Fairness in Data Science: Criteria, Algorithms and Open Problems

---

16th annual Innovations in Design, Analysis, and Dissemination (IDAD):

Frontiers in Biostatistics and Data Science meeting

Razieh Nabi, Ph.D.
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
razieh.nabi@emory.edu

Agenda

Part I

- Introduce alg fairness considerations via a series of examples
- Statistical fairness criteria
- Issues with statistical fairness criteria

Part II

- Introduce causal inference
- Relevant causal concepts, e.g., mediation and path-specific effects
- General causal perspective on algorithmic fairness constraints

Part III

- Imposing causal fairness constraints via constrained optimization
- Example application

Part I

Intro to alg fairness and statistical fairness criteria

Example 1/4: health risk screening algs

- Obermeyer et al. (2019) examine a commercial risk prediction algorithm used to manage health decisions for millions of hospital patients.
  - The algorithm's stated goal is to predict complex health needs for the purpose of targeting an intervention that manages those needs.
- Each patient is assigned risk score R by the alg: prediction of medical expenditures Y based on claims data from previous year X

$$R := \mathbb{E}[Y \mid X]$$

- X did not include race, and the score was approximately calibrated by race:

$$\mathbb{E}[R \mid Y, \text{Black}] \approx \mathbb{E}[R \mid Y, \text{White}]$$

Example 1/4: health risk screening algs ctd.

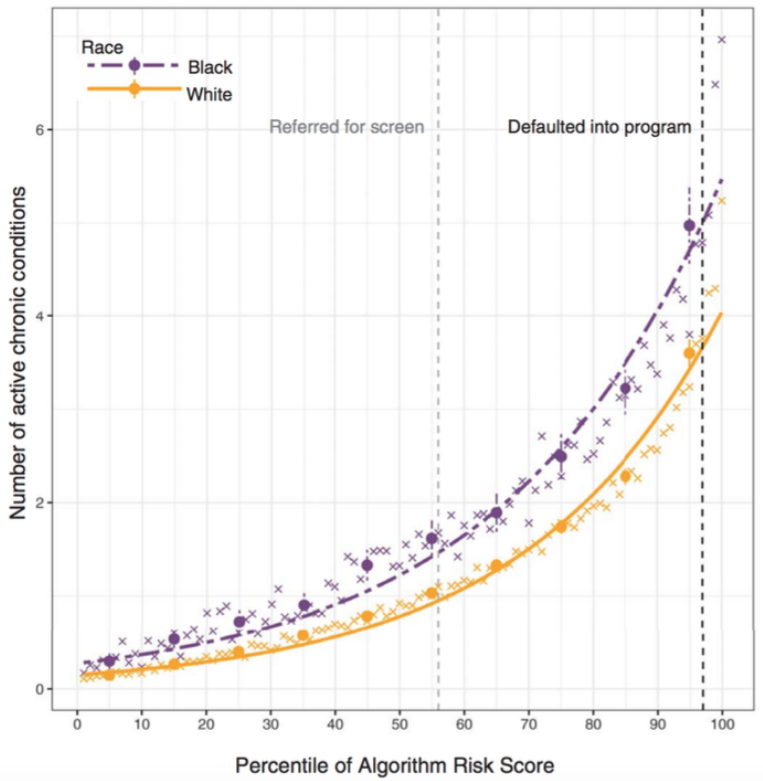
Are black and white patients with same predicted risk equally healthy? I.e.,

$$\mathbb{E}[H \mid R, \text{White}] \stackrel{?}{=} \mathbb{E}[H \mid R, \text{Black}]$$

- $\mathbb{E}[H \mid R, \text{White}] > \mathbb{E}[H \mid R, \text{Black}]$

  Across various definitions of "healthy," less-healthy Blacks scored at similar risk scores to more-healthy Whites.

- Scores are used to screen patients for a care management program, so Black patients are systematically under-enrolled.



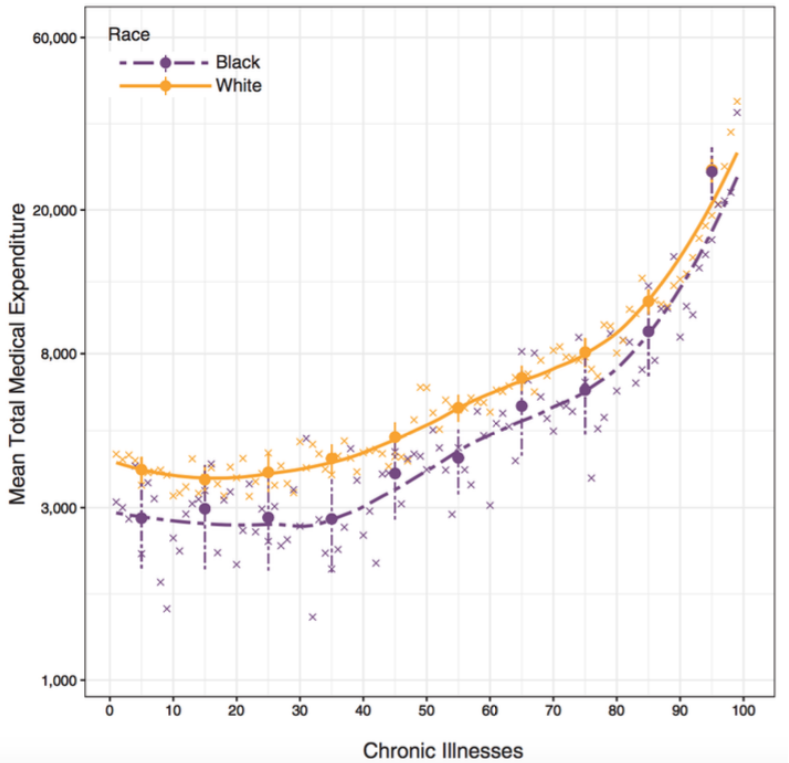
- What's going on here, and how can it be fixed?

Example 1/4: health risk screening algs ctd.

- $\mathbb{E}[Y \mid H, \text{White}] > \mathbb{E}[Y \mid H, \text{Black}]$

  Medical expenditures Y differed systematically by race.

- Diagnosis:
  "Health costs $\neq$ health needs"
  & socially unequal society has racial disparities in health costs.



Key takeaways:

- Target of prediction (label) can be a bad proxy for the underlying quality of interest, disparities can be "built in" to the outcome Y.
- Race info was not used in building the algorithm, so direct use of race is neither necessary nor sufficient for disparities to arise.
- In this case, $\mathbb{E}[H \mid R, \text{White}] = \mathbb{E}[H \mid R, \text{Black}]$ was used as a criterion to diagnose a problem.

Example 2/4: financial lending algs

- Clients apply to bank for loans. Banks make decisions based on default risk: won't give loan to "risky" clients. Roughly:
  - Based on historical data, model predicts timely repayment vs. non-repayment ($Y = 1$ vs 0)
  - For new client, based on their characteristics use model to estimate $P(Y = 1)$
  - Use some threshold to differentially offer loans to clients on the basis of these predictions

Example 2/4: financial lending algs ctd.

- Legal concerns may prevent bank from using protected attribute "race" directly in this model.
- However, strong correlations btw race and other vars (zipcode, neighborhood SES, home ownership, parental education, ...) may lead to very different loan rates across race groups even in "race-blind" model.
- Exclusion of race makes little practical difference to loan decisions.

Key takeaways:

- "Proxy correlations" illustrate how simply excluding race will fail to address equity
- Empirical studies suggest that more accurate algorithms my exacerbate disparities (not eliminate)
  - Why? detection of nonlinear relationships among race, outcomes, and other vars

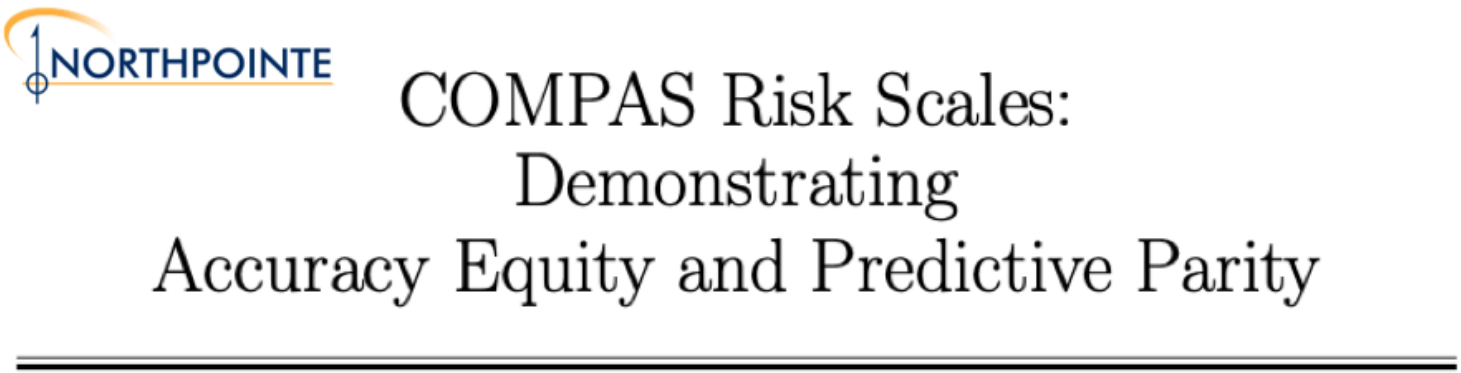
Example 3/4: automated resume screening for hiring

- Many institutions use algorithmic tools to automatically screen (or rank) resumes of job applicants.
- Infamous example: Amazon developed (but supposedly never used) a resume screening tool that was found to favor male job applicants. According to Reuters report:
  - Penalized resumes that included the word "women's," as in "women's chess club captain." Downgraded graduates of two all-women's colleges.
  - Favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured."

Example 4/4: recividism risk prediction

- Algorithms for recidivism risk prediction have been used in various criminal justice contexts: pretrial release conditions, bail determinations, etc.
- COMPAS is a tool from the company Northpointe that has been at the center of much attention since ProPublica published its critical analysis in 2016.
- ProPublica's analysis mostly focused on differences in error rates:
  - "Black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism", higher false positive rate for Blacks.
  - "White defendants were more likely than black defendants to be incorrectly flagged as low risk", lower false negative rate for Whites.
  - Authors argued the above amounts to discrimination against Blacks.

Example 4/4: recividism risk prediction ctd.



- Northpointe published a long response disputing the ProPublica's findings and claiming that "Propublica wrongly defines measures of discrimination."
  - ProPublica: focus on FPR and FNR, are these equal across groups?
  - Northpointe: focus on PPV, is probability of recidivating, given a high risk score, similar for blacks and whites?

Algorithmic fairness

These examples prompt two main questions:

Q1. Why *do* algorithms introduce unfair biases?

Q2. What *is* a good measure of unfairness?

- Algorithms introduce/reproduce/perpetuate disparities
  - Reliance on humans on every step of their development.
  - Reflecting the socially stratified, disparate, and unfair reality behind data.
  - Sensitive features may include: race, gender, sexual orientation, religion, etc.
- Ignoring the sensitive attributes is neither necessary nor sufficient.
  - E.g., Northpointe doesn't directly use race, but uses other factors like area code that act as potential proxies for race; the two are highly correlated in segregated neighborhoods.

Shared structure of prediction tasks

Most typical framing in alg fairness work has been as a (supervised) prediction problem, with subsequent decision as a simple function (e.g., thresholding) of predicted value.

$$\widehat{Y} = \mathbb{E}[Y \mid X]$$

- Assume decision $D = f(\widehat{Y})$, e.g., $D = \mathbb{I}(\widehat{Y} \geq \tau)$.
- Often Y is an imperfect proxy for a latent attribute ("credit-worthiness," "academic success").
- Alternative tasks include: rankings/recommendations, unsupervised learning (e.g. clustering by attributes), or optimal decision-rule learning problems.

Many researchers have focused on modifying algorithms to respect "fairness constraints."

Statistical fairness criteria

- **Disparate impact:** Decision $\perp\!\!\!\perp$ Group

$$\frac{p(\widehat{Y} = 1 \mid S \neq 1)}{p(\widehat{Y} = 1 \mid S = 1)} \geq 1 - \epsilon$$

- **Demographic/statistical parity:** Decision $\perp\!\!\!\perp$ Group

$$p(\widehat{Y} = 1 \mid S = 1) - p(\widehat{Y} = 1 \mid S \neq 1) \leq \epsilon$$

- **Equalized odds:** Decision $\perp\!\!\!\perp$ Group $\mid Y$

$$p(\widehat{Y} = 1 \mid S = 1, Y = 0) - p(\widehat{Y} = 1 \mid S \neq 1, Y = 0) \leq \epsilon,$$
$$p(\widehat{Y} = 1 \mid S = 1, Y = 1) - p(\widehat{Y} = 1 \mid S \neq 1, Y = 1) \leq \epsilon$$

- **Equal opportunity:** Decision $\perp\!\!\!\perp$ Group $\mid Y = 0$

$$p(\widehat{Y} = 1 \mid S = 1, Y = 0) - p(\widehat{Y} = 1 \mid S \neq 1, Y = 0) \leq \epsilon$$

- **Calibration:** $Y \perp\!\!\!\perp$ Group $\mid$ Decision

$$p(Y = 1 \mid S = 1, \widehat{Y}) - p(Y = 1 \mid S \neq 1, \widehat{Y}) \leq \epsilon$$

Conflicts in statistical fairness criteria

- Chouldechova (2017) shows that *so long as base rates differ across groups* (e.g., diff recividism rates), then "equalized odds" and "calibration" **cannot be both satisfied**.
- Kleinberg et al. (2016) prove a similar incompatibility result for balance in the positive class, balance in the negative class, and calibration within groups.
- Barocas et al. (2019) also derive similiar conflicts between "statistical parity" and the other two, when base rates are not equal.

- This presents a problem: if associational fairness criteria are each seemingly plausible but mutually incompatible in real problems, which should we sacrifice?

Limitations of statistical fairness criteria

Associative/statistical measures of (un)fairness:

- Tailored for classification problems
- Ignore the causal relations among variables
- Not adaptable to use context-specific information

Desirable definition of (un)fairness should:

- Use context-specific information
- Listen to causal relations
- Be nonparametric

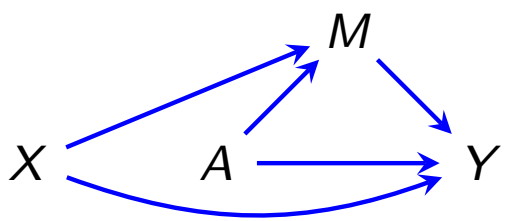
Possible causal structures underlying the data

Directed Acyclic Graphs (DAGs) are often used to encode causal relations.

- Arrows in the DAG represent (possible) causal relations among variables (e.g., that "skill level is a cause of job performance")
- Conditional independence relations among variables can be read from the DAG looking at "blocked" paths (d-separation criterion)

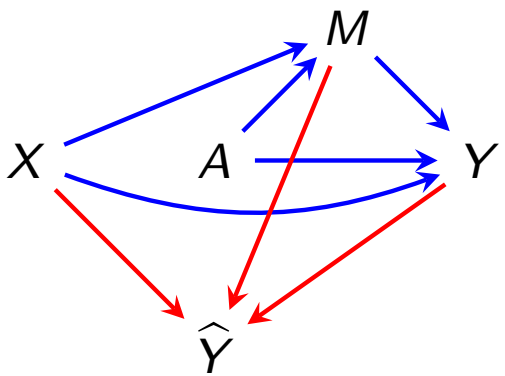
As an example, let:

- A denote relevant group membership
- Y denote outcome of interest (health, loan repayment, recidivism)
- M denote variables causally dependent on group membership
- X denote other covariates



In this case we assume A and X are independent, though can also allow that A and X are associated somehow (e.g., selection into sample)

Possible causal structures underlying the data



- Even if prediction alg does not use A, as long as there is a mechanism connecting A to any variable that determines $\widehat{Y}$, statistical parity will be violated: $\widehat{Y} \not\perp\!\!\!\perp A$
- Also guaranteed to violate equalized odds: $\widehat{Y} \not\perp\!\!\!\perp A \mid Y$ by d-separation in DAGs due to "collider" at Y
- If $A \to \cdots \to Y$, will not have calibration generally, $Y \not\perp\!\!\!\perp A \mid \widehat{Y}$

Main methodological questions

• These considerations have led to various **causality-informed perspectives** on algorithmic fairness.

• In the remaining of this short course, we would like to answer three main methodological questions:

1. How to express fairness principles mathematically?
   (using causal and counterfactual reasoning)
2. How to modify statistical procedures to reduce unfair effects?
   (constrained learning)
3. How to generalize and deploy these modified algorithms?

- RN and I. Shpitser, *Fair Inference on Outcomes,* AAAI 2018.
- RN, D. Malinsky, and I. Shpitser, *Learning Optimal Fair Policies,* ICML 2019.
- RN, D. Malinsky, and I. Shpitser, *Optimal Training of Fair Predictive Models,* CLeaR 2022.

Part II

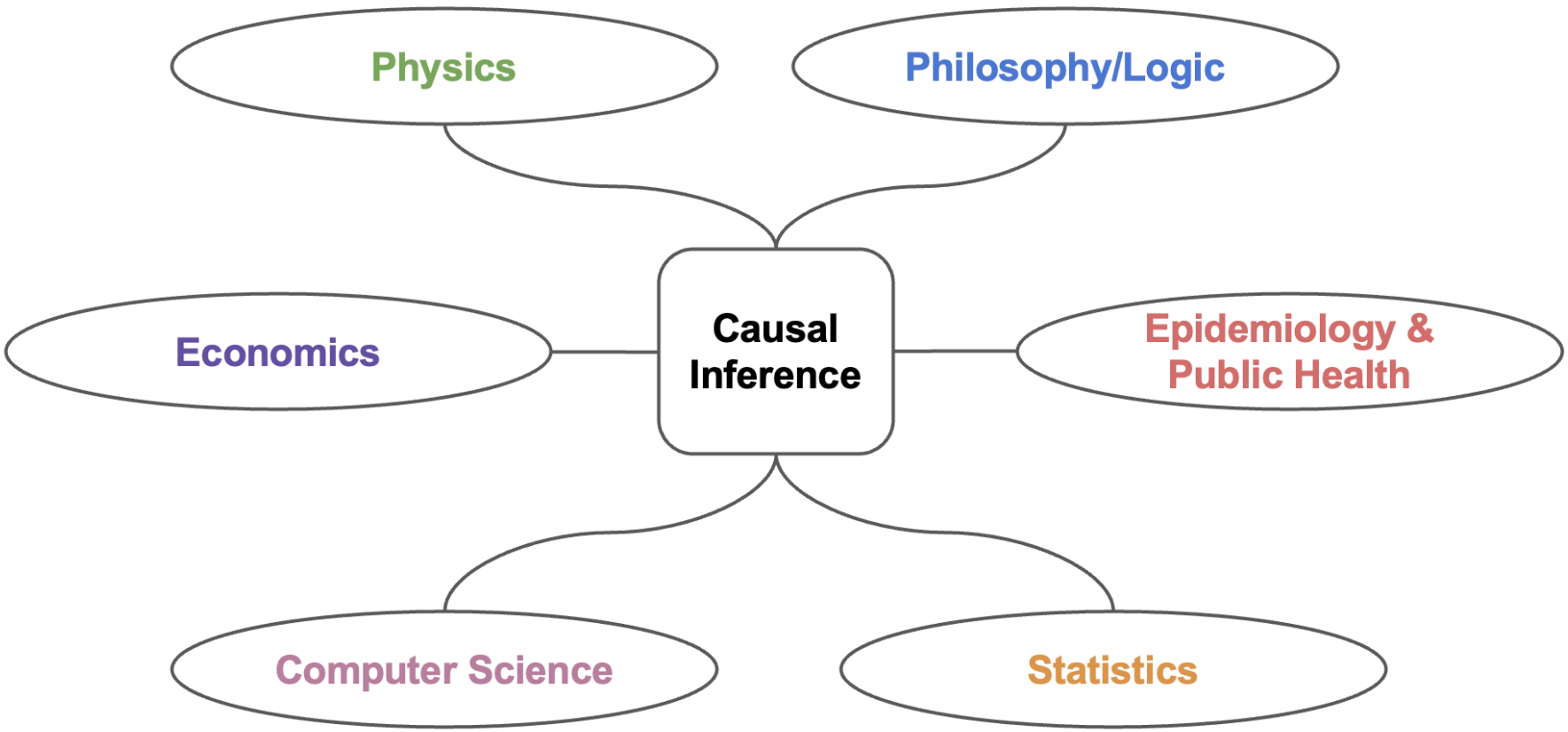
Intro to causal inf and causal fairness criteria

Overview of Part II

1. Basics of causal inference
   - Potential outcomes and graphs: the mathematical language of causation.
   - Causal parameter: average causal effect (ACE)
   - Identification and estimation (briefly).
2. Mediation analysis
   - Direct and indirect effects.
   - Identification and estimation (briefly).
3. Causal fairness criteria
   - Path-specific effects in predictive decision support (Nabi et al., 2018)
   - Path-specific effects in sequential decision making (Nabi et al., 2019)

Basics of causal inference

Causal inference: a multidisciplinary area of research



Causality in philosophy

David Hume (1711-1776)

> *We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second, . . . where, if the first object had not been the second never had existed.*

- Imperfect regularities: smoking does not always give you cancer.
- Irrelevance: yelling magic words before throwing salt into water does not cause the salt to dissolve.
- Spurious regularities: low pressure systems can create storms and they also cause the pressure reading in a barometer to drop. So we will always observe the reading in the barometer to drop before a storm, but this is clearly not causing the storm!
- Encoding uncertainty via probabilities gives us a way of addressing some (but not all) of these issues

Causal deniers

Bertrand Russell (1872-1970)

> *All philosophers [. . .] imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word "cause" never occurs. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.*

Causality in statistics (using data)

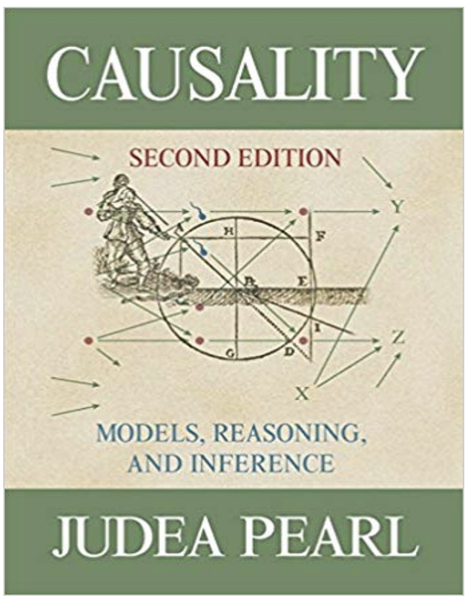
James Lind (1716-1794)

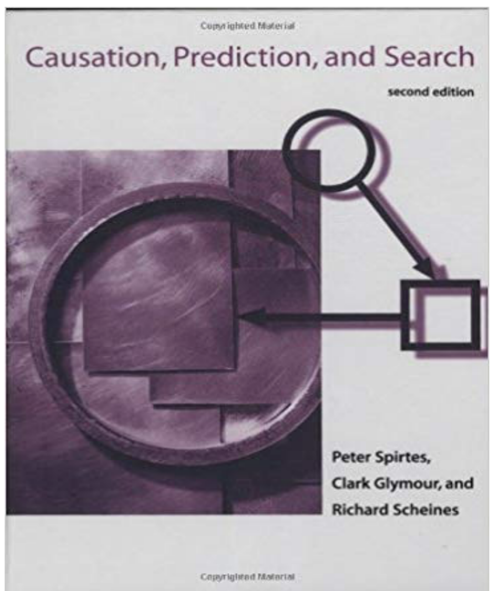
- First reported Randomized Controlled Trial

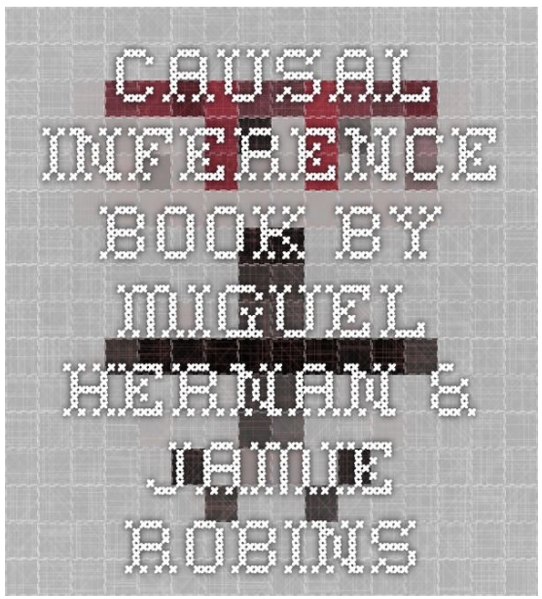
- How to treat scurvy? (James Lind,[1] 1747)
  - 12 scorbutic sailor treated with different acids, e.g. vinegar, cider, and lemon but otherwise treated them exactly the same.
  - Only the condition of the sailor treated by lemon improved.
  - Scurvy results from a lack of vitamin C.

[1] A treatise of the scurvy: `http://inspire.stat.ucla.edu/unit_04/scurvy.pdf`

Modern view of causal inference







Association vs causation

Most scientific inquiry/data analyses fall into one of the two paradigms:

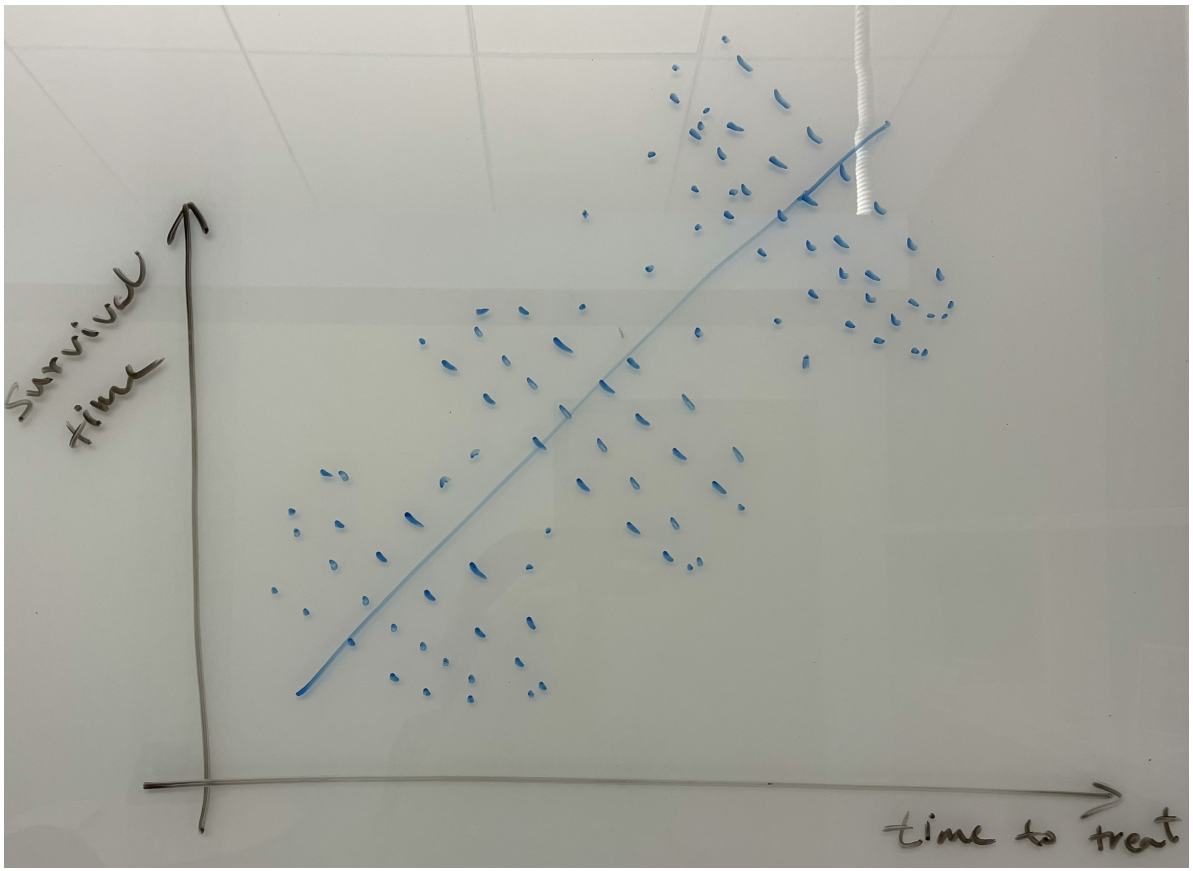
- Association (seeing, observing)
  - Associational paradigms: (Un)Supervised learning, Reinforcement learning
- Causation (doing, intervening, retrospection, understanding)
  - Causal paradigms: Effect quantification, Causal discovery, Decision making

Association vs causation ctd.

The following two statements can simultaneously be true:

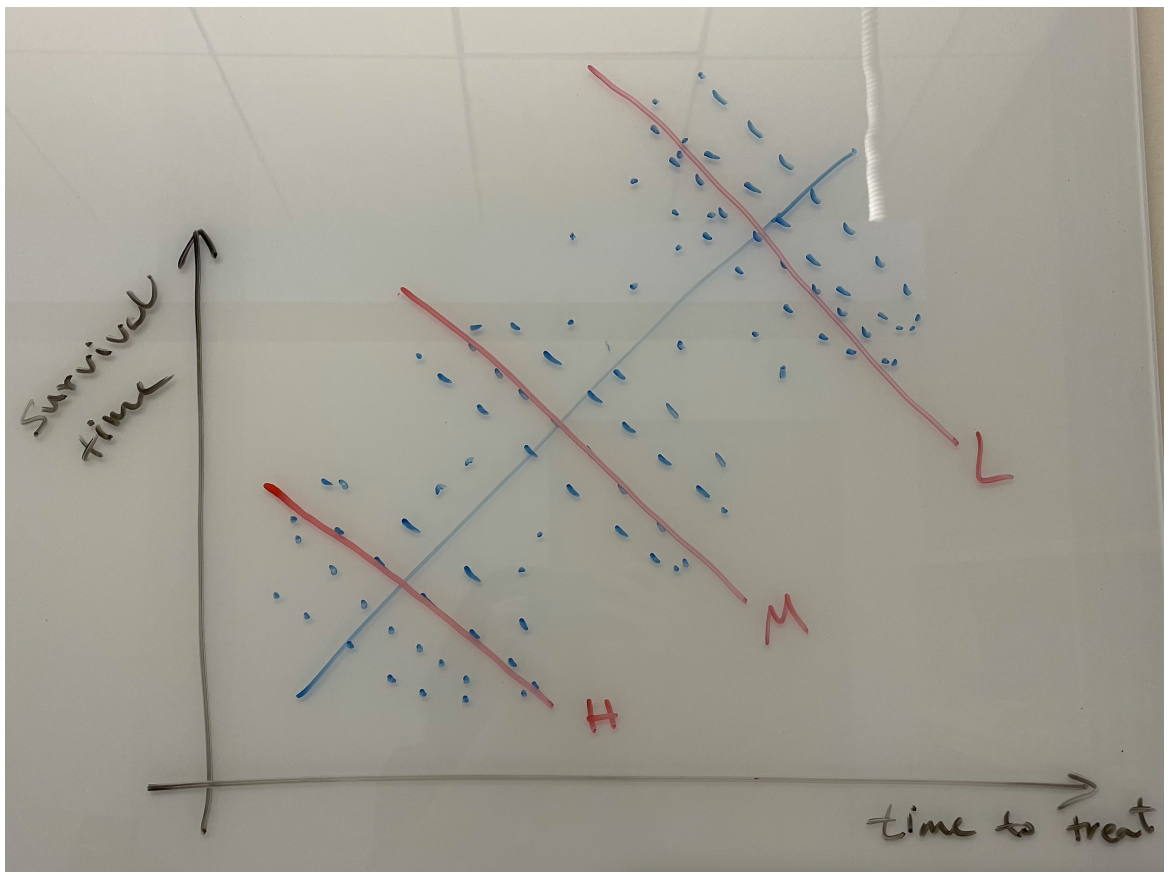
- (Predictive) Those who receive HIV treatment immediately upon diagnosis have shorter survival time, on average, than those who wait.
  - Waiting to get treated sounds more effective!
- (Causal) Given the choice to treat immediately or wait until symptoms develop, treating immediately will lead to longer survival on average.
  - Waiting to get treated sounds less effective!
- A classic description of *confounding* that teases apart prediction from causation.

Example: survival time as a function of treatment initiation



Example: survival time as a function of treatment initiation

Sub-populations with varying levels of disease severity (Low, Medium, High)



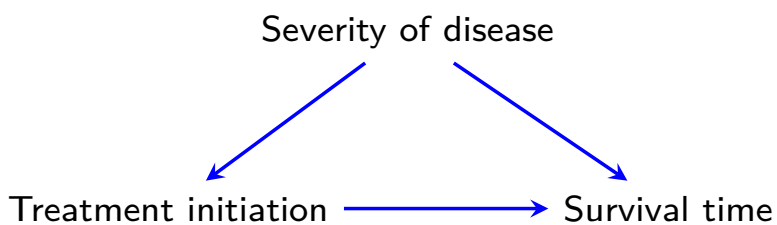
Simpson's paradox

Simpson's paradox:

- Direction of correlation changes when analyzing sub-populations vs population as a whole.

A causal explanation: treatment decisions are made based on patients' characteristics and clinical history.

- Patient's severity of disease affects treatment initiation.
- Severity of disease affects the survival time.
- Severity of disease is a **confounder**.



Takeaways:

- Machine learning and statistical methods search for patterns, and they often find spurious correlations.
- **To take actions and make decisions, we need more than mere correlations!**

Causal workflow

1. Defining causal quantities, this will be done in terms of counterfactuals.
2. Defining a causal model that links counterfactuals to factual variables.
   - Encoding assumptions necessary to identify causal quantities.
   - Identifying the causal estimands as a function of observed data in this model.
3. Defining a statistical model to deal with the curse of dimensionality.
   - Performing statistical inference which includes testing and estimating the magnitude of a causal effect given the observed data.
4. Assessing assumptions (sensitivity analysis).

Counterfactuals

- **Treatment**: An intervention or exposure; investigators wish to assess the effect of treatment compared to no treatment.
- Suppose you're contemplating taking an aspirin for your headache, and the outcome Y denotes whether or not you're headache free say in an hour.
- As a thought experiment, you may think of two potential outcome variables either of which may be observed depending on whether or not you decide to take the aspirin.
  - $Y(0)$: headache status after not taking aspirin
  - $Y(1)$: headache status after taking aspirin
- $Y(a)$ is the outcome that you would observe if, possibly countering to fact, you followed treatment $a \in \{0, 1\}$.
  - Y^a is referred to as a **potential outcome** or a **counterfactual**.
  - Different notations in the literature: Y^a, Y_a, $Y(a)$, and do-calculus notation of Pearl: $Y \mid \text{do}(A = a)$
- Observed and potential outcomes share the same domain space $Y, Y(a) \in \mathcal{Y}$

Establishing causality

- We can establish causality by comparing potential outcomes:
  - $Y(a) = Y(0)$, aspirin has no effect on my headache outcome
  - $Y(1) > Y(0)$, aspirin has a beneficial effect on my headache outcome
  - $Y(1) < Y(0)$, aspirin has a harmful effect on my headache outcome

- **Individual-level treatment effect** (ITE):

$$\text{ITE} := Y_i(a = 1) - Y_i(a = 0)$$

- **Average treatment effect** (ATE):

$$\text{ATE} := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \sum_y y \times p(y(1)) - \sum_y y \times p(y(0))$$

- "Oracle table": contains potential outcomes for every individual

ID	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	1	0	1 (protected)
2	1	1	0 (immune)
3	0	1	-1 (harmed)
4	0	0	0 (doomed)

Fundamental problem of causal inference

- Fundamental problem of causal inference is that we only observe one of the two potential outcomes.
- It means, if in the data sample, you happen to be a person with $A = 1$, then $Y(1)$ is observed and $Y(0)$ is missing, and vice versa for a person with $A = 0$.

ID	A	Y	$Y(1)$	$Y(0)$
1	1	0	0	?
2	1	1	1	?
3	0	1	?	1
4	0	0	?	0

- **It is impossible to evaluate individual-level causal effects.**
- This is fundamentally a missing data problem. The only difference is that the full data is never observed with probability one.
- All is not lost! Under some assumptions, we can still say something about population-level causal effects.

The task of identification

- Need assumptions to link the counterfactual distribution of $p(Y(a))$ to the observed data distribution. This link leads to identification arguments.
- A parameter is said to be **identified** under a particular collection of assumptions if it can be expressed as a unique function of the distribution (law) of the observed variables.
- Example: under *consistency*, *conditional ignorability*, and *positivity* assumptions, the ACE is identified. Let's see what these assumptions mean.

Identification assumptions

1. Consistency: the observed outcome is the same as the potential outcome where we set the individual's treatment to the same value they actually received. That is, $Y^A = Y$ or in other words:

$$Y = A \times Y(1) + (1 - A) \times Y(0).$$

2. Conditional ignorability: assume X is a rich vector of covariates that contains **all common causes** of A and Y (that is all risk factors for Y that also determine A),

$$Y(a) \perp\!\!\!\perp A \mid X, \text{ for all } a \in 0, 1.$$

   This means within levels of X, the data mimics a randomized trial with the randomization probabilities now allowed to depend on X.

3. Positivity:

$$\text{If } p(X = x) > 0, \quad \text{then} \quad p(A = a \mid X = x) > 0.$$

Identification of ATE

If consistency, positivity, and conditional randomization assumptions hold:

$$
\begin{aligned}
\mathbb{E}[Y(a)] &= \sum_{x,y} y \times p(y(a) \mid x) \times p(x) && \textit{(by definition)} \\
&= \sum_{x,y} y \times p(y(a) \mid x, A = a) \times p(x) && \textit{(by conditional randomization)} \\
&= \sum_{x,y} y \times p(y \mid x, A = a) \times p(x) && \textit{(by consistency)} \\
&= \sum_{x} \mathbb{E}[Y \mid x, A = a] \times p(x) && \textit{(by definition)} \\
&= \mathbb{E}\Big[\mathbb{E}[Y \mid X, A = a]\Big]
\end{aligned}
$$

This functional is known as **adjustment functional** or **g-formula**.[2]

Therefore ACE of A on Y is identified as follows:

$$\text{ATE} := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}\Big[\mathbb{E}[Y \mid X, A = 1] - \mathbb{E}[Y \mid X, A = 0]\Big].$$

[2] James Robins. "A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect", In Mathematical Modeling, 1986. [link]

Estimation of ATE

Under consistency, positivity, and conditional randomization, we have:

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \sum_x \Big\{ \mathbb{E}[Y \mid x, A = 1] \times -\mathbb{E}[Y \mid x, A = 0] \Big\} \times p(x).$$

Given a sample of n i.i.d. units $\{X_i, A_i, Y_i\}$, the statistical task is an inference on the identifying functional of the ATE.

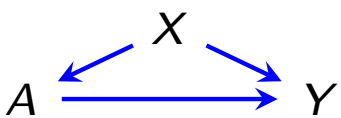
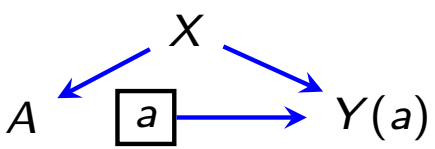
We now discuss the following methods:

- Nonparametric g-computation (nonparametric plugin)
- Nonparametric IPW
- Parametric g-computation (parametric plugin)
- Parametric IPW
- Doubly robust estimator (augmented IPW, AIPW)

Introducing causal graphs

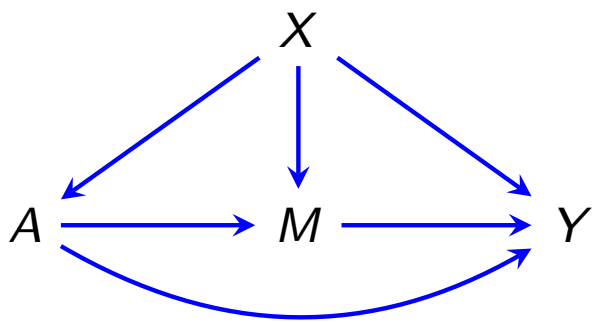
- Causal relationships in multivariate systems can be very complex.
- Will think about causal relationships using graphs, which are a helpful way to visualize complex causal models.
  - Nodes are variables, $\rightarrow$ means "directly causes."
  - Absences of nodes and edges are important.
- General method (d-separation) for reading off independences via paths:
  - A d-separated from B if all paths are "blocked."
  - A path is blocked if it has a blocking triplet.
  - Blocking triplets:
    - $V_1 \rightarrow V_2 \rightarrow V_3$ (V_2 conditioned on).
    - $V_1 \leftarrow V_2 \rightarrow V_3$ (V_2 conditioned on).
    - $V_1 \rightarrow V_2 \leftarrow V_3$ (neither V_2 nor any descendant of V_2 is conditioned on).

Introducing causal graphs ctd.

- We can read off independencies between counterfactuals and factuals by constructing Single World Intervention Graphs (SWIGs)
- Randomization example (one treatment A, one outcome Y)
  - Observed situation:  $A \not\perp\!\!\!\perp Y$
  - Hypothetical situation:  $A \perp\!\!\!\perp Y(a)$
- Conditional randomization example
  - Observed situation:  $A \not\perp\!\!\!\perp Y \mid X$
  - Hypothetical situation:  $A \perp\!\!\!\perp Y(a) \mid X$

Mediation analysis

Overview of mediation analysis



- Decompose effect of T on Y along different causal pathways

$$\text{ACE} = \underbrace{\text{Direct Effect}}_{A \to Y} + \underbrace{\text{Indirect Effect}}_{A \to M \to Y}$$

- $Y(a)$: potential (counterfactual) outcome Y had $A = a$
- $M(a)$: potential (counterfactual) mediator M had $A = a$
- $Y(a, m)$: potential (counterfactual) outcome Y had $A = a$ and $M = m$
- $Y(a, M(a'))$: potential outcome Y had $A = a$ and M behaving as if $A = a'$

- Direct Effect $= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$

Direct and indirect effects

- **Natural direct effect (NDE):** comparing outcome response to $A = 1$ and $A = 0$ while M takes on the natural value under $A = 0$ (i.e., $M(0)$)

$$\text{NDE} = \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]$$

  - *Example: what's the effect of genetic variant on lung cancer if for each individual we set the smoking to whatever value they would naturally smoke had they not had the variant?*

- **Natural indirect effect (NIE):** comparing outcome response to $M(1)$ and $M(0)$ while treatment is fixed to $A = a$

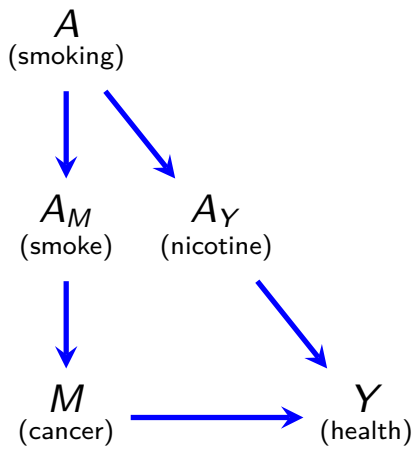
$$\text{NIE} = \mathbb{E}[Y(1, M(1)) - Y(1, M(0))]$$

  - *Example: if everyone had the variant, then how would the lung cancer rate change if level of smoking would change from the value it would have naturally arise under presence of variant vs absence of variant?*

- Effect decomposition:

$$\begin{aligned}\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] &= \mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(0, M(0))] \\ &= \mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(1, M(0))] + \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0, M(0))] \\ &= \text{NIE} + \text{NDE}\end{aligned}$$

Example of a nested counterfactual $Y(a, M(a'))$: smoking



	A_Y **nicotine**	A_M **smoke**	Intervention
$Y(1, M(0))$	1	0	nicotine patch
$Y(0, M(1))$	0	1	nicotine-free cigarettes
$Y(1, M(1)) = Y(1)$	1	1	smokers
$Y(0, M(0)) = Y(0)$	0	0	non-smokers

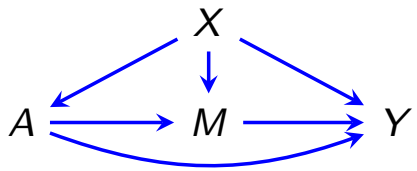
- **Direct effect**: nicotine patch (compared to no smoking)

$$\text{NDE} = \mathbb{E}[Y(1, M(0)) - Y(0)]$$

- **Indirect effect**: nicotine-free cigarettes (compared to smoking)

$$\text{NIE} = \mathbb{E}[Y(1) - Y(1, M(0))]$$

Identification assumptions for $\mathbb{E}\left[Y\left(a, M(a')\right)\right]$



1. **Conditional ignorability:**

   1.1 $Y(a) \perp\!\!\!\perp A \mid X$ (same as in CDE)

   i.e., conditioning on X suffices to deal w confounding btw A and Y.

   1.2 $Y(m) \perp\!\!\!\perp M \mid A, X$ (same as in CDE)

   i.e., conditioning on A, X suffices to deal w confounding btw M and Y.

   1.3 $M(a) \perp\!\!\!\perp A \mid X$

   i.e., conditioning on X suffices to deal w confounding btw A and M.

2. **Cross-world assumption:**

   2.1 $Y(a, m) \perp\!\!\!\perp M(a') \mid X$

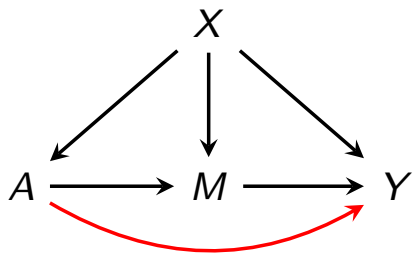
   i.e., within levels of X, causal mechanisms for M and Y have independent sources of noise, even if treatments "mismatch."

3. **Positivity:** As stated before. (same as in CDE)

4. **Consistency:** As stated before. (same as in CDE)

Identification and estimation of the *direct effect*

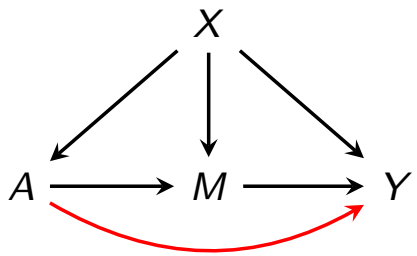
$$\text{Direct effect} = \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$$



$$\mathbb{E}[Y(1, M(0))] = \mathbb{E}\Big[\sum_m \mathbb{E}[Y \mid x, m, A = 1] \times p(M = m \mid x, A = 0)\Big] = g(P_Z)$$

Identification and estimation of the *direct effect*

Direct effect $= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$



$$\mathbb{E}[Y(1, M(0))] = \mathbb{E}\Big[\sum_m \mathbb{E}[Y \mid x, m, A = 1] \times p(M = m \mid x, A = 0)\Big] = g(P_Z)$$

Plugin estimator: $\mathbb{P}_n\Big(\sum_m \widehat{\mathbb{E}}[Y \mid x_i, m, A = 1] \times \widehat{p}(M = m \mid x_i, A = 0)\Big)$

Inverse probability weighting: $p(A \mid X),\ p(M \mid X, A)$

Mixed estimator: $p(A \mid X),\ \mathbb{E}[Y \mid X, A, M]$

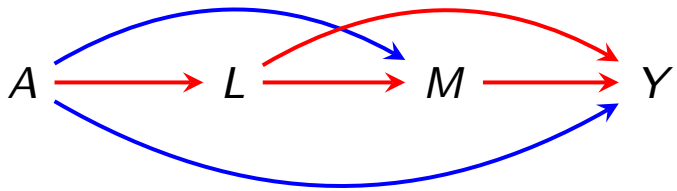
Augmented IPW: $p(A \mid X),\ p(M \mid X, A),\ \mathbb{E}[Y \mid X, A, M]$ (triply robust)

(Tchetgen Tchetgen and Shpitser, 2012)

Path-specific effects: multiple mediators

- Direct effect: effect along the direct arrow.
- Indirect effect: effect along all other arrows.
- Maybe we want effect along a specific path or a bundle of paths.
- Effect along a specified set of paths is called a **path-specific effect**, where
  - The treatment assignment is fixed along the paths that we are not interested in, and
  - It changes along the specified paths.

An example of a path-specific effect



What is the effect of A on Y through L?

That is the effect along the following paths:

- $A \rightarrow L \rightarrow Y$
- $A \rightarrow L \rightarrow M \rightarrow Y$

It's a counterfactual contrast of the following form:

$$\mathbb{E}\left[Y\left(a, L\left(a'\right), M\left(a, L(a')\right)\right)\right] - E[Y(a)]$$

Causal fairness criteria

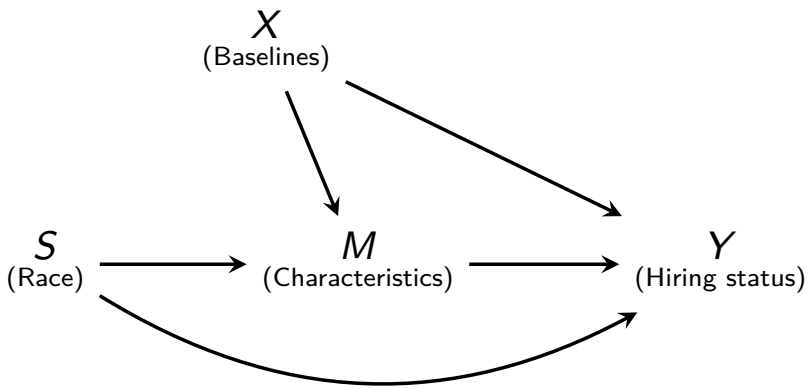
Legal opinions on anti-discrimination

7th circuit court case (Carson vs Bethlehem Steel Corp, 1996):

"The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (sex, national origin, etc.) and everything else had remained the same."

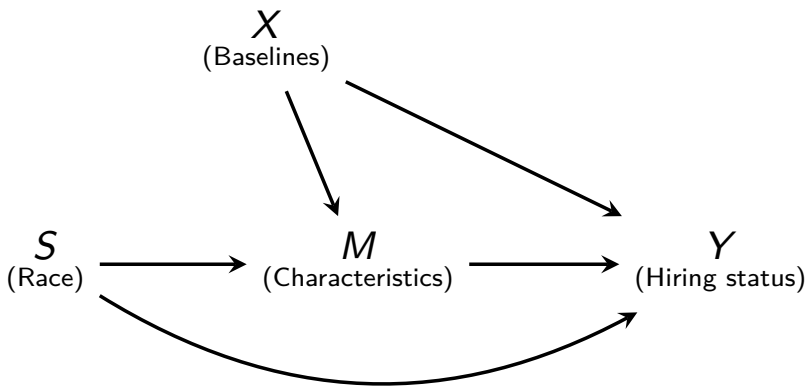
Mathematical expression of a legal quote

Would the employer have taken the same action had the employee been of a different race and everything else had remained the same?



Mathematical expression of a legal quote

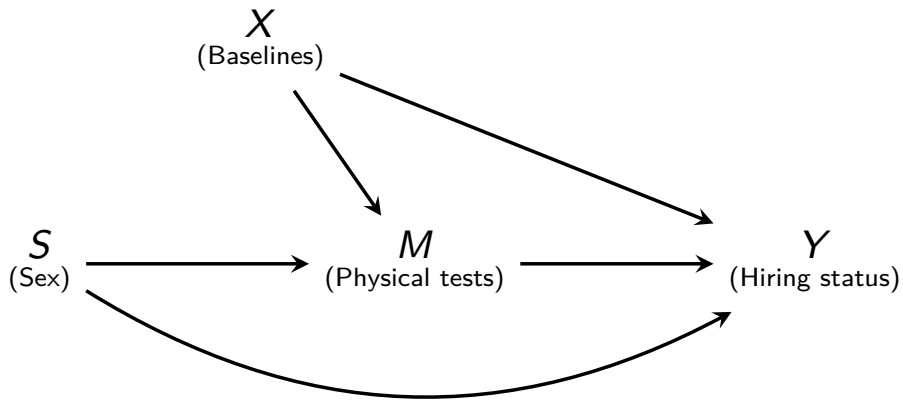
Would the employer have taken the same action had the employee been of a different race and everything else had remained the same?



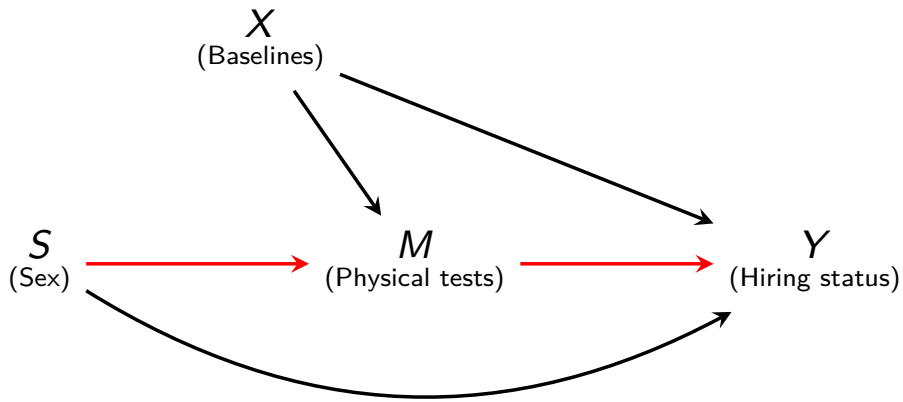
Name-swapping experiments to evaluate racism in hiring:

- African American: $S = 1$, Caucasian: $S = 0$,
- $Y(1, M(0))$: hiring a Caucasian with an African American sounding name
- $Y(0)$: hiring a Caucasian
- Direct effect: $\mathbb{E}\Big[Y\Big(1, M(0)\Big)\Big] - \mathbb{E}\Big[Y(0)\Big]$

Is unfairness always about the direct effect of S on Y?

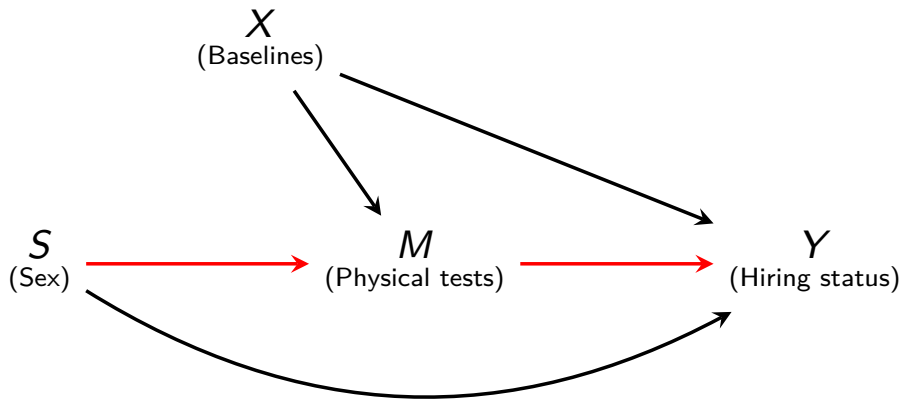


Is unfairness always about the direct effect of S on Y?



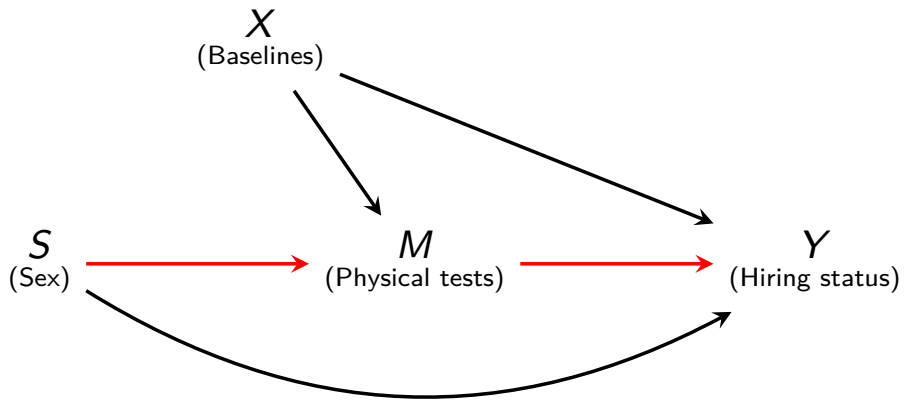
- Y : hiring a fire fighter
  - $S \rightarrow M \rightarrow Y$ ✓

Is unfairness always about the direct effect of S on Y?



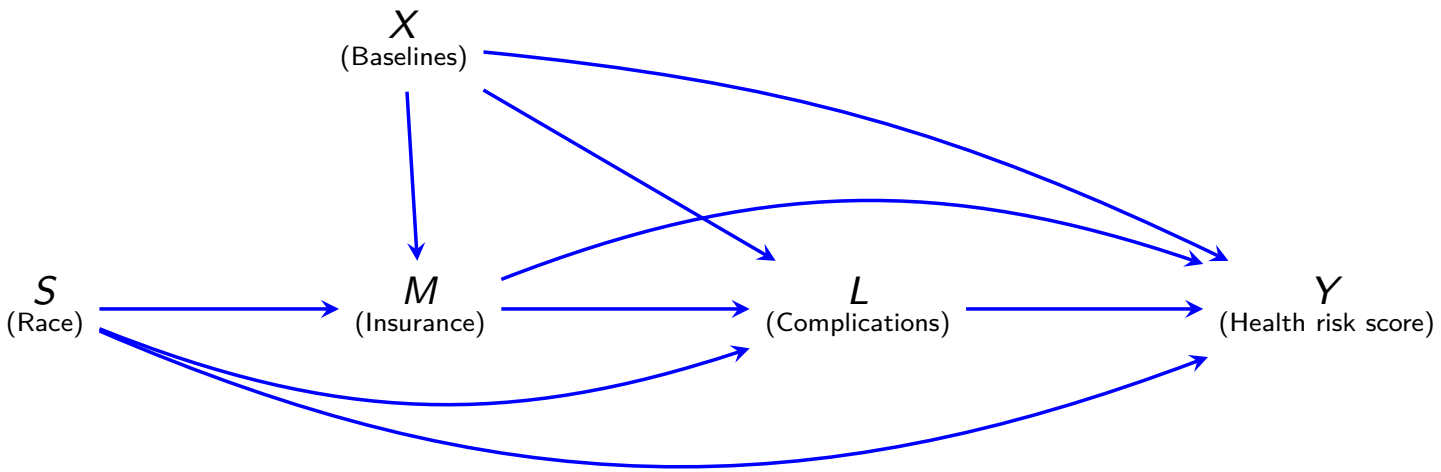
- Y : hiring a fire fighter
  - $S \rightarrow M \rightarrow Y$ ✓
- Y : hiring an accountant
  - $S \rightarrow M \rightarrow Y$ ✗

Is unfairness always about the direct effect of S on Y?

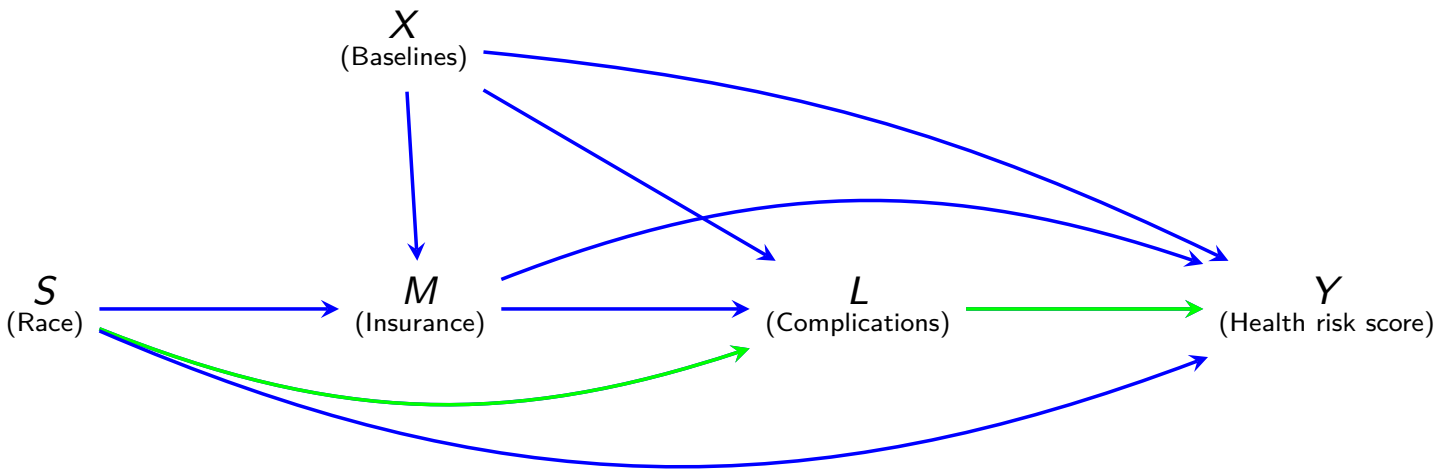


- Y : hiring a fire fighter
  - $S \rightarrow M \rightarrow Y$ ✓
- Y : hiring an accountant
  - $S \rightarrow M \rightarrow Y$ ✗
- So the answer is NO! Definition should be context-specifific.

From mediation to arbitrary path-specific effects

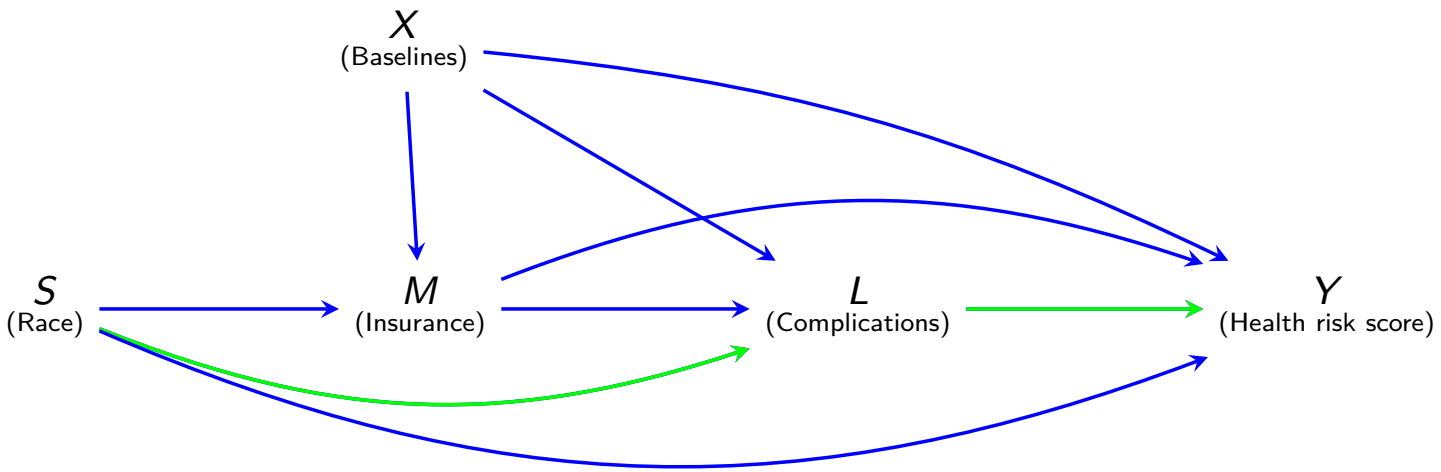


From mediation to arbitrary path-specific effects



- $S \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow L \rightarrow Y$ ✗
- $S \rightarrow L \rightarrow Y$ ✓

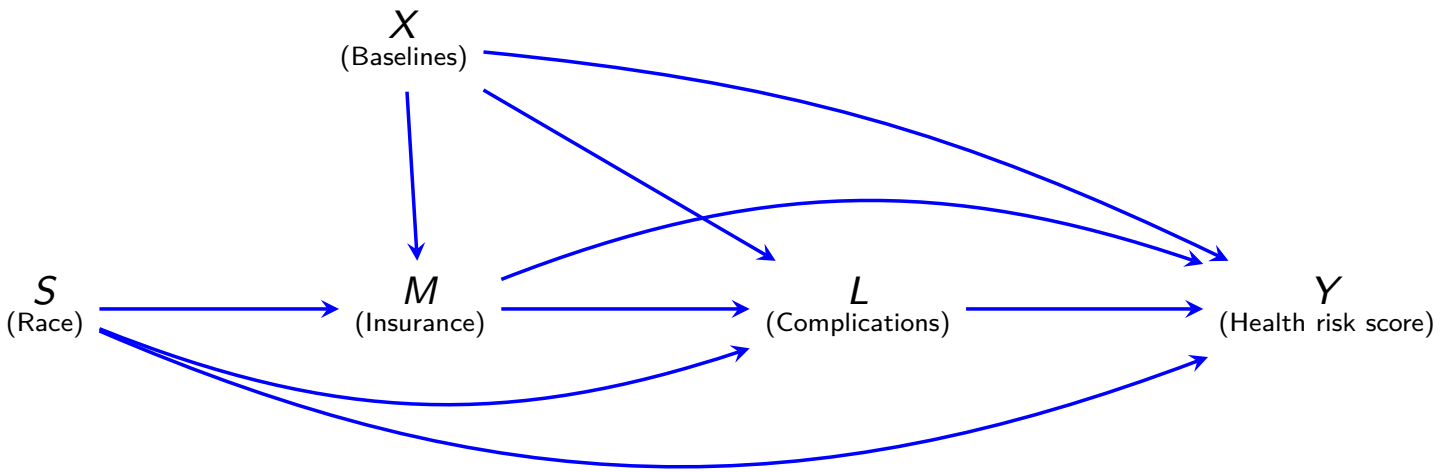
From mediation to arbitrary path-specific effects



- $S \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow L \rightarrow Y$ ✗
- $S \rightarrow L \rightarrow Y$ ✓

$$\mathbb{E}\left[Y\left(s, M(s), L\left(s', M(s)\right)\right)\right] = g(P_Z)$$

From mediation to arbitrary path-specific effects



- Path-specific effect (PSE):
  - Along pathways of interest, all nodes behave as if $S = s$,
  - Along all other pathways, nodes behave as if $S = s'$.

- Identification and estimation of PSEs:
  (Works by Shpitser, Tchetgen Tchetgen, VanderWeele, Avin, Pearl, Robins, Richardson, Malinsky, Miles, Diaz, and more)

Our definition of fairness

- $\text{ACE} = \text{PSE}^{\text{unfair}} + \text{PSE}^{\text{fair}}$
- $\text{PSE}^{\text{unfair}}$: effect of S on Y along unfair causal pathways

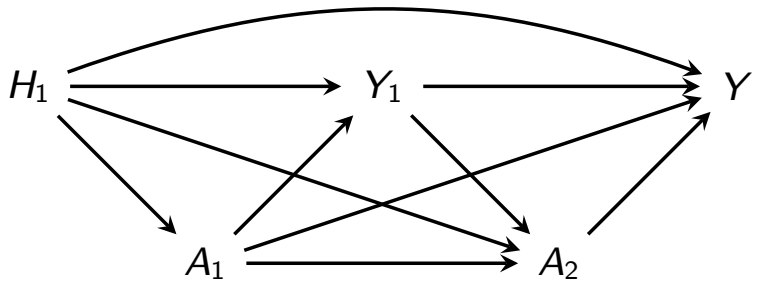
  (RN and Shpitser, Fair Inference on outcomes, AAAI, 2018.)
- Determining unfair pathways is a domain specific issue
  - This is a feature not a bug.

Context: sequential decision making

Decision rule: $f_{A_i} : \mathcal{H}_i \mapsto \mathcal{A}_i$

Policy: $f_A = \{f_{A_1}, f_{A_2}\}$

(dynamic treatment regimes)

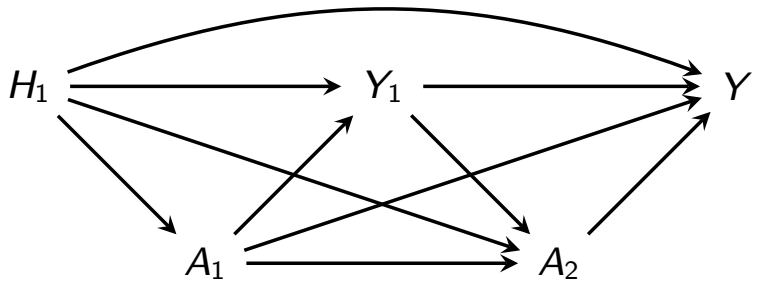


Context: sequential decision making

Decision rule: $f_{A_i} : \mathcal{H}_i \mapsto \mathcal{A}_i$

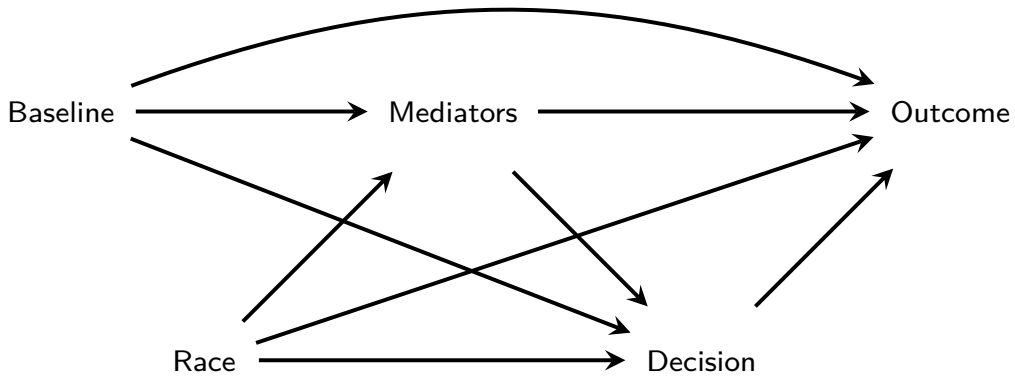
Policy: $f_A = \{f_{A_1}, f_{A_2}\}$

(dynamic treatment regimes)



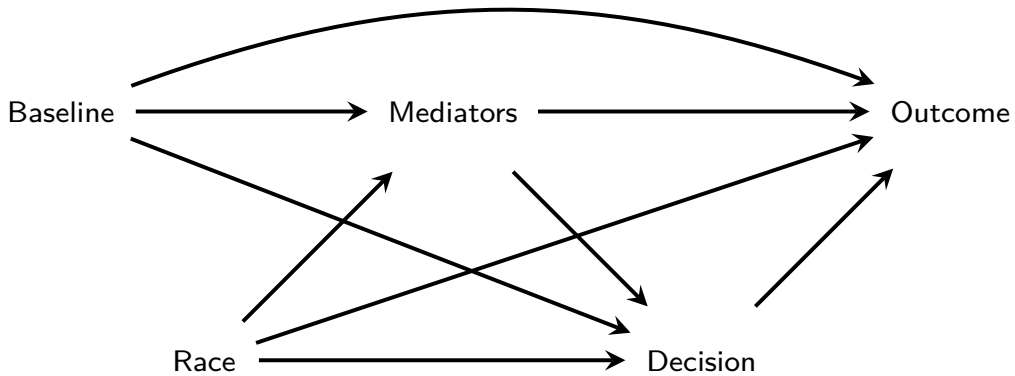
- Counterfactual response under f_A is denoted by $Y(f_A)$
- Optimal policy: $f_A^* := \arg\max_{f_A} \mathbb{E}[Y(f_A)]$
- Fairness concerns arise since $H_1 = \{X, S, M\}$

Example: child welfare



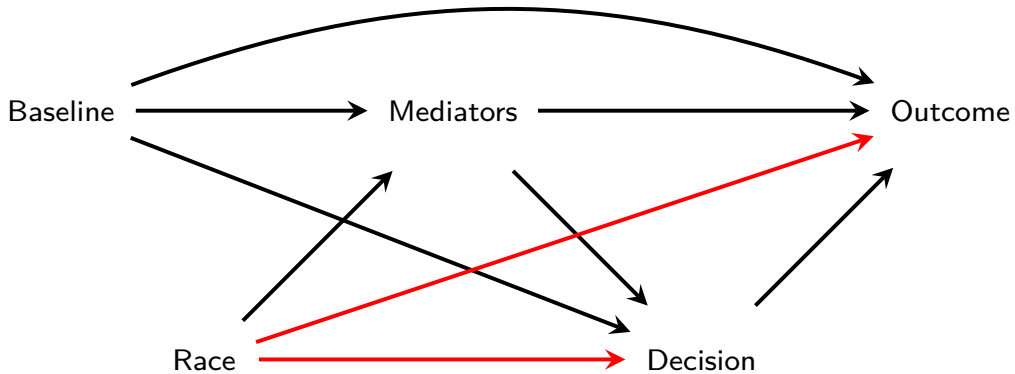
- Decision to dispatch case-worker may depend on all available information, and optimal decision would minimize negative outcomes (e.g. child separation and/or hospitalization).
- Unconstrained optimal decision-making may lead to unacceptable racial disparities.

Example: child welfare



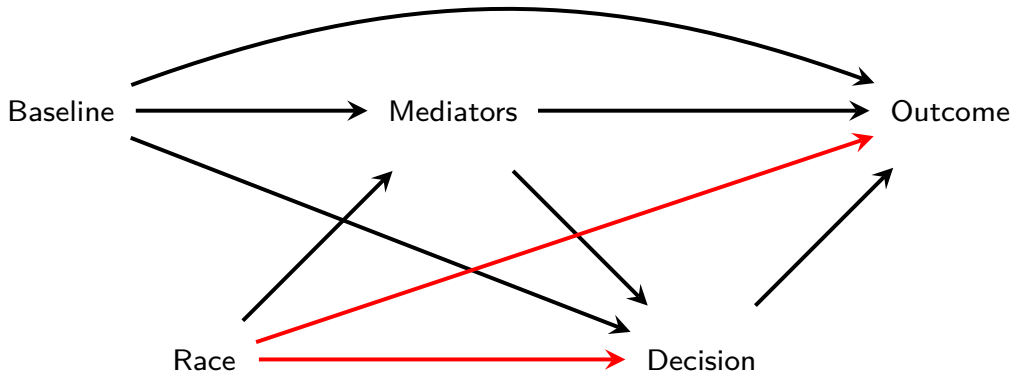
- Decision to dispatch case-worker may depend on all available information, and optimal decision would minimize negative outcomes (e.g. child separation and/or hospitalization).
- Unconstrained optimal decision-making may lead to unacceptable racial disparities.
- Ignoring race information is insufficient: dependence due to proxies
- Is it sufficient to define fairness in automated descision making the same way as we did in doing fair predictions?

A causal perspective



- In a "fairer world," certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects (RN and Shpitser, 2018).

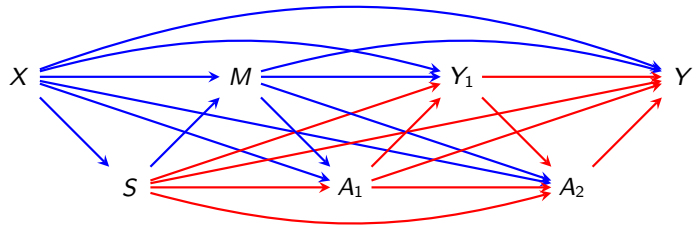
A causal perspective



- In a "fairer world," certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects (RN and Shpitser, 2018).
- Approximate the "nearest fair world" and learn optimal policies there (RN, Malinsky, Shpitser, 2019)
- Must sacrifice some optimality to make decisions fairly.

Fairness in automated decision making

Fairness in automated decision making



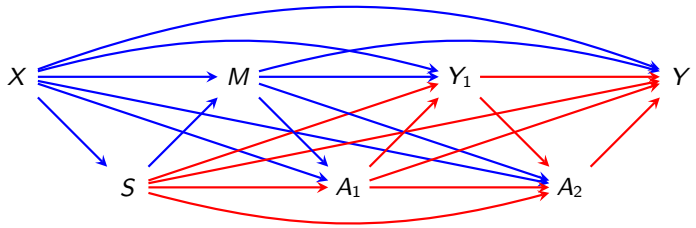
- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

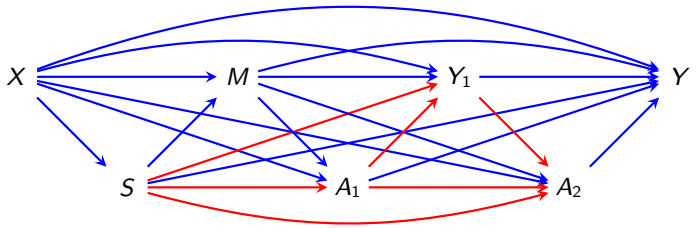
Example: unfair paths from S to Y:
$\{S \to Y,\ S \to Y_1 \to \cdots \to Y,$
$S \to A_1 \to \ldots \to Y,$
$S \to A_2 \to \ldots \to Y\}.$

$$\text{PSE}^{sy} = g_1(P_Z)$$

$$Z = \{X, S, M, A_1, ..., A_K, Y_1, ..., Y_K\}$$

Fairness in automated decision making





- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

  Example: unfair paths from S to Y:
  $\{S \to Y,\ S \to Y_1 \to \cdots \to Y,$
  $S \to A_1 \to \ldots \to Y,$
  $S \to A_2 \to \ldots \to Y\}.$

- **Prospective bias**:
  functional form of policy depends on sensitive features.

  Example: unfair paths from S to A_1, A_2:
  $\{S \to A_1\},$
  and
  $\{S \to A_2, \quad S \to A_1 \to \ldots \to A_2\}$

$$\text{PSE}^{sy} = g_1(P_Z)$$

$$\text{PSE}^{sa_k} = g_k(P_Z)$$

$$Z = \{X, S, M, A_1, ..., A_K, Y_1, ..., Y_K\}$$

Part III

Constrained learning, example application

Two types of questions

- Let $O = (X, S, Y) \sim P \in \mathcal{M}$
- Assume $\Psi : \ P \in \mathcal{M} \mapsto \Psi(P) \in \bar{}$, e.g.,
  - Supervised learning:
    $$\Psi(P) = \mathbb{E}_P[Y \mid X, S] \text{ or } \Psi(P) = P(Y = 1 \mid X, S)$$
  - Dynamic treatment regime:
    $$\Psi(P) = \arg\max_{f_a \in \mathcal{F}} \ \mathbb{E}[Y(f_a)], \text{ where } f_a : \mathcal{H} \mapsto a \in \mathcal{A}$$
- Given $\Psi(P)$:
  1. Does $\Psi(P)$ encode unfair biases?
     - Various notions of fairness were discussed earlier

       $$\Theta_\Psi : P \in \mathcal{M} \mapsto \Theta_\Psi(P) \in \hat{}$$
  2. **How to mitigate unfair biases $\Theta_\Psi(P)$ from $\Psi(P)$?**

Overview of Part III

How to mitigate unfair biases $\Theta_\Psi(P)$ from $\Psi(P)$?

1. Predictive decision support
   (make decisions only based on model outputs)

2. Sequential decision making
   (make decisions based on optimizing a utility function)

3. Data application
   (using COMPAS data)

Predictive decision support

Let $\Psi(P) = P(Y = 1 \mid X, S)$

a. Pre-process the observed data O

b. Post-process the statistical output $\Psi(P)$

c. Re-train $\Psi(P)$ subject to fairness constraints

Pre-process data O

- Let $\Psi(P) = P(Y = 1 \mid X, S)$, and
- Let $\widehat{Y} = 1$ if $\Psi(P) > \delta$, and 0 otherwise (dependent on Ψ, δ)

- **Constraint** $\Theta_\Psi(P)$: $\widehat{Y} \perp\!\!\!\perp S$
  - "Independence", demographic/statistical parity

- **Approach**: representation learning (Zemel et al.; 2013)

X, S ⇝ representation Z ⇝ $\Psi(P) =$ only a function of Z

Find data representation Z by:

- Maximizing mutual information between $\{X, Z\}$, and
- Minimizing mutual information between $\{S, Z\}$.

- Train $\Psi^*(P) = P(Y = 1 \mid Z)$

Post-process $\Psi(P)$

- Let $O = \{C, S, M, Y\}$ and $\Psi(P) = \mathbb{E}_P[Y \mid C, S, M]$

- **Constraint**: unfair path-specific effects (N, S; 2018)

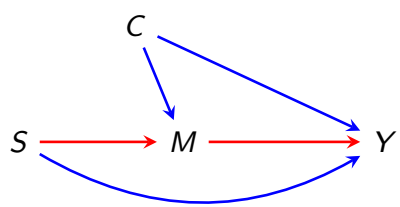
  Example: assume indirect effect of S on Y is unfair:

$$\Theta_\Psi(P) := \mathbb{E}[Y(0, M(1))] - \mathbb{E}[Y(0)]$$

  Approach: Chiappa (2019) suggests to learn $\Psi^*(P)$ by:

  - "Correcting" all descendants of S along unfair pathways

$$m_i^{\text{mod}} = \theta^m + \theta_c^m c_i + \cancel{\theta_s^m} s_i + \epsilon_i^m$$
$$y_i^{\text{mod}} = \theta^y + \theta_c^y c_i + \theta_s^y s_i + \theta_m^y m_i^{\text{mod}} + \epsilon_i^y$$



$$\Theta_\Psi(P) = \theta_s^m \times \theta_m^y$$

Constrained learning

- Impose the pre-specified constraint at the training time
  - Solving a constrained optimization problem
- In the remaining, we discuss approaches outlined in:
  1. N & S. “Fair inference on outcomes." AAAI 2018.
  2. N, M, & S. “Learning optimal fair policies." ICML 2019.
  3. N, M, & S. “Optimal training of fair predictive models." CLeaR 2022.

Defining a fair world/distribution

- **Idea**: move the statistical task from $P(O)$ to $P^*(O)$
- $P(O)$: observed (unfair) data distribution
- $\Theta(P)$: pre-specified notion of fairness
- $P^*(O)$: fair distribution
  - The closest distribution to $P(O)$ where $\Theta(P)$ is satisfied

Definition: fair world $P^*(O)$

Let ϵ_l, ϵ_u denote lower/upper tolerance bounds on $\Theta(P)$.

$$P^*(O) \equiv \arg\min_{Q} \; D_{KL}(P \,||\, Q),$$
$$\text{s.t.} \quad \epsilon_l \leq \Theta(P) \leq \epsilon_u.$$

Approximating the fair world

- Assume we observe n i.i.d. copies of $O \sim P_\alpha \in \mathcal{M}^{\text{par}}$
  - $\mathcal{M}^{\text{par}}$ denotes a parametric model
  - α is a finite set of parameters that index a distribution
- Let α^* be a finite set that indexes the fair distribution P^*
- Estimate α^* via solving:

$$\widehat{\alpha^*} = \arg\max_{\alpha} \mathcal{L}_n(O;\alpha)$$
$$\text{subject to } \epsilon_l \leq \widehat{\Theta}_n(P_\alpha) \leq \epsilon_u,$$

  where $\mathcal{L}_n(O;\alpha)$ denotes the likelihood of observed data, and $\widehat{\Theta}_n(P_\alpha)$ is an estimator for $\Theta(P_\alpha)$
- $\Psi(P_{\alpha^*})$ is the "fair version" of $\Psi(P_\alpha)$
- There are three main discussion points

#1 Multiple fair worlds

- We might have multiple candidates for $\widehat{\Theta}_n(P_\alpha)$
- Assume $O = (X, S, M, Y)$ and $\Theta(P)$ is the direct effect of S on Y

$$\Theta(P) = \mathbb{E}_{P_x}\left[\sum_{m,y} y \times \{P(y \mid S = 1, X, m) - P(y \mid S = 0, X, m)\} \times P(m \mid S = 1, X)\right]$$

- Candidate estimators of $\Theta(P)$ use different parts of $P(O)$:
  - Plugin estimator: $P(M, Y \mid X, S)$

    $$P_1^*(O) = P(X) \times P(S \mid X) \times P^*(M \mid X, S) \times P^*(Y \mid X, S, M)$$

  - Efficient influence function: $P(S, M, Y \mid X)$

    $$P_2^*(O) = P(X) \times P^*(S \mid X) \times P^*(M \mid X, S) \times P(Y \mid X, S, M)$$

- How are P_1^* and P_2^* compared to P?

Comparing fair worlds

Theorem (Nabi et al., 2020)

Let $Z_1, Z_2 \subseteq O$. Let P_1^* constrain $P_{Z \setminus Z_1}$ and P_2^* constrain $P_{Z \setminus Z_2}$.

$$P_1^*(O) = \arg\min_Q \; D_{KL}(P \;||\; Q) \quad \text{s.t.} \;\; \epsilon_l \leq \Theta(P) \leq \epsilon_u \;\; \text{and} \;\; Q(Z_1) = P(Z_1),$$

$$P_2^*(Z) = \arg\min_Q \; D_{KL}(P \;||\; Q), \quad \text{s.t.} \;\; \epsilon_l \leq \Theta(P) \leq \epsilon_u \;\; \text{and} \;\; Q(Z_2) = P(Z_2).$$

If $Z_2 \subseteq Z_1 \subseteq Z$, then $D_{KL}(P \;||\; P_2^*) \leq D_{KL}(P \;||\; P_1^*)$.

"Optimal" fair world

- Can constrain the entire $P(O)$ to be fair.
- Empirical likelihood methods particularly appealing for constraining $P(X)$ (Owens; 2001)
- Let's assume $\Theta(P) = \mathbb{E}[m(X;\alpha)]$ is a path-specific effect.
- Maximize the hybrid likelihood:

$$\arg\max_{P_i,\alpha} \prod_{i=1}^{n} \overbrace{P_i}^{non-parametric} \overbrace{P(Y|m_i, s_i, x_i; \alpha_y)\ P(M|s_i, c_i; \alpha_m)\ P(S|x_i; \alpha_a)}^{parametric}$$

$$\text{such that} \quad \sum_{i=1}^{n} P_i = 1, \quad \sum_{i=1}^{n} P_i\ m(X_i;\alpha) = 0,$$

where $O = \{X, S, M, Y\}$.

- Can solve this via Lagrange multiplier methods. (Empirical likelihood literature)

#2 Fairness constraints and distributional shifts

- **Idea:** move the statistical problem from P to P^*
- **Issue:** BUT samples are drawn from P and not P^*
- **Suggestion:** use unconstrained knowledge btw P and P^*

- Example: given i^{th} individual $O_i = (X_i, S_i, M_i, Y_i = ?)$
  - Fair world:

$$P^*(O) = \underbrace{P(X, S)}_{\text{unconstrained}} \times P^*(M \mid X, S) \times P^*(Y \mid X, S, M)$$

  - Fair prediction:

$$\mathbb{E}^*[Y_i \mid X_i, S_i] = \sum_m \mathbb{E}^*[Y_i \mid X_i, S_i, m] \times P^*(m \mid X_i, S_i)$$

#3 Complex optimization problem

- Typical learning problem of the form:

$$\begin{aligned} \alpha^* &= \arg\max_{\alpha} \quad \mathcal{L}_n(O;\alpha) \\ &\text{subject to} \quad \widehat{\Theta}_n(P_\alpha) = 0. \end{aligned}$$

- This is very hard in general, mainly because $\Theta(P)$ is often a complex functional of observed data.

- Alternative, use structural nested model ideas (Robins; 99).
- Reparameterize the likelihood.

Likelihood re-parameterization

$$\mathbb{E}[Y \mid X, S, M] = \underbrace{\mathbb{E}[Y \mid X, S, M] - \mathbb{E}[Y \mid S, X = 0, M = 0]}_{f(X,S,M)}$$
$$- \sum_{X,M} f(X, S, M) \times P(M \mid S = 0, X) \times p(X)$$
$$+ \underbrace{\sum_{X,M} \mathbb{E}[Y \mid X, S, M] \times P(M \mid S = 0, X) \times P(X)}_{\phi(S) \;=\; w_0 + w_s \times S}.$$

The coefficient w_s corresponds to the direct effect, since

$$\begin{aligned}
\text{NDE} &= \sum_{X,M} \Big\{ \mathbb{E}[Y \mid X, S = 1, M] - \mathbb{E}[Y \mid X, S = 0, M] \Big\} \; P(M \mid S = 0, X) \; P(X) \\
&= \phi(S = 1) - \phi(S = 0) \\
&= w_s
\end{aligned}$$

Likelihood re-parameterization ctd.

Theorem *Assume the observed data distribution $p(Y, Z)$ is induced by a causal model where $Z = \{X, A, M\}$ includes pre-treatment measures X, binary treatment A, and post-treatment pre-outcome mediators M. Let $p(Y(\pi, a, a'))$ denote the potential outcome distribution that corresponds to the effect of A on Y along proper causal paths in π, where π includes the direct edge $A \to Y$, and let $p(Y_0(\pi, a, a'))$ denote the identifying functional for $p(Y(\pi, a, a'))$ obtained from the edge g-formula, where the term $p(Y|Z)$ is evaluated at $\{Z \setminus A\} = 0$. Then $\mathbb{E}[Y|Z]$ can be written as:*

$$\mathbb{E}[Y|Z] = f(Z) - \big(\mathbb{E}[Y(\pi, a, a')] - \mathbb{E}[Y_0(\pi, a, a')]\big) + \phi(A),$$

where $f(Z) := \mathbb{E}[Y|Z] - \mathbb{E}[Y|A, \{Z \setminus A\} = 0]$ and $\phi(A) = w_0 + w_a A$. Furthermore, w_a corresponds to π-specific effect of A on Y.

(N, M, S. “Optimal training of fair predictive models." CLeaR 2022)

Constrained learning: a Bayesian approach

- Described methods so far are fundamentally frequentist
- Bayesian methods can be adapted
  - Sample the posterior using Markov chain Monte Carlo approaches
  - Use the sample to compute any function of the posterior distribution
- E.g., BART, a popular Bayesian random forest method (Chipman et al.; 2010)
  - Construct a distribution over a forest of regression trees (with a prior that favors small trees)
  - Sample the posterior using Gibbs sampling
  - Reject all draws that violate the constraint
  - Gibbs sampler will generate samples from a constrained posterior directly (Gelfand et al., 1992)
- Finding novel ways to solve the constrained optimization is an open area of research

Sequential decision making

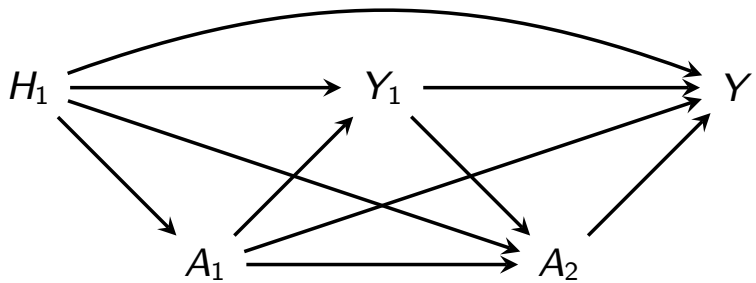
More complex statistical targets

Example: **Sequential decision making**

Decision rule: $f_{A_i} : \mathcal{H}_i \mapsto \mathcal{A}_i$

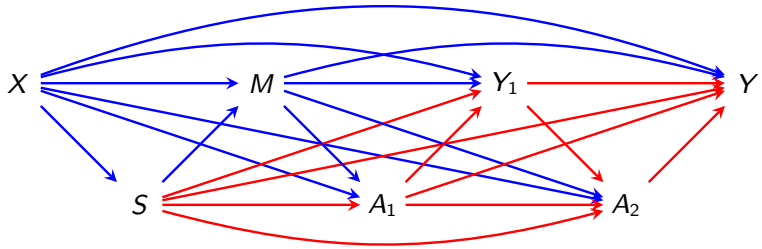
Policy: $f_A = \{f_{A_1}, f_{A_2}\}$

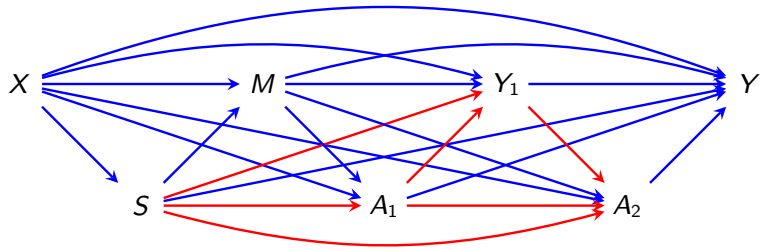
(dynamic treatment regimes)



- Counterfactual response under f_A is denoted by $Y(f_A)$
- Optimal policy: $f_A^* := \arg\max_{f_A} \mathbb{E}[Y(f_A)]$
- Fairness concerns arise since $H_1 = \{X, S, M\}$

Sources of bias in policy learning





- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

Example: unfair paths from S to Y:

$$\{S \to Y,\ S \to Y_1 \to \cdots \to Y,\\ S \to A_1 \to \ldots \to Y,\\ S \to A_2 \to \ldots \to Y\}.$$

$$\text{PSE}^{sy} = \Theta_1(P)$$

- **Prospective bias**:
  functional form of policy depends on sensitive features.

Example: unfair paths from S to A_1, A_2:

$$\{S \to A_1\},\\ \{S \to A_2,\quad S \to A_1 \to \ldots \to A_2\}$$

$$\text{PSE}^{sa_k} = \Theta_k(P)$$

Defining a fair world/distribution

- $P(O)$: observed (unfair) distribution
- A set of identified unfair PSEs denoted by $\Theta_j(P)\ \forall j \in \{1, ..., J\}$
- $P^*(O)$: fair distribution
  - Close to $P(O)$ via Kullback-Leibler divergence
  - A distribution where unfair effects are null
- Give *lower/upper tolerance bounds* $\epsilon_j^-, \epsilon_j^+$, $P^*(O)$ is defined as:

$$P^*(O) \equiv \arg\min_Q \ D_{KL}(P \mid\mid Q)$$

$$\text{such that} \ \ \epsilon_j^- \leq \Theta_{j,n}(P) \leq \epsilon_j^+, \ \ \forall j \in \{1, ..., J\}$$

Approximating the fair world with finite samples

- Assume n iid copies of $O \sim P_\alpha \in \mathcal{M}^{\text{par}}$
- Likelihood function: $\mathcal{L}_n(O; \alpha)$
- Let $\widehat{\Theta}_n(P_\alpha)$ denote the estimator for $\Theta(P_\alpha)$
- Let α^* denote the set of parameters that index $P^*(O)$
- Estimate α^* via solving:

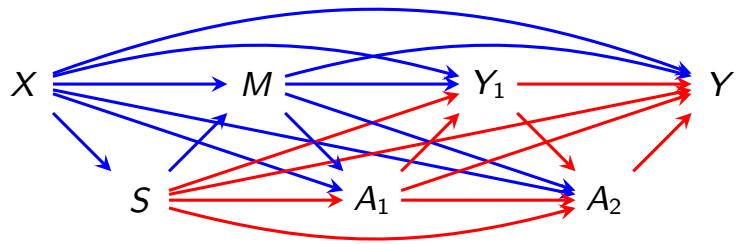
$$\widehat{\alpha^*} = \arg\max_{\alpha} \quad \mathcal{L}_n(O; \alpha)$$
$$\text{such that} \quad \epsilon_j^- \leq \widehat{\Theta}_{j,n}(P_\alpha) \leq \epsilon_j^+, \; j = 1, \ldots, J.$$

Example: a two-stage decision point

Approximate $P^*(O)$ by solving:

$$\widehat{\alpha^*} = \arg\max_{\alpha} \quad \mathcal{L}_n(O;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{\Theta}_{j,n}(P_\alpha) \leq \epsilon_j^+, \; j = 1,2,3.$$



Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{\Theta}_{sy}(P) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{P(S_n|X_n)}\,\frac{P(M_n|s',X_n)}{P(M_n|s,X_n)} - \frac{\mathbb{I}(S_n = s')}{P(S_n|X_n)}\right\} Y_n,$$

$$\widehat{\Theta}_{sa_1}(P) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{P(S_n|X_n)}\,\frac{P(M_n|s',X_n)}{P(M_n|s,X_n)} - \frac{\mathbb{I}(S_n = s')}{P(S_n|X_n)}\right\} A_{1n},$$

$$\widehat{\Theta}_{sa_2}(P) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{P(S_n|X_n)}\,\frac{P(M_n|s',X_n)}{P(M_n|s,X_n)} - \frac{\mathbb{I}(S_n = s')}{P(S_n|X_n)}\right\} A_{2n}.$$

Constraints involve $P(S \mid X;\alpha_s)$ and $P(M \mid S,X;\alpha_m)$ models.

Breaking the cycle of injustice

- Let $P^*(M \mid S, X; \alpha_m)$ and $P^*(S \mid X; \alpha_s)$ be the constrained models chosen to satisfy $\text{PSE}^{sy} = \text{PSE}^{sa_1} = \text{PSE}^{sa_2} = 0$
- Let $\widetilde{P}(O)$ be the joint distribution induced by $P^*(M|S, X; \alpha_m)$ and $P^*(S|X; \alpha_s)$:

$$\widetilde{P}(O) \equiv P(X)P^*(S|X; \alpha_s)\ P^*(M|S, X; \alpha_m) \prod_{k=1}^{K} P(A_k|H_k)P(Y_k|A_k, H_k).$$

- Then PSE^{sy} and PSE^{sa_k} taken wrt $\widetilde{P}(O)$ are also zero.

  $\implies$ constraining the S and M models induces a "fair distribution" no matter how A_k or Y_k are determined.

Three strategies for policy estimation

We consider three strategies for estimating the optimal policy:

- Q-learning
- Value search
- G-estimation

In each case, we must modify these procedures to operate wrt the fair distribution.

As an example, let's look at value search.

Optimal fair policy: Value search

- **Optimal policy**: $f_A^* = \arg\max_{f_A} \mathbb{E}[Y(f_A)]$
- **Unfair world**: expectations wrt to $P(O)$

$$\mathbb{E}[Y(f_A)] = \mathbb{E}\left[\frac{\mathbb{I}(A_1 = f_{A_1}(H_1))}{P(A_1 \mid H_1; \psi)} \times \frac{\mathbb{I}(A_2 = f_{A_2}(H_2))}{P(A_2 \mid H_2; \psi)} \times Y\right],$$

- **Fair world**: expectations wrt to $P^*(O)$

$$\widetilde{\mathbb{E}}[Y(f_A)] = \frac{1}{Z} \sum_{m,s} \mathbb{E}[Y(f_A)] \times P^*(m \mid X, s; \alpha_m) \times P^*(s \mid X; \alpha_s)$$

Data application

Data application: COMPAS

- ProPublica: 2 years worth of COMPAS scores
- Broward county sheriff's office in Florida
- Total of 5278 individuals scored in 2013 and 2014

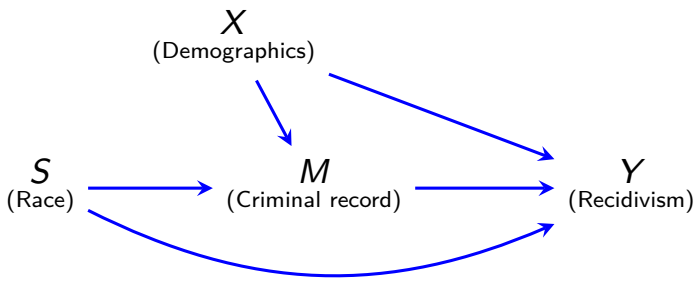
- Race: African Americans (60%) and Caucasians (40%)
- Demographics: gender and age
- Criminal record: binary indicator of crime counts $>$ one
- Recidivism: binary indicator
- COMPAS scores

Task: prediction

Q. Is there any **bias in the data** wrt race in predicting recidivism?

Unfair mechanisms:

$$S \rightarrow Y$$

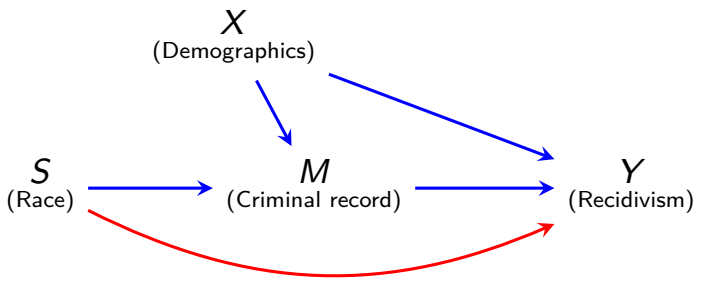


Task: prediction

Q. Is there any **bias in the data** wrt race in predicting recidivism?

Unfair mechanisms:

$S \rightarrow Y$



1. Fairness notion $\Theta(P)$

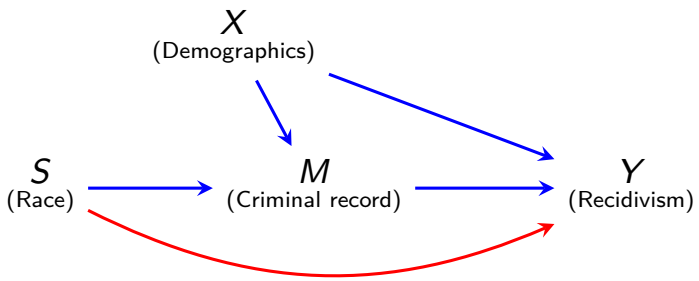
- $\Theta(P)$: direct effect of S on Y
- Let $\{\text{Blacks: } S = 1\}$ and $\{\text{Whites: } S = 0\}$
  - $\mathbb{E}[Y(1, M(0))]$: risk of recidivism had individuals been Black and everything else had been as if they were White
  - $\mathbb{E}[Y(0)]$: risk of recidivism had individuals been White
- Let $\Theta(P)$ be an odds ratio comparison between $\mathbb{E}[Y(1, M(0))]$ and $\mathbb{E}[Y(0)]$

Task: prediction

Q. Is there any **bias in the data** wrt race in predicting recidivism?

Unfair mechanisms:

$$S \rightarrow Y$$



2. Compute $\Theta(P)$

- $\Theta(P)$ is identified from $P(Y, M, S, X)$ as follows:

$$\mathbb{E}[Y(1, M(0))] = \mathbb{E}\left[\frac{\mathbb{I}(S=0)}{P(S=0)} \times \mathbb{E}[Y \mid S=1, X, M]\right]$$

$$\mathbb{E}[Y(0)] = \mathbb{E}\Big[\mathbb{E}[Y \mid S=0, X, M]\Big]$$

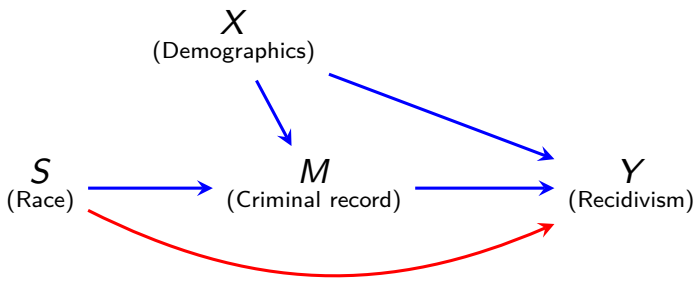
- Use BART to fit the outcome model $\mathbb{E}[Y \mid S, X, M]$
- $\mathbb{E}[Y(1, M(0))] = 0.47, \quad \mathbb{E}[Y(0)] = 0.40$
- $\Theta(P) = 1.3$ $(1.01, 1.45)$ (odds ratio scale)

Task: prediction

Q. Is there any **bias in the data** wrt race in predicting recidivism?

Unfair mechanisms:

$S \rightarrow Y$

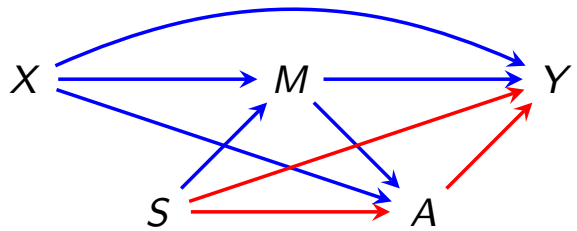


3. Remove $\Theta(P)$ from $\Psi(P) = P(Y = 1 \mid S, X, M)$

	$\Theta(P)$ (odds ratio scale, null $= 1$)	Accuracy %
Unfair world $P(O)$	$1.3\ (1.01, 1.45)$	67.8
Fair world $P^*(O)$	$0.95 \leq \Theta(P) \leq 1.05$	66.4

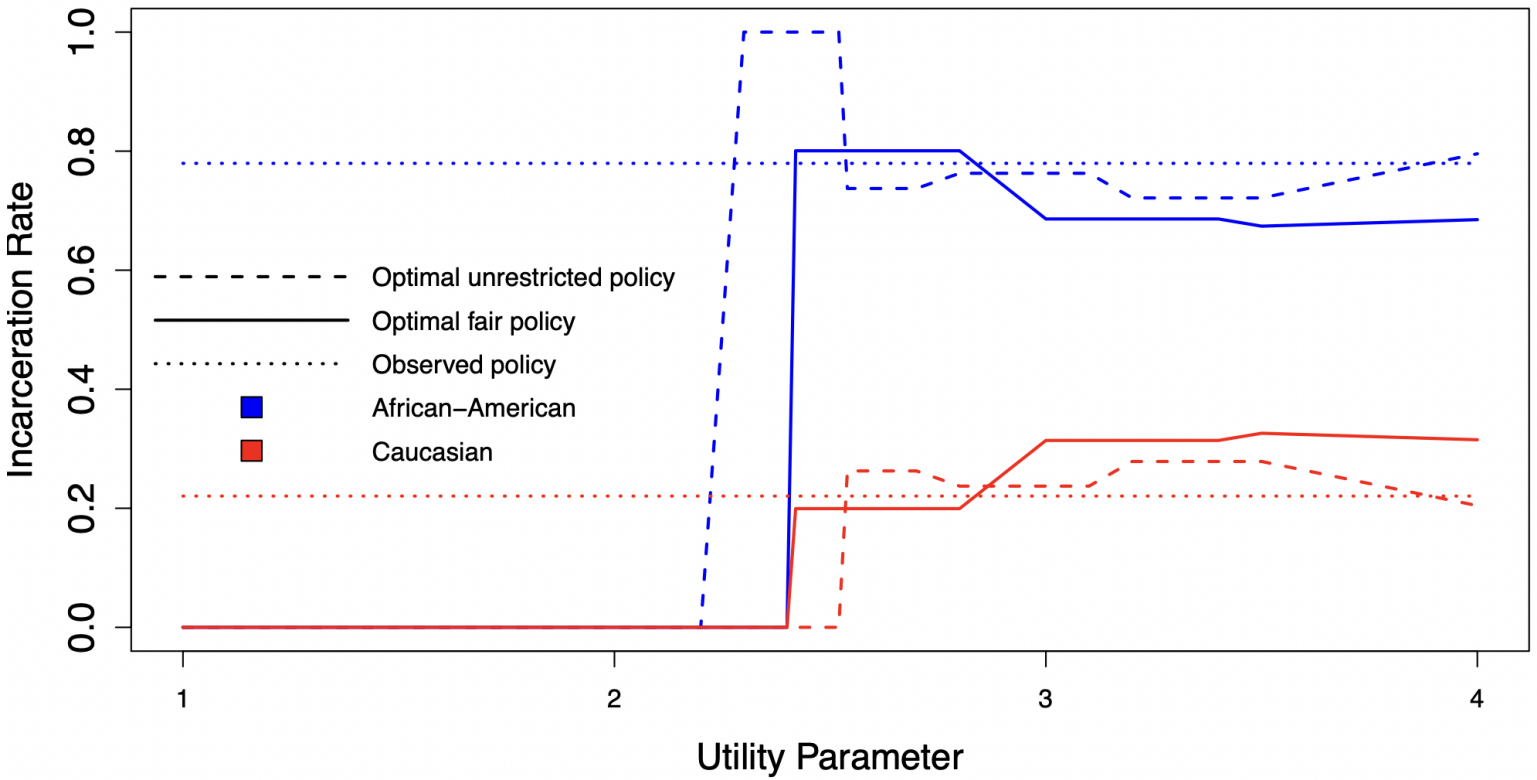
- BART and constrained MCMC to obtain the fair distribution
- Less than 2% relative change in out of sample performance

Task: decision making



- S: race, X: other demographics, M: prior convictions
- A: incarceration (based on risk of recidivism)
- Heuristic utility: $Y \equiv (1 - A) \times \{\theta R + (1 - R)\} - A$
  - R: whether or not recidivism occurred in a span of two years
  - Negative utility (social, economical costs) associated with incarceration $A = 1$.
  - Some cost to releasing individuals who go on to reoffend (i.e., for whom $A = 0$ and $R = 1$) controlled by θ
  - Positive utility associated with releasing individuals who do not go on to recidivate (i.e., for whom $A = 0$ and $R = 0$)

Data application ctd.



Question: What would be the resulting difference in pre-trial incarceration rate under a "fair" vs. unconstrained optimal policy?

Result: "fair" vs. unconstrained policies differ, and incarceration rates depend crucially on the utility function.

Data application ctd.

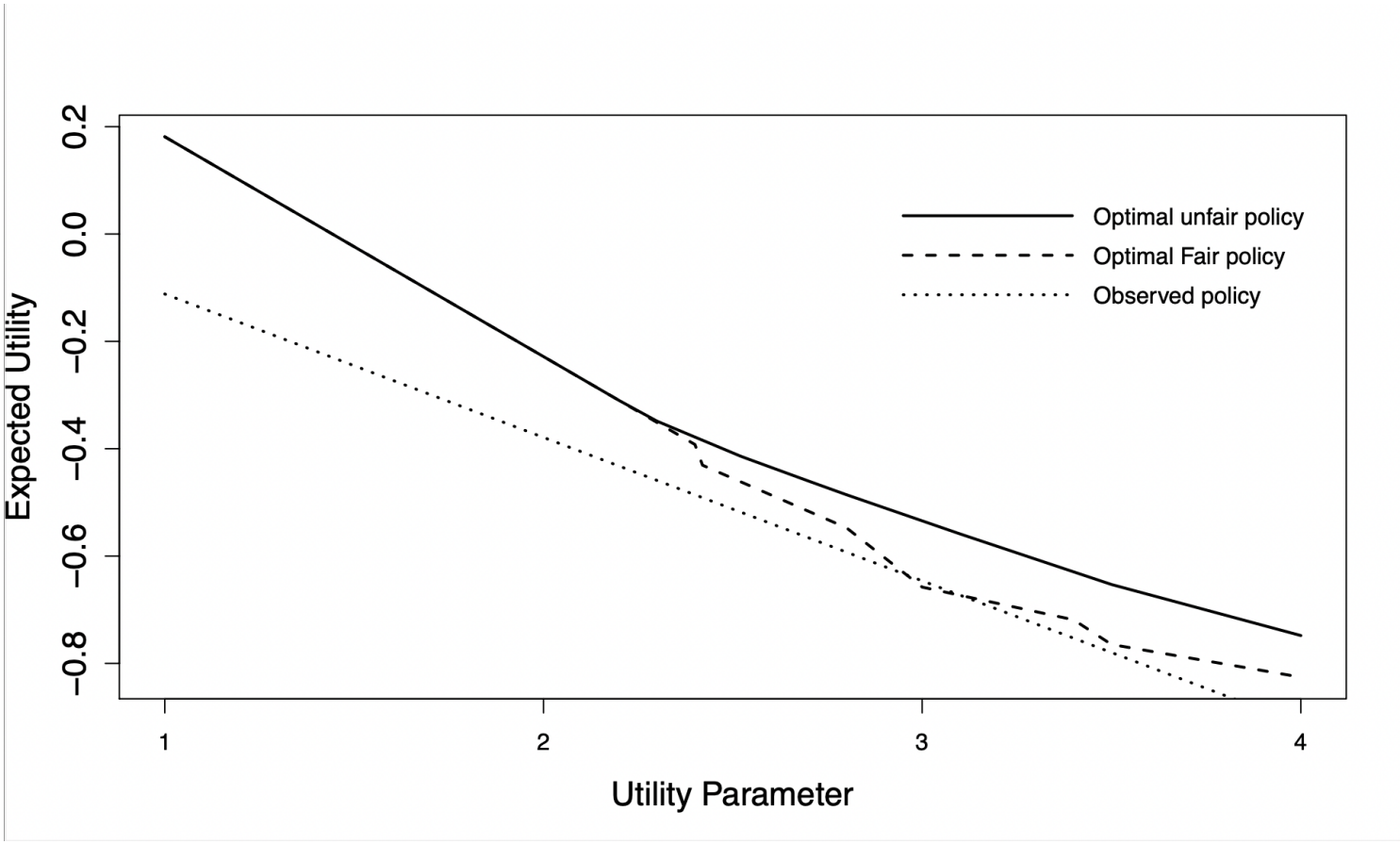


Figure: Relative expected utilities for policies as function of θ

Assessing disparities in cardiac surgical outcomes

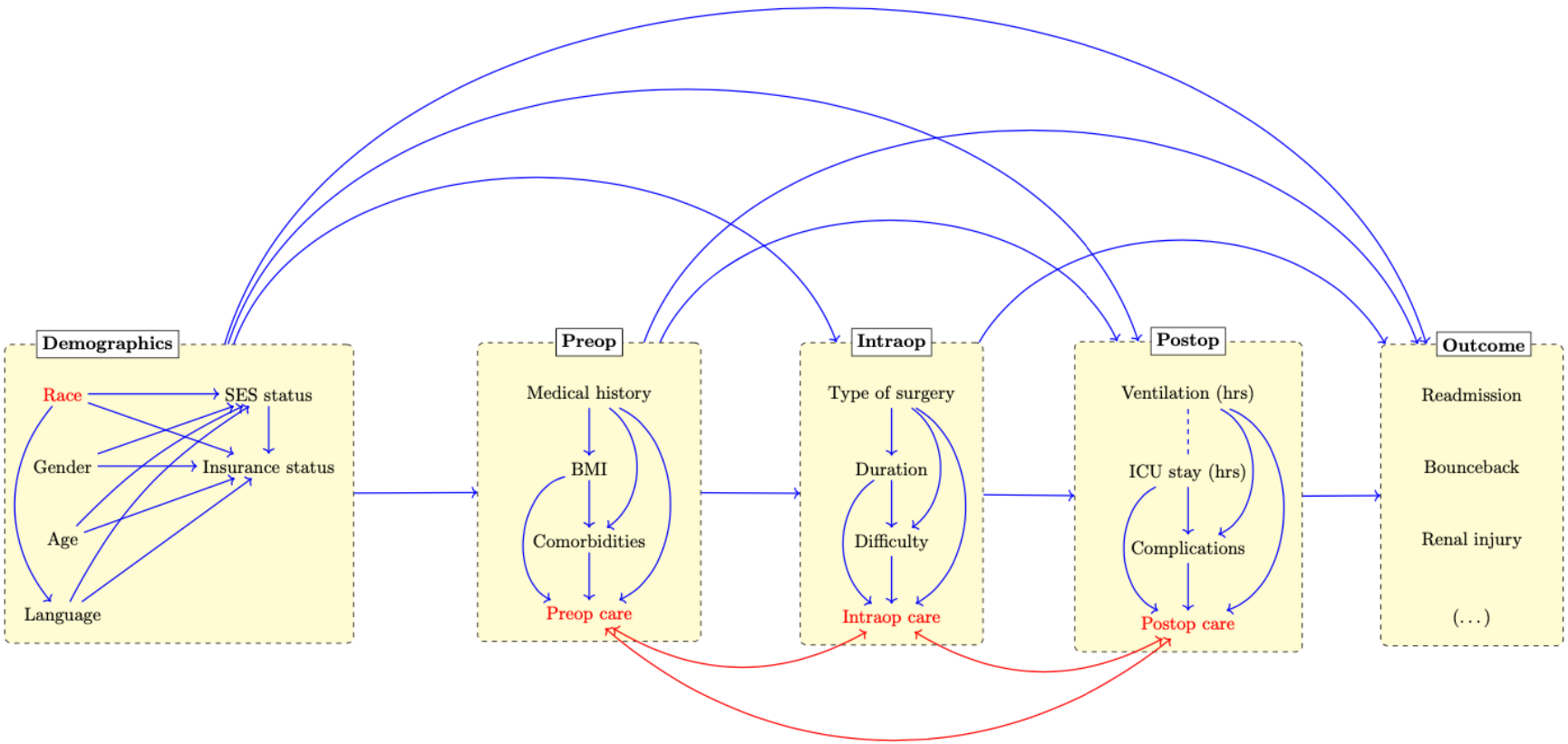
- S : race, Y : readmission, bounceback
- Decomposition of effects along different causal pathways

$S \rightarrow$ SES status $\rightarrow Y$

$S \rightarrow$ Access to high-quality cardiac surgeons $\rightarrow Y$

$S \rightarrow$ Differences in care $\rightarrow Y$

Assessing disparities in cardiac surgical outcomes



Concluding remarks

Recall the three methodological questions

1. **How to express fairness principles mathematically?**
   - The approach we take requires substantive *ethical* input from experts and/or the public.
   - We also require specifying a causal model (based on domain knowledge or causal structure learning).
   - Dealing with unidentified causal effects (use of bounds)

2. **How to modify statistical procedures to reduce unfair effects?**
   - Constrained MLE (hybrid likelihood)
   - Developing more robust constrained optimization methods to use data as efficiently as possible

3. **How to generalize and deploy these modified algorithms?**
   - Be mindful of the fact that samples are collected in p and not p^*
   - Find more effective approaches to map instances between p and p^*

Individual Versus Group Level Fairness

- Population level causal parameters, e.g. $\mathbb{E}[Y(1) - Y(0)]$ are much easier to identify, hence much easier to work with population level fairness criteria.
- Individual level criteria, e.g. $Y_i(1) = Y_i(0)$ are perhaps more desirable, but much more difficult to ensure.
- Important note: conditioning on covariates leads to a population level criterion (non-withstanding causal machine learning terminology).
- In other words, $\mathbb{E}[Y(1) - Y(0)|\vec{X} = \vec{x}] = 0$ reads "the causal effect in a population of people for whom $\vec{X} = \vec{x}$ is 0."
- Whereas, $Y_i(1) - Y_i(0)$ reads "the causal effect for a specific person i (Alice) is 0."
- The former is sometimes identified, the latter is almost never identified.

Race As A Predictor

- Should race be used in risk scores or diagnostic tools?
- Recent work argues against (Lesley et al, 2021), (Diao et al, 2021):
  - Race is a social, not a biological construct.
  - Concerns regarding "algorithmic bias": laundering in structural unfairness behind seeming impartiality of algorithms.
  - Race identity should be about the patient's choice, not the investigator's.
- Arguments for (possibly controversial!):
  - True important predictors may be unavailable, and "race" (as coded) may be the best imperfect proxy.
  - Eliminating race as a predictor may reduce overall model performance, even if many other predictors are introduced (Hsu et al, 2021).
  - In some cases, eliminating race doesn't address the issue: which is unfair causal pathways.
- My own view: inclusion of race is not a priori bad, nor is exclusion a priori good. Discussion has to be tied to a specific fairness criterion we wish to satisfy in a specific application.

Concluding Remarks

- The algorithmic fairness literature is vast, and quickly growing.
- Criteria are often not motivated by use cases.
- Lots of unsurprising negative results (optimality and fairness not jointly achievable, multiple criteria not jointly achievable).
- Causal criteria often have strong motivations, but come with their own challenges (identifiability, need for a causal model).
- Ethics debates are very old, and often intractable.
- The purposes of data scientists in making algorithms fair is clarifying and formalizing legal and political desiderata.
- There is no substitute for a vigorous debate in the public square!.

References for Part I

- Angwin et al. (2016) Machine bias. ProPublica.
- Barocas et al. (2019) Fairness and machine learning. www.fairmlbook.org
- Bao et al. (2021). It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. arXiv:2106.05498.
- Chouldechova (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163. Dieterich et al. (2016) COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical Report by Northpointe.
- Deshpande et al. (2020) Mitigating demographic bias in AI-based resume filtering. In ACM Conference on User Modeling, Adaptation and Personalization.
- Glymour & Herington (2019) Measuring the Biases that Matter. In Proceedings of FAccT.
- Lee & Floridi (2021) Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. Minds and Machines, 31:165-191.
- Kleinberg et al. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In Proceedings of ITCS.
- Obermeyer et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
- Pfohl et al. (2021). An empirical characterization of fair machine learning for clinical risk prediction. Journal of Biomedical Informatics, 113, 103621.
- Raghavan et al. (2020) Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of FAccT.

References for Part II

- "Counterfactual Fairness." Kusner, M., Loftus, J., Russell, C., and Ricardo, S. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.
- "Avoiding Discrimination Through Causal Reasoning." Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.
- Nabi and S. "Fair Inference On Outcomes." In Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence (AAAI-18).
- Nabi, Malinsky and S. "Learning Optimal Fair Policies." In Proceedings of the Thirty-Sixth International Conference on Machine Learning (ICML-19).
- Nabi, Malinsky, and S. "Optimal Training of Fair Predictive Models for Decision Support." In the First Conference on Causal Reasoning and Learning (CLEAR).
- Mitchell, Potash and Barocas. "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions." `https://arxiv.org/abs/1811.07867`

References for Part III

- Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. "Learning fair representations." In International conference on machine learning, pp. 325-333. PMLR, 2013.
- Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).
- Woodworth, Blake, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. "Learning non-discriminatory predictors." In Conference on Learning Theory, pp. 1920-1953. PMLR, 2017.
- Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics 4, no. 1 (2010): 266-298.
- Gelfand, Alan E., Adrian FM Smith, and Tai-Ming Lee. "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling." Journal of the American Statistical Association 87, no. 418 (1992): 523-532.
- Art Owen. Empirical Likelihood. Chapman & Hall, 2001.
- Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." Advances in neural information processing systems 30 (2017).
- Chiappa, Silvia. "Path-specific counterfactual fairness." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7801-7808. 2019.
- Nabi, Razieh, and Ilya Shpitser. "Fair inference on outcomes." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1. 2018.
- Nabi, R., Malinsky, D. and Shpitser, I., 2019, May. Learning optimal fair policies. In International Conference on Machine Learning (pp. 4674-4682). PMLR.
- Nabi, Razieh, Daniel Malinsky, and Ilya Shpitser. "Optimal training of fair predictive models." In Conference on Causal Learning and Reasoning, pp. 594-617. PMLR, 2022.

Sincerely,

Razieh Nabi, PhD (she/her/hers)
Rollins Assistant Professor
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

@raziehnabi
razieh.nabi@emory.edu
https://raziehnabi.com