Fairness in Machine Learning for Healthcare #FairMLHealth

Muhammad Aurangzeb Ahmad[1,2], Carly Eckert MD MPH[1,2], Christine Allen[1], JuHua Hu[2], Vikas Kumar[1], Arpit Patel MD[3], Ankur Teredesai[1,2]

1) KenSci Inc.
2. Department of Computer Science and Systems, University of Washington Tacoma

AUGUST 23-27
KDD2020

#FairMLHealth Tutorial Overview

- Foundations: Fairness in Healthcare ML
- Measurement & Mismeasurement of Fairness
- Operationalizing Fairness in Healthcare ML
- Domain Challenges in Healthcare ML
- Fairness in Healthcare ML in Action
- Best Practices
- Library Demo
- Conclusion

Foundations: Fairness in Healthcare ML

Audience Poll

How many people in the audience?

1. Are Physicians/MDs
2. Work in the healthcare domain
3. Have built a machine learning model
4. Work with healthcare data
5. Plan to work with applied AI/ML in healthcare in the near future

Elements of Responsible AI in Healthcare

Explainability
& Transparency

Fairness &
Unbiased

Robustness

Privacy &
Security

Elements of Ethical ML in Healthcare

Explainability & Transparency

Fairness & Unbiased

Robustness

Privacy & Security

2018 KDD

AUGUST 23-27th KDD2020 Virtual Conference

Fairness in Machine Learning in General

- Classification
- Regression
- Ranking
- Recommendation
- Bandit Learning
- Reinforcement Learning
- NLP
- Clustering
- Representational Learning
- Causal Inference

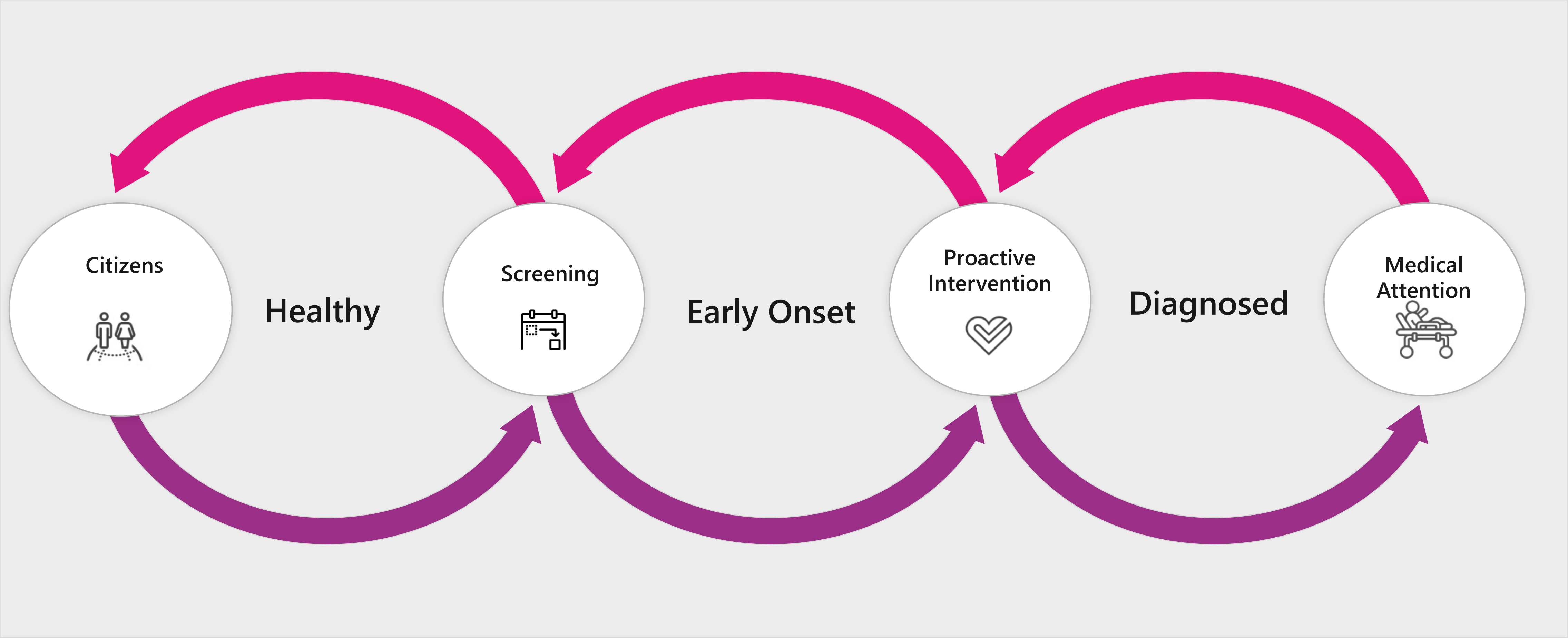
Fairness in Machine Learning in General

- Classification
- Regression
- Ranking
- Recommendation
- Bandit Learning

- Reinforcement Learning
- NLP
- Clustering
- Representational Learning
- Causal Inference

The Spectrum is not Discrete – AI across the Healthcare Continuum

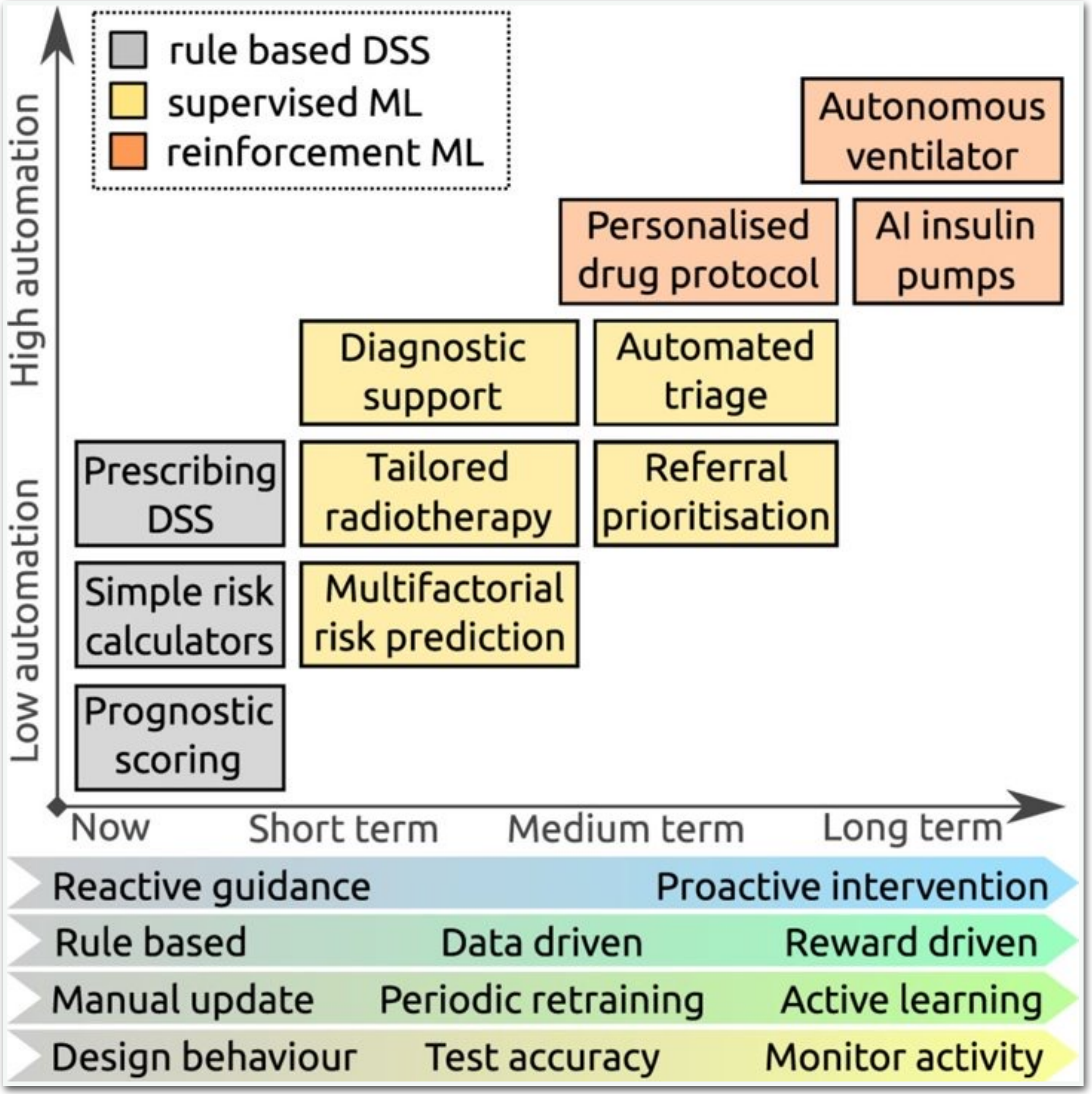
Early Detection and Proactive Intervention to Keep Citizens Healthy



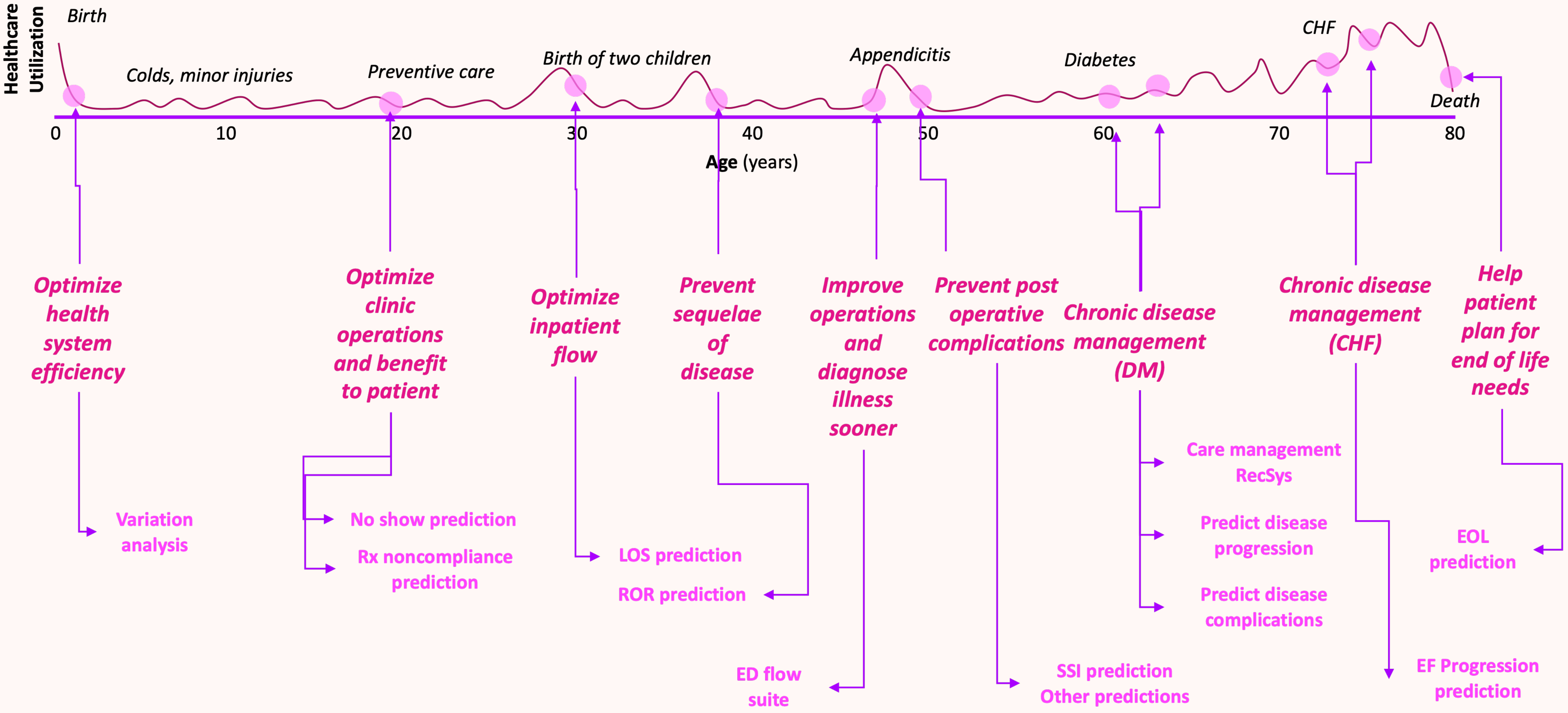
Assistive Intelligence

Expected trends in machine learning (ML) research: boxes show representative examples of decision support tasks that are currently offered by rule-based systems grey), and hypothetical applications of ML systems in the future (yellow and orange), demonstrating increasing automation.

Robert Challen et al. BMJ Qual Saf 2019;28:231-237



Continuum of Care & AI



Optimize health system efficiency

Variation analysis

Optimize clinic operations and benefit to patient

No show prediction

Rx noncompliance prediction

Optimize inpatient flow

LOS prediction

Prevent sequelae of disease

ROR prediction

Improve operations and diagnose illness sooner

ED flow suite

Prevent post operative complications

SSI prediction
Other predictions

Chronic disease management (DM)

Care management RecSys

Predict disease progression

Predict disease complications

Chronic disease management (CHF)

EF Progression prediction

Help patient plan for end of life needs

EOL prediction

[A. Teredesai w Eckert et al, w Ahmad et al, w Frichman et al, w Basu Roy et al]

“We are concerned about the constant use of federal funds to support this most notorious expression of segregation. Of all the forms of inequality, injustice in health is the most shocking and the most inhuman because it often results in physical death.”

- Reverend Martin Luther King Jr. (Chicago, March 25, 1966)

Source: Multiple newspaper article reports. https://quoteinvestigator.com/2015/10/22/mlk-health/

Motivation

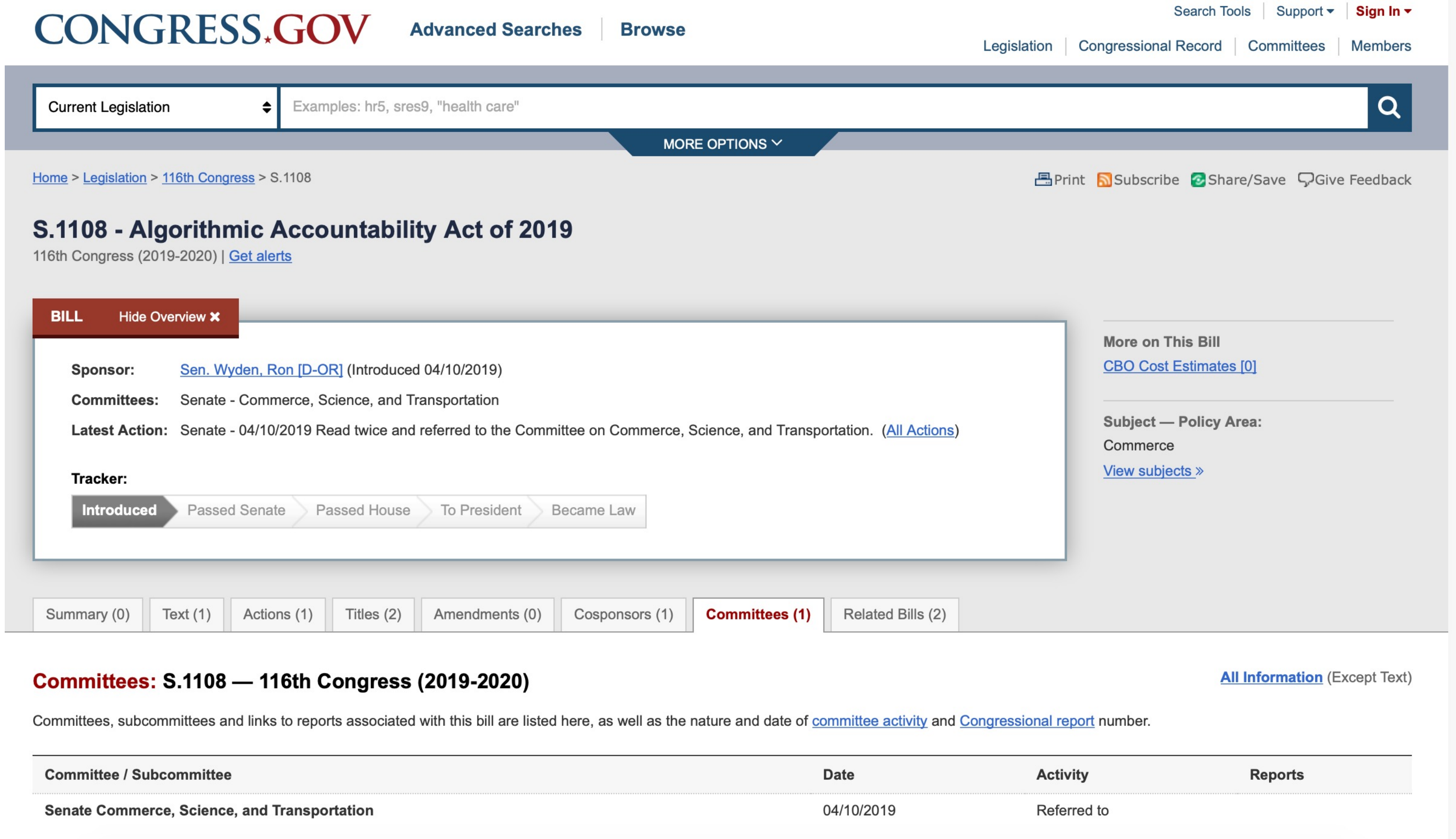
Bias, Discrimination and Unfair practices in healthcare is centuries old

With the integration of AI + Healthcare the potential to discriminate and perpetuate unfair and biased practices in healthcare increases many folds

The problem healthcare AI is a multi-faceted systems level problem that necessitates careful consideration of different notions of fairness in healthcare to different conceptions of Fairness concepts in AI

The Algorithmic Accountability Act

“Under the bill, the FTC could require companies to perform “impact assessments” on their own algorithmic decision-making systems. Such assessment would assess potential consequences for “**accuracy, fairness, bias, discrimination, privacy and security**” within automated systems and companies would be required to correct any issues they uncovered during the process.”



Home > Legislation > 116th Congress > S.1108

S.1108 - Algorithmic Accountability Act of 2019

116th Congress (2019-2020) | Get alerts

Sponsor: Sen. Wyden, Ron [D-OR] (Introduced 04/10/2019)

Committees: Senate - Commerce, Science, and Transportation

Latest Action: Senate - 04/10/2019 Read twice and referred to the Committee on Commerce, Science, and Transportation. (All Actions)

Tracker: Introduced | Passed Senate | Passed House | To President | Became Law

More on This Bill
CBO Cost Estimates [0]

Subject — Policy Area:
Commerce
View subjects »

Committees: S.1108 — 116th Congress (2019-2020)

All Information (Except Text)

Committees, subcommittees and links to reports associated with this bill are listed here, as well as the nature and date of committee activity and Congressional report number.

Committee / Subcommittee	Date	Activity	Reports
Senate Commerce, Science, and Transportation	04/10/2019	Referred to	

Bias & Discrimination in Healthcare: History

The seminal figures of modern medicine (Anton van Leeuwenhoek$^{(1632-1723)}$, the Father of Microscopy, Marcello Malphighi$^{(1628-1694)}$, the Father of Histology, Carl Linnaeus$^{(1707-1778)}$ the Father of Biological Classification) held racial and biased beliefs that greatly influenced modern medicine and healthcare (Byrd 2001)

Education: Many Western physicians assumed poor health as normal for Black ("Negro Diseases”). In Medical Schools syllabus until the 1960s in the US

Medical Profession: With few exceptions Blacks were not represented in the medical profession in the US until the last 19th century and the percentage in the profession remained at 2% from 1900 to 1980

Sterilization: A third of Puerto Rican women of childbearing age were sterilized under coercion from 1930s to 1970s. Many Mexican and Native American women were also sterilized (de Malave 1999). International examples of ethnic sterilization are also plenty (India ‘70s)

Fatality: Higher cases of death during childbirth and lower birth weight among Black Women (Randall 1995)

Bias & Discrimination in Healthcare: History

Tuskegee Experiments: From 1932 to 1972, the US government tracked and lied to 600 hundred low-income African-American men in Tuskegee, AL on a study where sham treatments were given for Syphilis. Many men needlessly passed the disease to family, suffered and died (Thomas and Quinn 1991)

Sickle cell disease: which mostly affects African-Americans, received less attention in research than other prominent diseases, mainly because its disproportionately affected African-Americans (Wailoo 2017)

Bypass Surgery: Black men are significantly less likely to be recommended for bypass surgery than White men. Mainly because of the incorrect perception Black patients were less well-educated and less likely to engage in physical activity after the surgery. Thus, the physicians concluded that they were poorer candidates for the surgery (Malat and Griffin 2006)

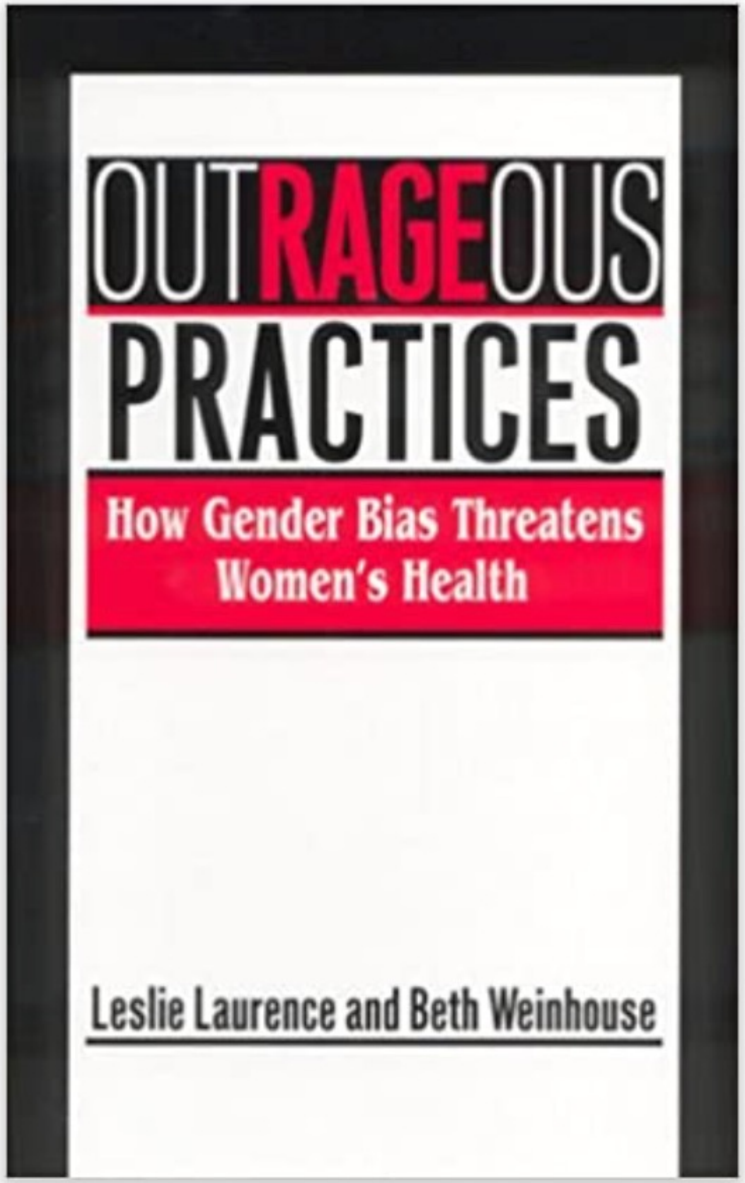
Bias in Healthcare: Examples

- Rockefeller University’s NIH supported study on how obesity affected breast and uterine cancer did not enroll any woman (Simkin 1995)

- It was found that older women were less likely to be given lifesaving interventions as compared to men (Bierman 2007)

- It has been observed that women are less likely to be given analgesia (Chen 2008)

- The 1982 Multiple Risk Factor Intervention Trial aimed at exploring whether dietary change and exercise could help prevent heart disease included no women out of trial size of 13,000



OUTRAGEOUS PRACTICES

How Gender Bias Threatens Women’s Health

Leslie Laurence and Beth Weinhouse

Source: https://nypost.com/2018/04/21/medical-research-has-a-woman-problem/

Bias in Healthcare: Examples

- For most of the 15 leading causes of death in the US including heart disease, cancer, stroke, diabetes, kidney disease, hypertension, liver cirrhosis and homicide, Blacks have higher death rates than whites (Kung et al. 2008)
- These elevated death rates exist across the life-course with African Americans and American Indians having higher age-specific mortality rates than whites from birth through the retirement years [Williams 2005]
- Experiencing racist treatment is also a social determinant of health. Experience of interpersonal racism has been observed as a mechanism that partially explains differences between Aboriginal and non-Aboriginal people’s health [Larson et al 2007]
- There is a long history of unfair diagnosis of psychological conditions in minorities and women [Gard et al 1997]

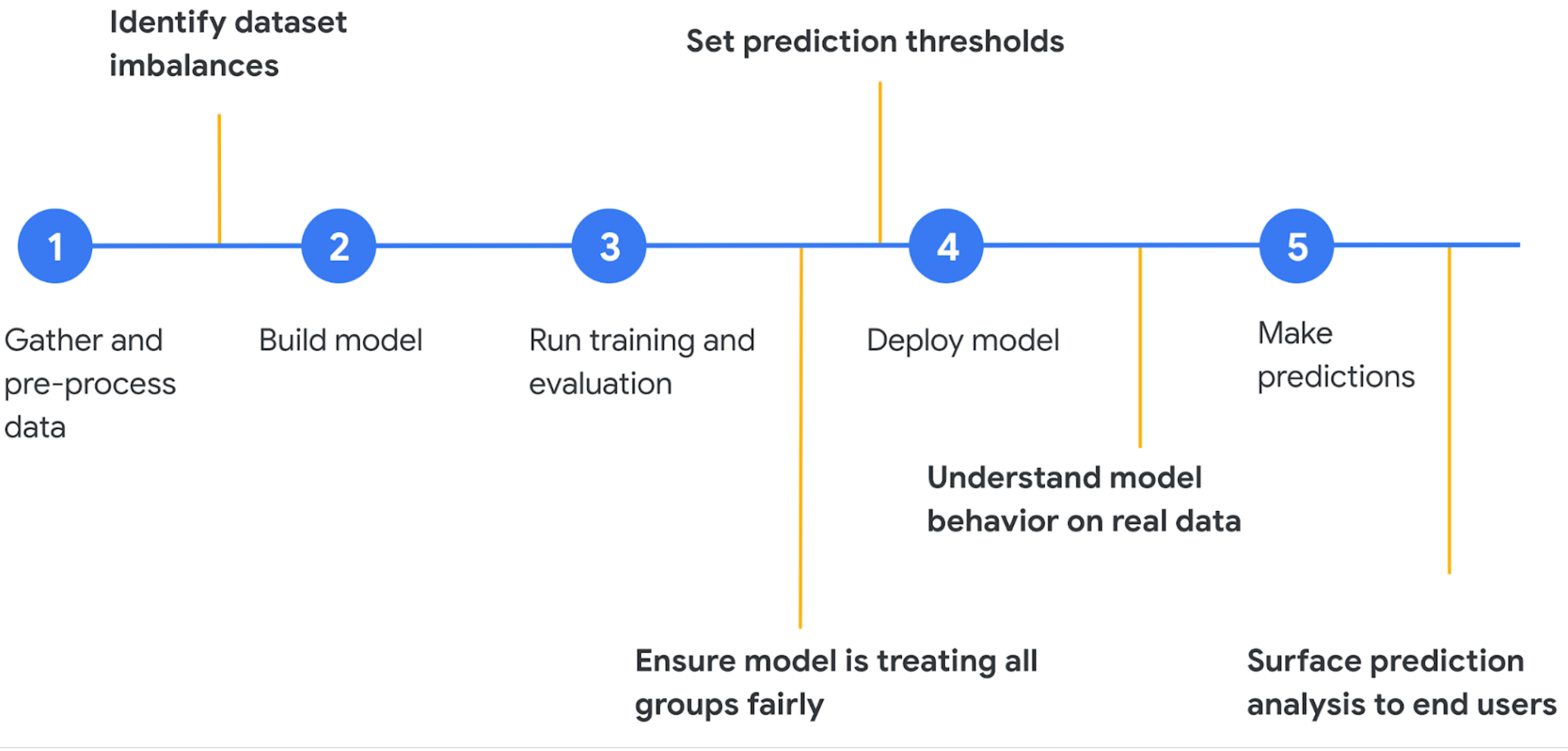
Bias in Healthcare AI: History

- One of the earliest examples (1970s) of algorithmic discrimination comes from healthcare where an algorithm employed by St. George's Hospital Medical School in the UK was discriminating based on race and gender in making initial screening decisions for applicants to medical school
- In 1976 Joseph Weizenbaum was one of the first computer scientists to raise the question of algorithmic bias (Weizenbaum 1976)
- Clinicians are more likely to believe AI that supports current practices and thus perpetuate implicit biases (Parikh 2019)
- Among women with breast cancer, Black women have a lower likelihood of being tested for high risk mutations compared with Caucasian women. An AI model that uses genetic tests is more likely to mischaracterize the risk of breast cancer, although the risk is the same for both (Parikh 2019)

Bias in Healthcare AI: History

- Idahoans with cognitive/learning disabilities had their healthcare benefits reduced by $20—30K based on AI without any explanation which led to a lawsuit by ACLU that revealed that the decisions were made by an AI [Stanley 2017]

Fairness in Machine Learning is more than imbalanced datasets!


Identify dataset imbalances
Set prediction thresholds
1
2
3
4
5
Gather and pre-process data
Build model
Run training and evaluation
Deploy model
Make predictions
Understand model behavior on real data
Ensure model is treating all groups fairly
Surface prediction analysis to end users

[Google Cloud 2018]

Bias in Healthcare AI: Is it just a data problem?

Generalizability and representativeness are also important considerations when interpreting randomized clinical trials. The generalizability of AI algorithms across subgroups is critically dependent on factors like representativeness of included populations, missing data, and outliers.

- EHRs are observational databases, the data reflects not just the health of the patients but also their interactions with the healthcare system e.g., the date of a code for diabetes is when the physician made the diagnosis, not when the patient first developed the disease (Agniel 2018)
- The billing code used for an office visit might be influenced more by reimbursement policies than the original reason for the visit
- Practices regarding how data is recorded may change over time e.g., reporting patient falls, opioid prescribing increased from 2005-12, but at rates that differed by practice and patient population [McLintock 2019]
- Data as a signal. Lab tests are ordered more often for sick patients (Agniel 2018)

Fairness in ML as a Systems Problem

AI Solution

Data

Algorithm

Model

Intervention

Each constituent element can contribute towards making the solution unfair or biased

User

User

Use Case

Context

Legal 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 status (Immigration Reform and Control Act 1965)
- Age (Age Discrimination in Employment Act of 1967)
- Familial status (Civil Rights Act of 1968)
- Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- Pregnancy (Pregnancy Discrimination Act 1978)
- Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act of 1994)
- Genetic information (Genetic Information Nondiscrimination Act of 2008)

[Solon Barocas and Moritz Hardt 2017]

Fairness in the age of COVID19

- **Healthcare rationing:** Due to stresses caused by the COVID-19 pandemic on national healthcare systems globally
- When the limited resources in acute medical settings cannot be accessed by all patients who need them healthcare rationing is unavoidable.
- Real Use Case: What happens when ICU demands exceeds the treatment facilities available? How should doctors decide between which patients to treat?

Fairness is Stakeholder Dependent

Physician: Of the patients that are labeled high risk for diabetes, how many are likely to be high risk?

Patient: What is the probability that I will be incorrectly labeled as low risk? Given that I am from a protected class, will I be given the same clinical services according to best evidence

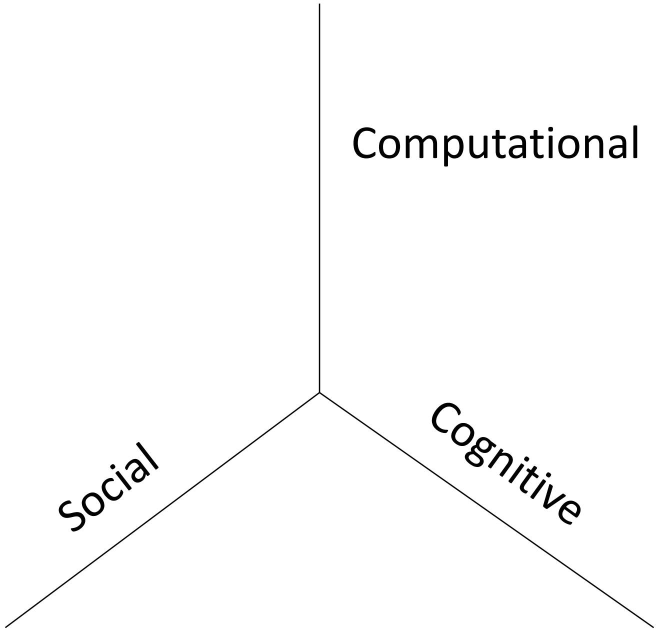
Societal (Group Fairness): Are the risks balanced across all protected classes?

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

[Narayanan 2018]

Dimensions of Fairness in Healthcare AI

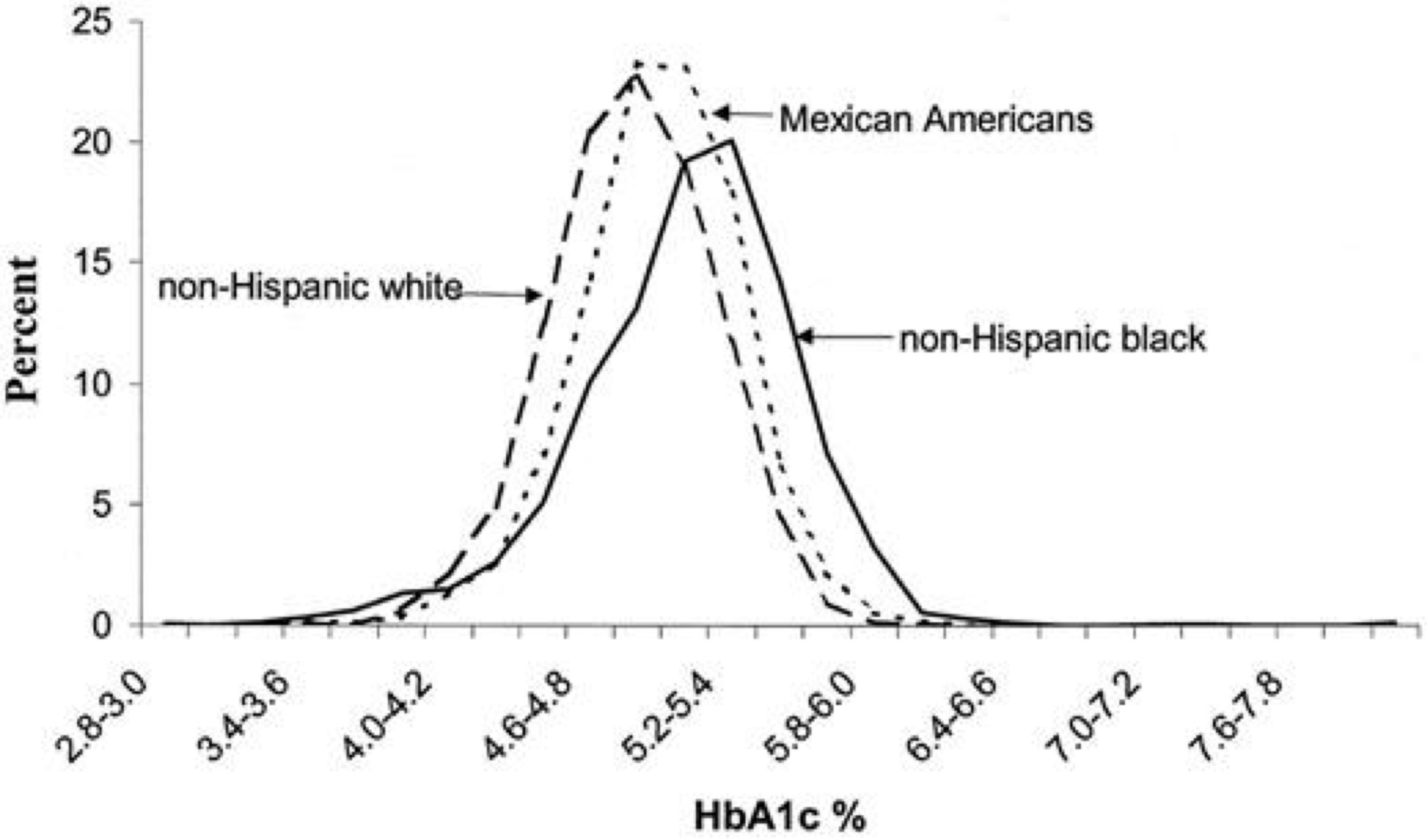
- **Computational**
  - **Data Bias**
  - **Model Bias**
  - **Loss Function Bias**
  - **Post-Hoc optimization**
- Social
  - Structural Bias
  - Embedded Practices
- Cognitive
  - Automation Bias
  - Automation Complacency
  - Delivery Bias



Sources of Bias in Healthcare AI

Bias in the ML Cycle	Data Bias	Non-Data Biases • **Model Bias** • **Loss Function Bias** • **Post-Hoc optimization**	Bias in Delivery • Cognitive Biases • Social Biases
Sources of Bias	• Selection/sample bias • Response bias • Publication bias • Prejudicial bias • Measurement bias • Hawthorne effect • Social desirability bias • Self-reporting bias	• Algorithmic Bias • Loss Function Bias • Post-Hoc Optimization	• Outcome Fairness • Lack of Understanding • Explainability • Lack of understanding/ Assume model is fair • Don't care
Bias Mitigation	• Equal representation	• Bias mitigation Algorithms • Fairness metrics • Explainable AI	• Acknowledgement & Explanation of bias during model delivery

AI needs FAT



HbA1c distribution by ethnicity in U.S. children and young adults ages 5–24 yr (NHANES-3, 1988–1994) [Saaddine et al., 2002]

- Hemoglobin A1c (HbA1c): widely used as a measure of risk for the development of diabetic complications [Herman et al., 2012, Edelman et a

Example: Differential Treatment by Race

- James, a 65-year-old Black male and David, a 65-year-old white male, both have coronary artery disease. On Sunday afternoon, both men experience chest pain and shortness of breath and are rushed to the ED by their spouses.
- Both men are seen by the same ED physician and are both diagnosed as having an acute myocardial infarction (a heart attack). Yet the clinical recommendations and interventions offered are different *and James is treated less aggressively*
- How do we determine that the two patient are being treated fairly? [Arora et a. 2018]

Measurement & Mismeasurement of Fairness

How we are categorized through data affects how we will be treated

Frank Pasquale in *The Black Box Society*

Discrimination: Treatment vs. Impact

Discrimination: The unjust or prejudicial treatment of different categories of people or things, especially on the grounds of race, age, or sex (Oxford Dictionary)

Disparate Treatment: The treatment depends on class membership
Example: implicit bias leading to differences in treatment in acute coronary syndrome

Disparate Impact: The treatment appears to be neutral, but it impacts the protected class
Example: hospital relocation and access to care for minority classes

[Barocas, S. and Selbst 2016]

Fairness & Predictive Performance

Fairness Measurement
What are the different ways to measure Fairness

Predictive Performance
How well is the model performing

Calibration
How good is the model calibration

Intervention & Allocation
How are the insights from the model being used to intervene

Protected Classes & Proxy Variables

- Many variables of interest correlate with protected class.

- Not all are considered illegitimate to use in decision making.
[e.g., educational qualifications in hiring decisions.]

- Many papers have proposed methods to identify and mitigate “proxy discrimination”.
[Based on correlations or causal paths in DAGs]

Protected Classes & Potential Surrogates
Age
Sex
Race / ethnicity
Insurance status
Disability, functional status
Zip code / census tract
Costs of care / utilization
Marital status
Disease conditions: HIV, mental health
Genetic results: BRCA

Model Performance and Fairness

Differences in performance

- Limited features
- Skewed distributions
- Limited data availability

Distribution of Error across sub-populations

- Different models with the same reported accuracy can have a very different distribution of error across population

Understanding disparities in predicted outcome

- Skewed Proxies
- External processes not captured in data

Example: Differential Treatment by Race

- James, a 65-year-old Black male and David, a 65-year-old white male, both have coronary artery disease. On Sunday afternoon, both men experience chest pain and shortness of breath and are rushed to the ED by their spouses.
- Both men are seen by the same ED physician and are both diagnosed as having an acute myocardial infarction (a heart attack). Yet the clinical recommendations and interventions offered are different *and James is treated less aggressively*
- How do we determine that the two patient are being treated fairly? [Arora et a. 2018]

Type	Description	Formulation	Motivation	Flaws
Unawareness	Do not include the sensitive attribute as a feature in the training data	$C=c(x, A) = c(X)$	Intuitive, easy to use and legal support (disparate treatment)	There can be many highly correlated features(e.g. neighborhood) that are proxies of the sensitive attribute(e.g. race)
Demographic Parity / Independence / Statistical Parity	The outcomes must be equal		Legal Support: “four-fifth rule” prescribes that a selection rate for any disadvantaged group that is less than four-fifths of that for the group with the highest rate.	Ignores any possible correlation between Y and A e.g., rules out perfect predictor C=Y when base rates are different (i.e. $P_0 [Y=1] \neq P_1 [Y=1]$) laziness: if we hire the qualified from one group and random people from the other group, we can still achieve parity
Equalized odds / Separation / Positive Rate Parity	Different groups deal with similar odds	C is independent of A conditional on Y: $P_0 [C = r \mid Y = y] = P_1 [C = r \mid Y = y] \; \forall \, r, y$	Optimality compatibility: C=Y is allowed. Penalize laziness: it provides incentive to reduce errors uniformly in all groups.	It may not help closing the gap between two groups
Predictive Rate Parity / Sufficiency	The performance of the predictive model should be the same for different groups	Y is independent of A conditional on C: $P_0 [Y = y \mid C= c] = P_1 [Y = y \mid C= c] \; \forall \, y, c \in \{0,1$	Optimality compatibility: C=Y satisfies Predictive Rate Parity. Equal chance of success(Y=1) given acceptance(C=1)	It may not help closing the gap between two groups
Individual Fairness	similar individuals should be treated similarly	$D(M(X),M(X')) \leq d(X,X')$	Rather than focusing on group, as individuals, we tend to care more about the individuals. Besides, individual fairness is more fine-grained than any group-notion fairness	It is hard to determine what is an appropriate metric function to measure the similarity of two inputs
Counterfactual Fairness	How do the outcome change if the values of the sensitive variables change	$P[C_{A\leftarrow 0}=c \mid X, A=a]=P[C_{A\leftarrow 1}=c \mid X, A=a]$	Counterfactual fairness provides a way to check the possible impact of replacing only the sensitive attribute	The idea is very ideal. In practice, it is hard to reach a consensus in terms of what the causal graph should look like and it is even harder to decide which features to use

The Impossibility Theorem of FAIRNESS

Demographic Parity, Predictive Rate Parity and Equalized Odds are mutually exclusive theoretically

Drew Fudenberg, David K. Levine, Fairness, risk preferences and independence: Impossibility theorems, Journal of Economic Behavior & Organization, Volume 81, Issue 2, 2012, Pages 606-612, https://doi.org/10.1016/j.jebo.2011.09.004.

Data Bias

Population bias: Are there differences between the data population’s demographics [...] and the target population?

Behavioral bias: Are there differences in user behavior across platforms (mobile, voice?) or contexts (work, party, family) [...]

Temporal bias Are there differences in populations or behaviors over time?

Redundancy Are there data items that appear in multiple copies, or are near duplicates, or happen artificially often (bots)?

Content production bias Are there lexical, syntactic, semantic, or structural differences in how content is produced vs the content that you want to surface?

Linking bias Are there differences in the attributes of networks, or user connections that affect your data?

Interface Bias Are there biases that result from UI design or presentation? (e.g. position/ranking bias)

Sampling Biases: Are there any biases resulting from data sampling choices?

Self-Selection Bias: Who would *not* participate in this product?

WHAT ARE THE ODDS THAT YOU MADE THIS COMPLICATED SPREAD-SHEET WITHOUT ANY CRITICAL ERRORS?
@ScottAdamsSays
Dilbert.com
DOES IT MATTER, AS LONG AS IT GIVES ME THE ANSWER I WANT?
© 2016 Scott Adams, Inc. /Dist. by Universal Uclick
1-7-16
IT SHOULD.
BUT ASK YOURSELF IF IT DOES.

[Crawford 2017; Gebru et al 2018]

Algorithm & Composition Bias

Algorithmic Bias

- What are the downstream consequences of model choice or even hyperparameter choice?
- Do algorithm assumptions lead towards biased models e.g., Naïve Bayes

Composition/Team Bias

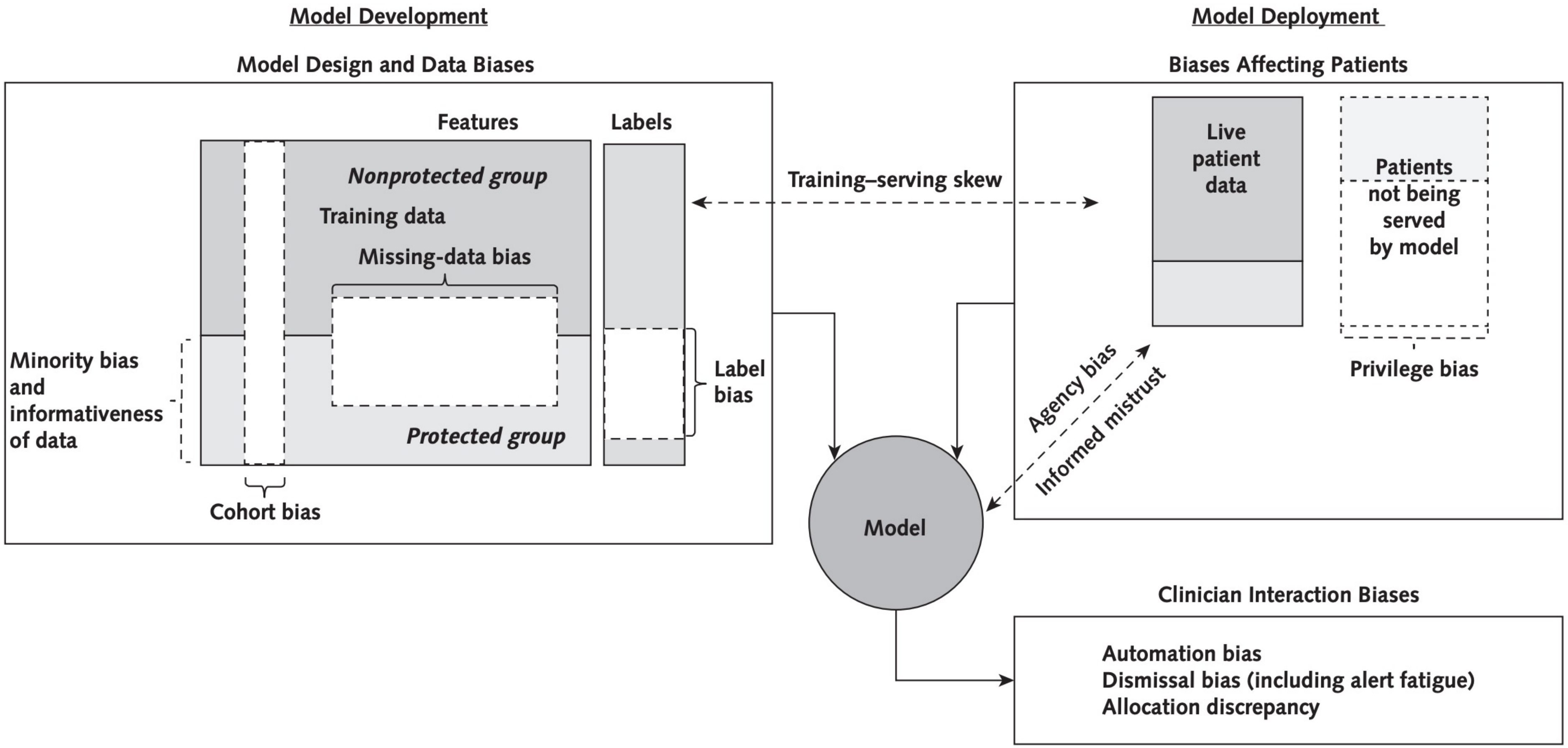
- Knowledge gaps in team
- Consultation with stakeholders and domain experts
- Representation of people affected

Bias: Outcomes & Clinical Perspective

- Model Performance
- Allocation of Services
- Clinical Outcomes

[Rajkomar et al 2018]

How Biases in Healthcare are interrelated



[Rajkomar et al 2018]

Trade-off in Fairness in ML

Performance vs. Fairness
The performance of a model may decrease as it becomes more fair

Fairness vs. Explainability
Since explainability and performance often have an inverse relationship, a similar relationship is observed for fairness

Fairness vs. Performance

Trade-off

The predictive performance of a model depends on the dataset, fairness criteria and the algorithm

In general, however, fairness negatively impacts performance because it diverts the objective from accuracy only to both accuracy and fairness. Therefore, a trade-off is needed

Fairness/Performance Trade-off & Beneficence

- ***Beneficence:*** An ethical principle that providers must do everything they can to benefit the patient

- Since the removal/reduction of bias could possibly reduce predictive performance e.g., adversarial training could increase fairness, it could also compromise overall prediction accuracy (especially the accuracy for non-protected groups). Thus undermining the principle of beneficence

- **Challenge:** How do we simultaneously reduce bias and maintain satisfactory model prediction performance?

Trade-offs: Fairness vs. Explainability

- The relationship between interpretability and fairness is complex and follow four different trends depending on the correlations between protected, non-protected attributes and class labels
- Interpretability-fairness trade-offs do not depend on group imbalance

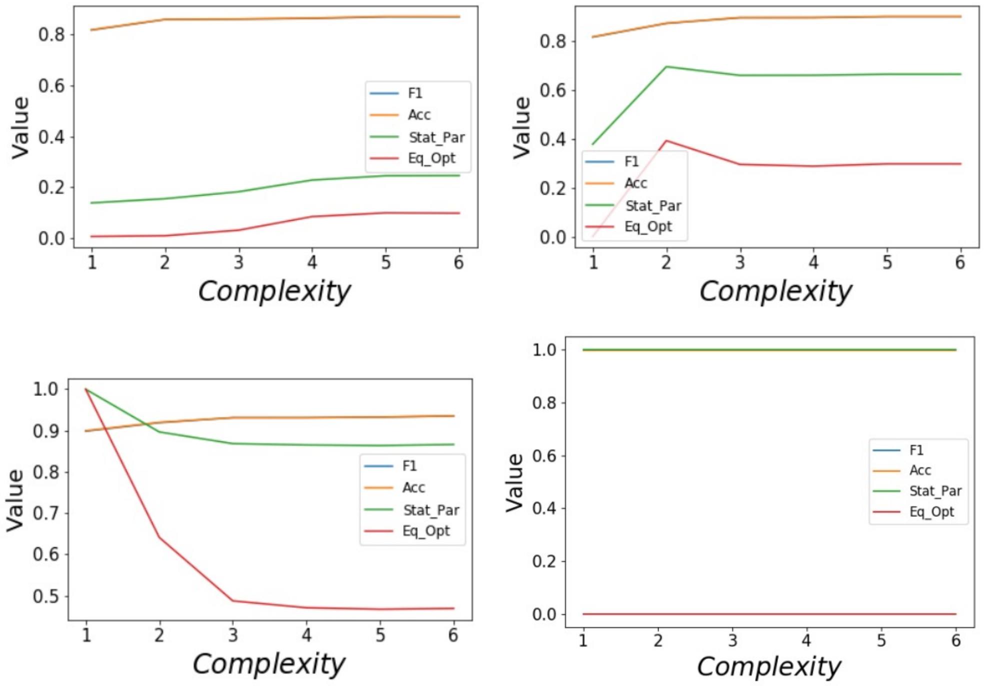


Figure 2. The effect of increasing the predictive power of the protected attribute p. C = logistic regression. $\sigma^2 = 10$ and $r = 2$. $p = 0.6$ (upper left), $p = 0.8$ (upper right), $p = 0.9$ (lower left) and $p = 0.999$ (lower right).

[Jabbari et al 2020]

Fairness/Explainability Trade-off

- Explainability of ML models is supposed to bring about greater scrutiny of models and thus the possibility of fair and equitable models
- However simplification of models may also bring about performance degradation as well as less fair models [Kleinberg and Mullainathan 2019]
- The trade-off is thus three way: Fairness vs. Performance vs. Explainability
- Domain specific guidance should be used to help navigate this complex trade-off landscape

Long Term Trade-offs?

- Under what circumstances does fairness criteria do indeed promote the long-term well-being of protected groups over time
- In standard classification setting such scenarios are not considered
- What if we introduce a one-step feedback model that exposes how decisions change the underlying population over time?

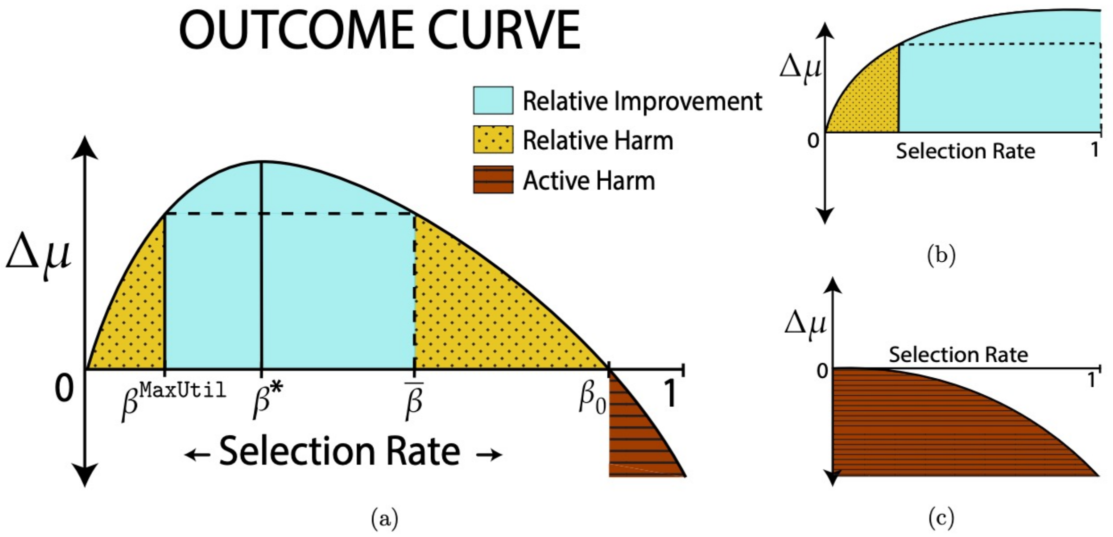


Figure 1: The above figure shows the *outcome curve*. The horizontal axis represents the selection rate for the population; the vertical axis represents the mean change in score. (a) depicts the full spectrum of outcome regimes, and colors indicate regions of active harm, relative harm, and no harm. In (b): a group that has much potential for gain, in (c): a group that has no potential for gain.

[Liu et al 2020]

Data Collection: Disability and Fairness

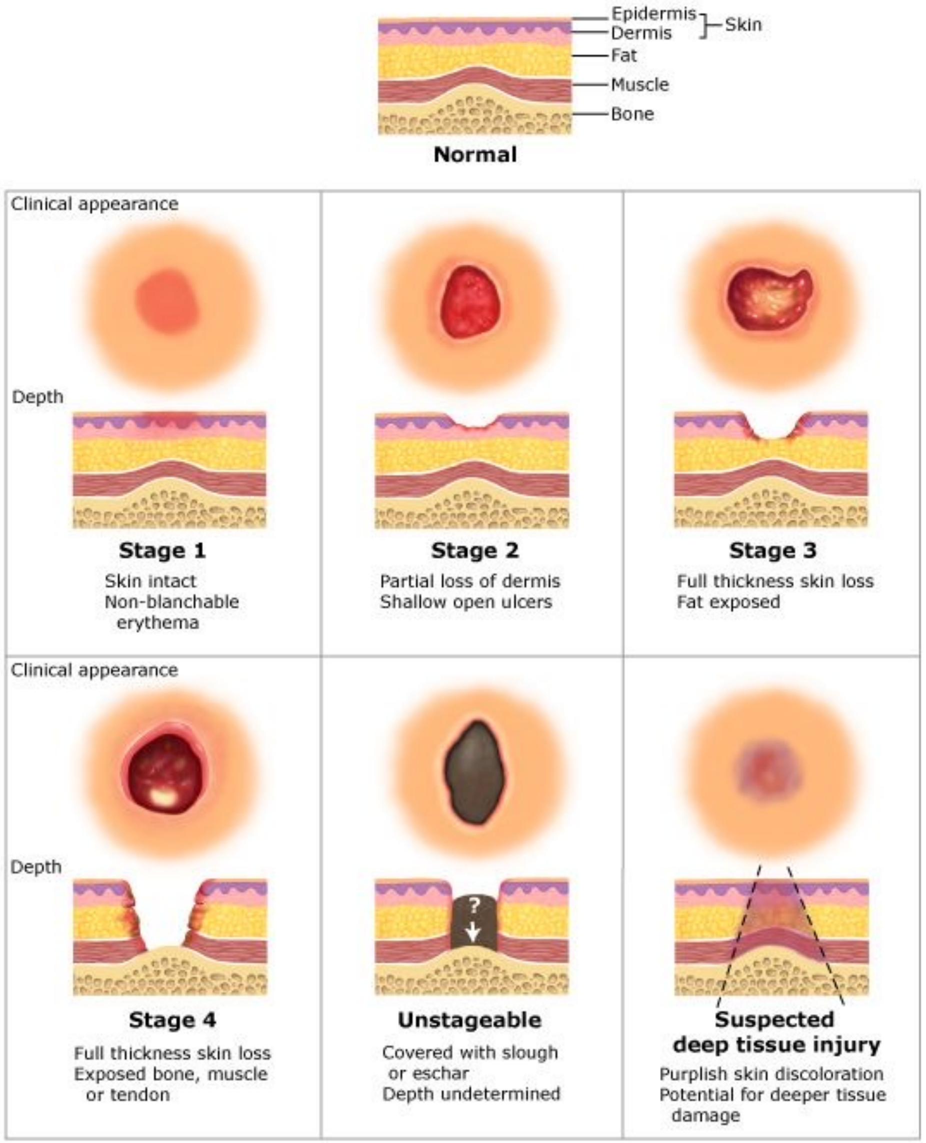
- **Problem:** For better predictability, it is critical to gather data that include people with disabilities and to ensure that these data are not completely subsumed by data from presumed “normative” populations
- **Conundrum:** Data collection in this context would also lead to issues related to confidentiality and privacy e.g., potentially dangerous for subjects
- **Multi-dimensional Problem:** Disability status has multiple dimensions, varies in intensity and impact, and varies changes over time. The simple protected vs. other class framework may not suffice

Data: Trouble with Labels

- In healthcare the ground truth can be subjective in nature
- Mental health evaluations, psychological assessments, pain assessment by clinicians vs. patients, patient reported outcomes
- Racial and sex/gender biased disparities have been observed for pain assessment across multiple studies
- Optimizing for the wrong label can lead to biased outcomes

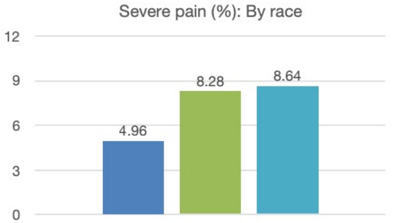
Data: Trouble with Labels: Pressure Injury

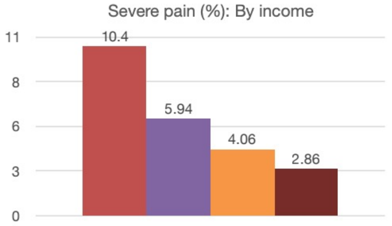
- “Localized damage to the skin and underlying soft tissue, usually over a bony prominence or related to a medical or other device.” NPUAP
- Multiple risk assessment exist for pressure injury (PI). The **Braden Scale** is the most widely used scale
- All measurement scales for PI are highly subjective in nature

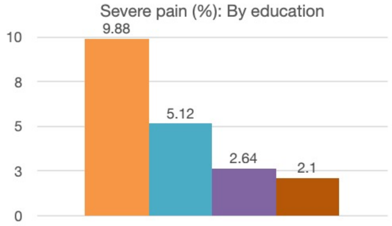


Data: Trouble with Labels: Knee Pain

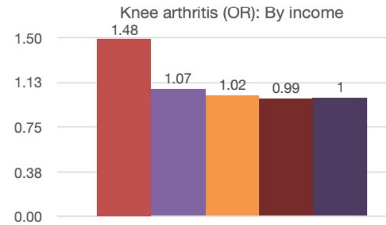
- Large **pain gradients**
  - Race
  - Income
  - Education

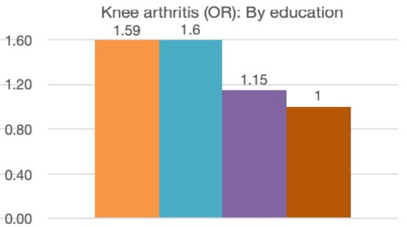

Severe pain (%): By race


Severe pain (%): By income

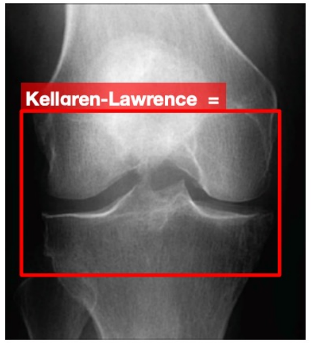

Severe pain (%): By education

- Higher prevalence of **painful conditions**
  - By income
  - By education

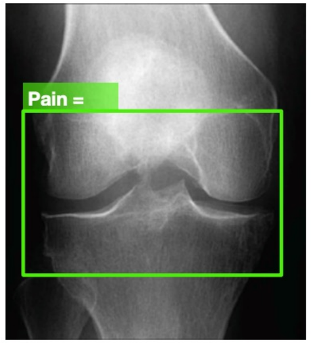

Knee arthritis (OR): By income

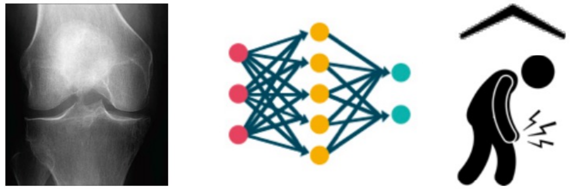

Knee arthritis (OR): By education

What if instead of **learning from the radiologist**…



We trained the algorithm to **listen to the patient?**





Simulation: Who would get surgery…
if the algorithm were in charge, not the doctor?

- Identify patients with severe pain and
  - High disease severity **according to human**
  - High disease severity **according to algorithm**

More - Black knees eligible for surgery

Less - Black knees, severe pain but ineligible for surgery

Severe pain + no surgery + high algorithm score = most likely to be on oral pain medicine ***incl. opiates***

Slide courtesy - Ziad Obermeyer from https://blogs.worldbank.org/impactevaluations/machine-learning-pain-relief
Grol-Prokopczyk, *Pain* 2017, Baldassari et al., *Osteoarthritis and Cartilage* 2014
Pierson, Emma, et al. "Using machine learning to understand racial and socioeconomic differences in knee pain" Under Review at JAMA 2019.

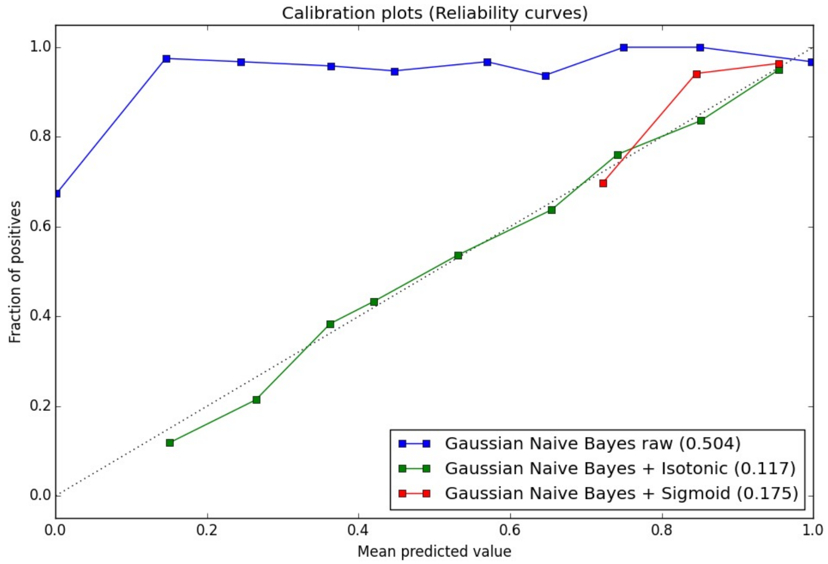
Data: Generalization

- In sub-Saharan Africa, women are diagnosed with breast cancer younger, on average, than are their peers in developed countries, and their disease is more advanced at diagnosis. Diagnostic AI tools trained on mammograms from Europe are primed to identify disease in its early stages in older women [Nordling 2019]
- Data security and access concerns have been raised about allowing developers to access such data from low-income countries

Fairness and Calibration

- **Calibration:** If we look at the set of people who receive a predicted probability of p, we would like a p fraction of the members of this set to be positive instances of the classification problem [Dawid 1982]
- If we are concerned about fairness between two groups G1 and G2 (e.g. African-American and white patients) then we would like this calibration condition to hold simultaneously for the set of people within each of these groups as well [Flores et al 2016]
- It is not feasible for certain notions of fairness

[Kleinberg et al 2016; Pleiss et al 2017]



The ‘Other’ Impossibility Theorem

Three notions of Calibration and Fairness

- **Group Calibration:** For each group *t*, and each bin *b* with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to *b*
- **Negative class Balance:** Requires that the average score assigned to people of across groups who belong to the negative class should be the same
- **Positive class Balance:** Requires that the average score assigned to people of across groups who belong to the negative class should be the same
- *Main Result:* It is not possible to satisfy all three conditions of calibration and fairness simultaneously

[Kleinberg et al 2016]

Deontic Justice & Fairness in Healthcare ML

- **Deontic Justice:** It is not just the state of affairs of unfairness that matters but also what were the conditions that led to that state of affairs [Binns 2018]
- This however requires integrating a perspective from philosophy, history, economics, sociology etc. This becomes a non-trivial problem
- Once identified, where should the locus of responsibility be; focus on improving outcomes
- When is a particular mistreatment of a protected group worse than the mistreatment of the protected group
- Modeling strongly coupled complex systems is hard!

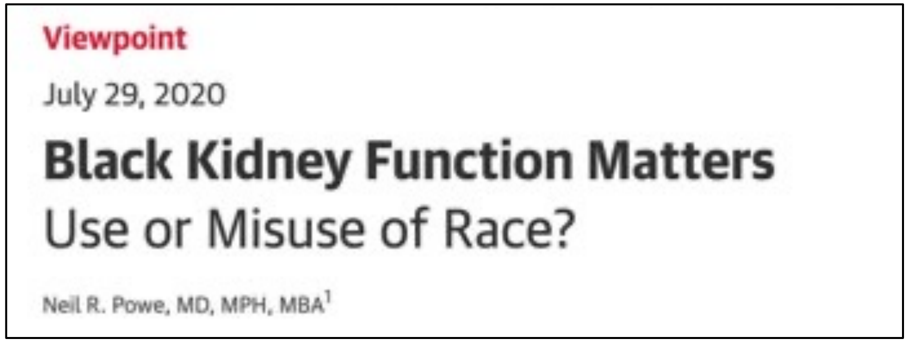
Luck Egalitarianism & Fairness in Healthcare

- What type of inequalities are acceptable?
- **Luck Egalitarianism:** Allow inequality in cases which result from people's efforts and risk taking and do not allow it in cases where it is because of brute luck (skin color, born with debilitating health condition) [Arneson 1989]
- **Coupled nature of social units:** People choices may be limited because they choose to take of sick, elderly, young family members
- Free choice are not always free; need to audit systems to determine how stakeholders are being affected?

Representational Fairness

- ***Distributive vs representative harms***

- stereotyping – the tendency to assign characteristics to all members of a group based on stereotypical features shared by a few [Abassi 2019]



Viewpoint

July 29, 2020

Black Kidney Function Matters

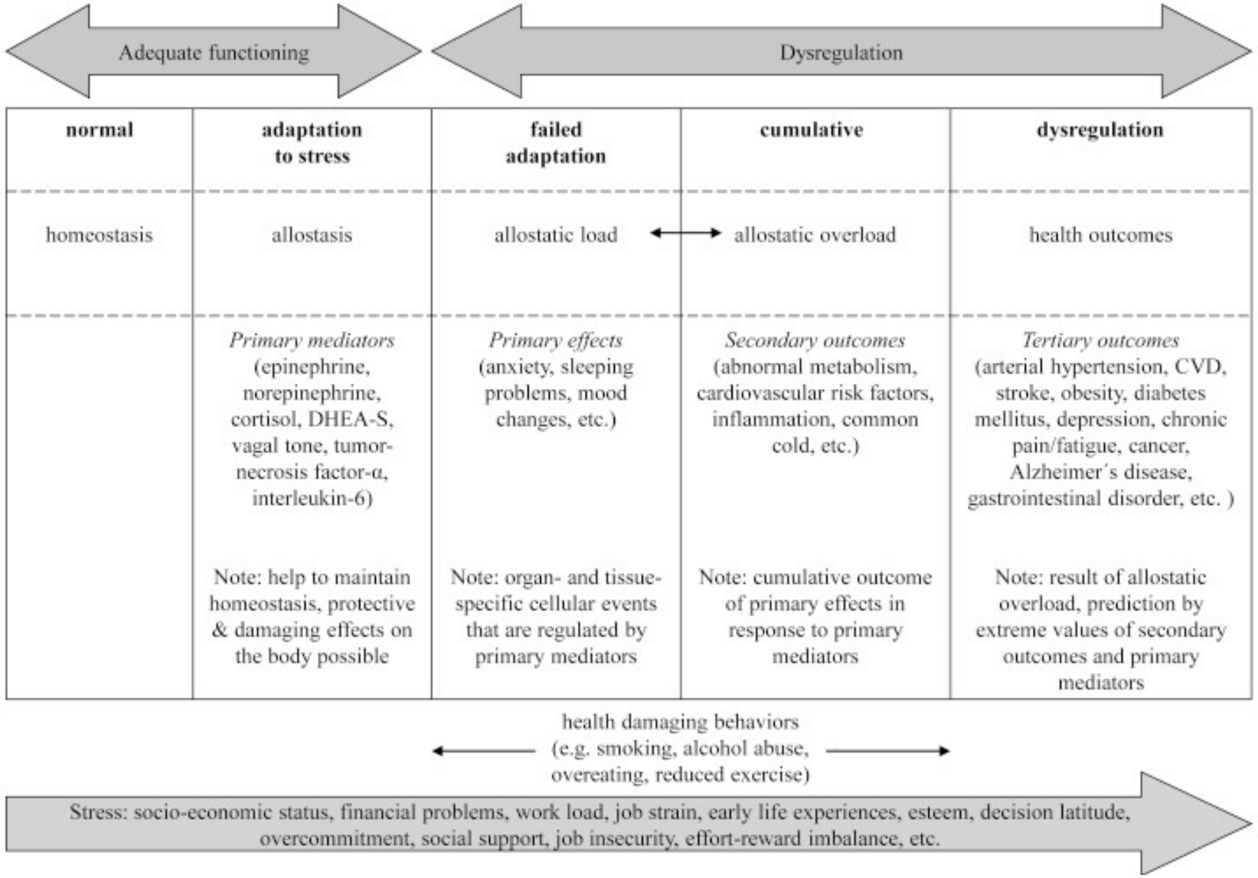
Use or Misuse of Race?

Neil R. Powe, MD, MPH, MBA[1]

- Such notions of representational fairness capture many of the most high-profile controversial examples of algorithmic bias [Binns 2018]

A Lifecycle view of Inequity

- What is the cumulative effect of the discrimination faced by a person over the course of a lifetime?
- What are the physiologic effects of chronic stressors related to disequity?



normal	adaptation to stress	failed adaptation	cumulative	dysregulation
homeostasis	allostasis	allostatic load ↔	allostatic overload	health outcomes
	Primary mediators (epinephrine, norepinephrine, cortisol, DHEA-S, vagal tone, tumor-necrosis factor-α, interleukin-6)	*Primary effects* (anxiety, sleeping problems, mood changes, etc.)	*Secondary outcomes* (abnormal metabolism, cardiovascular risk factors, inflammation, common cold, etc.)	*Tertiary outcomes* (arterial hypertension, CVD, stroke, obesity, diabetes mellitus, depression, chronic pain/fatigue, cancer, Alzheimer´s disease, gastrointestinal disorder, etc.)
	Note: help to maintain homeostasis, protective & damaging effects on the body possible	Note: organ- and tissue-specific cellular events that are regulated by primary mediators	Note: cumulative outcome of primary effects in response to primary mediators	Note: result of allostatic overload, prediction by extreme values of secondary outcomes and primary mediators

health damaging behaviors (e.g. smoking, alcohol abuse, overeating, reduced exercise)

Stress: socio-economic status, financial problems, work load, job strain, early life experiences, esteem, decision latitude, overcommitment, social support, job insecurity, effort-reward imbalance, etc.

[Mauss et al. 2015]

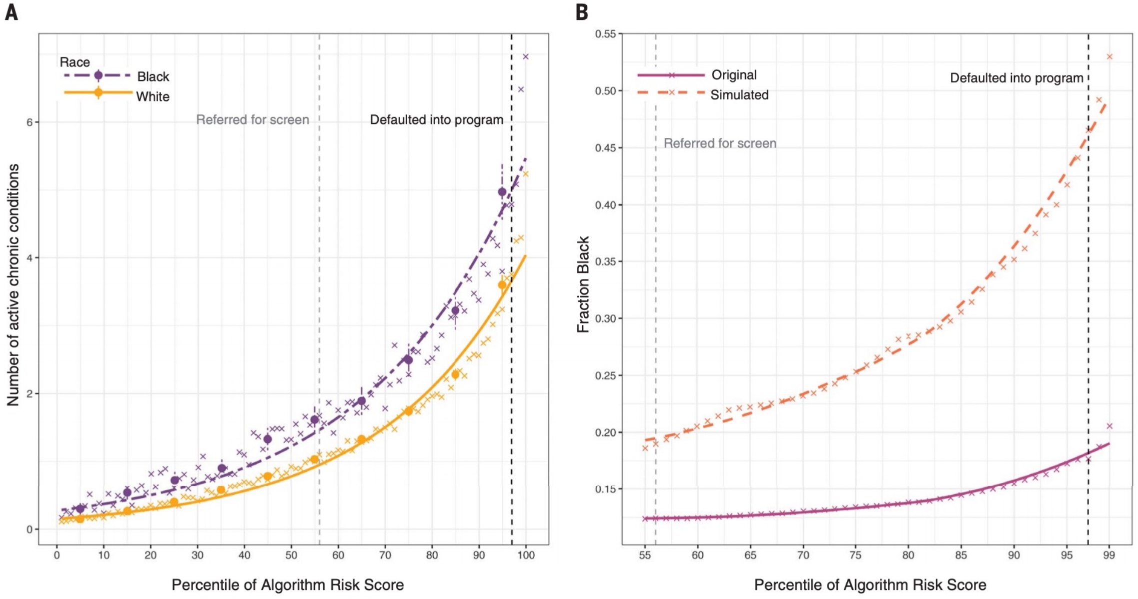
Tensions between disparate treatment and disparate impact

- Different groups have to be treated differently to maintain fairness.
- Humans (clinicians) deal with it on case-by-case basis. But this is not scalable for algorithmic decision making. [Narayanan 2018]
- Patient "no show" prediction

Operationalizing Fairness in Healthcare ML

Treatment Effect is not monotonic

- The predicted risk of some future outcome e.g., healthcare needs is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk. This is however not always true
- At the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites



Calculate an overall measure of health status, the number of active chronic conditions [or “comorbidity score,” a metric used extensively in medical research to provide a comprehensive view of a patient’s health] by race, conditional on algorithmic risk score.

[Obermeyer et al 2019]

Healthcare Needs ≠ Healthcare Costs

- Algorithm scores are a key input to decisions about future enrollment in a care coordination program
- If less-healthy Blacks scored at similar risk scores to more-healthy Whites, leading to substantial disparities in program screening
- The algorithm’s prediction on health needs is a prediction on health costs
- At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than Whites—on average, $1801 less per year, holding constant the number of chronic illnesses

[Obermeyer et al 2019]

Healthcare Needs ≠ Healthcare Costs

- Black patients generate very different kinds of costs: for example, fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis
- “These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities”

[Obermeyer et al 2019]

Long Term Impact of Fairness

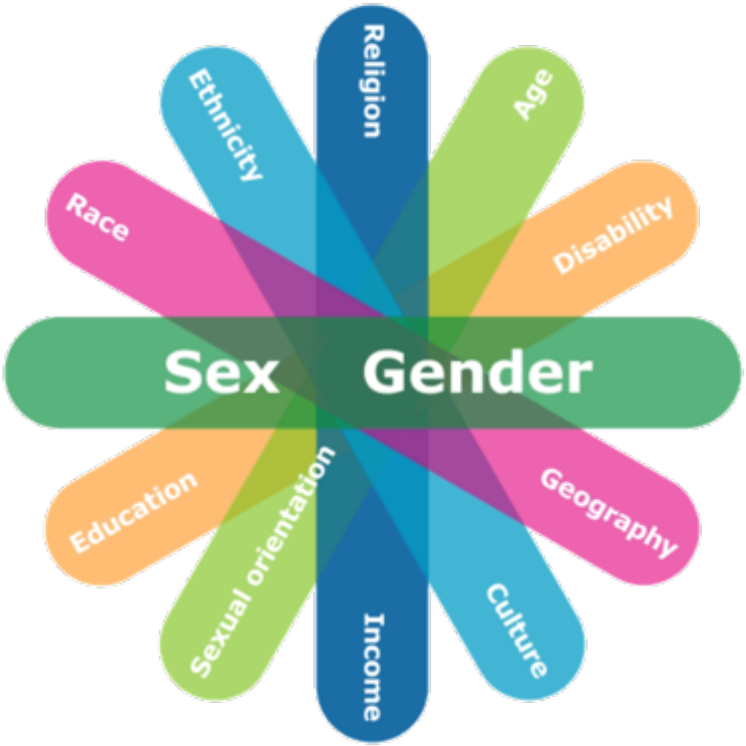
In socio-technical systems, we must consider how algorithms dynamically effect their environment, and the incentives of humans over time.

These kinds of effects are not considered when considering either statistical or individual notions of fairness in one-shot learning settings

Risk of readmission models and different incentive structures and programs

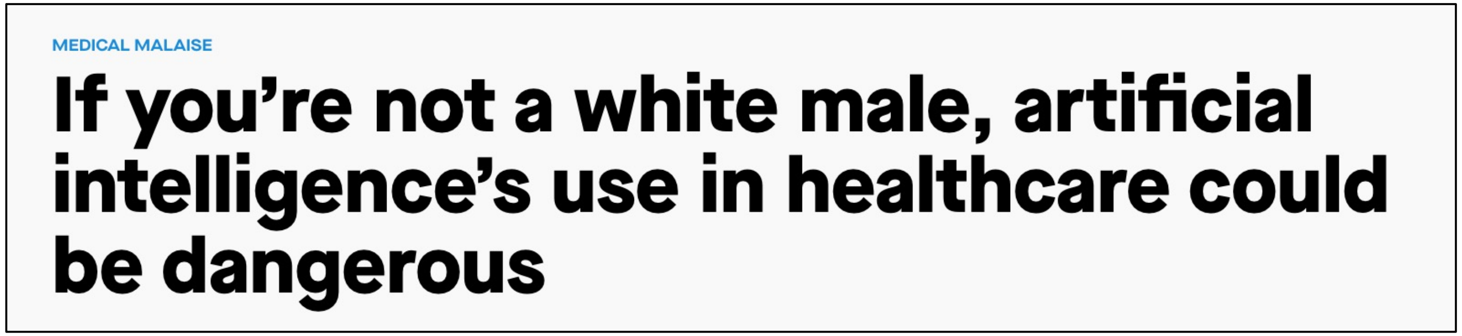
Fairness Gerrymandering (Intersectionality)

- **Intersectionality:** the interconnected nature of social categorizations such as race, class, and gender as they apply to a given individual or group, regarded as creating overlapping and interdependent systems of discrimination or disadvantage [Oxford Dictionary]
- Intersectionality is susceptible to (intentional or inadvertent) **fairness gerrymandering** where a classifier appears to be fair on each individual group, but not for subgroups



ML Problem: Intersectional Fairness

- The investigation of intersectional fairness, i.e., combination of multiple sensitive attributes, is relatively lacking in current research [47], [48]. Take bias mitigation for example, current work generally focus on one kind of bias. Although this may increase model fairness in terms of a specific bias, it is highly possible that the model is still biased from the intersectional perspective.



MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

Fairness Gerrymandering (Intersectionality)

- Statistical notions of fairness across exponentially (or infinitely) many subgroups, defined by a structured class of functions over the protected attributes
- This interpolates between statistical definitions of fairness, and recently proposed individual notions of fairness, but it raises several computational challenges. It is no longer clear how to even check or audit a fixed classifier to see if it satisfies such a strong definition of fairness
- The Computational problem of auditing subgroup fairness for both equality of false positive rates and statistical parity is equivalent to the problem of weak agnostic learning (Computationally hard in the worst case)
- However, it also suggests that common heuristics for learning can be applied to successfully solve the auditing problem in practice [Kearns et al 2017]

Multiaccuracy

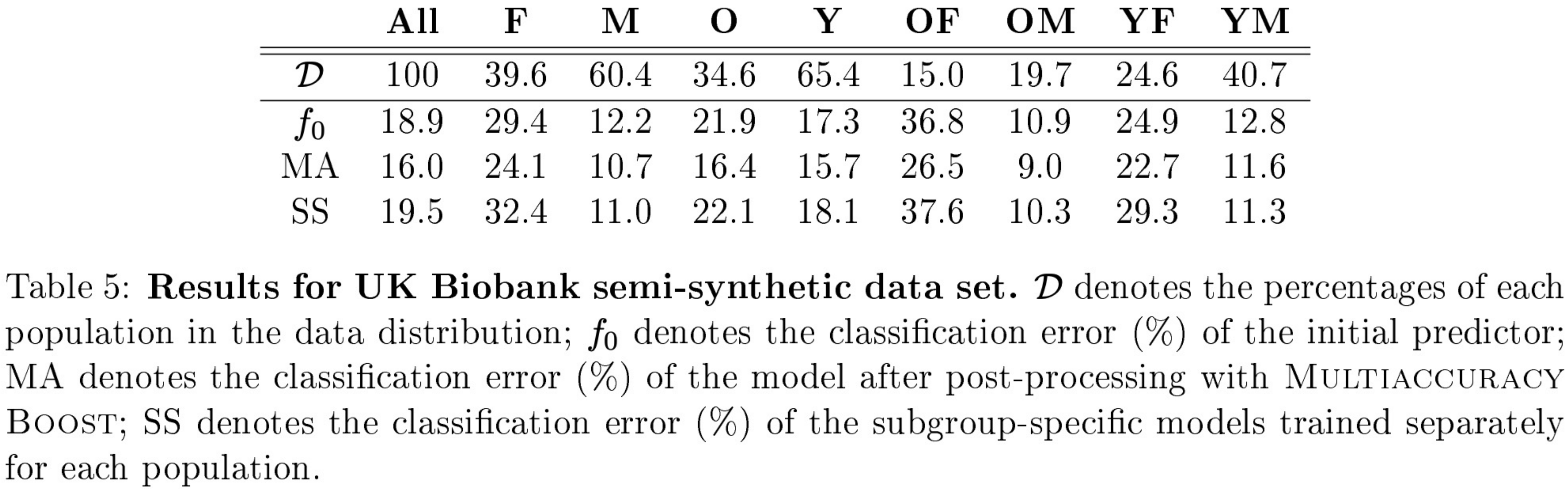
- **Multiaccuracy:** A strong notion of subgroup fairness. Models should be unbiased, overall as well as on but on every identifiable subpopulation
- Given: Black-box access to a classier *C*, and a relatively small validation set drawn from some representative distribution D
- Audit *C* to determine whether the predictor satisfies multiaccuracy.
- If auditing reveals that the predictor does not satisfy multiaccuracy, one could aim to post-process *C* to produce a new classier *C′* that is multiaccurate, without adversely affecting the subpopulations where C was already accurate [Kim et al 2019]

Multiaccuracy: Illustration

- Even with a ‘good` classifier, it may still exhibit biases on significant subpopulations when evaluated on a different sample distribution
- Scenario: Minority populations are underrepresented in the distribution used to train C vs. testing with different distributions
- Example: A disease prediction task based on real individuals, where the phenotype to disease relation is designed to be different for different subgroups [Kim et al 2019]
- 40,000 patient sampled from the UK Biobank with 60 features

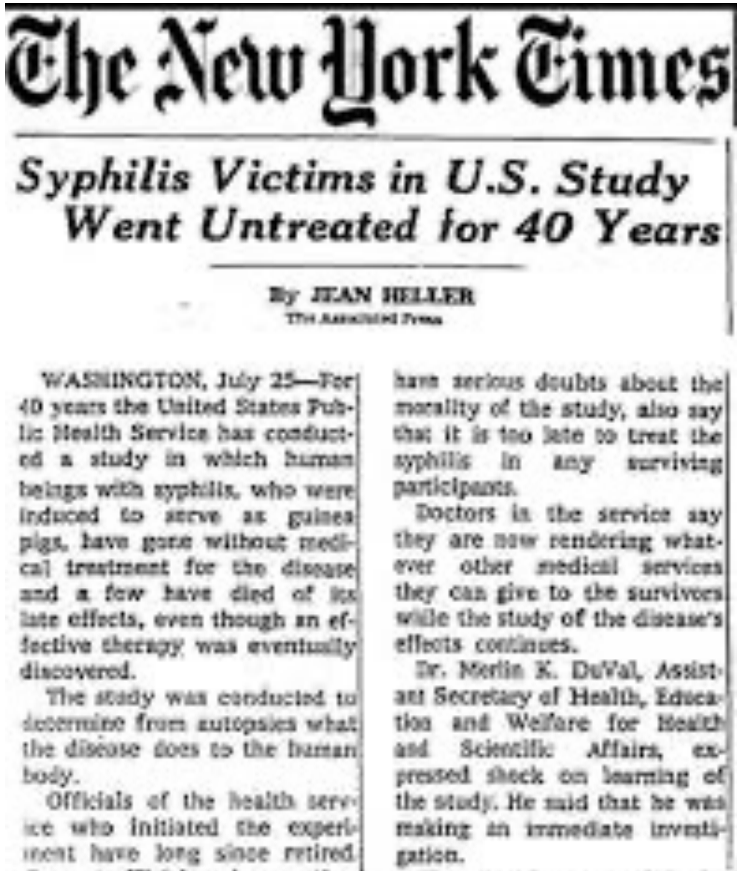
ML Problem: Multiaccuracy

- Generate a synthetic disease outcome for each subgroup, divide the data set into subgroups (Gender & Age)
- For each subgroup, create synthetic binary labels using a different polynomial function of the input features with different levels of difficulty



	All	F	M	O	Y	OF	OM	YF	YM
$\mathcal{D}$	100	39.6	60.4	34.6	65.4	15.0	19.7	24.6	40.7
f_0	18.9	29.4	12.2	21.9	17.3	36.8	10.9	24.9	12.8
MA	16.0	24.1	10.7	16.4	15.7	26.5	9.0	22.7	11.6
SS	19.5	32.4	11.0	22.1	18.1	37.6	10.3	29.3	11.3

Table 5: **Results for UK Biobank semi-synthetic data set.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; f_0 denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with MULTIACCURACY BOOST; SS denotes the classification error (%) of the subgroup-specific models trained separately for each population.



The New York Times

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea pigs, have gone without medical treatment for the disease and a few have died of its late effects, even though an effective therapy was eventually discovered.

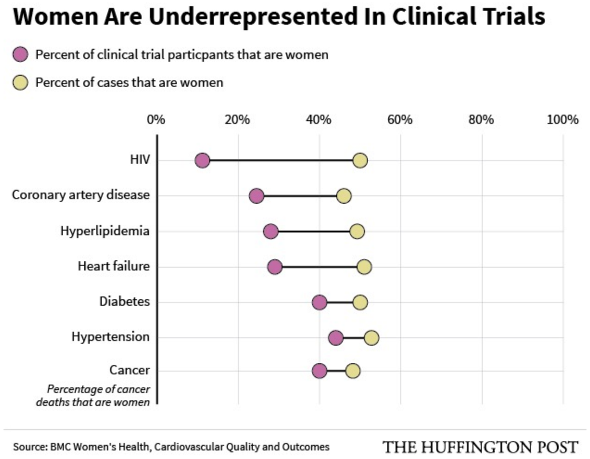
The study was conducted to determine from autopsies what the disease does to the human body.

Officials of the health service who initiated the experiment have long since retired.

have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say they are now rendering whatever other medical services they can give to the survivors while the study of the disease's effects continues.

Dr. Merlin K. DuVal, Assistant Secretary of Health, Education and Welfare for Health and Scientific Affairs, expressed shock on learning of the study. He said that he was making an immediate investigation.


Women Are Underrepresented In Clinical Trials

Source: BMC Women's Health, Cardiovascular Quality and Outcomes

THE HUFFINGTON POST

ML Problem: Exploration vs. Exploitation

Process Fairness vs. Outcome Fairness

- Process Fairness is ensuring that the process is fair and not just the outcome
- One way to measure it is by estimating the degree to which people consider the usage various features to be fair in a model (intuitive moral sense)
- Let *U* denote the set of all members of society, and *F* denote the set of all possible features that might be used in the decision-making process
- **Feature-Apriori Fairness:** Without a priori knowledge of how feature usage affects outcomes

$$\text{feature-apriori fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{\left|\bigcap_{f_i \in \mathcal{F}'} \mathcal{U}_{f_i}\right|}{|\mathcal{U}|}.$$

[Grgic-Hlaca et al 2016]

Process Fairness vs. Outcome Fairness

- **Feature-Accuracy Fairness:** Fair to use if it increases the accuracy of the classifier

$$\text{feature-accuracy fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Acc}_{f_i})|}{|\mathcal{U}|},$$

where

$$Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Acc}_{f_i}) = \begin{cases} \mathcal{U}_{f_i} \cup \mathcal{U}^{Acc}_{f_i}, & \text{if } Acc(\mathcal{C}_{\mathcal{F}'}) > Acc(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}) \\ \mathcal{U}_{f_i}, & \text{if } Acc(\mathcal{C}_{\mathcal{F}'}) \leq Acc(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}). \end{cases}$$

- **Feature-Disparity Fairness:** Fair to use even if it increases a measure of disparity (i.e. disparate impact or disparate mistreatment) of the classifier

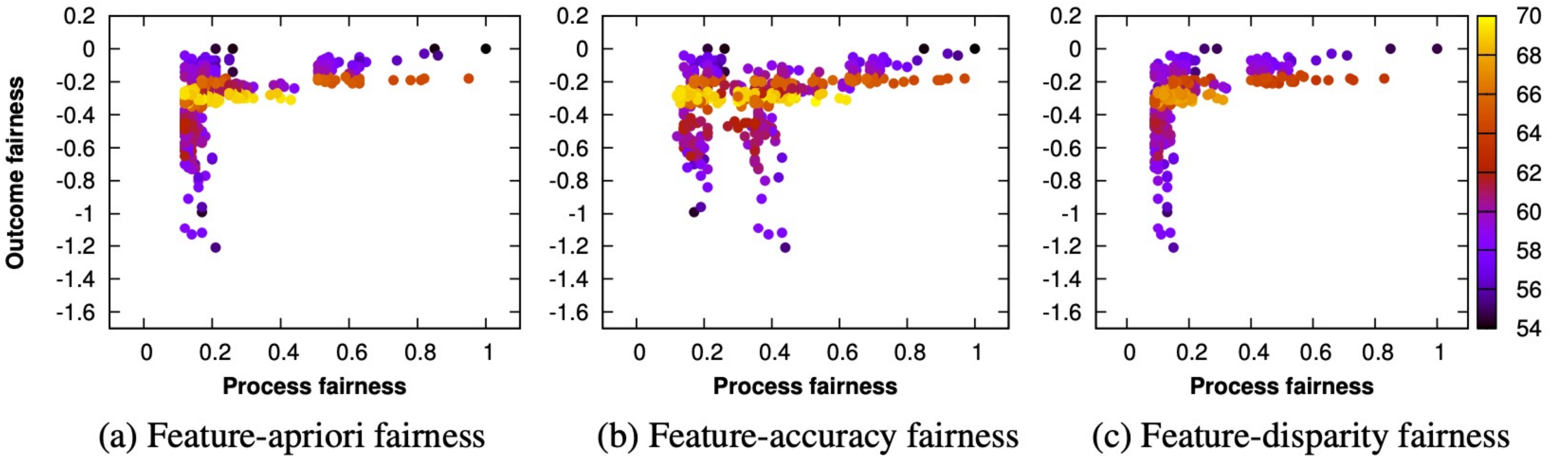
$$\text{feature-disparity fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Disp}_{f_i})|}{|\mathcal{U}|},$$

where

$$Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Disp}_{f_i}) = \begin{cases} \mathcal{U}_{f_i} \cup \mathcal{U}^{Disp}_{f_i}, & \text{if } Disp(\mathcal{C}_{\mathcal{F}'}) \leq Disp(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}) \\ \mathcal{U}^{Disp}_{f_i}, & \text{if } Disp(\mathcal{C}_{\mathcal{F}'}) > Disp(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}). \end{cases}$$

[Grgic-Hlaca et al 2016]

Process Fairness vs. Outcome Fairness



(a) Feature-apriori fairness (b) Feature-accuracy fairness (c) Feature-disparity fairness

Figure 2: Outcome fairness, measured as disparity in mistreatment, vs. different measures of process fairness for different classifiers. The color intensity of each point represents the accuracy of the corresponding classifier.

Process Fairness also exhibits Performance-Fairness trade-off

Decoupled Classifiers

- A model that ignores group membership may impose heterogenous trade-offs between groups
- **Decoupled classifiers:** Train a classifier for each group using data from that group
- Conditions: Each group should prefer their assigned model to (i) a pooled model that ignores group membership (rationality) and (ii) the model assigned to any other group (envy-freeness)

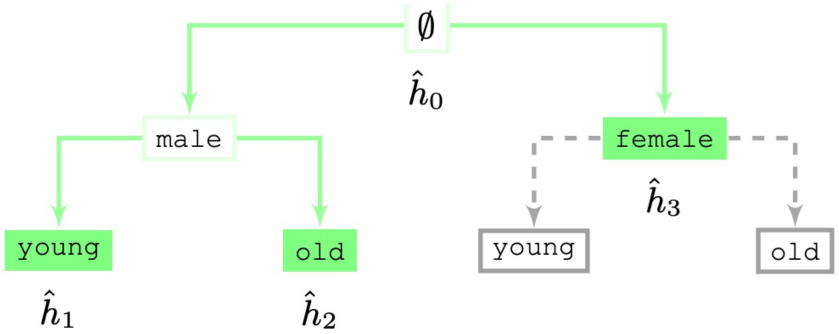


Figure 3. A set of decoupled classifiers assigned to 4 groups defined by 2 sensitive attributes $Z = (\texttt{male}, \texttt{female}) \times (\texttt{young}, \texttt{old})$. Here, we train the classifiers $H_T = \{\hat{h}_1, \hat{h}_2, \hat{h}_3\}$ using the data at the leaves $V_T = \{(\texttt{young}, \texttt{male}), (\texttt{old}, \texttt{male}), (\texttt{female})\}$. The tree structure ensures that decoupled classifiers are trained using the data pertaining to groups with shared sensitive attributes.

[Ustun et al 2018; Herbert-Johnson et al 2018]

Adversarial Debiasing

- Suppose we want to ensure that an adversary cannot infer the target variable
- For *Demography Parity*, the adversary gets the predicted label Yˆ . Intuitively, this allows the adversary to try to predict the protected variable using nothing but the predicted label
- For *Equality of Odds*, the adversary gets Yˆ and the true label Y
- For *Equality of Opportunity* on a given class y, we can restrict the training set of the adversary to training examples
- General, Model-Agnostic and Optimal (under certain conditions)

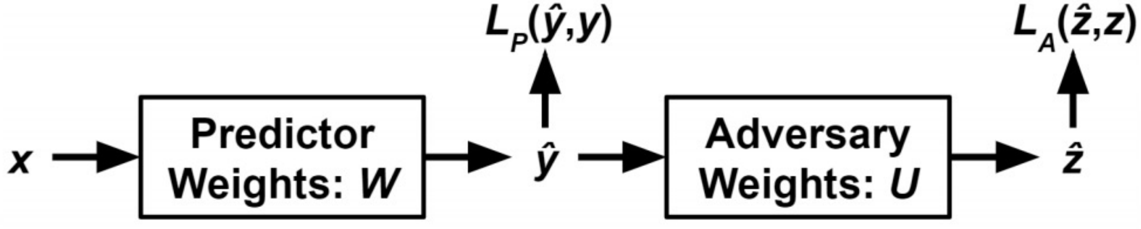
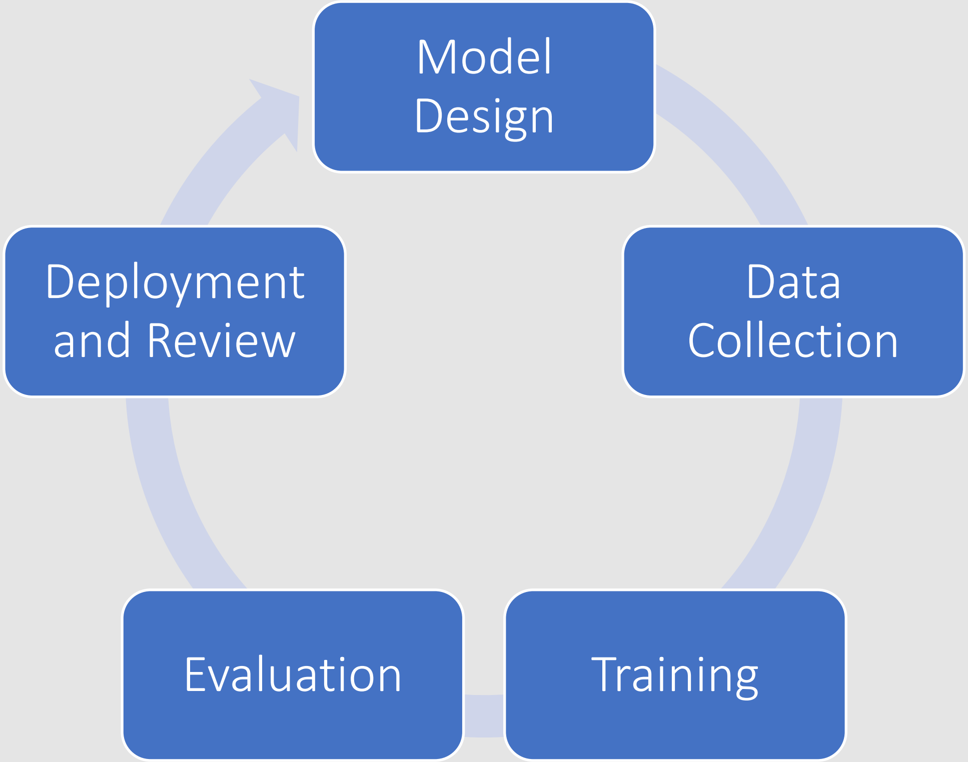


Figure 1: The architecture of the adversarial network.

[Beutel et al 2017; Zhang et al 2018]

Best Practices

Recommendations



[Rajkomar et al 2018]

Impossibility of Fairness in the real world

- Unfair practices do not exist in a vacuum but are embedded in the larger context of historical, social and political realities [Glymour et al 2019; Herington 2020]
- Measures of algorithmic bias assume that an algorithm which is fair in the abstract will be fair in the world.
- Centuries of injustice continue to permeate society and continue to be responsible for race- and gender-based inequality
- Implicit vs explicit biases can be difficult and / or impossible to adjust for and demand societal changes

Prediction & Policy

- Allocation of services, particularly those derived from outputs of machine learning models, must be continually evaluated for evidence of bias to ensure that services are delivered equally across protected groups
- The allocation of services will be determined by how clinicians or other end users interact with the model
  - Is there a disparate impact?
  - Is the clinical team subject to automation bias or dismissal bias? And how may that differentially affect patient groups?
  - Opportunity cost

Best Practices: Task Definition

- Clearly define the task & model’s intended effects
- Try to identify and document unintended effects & biases
- Clearly define any fairness requirements
- Involve diverse stakeholders & multiple perspectives
- Refine the task definition & be willing to abort
- [Cramer et al 2019]

Best Practices: Testing

- Check that test data matches deployment context
- Ensure test data has sufficient representation
- Continue to involve diverse stakeholders
- Revisit all fairness requirements
- Use metrics to check that requirements are met

[Cramer et al 2019]

Best Practices: Deployment

- Continually monitor – match between training data, test data, and instances you encounter in deployment – fairness metrics – user reports & user complaints
- Invite diverse stakeholders to audit system for biases
- Methods/tools to audit for shifts in population
- Methods/tools to determine whether a particular error is a one-off issue or is indicative of a systemic problem
- Audit existing system for biases (in collaboration with the teams that built the systems whenever possible)
  [Cramer et al 2019]

Best Practices: Feedback

- Continue to monitor – match between training data, test data, and instances you encounter in deployment – fairness metrics – user reports & user complaints
- Monitor users' interactions with system
- Consider prohibiting some types of interactions [Cramer et al 2019]

Challenges & Open Questions

- Which measures of fairness are most appropriate in a given context?
- Which variables are legitimate grounds for differential treatment, and why?
- When is disparity between groups acceptable and why?
- Should fairness consist of maximizing equal probability of obtaining some benefit, or minimizing the harms to the least advantaged?
- In making such tradeoffs, should the decision-maker consider only the harms and benefits imposed within the decision-making context, or also those faced by decision-subjects in other contexts?
- What relevance should past, future or inter-generational injustices have?

[Binns 2018]

Challenges & Open Questions

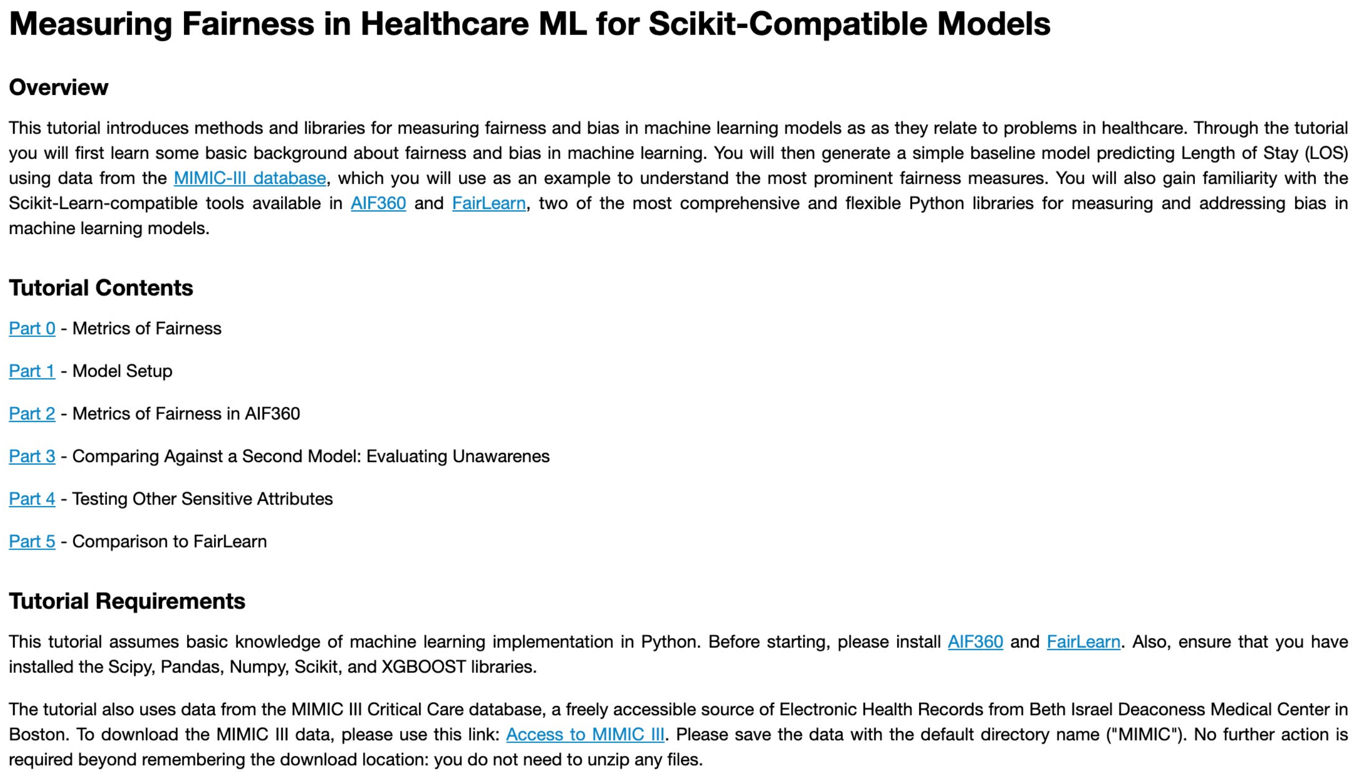
- Many aspects of fairness not captured by metrics or data; how do we address those?
- How to deal with Fairness Gerrymandering where there is insufficient data for modeling?

Library Demo

FairMLHealth Library

- Vision
  - An extensible Python library dedicated to fairness in machine learning specifically tailored for healthcare with domain knowledge integration
- Future Goals & Milestones
  - Measurement of Fairness in Healthcare Applications
  - Comparison of classifiers for Fairness and Performance Trade-offs
  - Arbitrary comparison of protected classes and intersectional classes
- Current Release: FairMLHealth 0.1: Alpha Release
  - Demonstration of measurement and comparison of fairness metrics for a publicly available dataset (MIMIC-3)

FairMLHealth Library



Measuring Fairness in Healthcare ML for Scikit-Compatible Models

Overview

This tutorial introduces methods and libraries for measuring fairness and bias in machine learning models as as they relate to problems in healthcare. Through the tutorial you will first learn some basic background about fairness and bias in machine learning. You will then generate a simple baseline model predicting Length of Stay (LOS) using data from the MIMIC-III database, which you will use as an example to understand the most prominent fairness measures. You will also gain familiarity with the Scikit-Learn-compatible tools available in AIF360 and FairLearn, two of the most comprehensive and flexible Python libraries for measuring and addressing bias in machine learning models.

Tutorial Contents

Part 0 - Metrics of Fairness

Part 1 - Model Setup

Part 2 - Metrics of Fairness in AIF360

Part 3 - Comparing Against a Second Model: Evaluating Unawarenes

Part 4 - Testing Other Sensitive Attributes

Part 5 - Comparison to FairLearn

Tutorial Requirements

This tutorial assumes basic knowledge of machine learning implementation in Python. Before starting, please install AIF360 and FairLearn. Also, ensure that you have installed the Scipy, Pandas, Numpy, Scikit, and XGBOOST libraries.

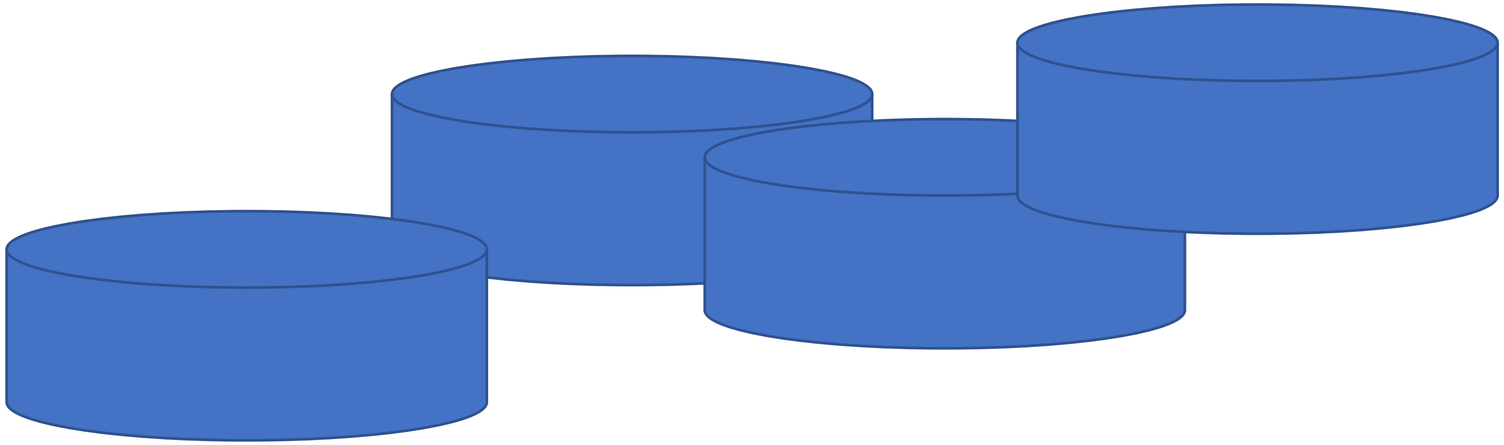
The tutorial also uses data from the MIMIC III Critical Care database, a freely accessible source of Electronic Health Records from Beth Israel Deaconess Medical Center in Boston. To download the MIMIC III data, please use this link: Access to MIMIC III. Please save the data with the default directory name ("MIMIC"). No further action is required beyond remembering the download location: you do not need to unzip any files.

Recap & Conclusion

Recap

- Foundations: Fairness in Healthcare ML
- Measurement & Mismeasurement of Fairness
- Operationalizing Fairness in Healthcare ML
- Domain Challenges in Healthcare ML
- Fairness in Healthcare ML in Action
- Best Practices
- Library Demo
- Conclusion

Call to Action: Datasets



- Deployment of Enterprise grade AI and ML models in healthcare at multiple locations in the US and internationally
- Fairness across multiple locations, settings and cohort

Call to Action: Partner with us!
It takes a village

- Deployment of Enterprise grade AI and ML models in healthcare at multiple locations in the US and internationally
- Fairness across multiple locations, settings and cohort

Call to Action: Resources: Websites

- AI Now Institute
- Algorithmic Justice League
- Berkman Klein Center for Internet and Society
- ML Healthcare Resources
- Partnership on AI

Call to Action: Resources: Libraries

Library	Creator	Metrics	Algorithms	Simulations
AIF 360	IBM	✅	✅	
Fairlearn	Hannah Wallach et al	✅	✅	
Fairness Comparison	Sorelle Friedler	✅		
Fairness Indicators	Tensorflow	✅		
ML Fairness Gym	Google			✅
Themis-ML	Niels Bantilan	✅	✅	

And now the FairMLHealth Library

References

- Byrd, W. Michael, and Linda A. Clayton. "Race, medicine, and health care in the United States: a historical survey." Journal of the National Medical Association 93, no. 3 Suppl (2001): 11S.
- Wailoo, Keith. "Sickle cell disease—a history of progress and peril." N Engl J Med 376, no. 9 (2017): 805-807.
- de Malave, Florita Z. Louis. Sterilization of Puerto Rican women: a selected, partially annotated bibliography. University of Wisconsin System, Women's Studies Librarian, 1999.
- Randall, Vernellia R. "Slavery, Segregation and Racism: Trusting the Health Care System Ain't Always Easy--An African American Perspective on Bioethics." . Louis U. Pub. L. Rev. 15 (1995): 191.
- King, M. L. Jr. (1966). National Convention for Medical Committee for Human Rights. Washington, DC. Excerpt from speech retrieved from http://www.goodreads.com/quotes/106932-of-all-the-forms-ofinequality-injustice-in-health-care
- Parikh, Ravi B., Stephanie Teeple, and Amol S. Navathe. "Addressing bias in artificial intelligence in health care." Jama 322, no. 24 (2019): 2377-2378.
- Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and bias in human use of automation: An attentional integration." Human factors 52, no. 3 (2010): 381-410.
- Agniel, Denis, Isaac S. Kohane, and Griffin M. Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study." Bmj 361 (2018).
- McLintock K, Russell AM, Alderson SL, et al. The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. BMJ Open2014;4:e005178. doi:10.1136/bmjopen-2014-005178. pmid:25142262
- Weizenbaum, Joseph (1976). Computer power and human reason : from judgment to calculation. San Francisco: W.H. Freeman
- Adelman, Larry. "Unnatural causes: Is inequality making us sick?." Preventing Chronic Disease 4, no. 4 (2007).
- Simkin RJ. Women's health: time for a redefinition. *CMAJ*. 1995;152(4):477-479.

References

- Almond, Amanda Lee. "Measuring racial microaggression in medical practice." Ethnicity & health 24, no. 6 (2019): 589-606.Barocas, Solon and Hardt, Moritz., Fairness in machine learning, NeurIPS Tutorial, 2017.
- Dovidio, John F., Susan Eggly, Terrance L. Albrecht, Nao Hagiwara, and Louis A. Penner. "Racial biases in medicine and healthcare disparities." TPM: Testing, Psychometrics, Methodology in Applied Psychology 23, no. 4 (2016).
- Agniel, Denis, Isaac S. Kohane, and Griffin M. Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study." Bmj 361 (2018).
- Parikh, Ravi B., Stephanie Teeple, and Amol S. Navathe. "Addressing bias in artificial intelligence in health care." Jama 322, no. 24 (2019): 2377-2378.
- Bierman, Arlene S. "Sex matters: gender disparities in quality and outcomes of care." Cmaj 177, no. 12 (2007): 1520-1521.
- Binns, Reuben. "Fairness in machine learning: Lessons from political philosophy." arXiv preprint arXiv:1712.03586 (2017).
- Chen, Esther H., Frances S. Shofer, Anthony J. Dean, Judd E. Hollander, William G. Baxt, Jennifer L. Robey, Keara L. Sease, and Angela M. Mills. "Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain." Academic Emergency Medicine 15, no. 5 (2008): 414-418.
- Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." arXiv preprint arXiv:1810.08810 (2018).
- Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and bias in human use of automation: An attentional integration." Human factors 52, no. 3 (2010): 381-410.

References

- McLintock K, Russell AM, Alderson SL, et al. The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. BMJ Open2014;4:e005178. doi:10.1136/bmjopen-2014-005178. pmid:25142262
- Weizenbaum, Joseph (1976). Computer power and human reason : from judgment to calculation. San Francisco: W.H. Freeman
- Adelman, Larry. "Unnatural causes: Is inequality making us sick?." Preventing Chronic Disease 4, no. 4 (2007).
- Almond, Amanda Lee. "Measuring racial microaggression in medical practice." Ethnicity & health 24, no. 6 (2019): 589-606.Barocas, Solon and Hardt, Moritz., Fairness in machine learning, NeurIPS Tutorial, 2017.
- Byrd, W. Michael, and Linda A. Clayton. "Race, medicine, and health care in the United States: a historical survey." Journal of the National Medical Association 93, no. 3 Suppl (2001): 11S.
- Wailoo, Keith. "Sickle cell disease—a history of progress and peril." N Engl J Med 376, no. 9 (2017): 805-807.
- de Malave, Florita Z. Louis. Sterilization of Puerto Rican women: a selected, partially annotated bibliography. University of Wisconsin System, Women's Studies Librarian, 1999.
- Randall, Vernellia R. "Slavery, Segregation and Racism: Trusting the Health Care System Ain't Always Easy--An African American Perspective on Bioethics." . Louis U. Pub. L. Rev. 15 (1995): 191.
- King, M. L. Jr. (1966). National Convention for Medical Committee for Human Rights. Washington, DC. Excerpt from speech retrieved from http://www.goodreads.com/quotes/106932-of-all-the-forms-ofinequality-injustice-in-health-care

References

Corbett-Davies, Sam., Goel, Sharad., Defining and Designing Fair Algorithms, Tutorials at EC 2018 and ICML 2018.

Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." arXiv preprint arXiv:1808.00023 (2018).

Cramer, Henriette., Holstein, Kenneth., Jennifer Wortman Vaughan et. al., Translation Tutorial: Challenges of incorporating algorithmic fairness into industry practice, FAT* Tutorial, 2019.

Crawford, Kate The Trouble with Bias, NeurIPS Keynote, 2017.

Dawes, Robyn M., David Faust, and Paul E. Meehl. "Clinical versus actuarial judgment." *Science* 243, no. 4899 (1989): 1668-1674.

Dresser, Rebecca. "Wanted single, white male for medical research." The Hastings Center Report 22, no. 1 (1992): 24-29.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. "A comparative study of fairness-enhancing interventions in machine learning." In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 329-338. 2019.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv preprint arXiv:1609.07236 (2016).

Garcia, M (2017): “Racist in the Machine: The Disturbing Implications of Algorithmic Bias” In World Policy Journal.

Gajane, Pratik, and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning." arXiv preprint arXiv:1710.03184 (2017).

Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." *IBM Journal of Research and Development* 63, no. 4/5 (2019): 4-1.

References

- Ghassemi, Marzyeh, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. "Opportunities in machine learning for healthcare." arXiv preprint arXiv:1806.00388 (2018).
- Grote, Thomas, and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare." Journal of Medical Ethics 46, no. 3 (2020): 205-211.
- Hajian, Sara., Bonchi, Francesco and Castillo, Carlos ., Algorithmic bias: From discrimination discovery to fairness-aware data mining, KDD Tutorial, 2016.
- Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-16. 2019.
- Hutchinson, Ben., and Mitchell, Margaret., "50 Years of Test (Un) fairness: Lessons for Machine Learning." In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 49-58. 2019.

Hutchinson, Ben., and Mitchell, Margaret., Translation Tutorial: A History of Quantitative Fairness in Testing, FAT* Tutorial, 2019.

Jensen, Arthur R. "Bias in mental testing." (1980).

Joseph, Matthew, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. "Rawlsian fairness for machine learning." arXiv preprint arXiv:1610.09559 1, no. 2 (2016).

Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." In Advances in Neural Information Processing Systems, pp. 4066-4076. 2017

Kamiran, Faisal, and Toon Calders. “Classifying Without Discriminating.” In *Proc. 2nd International Conference on Computer, Control and Communication*, 2009.

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding discrimination through causal reasoning." In *Advances in Neural Information Processing Systems*, pp. 656-666. 2017.

Krieger, Nancy. "Discrimination and health inequities." *International Journal of Health Services* 44, no. 4 (2014): 643-710.

References

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. “Delayed Impact of Fair Machine Learning.” In *Proc. 35th ICML*, 3156–64, 2018.

Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv preprint arXiv:1706.07269 (2017).

Nabi, Razieh, and Ilya Shpitser. "Fair inference on outcomes." In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Narayanan, Arvind 21 fairness definitions and their politics, FAT* Tutorial, ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*) 2018

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. "Ensuring fairness in machine learning to advance health equity." Annals of internal medicine 169, no. 12 (2018): 866-872.

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine learning in medicine." New England Journal of Medicine 380, no. 14 (2019): 1347-1358.

Rudin, Cynthia, and Berk Ustun. "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice." Interfaces 48, no. 5 (2018): 449-466.

Smedley BD, Stith AY, Nelson AR. Institute of medicine, committee on understanding and eliminating racial and ethnic disparities in health care. Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare Washington, DC: National Academies Press; 2003.

Vayena, Effy, Alessandro Blasimme, and I. Glenn Cohen. "Machine learning in medicine: addressing ethical challenges." PLoS medicine 15, no. 11 (2018).

Tamayo-Sarver, Joshua H., Susan W. Hinze, Rita K. Cydulka, and David W. Baker. "Racial and ethnic disparities in emergency department analgesic prescription." American journal of public health 93, no. 12 (2003): 2067-2073.

Thomas SB, Quinn SC. The Tuskegee Syphilis Study, 1932 to 1972: implications for HIV education and AIDS risk education programs in the black community. *Am J Public Health*. 1991;81(11):1498-1505. doi:10.2105/ajph.81.11.1498

Zhang, Junzhe, and Elias Bareinboim. “Fairness in Decision-Making — the Causal Explanation Formula.” In *Proc. 32nd AAAI*, 2018.

Appendix