse power laws to estimate the benefit of **more data** and reducing variance.
3. Using topic modeling, we identified **subpopulations** to gather more features to reduce noise.

● Asian ▲ Black ✚ Hispanic ■ Other ★ White

Machine Learning for Equitable Healthcare

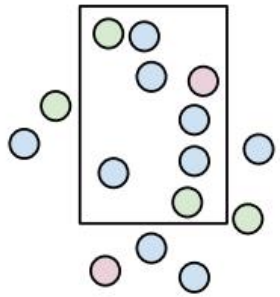
Problem Selection

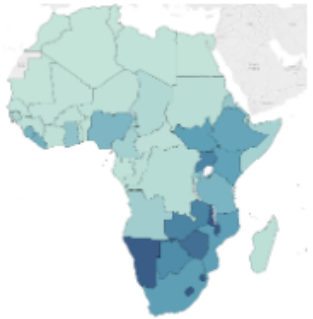




Disparities in funding and research study

Data Collection

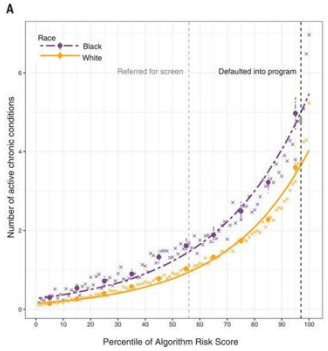




Search queries for African health needs

Outcome Definition





Predicting for cost vs. health need

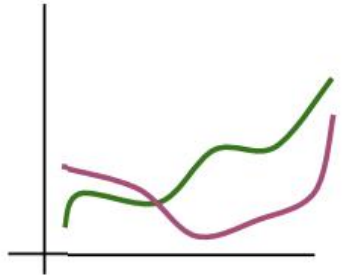
Algorithm Development

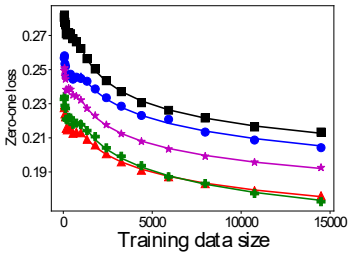




Metrics for fairness

Post-Deployment Considerations





Bias, variance, and noise for fairness audits

Discussion: Will AI make racial bias in healthcare better or worse?

About half of those who see racial or ethnic bias in health and medicine think the use of artificial intelligence would help address the issue

*Among those who say racial or ethnic bias is a **major/minor problem** in health and medicine, % who say that if artificial intelligence is used more, the issue of bias and unfair treatment based on a patient's race or ethnicity would ...*



Here are the main reasons they gave

% who give the following as a reason for their response above

Among those who say it will **get better**

36% said: Human bias is not involved, AI is neutral

"AI isn't going to have preconceived ideas about a patient's symptoms. It will just analyze the data."

28% said: AI is not biased toward a patient's characteristics

Among those who say it will **stay about the same**

28% said: Because people, training or data are still biased

"AI is programmed by humans and humans are biased."

8% said: A real person/doctor would be primarily treating people even if we adopted AI, so no change expected

Among those who say it will **get worse**

28% said: AI reflects human bias

"AI is trained on existing datasets which are already biased by the humans who made them."

10% said: Human judgment is needed in medicine

Pew Research Center, "60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care" https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/, Feb 2023.

Ethical ML for Healthcare discussion questions

- What incentive structures exist for encouraging / discoursing ethical ML for healthcare?
- How would you “evaluate” an entire machine learning pipeline?
- Who has ultimate responsibility for an ethical ML pipeline?


PUBLIC HEALTH
SERVICE -
- CLINIC HOURS
U.S. PHS STUDY
TUSKEGEE
STUDY
TUSKEGEE
STUDY

Summary

- ✓**Problem Selection** (10 mins)
- ✓**Data Collection** (10 mins)
- ✓**Outcome Definition** (5 mins)
- ✓**Algorithm Development** (10 mins)
- ✓**Post-Deployment Considerations** (5 mins)

How can we make Data 146 better for you?

Next class: AI policy and regulation