Fairness in Data Science: Criteria, Algorithms and Open Problems

Part I: Intro, examples, statistical fairness constraints

Daniel Malinsky, Razieh Nabi, Ilya Shpitser

Joint Statistical Meetings

August 6, 2022

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

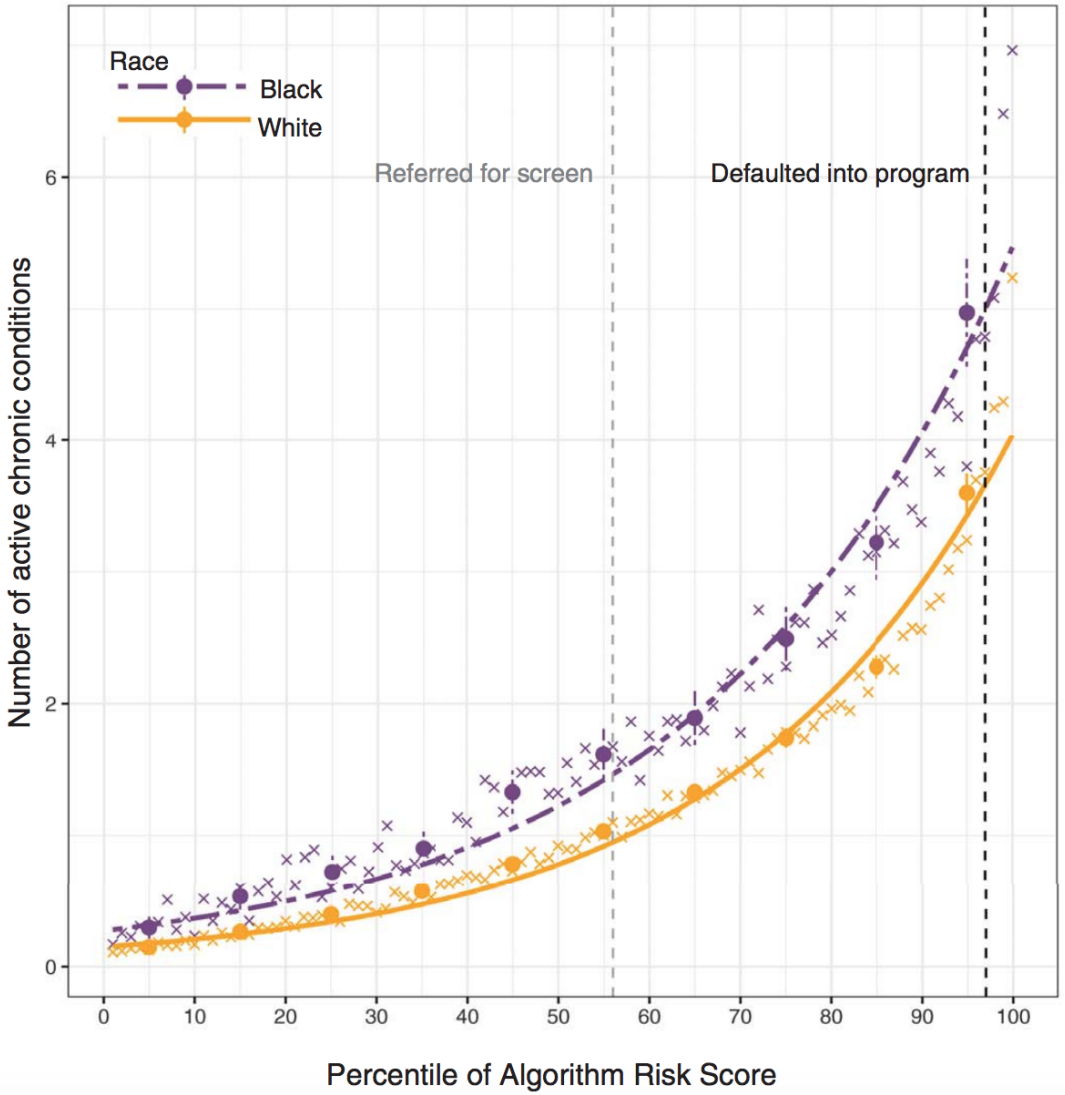
- Imposing causal fairness constraints via constrained optimization
- Example application

Example: racial bias in 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
- Are black and white patients with same predicted risk R actually equally healthy? I.e., is $\mathbb{E}[H \mid R, \text{ White}] = \mathbb{E}[H \mid R, \text{ Black}]$?

Example: racial bias in health risk screening algs

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

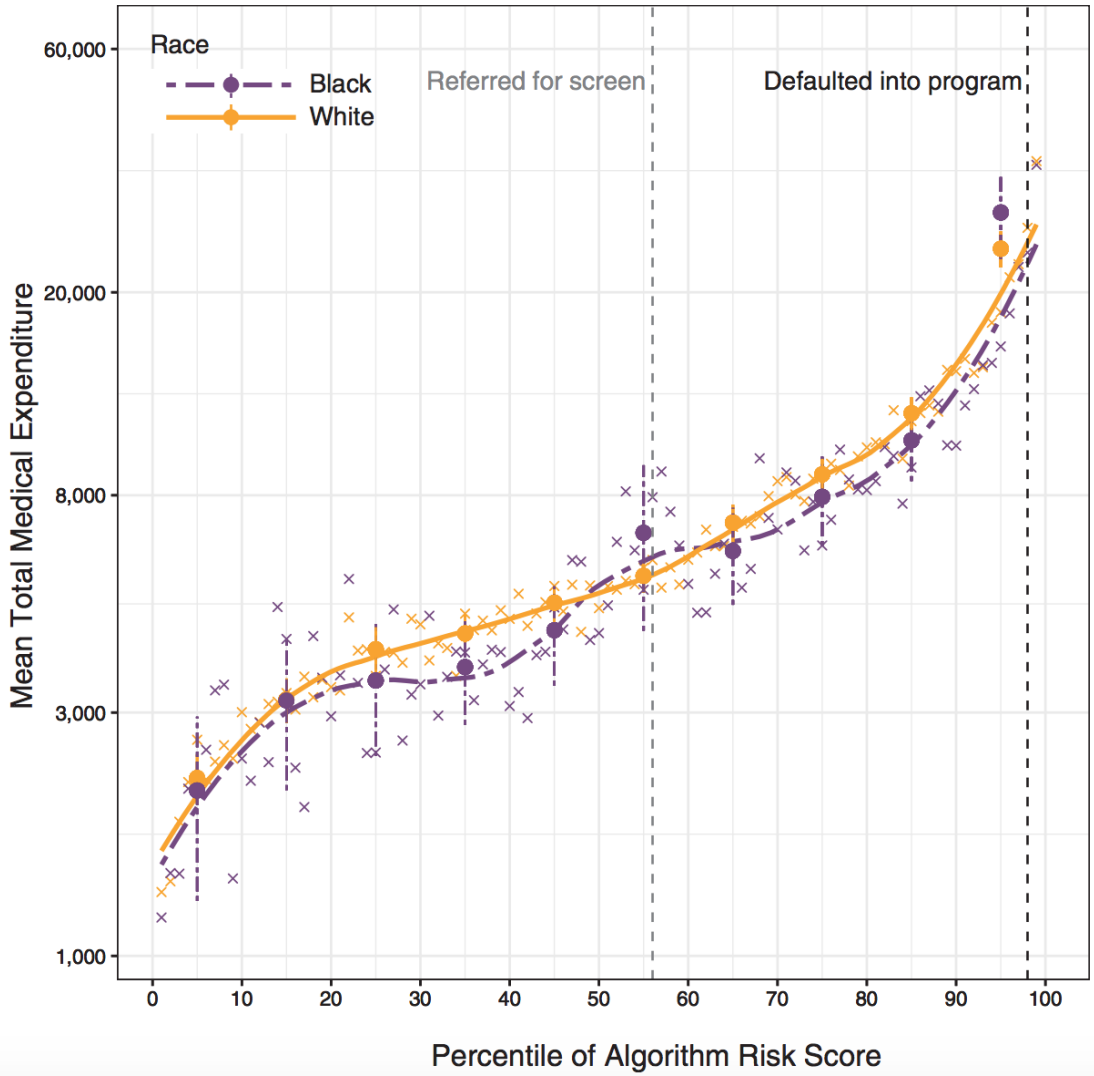


Example: racial bias in health risk screening algs

- Across various definitions of “healthy,” less-healthy Blacks scored at similar risk scores to more-healthy Whites.
- $\Rightarrow$ 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: racial bias in health risk screening algs

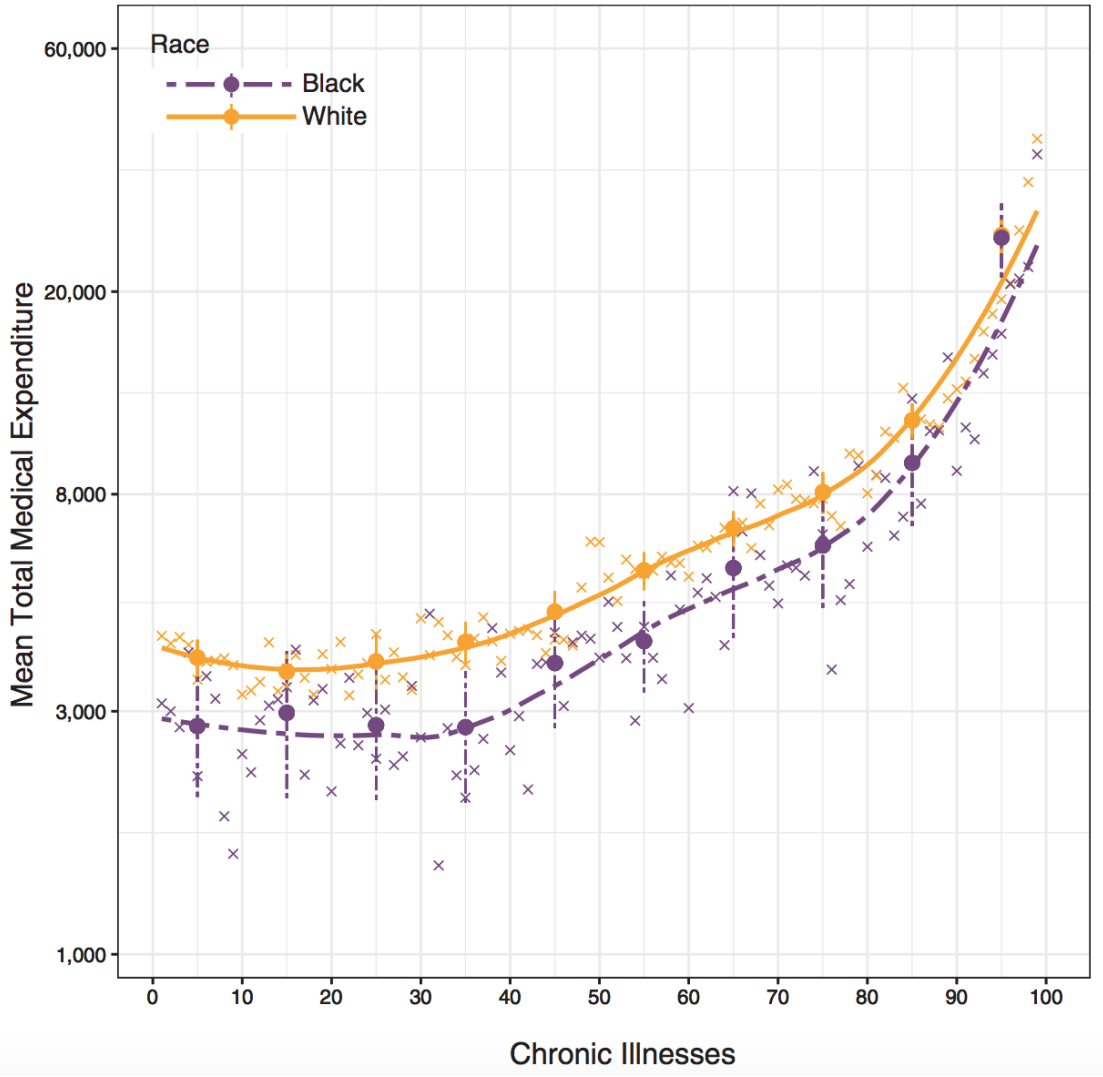
Note: estimated risk R accurately predicted Y, did not include race among the features X, and was approximately calibrated by race (higher $R \implies$ higher Y for both racial groups)



Example: racial bias in health risk screening algs

However, medical expenditures Y differed systematically by race.

More severely ill Black patients $\implies$ lower medical expenditures than healthier White patients



Example: racial bias in health risk screening algs

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

Authors propose alternative predictive algorithms (e.g., predicting chronic disease incidence) and see marked decrease in subsequent disparities in enrollment into the intervention

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 ("label choice bias")
- 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

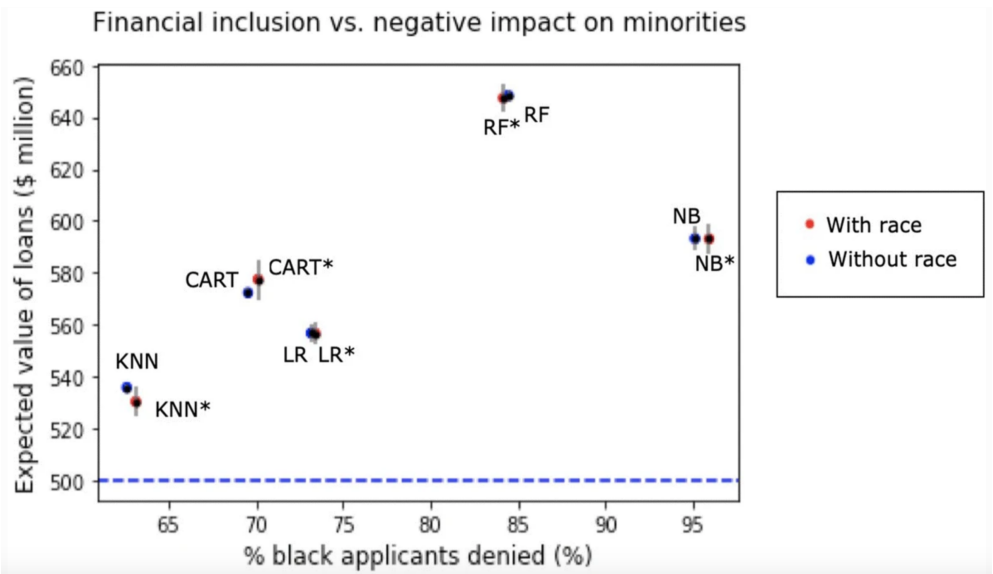
Example: financial lending algs

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: financial lending algs

- 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.



Lee & Floridi (2021) using US mortgage data, comparing several algorithms

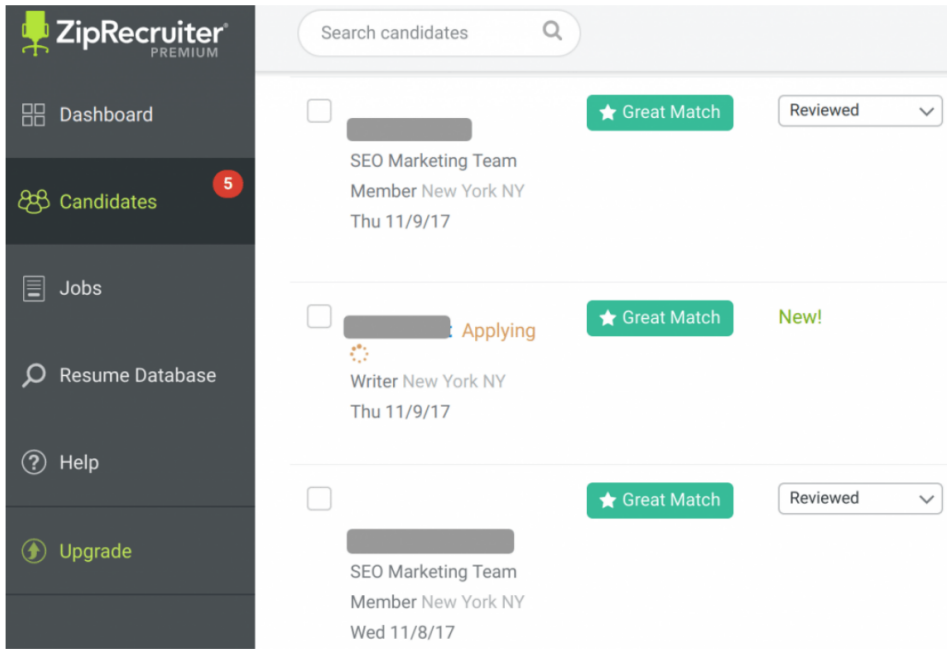
Example: financial lending algs

Key takeways:

- “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? “triangulation of race” by algorithms and/or detection of nonlinear relationships among race, outcomes, and other vars

Example: automated resume screening for hiring

- Many institutions use algorithmic tools to automatically screen (or rank) resumes of job applicants.
- The training data may be resume text, but also sometimes may include bespoke questionnaires, “video resumes,” or game play interactions.
- Targets of prediction may include performance reviews, sales numbers, or retention time based on past employees. May also include abstract scores of cognitive or behavioral traits. See Raghavan et al. (2020).



ZipRecruiter’s “Great Match” badges

Example: automated resume screening for hiring

- 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.”
- Some companies evaluate their software for bias, often including compliance with legal standards and UGESP “4/5 rule”
- A particular worry is “differential validity”: when an assessment is better at ranking members of one group than another

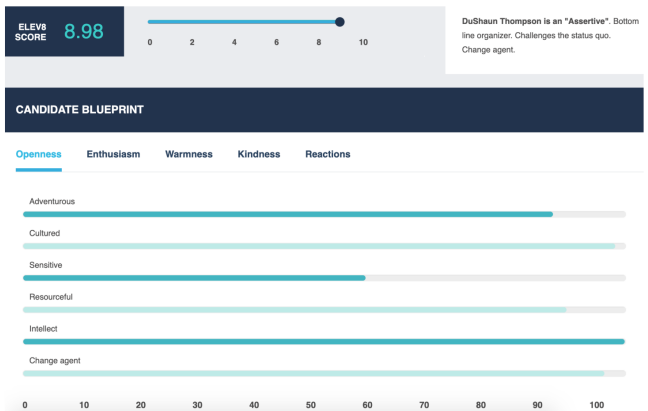


Figure 2: Part of a sample candidate profile from 8 and Above, based on a 30-second recorded video cover letter (screenshot from the 8 and Above website: https://www.8andabove.com/p/profile/blueprint/643)

Example: automated resume screening for hiring

- Deshpande et al. (2020) study nationality-related socio-linguistic bias in a document "similarity matching" tool for resume screening
- Data from Singapore, resumes of either Chinese, Malaysian, or Indian origin $\Rightarrow$ standard approach led to severe downweighting of Chinese applicants.
- Authors propose a modification that penalizes "matching keywords that are typical to one section of society while encouraging matching keywords that are common among all demographics." $\Rightarrow$ More even distribution of matched resumes from different nationalities.

Table 5: Top Words from China before and after de-biasing

Before de-biasing	After de-biasing
china	management
management	financial
financial	research
business	investment
investment	business
research	credit
university	finance
finance	company
company	university
kong	fund

Example: recidivism risk prediction

- Algorithms for recidivism risk prediction have been used in various criminal justice contexts: pretrial release conditions, bail determinations, probation eligibility, etc.
- COMPAS is a tool (and now widely used benchmark dataset) 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"
  - "White defendants were more likely than black defendants to be incorrectly flagged as low risk"

Example: recividism risk prediction

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

ProPublica analysis

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	41%	37%
Labeled Lower Risk, Yet Did Re-Offend	29%	35%

Northpointe analysis

Example: recidivism risk prediction

Table 3.3: AUC results for the General Recidivism Risk Scale (GRRS) decile scores and Violent Recidivism Risk Scale (VRRS) decile scores in the data analysis samples used for ProPublica's reverse logistic regression models (Sample A).

	Sample	n	events	base rate	AUC	Lower 95% CI	Upper 95% CI
GRRS							
	White	2103	822	0.39	0.693	0.670	0.716
	Black	3175	1661	0.52	0.704	0.686	0.722
	All	6172	2809	0.46	0.710	0.697	0.723
VRRS							
	White	1459	174	0.12	0.683	0.640	0.726
	Black	1918	404	0.21	0.708	0.680	0.737
	ALL	4020	652	0.16	0.719	0.698	0.741

Note. GRRS outcome is a misdemeanor or felony arrest. VRRS outcome is a violent misdemeanor or felony arrest.

Northpointe analysis

Example: recidivism risk prediction

Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402
*Note: *p<0.1; **p<0.05; ***p<0.01*	

ProPublica analysis

Example: recidivism risk prediction

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?

Example: recidivism risk prediction

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?

ProPublica also ask if race is conditionally associated with score values.

Northpointe also ask if overall accuracy (AUC) is similar across groups.

Warning: there are a lot of issues with the COMPAS data!

- Bias in Y is a big issue. "Re-arrest" $\neq$ "Re-offense" $\neq$ "failure to appear/pretrial flight"...
- Bias in covariates and bias in demographic labels (are these measuring appropriate things?)
- Complicated relationship between risk predictions and actual judicial decisions
- Important normative/ethical questions about role of risk assessment in criminal justice reform are typically ignored.

Table 1: Issues with the COMPAS Dataset

Issues with COMPAS Dataset	CS translation
Arbitrary choice of threshold for assessing classification disparities	Incorporate discretion
2-year follow-up period includes arrests after case closed (no longer pretrial)	Bias in outcome value Y
Differing follow-up periods for rearrested vs. not rearrested individuals	Bias in outcome value Y
Restricted range of risk scores	Issues with distribution of data D

See Bao et al. (2022)

Shared structure of prediction tasks

Hiring, admissions (e.g., university), financial lending, & targeted advertising are common examples in algorithmic fairness lit because they share similar structure: predict “good candidate” from available features in order to differentially distribute some valuable thing.

These tasks are also subject to legal restrictions that are complicated, often related to “disparate impact.”

Risk prediction in criminal justice and healthcare can be quite different, affecting distribution of both “good” and “bad” things (resources, punishment) depending how they’re used. No reason to expect one-size-fits-all answers.

Shared structure of prediction tasks

Nevertheless, 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 actually 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.

Data reflect the socially stratified, disparate, and unfair reality behind the data.

One kind of approach would be to intervene on the reality behind the data or how data is used. Another approach is to intervene on the algorithm.

Statistical fairness criteria

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

- Decision $\perp\!\!\!\perp$ Group ("statistical parity")
- Decision $\perp\!\!\!\perp$ Group $\mid Y$ ("equalized odds" or "equal error rates")
- $Y \perp\!\!\!\perp$ Group $\mid$ Decision ("equal positive/negative predictive values" or "calibration")

Statistical fairness criteria

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

- $\widehat{Y} \perp\!\!\!\perp$ Group ("statistical parity") [also $\widehat{Y} \perp\!\!\!\perp$ Group | Covariates]
- $\widehat{Y} \perp\!\!\!\perp$ Group | Y ("equalized odds" or "equal error rates")
- $Y \perp\!\!\!\perp$ Group | $\widehat{Y}$ ("equal positive/negative predictive values" or "calibration")

Statistical fairness criteria

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

- $\widehat{Y} \perp\!\!\!\perp$ Group ("statistical parity") [also $\widehat{Y} \perp\!\!\!\perp$ Group | Covariates]
- $\widehat{Y} \perp\!\!\!\perp$ Group | Y ("equalized odds" or "equal error rates")
- $Y \perp\!\!\!\perp$ Group | $\widehat{Y}$ ("equal positive/negative predictive values" or "calibration")

Note: these can go under other names in the literature...

Conflicts

Chouldechova (2017) shows that *so long as base rates differ across groups* (e.g., diff recividism rates), then "equal error rates" and "equal positive/negative predictive values" **cannot be both satisfied**.

Follows from:

$$\text{FPR} = \frac{p}{1-p}\frac{1-\text{PPV}}{\text{PPV}}(1-\text{FNR})$$

where p is prevalence (base rate).

Conflicts

Kleinberg et al. (2016) prove a similar incompatibility result for balance in the positive class, balance in the negative class,[1] and calibration within groups. (Let $\widehat{Y}$ be a score, Y binary.)

Balance in positive class: $\widehat{Y} \perp\!\!\!\perp \text{Group} \mid Y = 1$

Balance in negative class: $\widehat{Y} \perp\!\!\!\perp \text{Group} \mid Y = 0$

Calibration within groups: $P(Y = 1 \mid \widehat{Y}, \text{Group}) = \widehat{Y}$.

[1] These are actually typically defined using "mean scores": e.g., $\mathbb{E}[\widehat{Y} \mid Y = 1, \text{Group}_1] = \mathbb{E}[\widehat{Y} \mid Y = 1, \text{Group}_2]$

Conflicts

Kleinberg et al. (2016) prove a similar incompatibility result for balance in the positive class, balance in the negative class,[1] and calibration within groups. (Let $\widehat{Y}$ be a score, Y binary.)

Balance in positive class: $\widehat{Y} \perp\!\!\!\perp \text{Group} \mid Y = 1$

Balance in negative class: $\widehat{Y} \perp\!\!\!\perp \text{Group} \mid Y = 0$

Calibration within groups: $P(Y = 1 \mid \widehat{Y}, \text{Group}) = \widehat{Y}$.

Theorem: If there is a risk assignment satisfying balance in positive class, balance in negative class, and calibration within groups, then there must either be perfect prediction or equal base rates.

[1] These are actually typically defined using "mean scores": e.g., $\mathbb{E}[\widehat{Y} \mid Y = 1, \text{Group}_1] = \mathbb{E}[\widehat{Y} \mid Y = 1, \text{Group}_2]$

Conflicts

Barocas et al. (2019) also derive similiar conflicts between “statistical parity” and the other two, when base rates are not equal.

The proof for equalized odds is easy. Let A denote group membership. Assume $Y \not\perp\!\!\!\perp A$ (unequal base rates). Then statistical parity and equalized odds cannot both hold.

$\widehat{Y} \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A|\widehat{Y} \implies (Y, \widehat{Y}) \perp\!\!\!\perp A \implies Y \perp\!\!\!\perp A$.
Follows from contraction and decomposition properties of $\perp\!\!\!\perp$. So:

$Y \not\perp\!\!\!\perp A \implies \widehat{Y} \not\perp\!\!\!\perp A$ or $Y \not\perp\!\!\!\perp A|\widehat{Y}$

Conflicts

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

Imposing statistical fairness constraints

There are various approaches to training prediction algorithms to enforce such statistical constraints.

The most straightfoward is simply penalizing violations of the chosen constraint in the training procedure:

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}[\mathcal{L}(Y, f(X))] + \lambda R$$

where R is a non-negative regularizer indicative of the extent to which the fairness criterion is violated, λ is a non-negative tuning parameter, and $\mathcal{L}$ is a loss function.

Imposing statistical fairness constraints

There are various approaches to training prediction algorithms to enforce such statistical constraints.

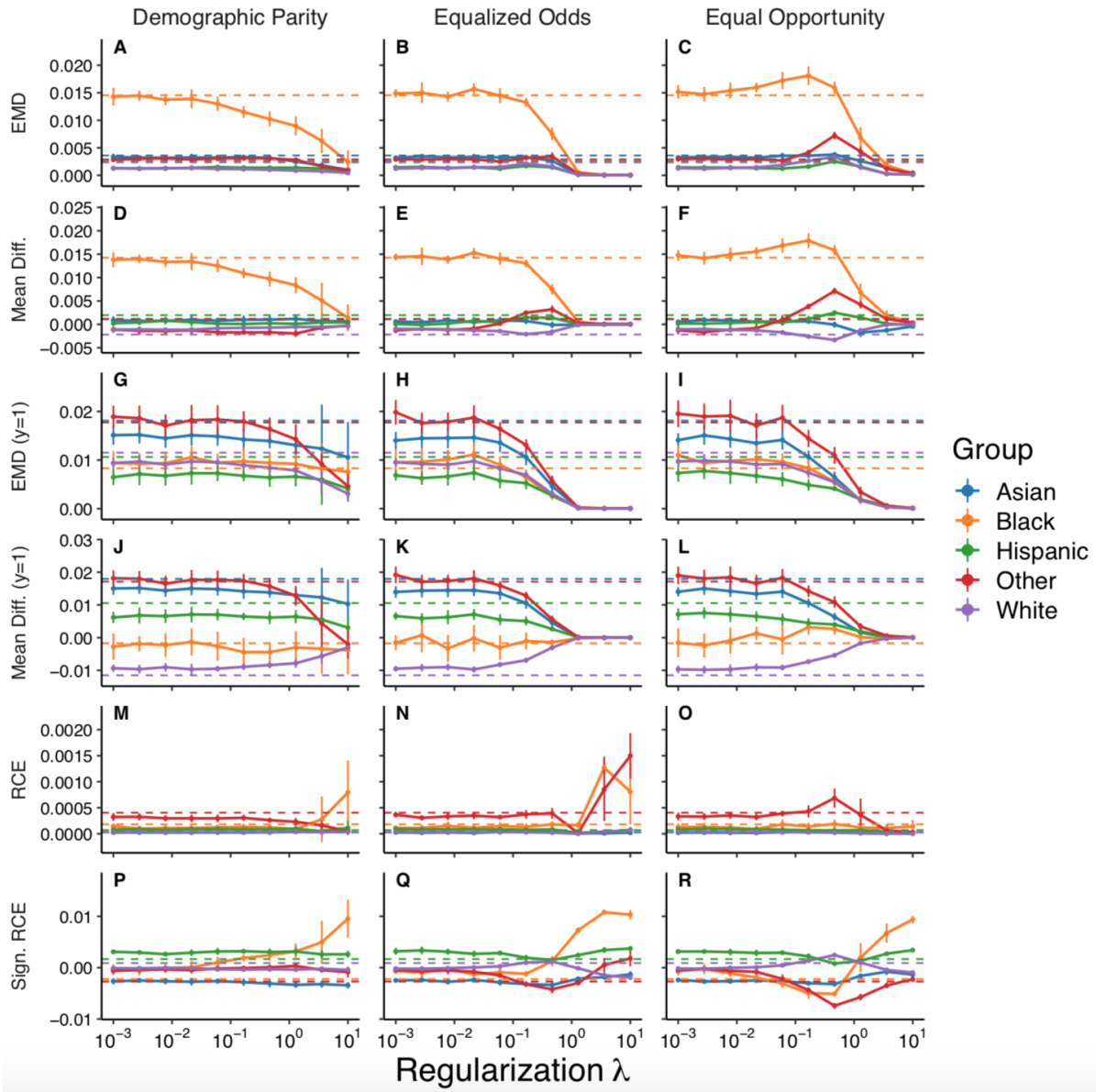
The most straightfoward is simply penalizing violations of the chosen constraint in the training procedure:

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}[\mathcal{L}(Y, f(X))] + \lambda R$$

where R is a non-negative regularizer indicative of the extent to which the fairness criterion is violated, λ is a non-negative tuning parameter, and $\mathcal{L}$ is a loss function.

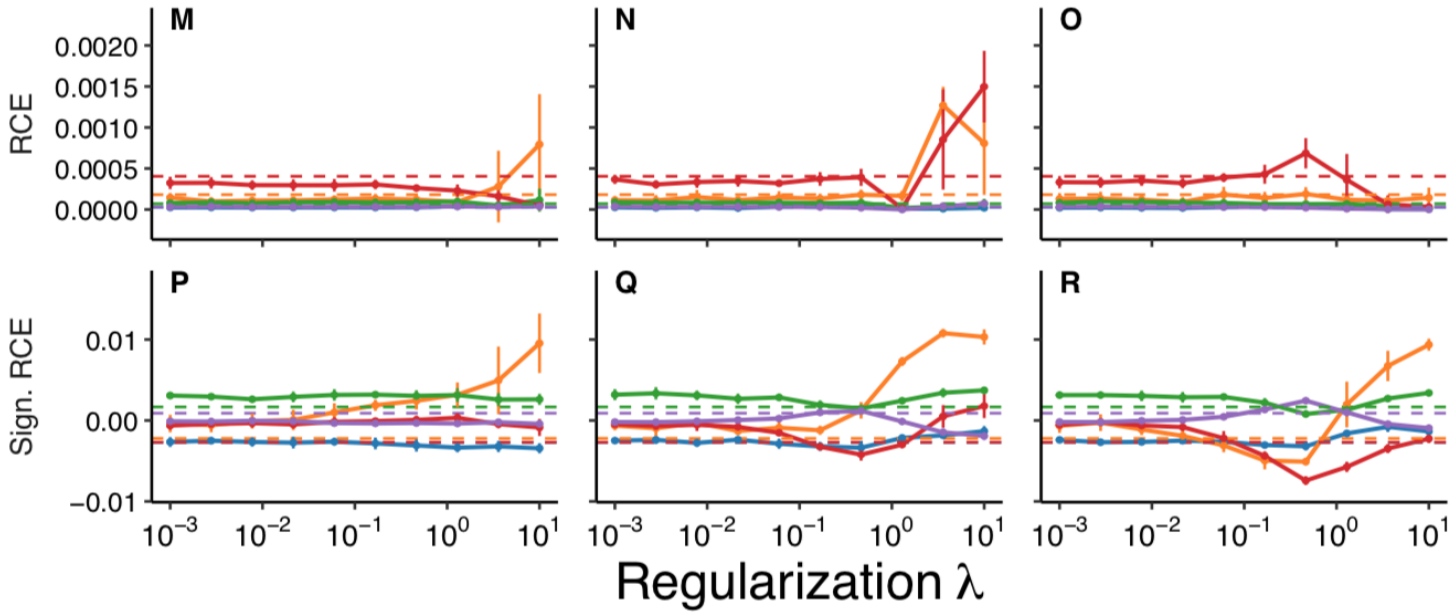
Ex: $R_{EqOdds} = \sum_y \sum_a D(P(f \mid A = a, Y = y) || P(f \mid Y = y))$
where $D(\cdot||\cdot)$ is some function that measures distance between distributions.

Consequences in applied problems



From Pfohl et al. (2021) in an application predicting risk of 30-day hospital readmission using STARR database

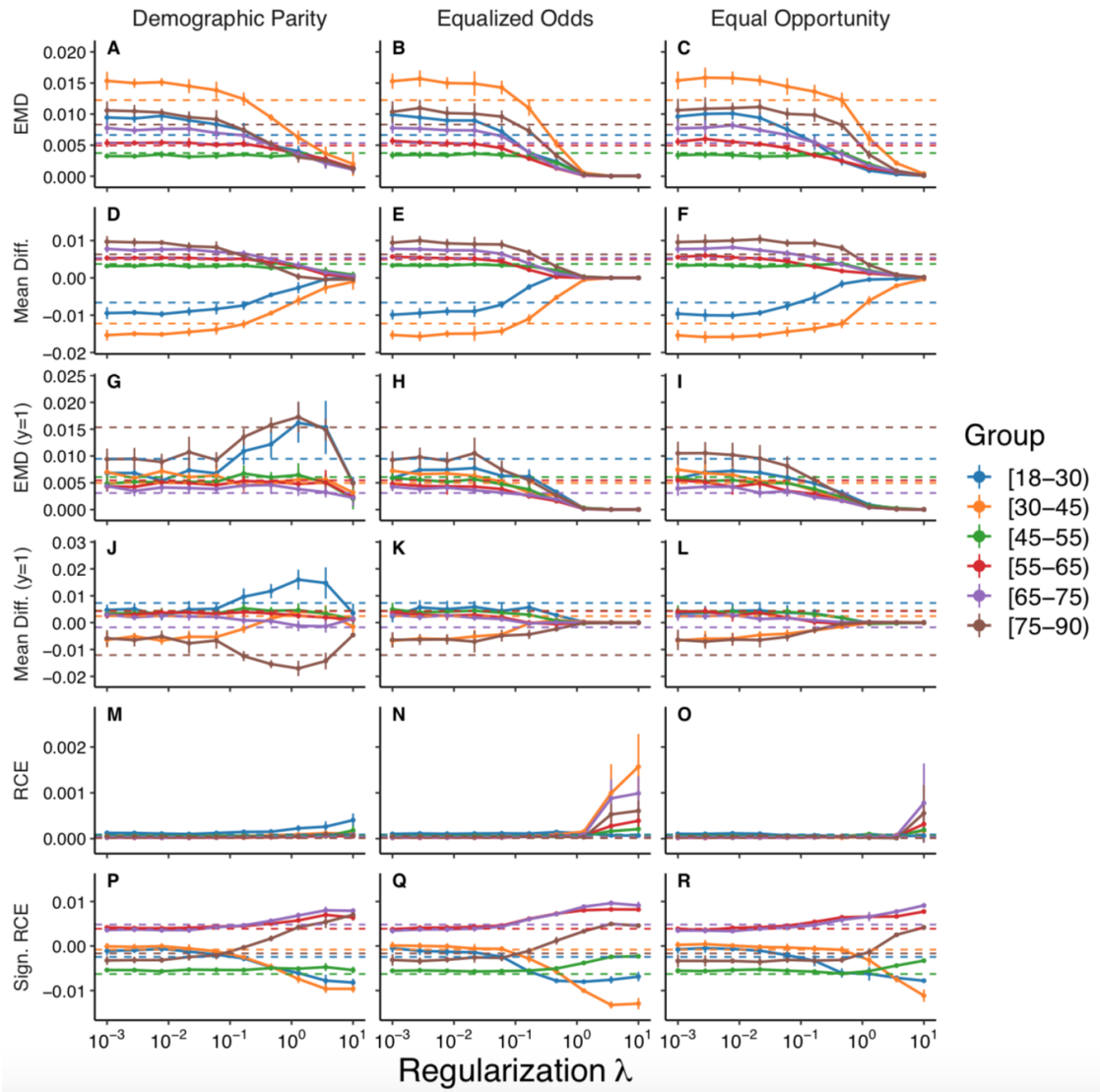
Consequences in applied problems



From Pfohl et al. (2021) in an application predicting risk of 30-day hospital readmission using STARR database

RCE denotes a measure of "relative calibration error" by group, i.e., how different is $P(Y = 1 \mid \widehat{Y}, A = a_j)$ from $P(Y = 1 \mid \widehat{Y})$ for each group a_j.

Consequences in applied problems



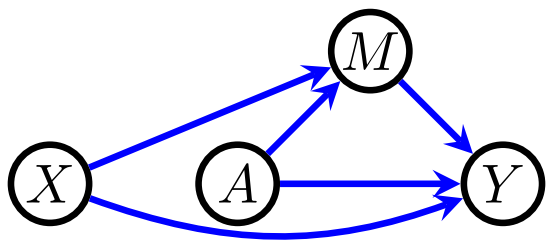
From Pfohl et al. (2021) in an application predicting risk of 30-day hospital readmission using STARR database

Note: these issues are not specific to these fairness constraints! Can ask same questions for other statistical and causal constraints.

Possible causal structures underlying the data

- 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)



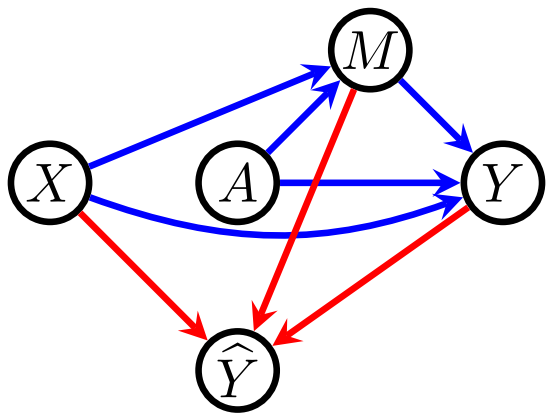
DAGs (briefly)

- 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 tbd later)
- If all independence relations among variables follow from d-separation in a DAG we say the distribution is “faithful" to the DAG. This implies there are no “accidental” or “non-structural” independence relations.

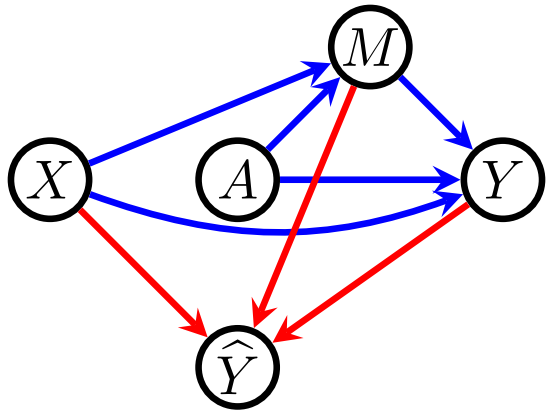
Possible causal structures underlying the data

- 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)

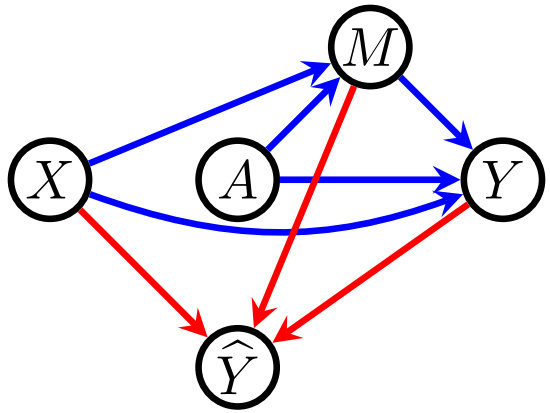


Possible causal structures underlying the data



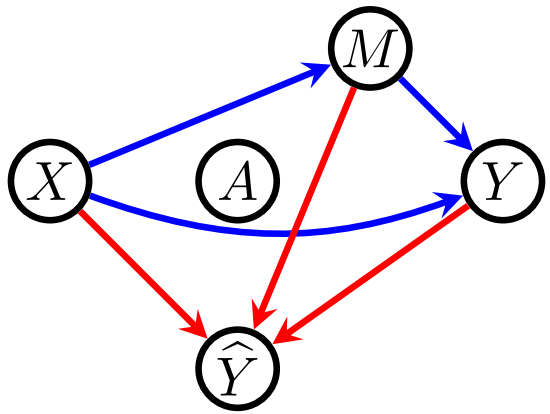
- Even if prediction alg does not use A, as long as there is a mechanism connecting A to any variable that determines $\widehat{Y}$, statistical parity will be violated: $\widehat{Y} \not\perp\!\!\!\perp A$

Possible causal structures underlying the data



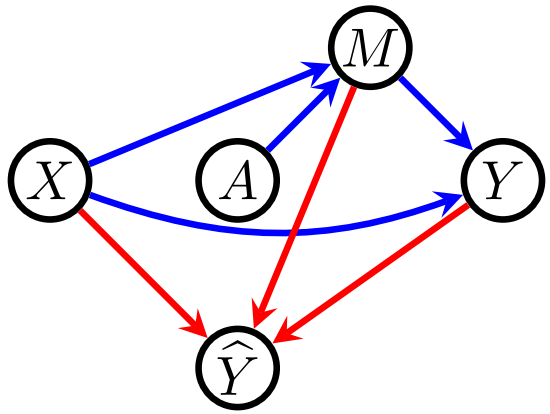
- Even if prediction alg does not use A, as long as there is a mechanism connecting A to any variable that determines $\widehat{Y}$, statistical parity will be violated: $\widehat{Y} \not\perp\!\!\!\perp A$
- Also guaranteed to violate equalized odds: $\widehat{Y} \not\perp\!\!\!\perp A \mid Y$ by d-separation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

Possible causal structures underlying the data



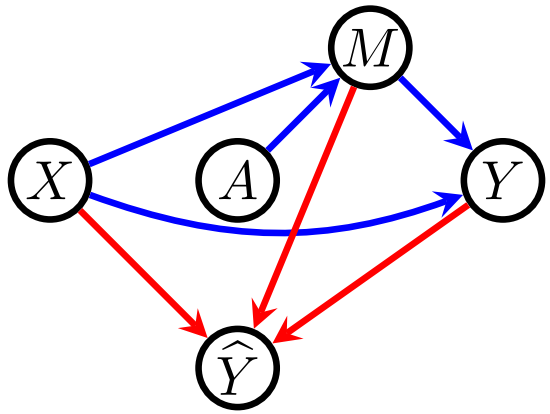
- If $A \to \cdots \to Y$, will not have calibration generally, $Y \not\perp\!\!\!\perp A \mid \widehat{Y}$
- Can achieve calibration either if A is causally inert (doesn't cause Y or M) or in "fringe" cases, e.g., $\widehat{Y} = Y$ or $\widehat{Y} = A$ or various combos of associations nearly cancel out ("unfaithfulness")

Possible causal structures underlying the data



- If $A \to \cdots \to Y$, will not have calibration generally, $Y \not\perp\!\!\!\perp A \mid \widehat{Y}$
- Can achieve calibration either if A is causally inert (doesn't cause Y or M) or in "fringe" cases, e.g., $\widehat{Y} = Y$ or $\widehat{Y} = A$ or various combos of associations nearly cancel out ("unfaithfulness")

Possible causal structures underlying the data



- If $A \to \cdots \to Y$, will not have calibration generally, $Y \not\perp\!\!\!\perp A \mid \widehat{Y}$
- Can achieve calibration either if A is causally inert (doesn't cause Y or M) or in "fringe" cases, e.g., $\widehat{Y} = Y$ or $\widehat{Y} = A$ or various combos of associations nearly cancel out ("unfaithfulness")
- Modifying algs to impose these fairness constraints *induce* unfaithfulness on the joint distribution (cf. Glymour & Herington 2019)

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

Up next: relevant background on causality

References

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.

Satisfying Causal Fairness Criteria: Algorithms and Open Problems

Part II: Introduction to Causal Inference, and Causal Fairness Criteria.

Daniel Malinsky, Razieh Nabi, Ilya Shpitser

Joint Statistical Meeting

August 7, 2022

Overview

1. Basics of causal inference
   - Potential outcomes and graphs: the mathematical language of causation.
   - Causal parameters (the average causal effect).
   - Identification and estimation (briefly).
2. Mediation Analysis
   - Motivation: discrimination and etiology.
   - Direct and indirect effects.
   - Identification and estimation (briefly).
3. Causal fairness criteria.
   - Causal versions of associative criteria.
   - Actual cause inspired counterfactual fairness (Kusner et al, 2017).
   - Causal fairness based on resolving or proxy variables on causal paths (Kilberus et al, 2017)
   - Constraining path-specific effects (Nabi et al).

Basics Of Causal Inference

Getting Causality From A Statistical Model

- Can learn model parameters from data.
- Would be great if we could interpret them causally.
- Example: large coefficient in linear regression – large causal effect (guns cause murders, alcohol causes accidents, etc.)
- But everyone knows: association does not imply causation:
  - People in hospitals tend to be sick. Therefore I should not go to the hospital – it will make me sick. (N.B.: not entirely unreasonable!)
  - People who own olympic gold medals in running tend to be fast runners! Therefore to become fast, I should buy a medal!
  - "Cargo cult" or superstitious behavior.
- When does association imply causation?

A Quote on Causality from the 1740s

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.

David Hume (1748)

- "First" = cause, "second" = effect:

  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.

- This is a counterfactual definition.
- Let's try to think about this formally.

Counterfactuals

- Will need outcome Y (like in a regression) and treatment or exposure A.
- Will define a potential outcome:

$$Y(a) \equiv \text{“}Y \text{ if } A\text{, possibly contrary to fact, had value } a\text{.”}$$

- What this is not (in general): Y conditional on $A = a$ $(Y \mid a)$.
- What this is:
  - A is input to a procedure in R or Stata, Y is the output. $Y(a)$: output of procedure if we stopped execution in debugger, and set input to a.
  - $p(Y \mid a)$: probability of rain if my lawn is wet.
  - $p(Y(a))$: probability of rain if I sprayed my lawn with a hose.

Encoding Hume's Definition

- $A = 1$: fire, $Y = 1$: smoke, $A = 0$: no fire, $Y = 0$: no smoke.
- Smoke follows fire: $Y(A = 1) = 1$.
- If there had been no fire, there would have been no smoke: $Y(A = 0) = 0$.
- Can establish causality by comparing $Y(a)$ for different a:

$$Y(A = 1) - Y(A = 0).$$

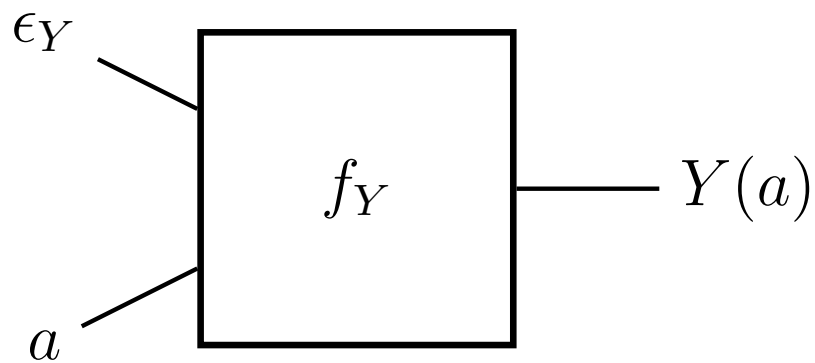
- This is called a causal effect or causal contrast.
- Shorthand: $Y(1) \equiv Y(A = 1)$, if A is understood.
- Pearl's do operator is equivalent: $Y(1)$ is the same as $Y|\text{do}(A = 1)$.

Linking The Counterfactual And The Factual

- Data records what actually happened.
- What we want is something that did not happen.
- We need to link counterfactuals and observed data.
- Standard assumption is called consistency: $Y(A) = Y$. Read:

  "Observed Y and Y if we were to set A to whatever value it was observed are the same variable."

- Intuition in terms of *an invariant causal mechanism* $Y(a) \leftarrow f_Y(a, \epsilon_Y)$:



- If A was observed to be a, we can get $Y(a)$ as Y. But what if A were something else?

Fundamental Problem of Causal Inference

For every row, only see the right two columns, in particular only one outcome (observed Y)!
Interested in middle column that's a function of the left two columns.

	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	A	Y
	1.1	2.3	-1.2	1	1.1
	1.8	0.3	1.5	0	0.3
	2.0	2.1	-0.1	0	2.1
	0.1	1.3	-1.2	1	0.1
mean	1.25	1.5	-0.25		

Consistency gives us half of the left two columns, since $Y(a) = Y$. Need assumptions to link table on A, Y and the rest of the table on $Y(1), Y(0)$.

Ignorability

- A determines what treatment people get.
- Intuition: want A not to depend on potential outcome.
- Example: flip a coin, if heads $A = 1$, if tails, $A = 0$.
- Results in "fair" assignment, any difference in $Y(A)$ has to do with the person, not the assignment mechanism.
- False if e.g. sick people get $A = 1$, healthy people get $A = 0$.
- Formally: $\{Y(1), Y(0)\} \perp\!\!\!\perp A$.
- Known as ignorability.

Consequences Of Ignorability

- Remember, want to compare, $Y(1)$ and $Y(0)$. Have data on Y and A.
- Assume we had *infinite* amount of data, in fact we knew the underlying distribution $p(Y, A)$.
- Assume $\{Y(1), Y(0)\} \perp\!\!\!\perp A$, and consistency. Then

$$p(Y(1)) = p(Y(1)|A = 1) = p(Y|A = 1)$$
$$p(Y(0)) = p(Y(0)|A = 0) = p(Y|A = 0)$$

- Ignorability (random treatment assignment) means association is causation(!)
- Basis of the causal validity of randomized controlled trials.
- Using causal model assumptions to express causal parameters in terms of the observed data distribution is the subject of *causal identification theory*. While the above identifying formula is quite simple, in general identifying formulas can get quite involved!

Example

- Assume ignorablity, and our table:

	$Y(1)$	$Y(0)$	$Y(1)-Y(0)$	A	Y^{obs}
	1.1	2.3	-1.2	1	1.1
	1.8	0.3	1.5	0	0.3
	2.0	2.1	-0.1	0	2.1
	0.1	1.3	-1.2	1	0.1
mean	1.25	1.5	-0.25		

- Then $\mathbb{E}[Y(1)] = \mathbb{E}[Y \mid A = 1]$, $\mathbb{E}[Y(0)] = \mathbb{E}[Y \mid A = 0]$.
- $\mathbb{E}[Y \mid A = 1] \approx 0.6$, $\mathbb{E}[Y \mid A = 0] \approx 1.2$, so

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \approx 0.6 - 1.2 = -0.6.$$

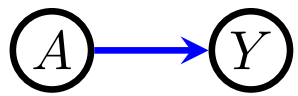
- Difference is called the average causal effect (ACE).
- May also want individual causal effect, e.g. $1.1 - 2.3 = -1.2$ for the first row.
- Important: individual causal effects are not identified (will return to this issue later). Most causal inference contends with population level parameters, such as the ACE.

Introducing: Causal Graphs

- Ignorability is not a realistic causal model, as treatment assignment is often *biased*.
- 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.
- For now: nodes are variables, $\rightarrow$ means "directly causes."
- Will make more precise as we go.
- Absences of nodes and edges are important.

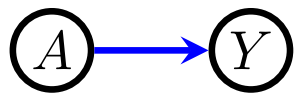
Introducing: Causal Graphs

- Ignorability is not a realistic causal model, as treatment assignment is often *biased*.
- 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.
- For now: nodes are variables, $\rightarrow$ means "directly causes."
- Will make more precise as we go.
- Absences of nodes and edges are important.
- Randomization example (one treatment A, one outcome Y).
- Observed situation:

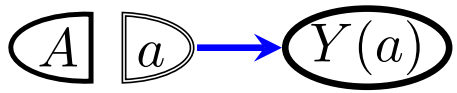


Introducing: Causal Graphs

- Ignorability is not a realistic causal model, as treatment assignment is often *biased*.
- 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.
- For now: nodes are variables, $\rightarrow$ means "directly causes."
- Will make more precise as we go.
- Absences of nodes and edges are important.
- Randomization example (one treatment A, one outcome Y).
- Observed situation:



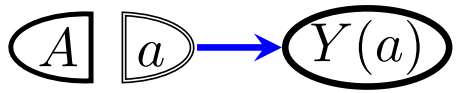
- Hypothetical situation (what if A were a):



- Think of hypothetical $A = a$ as setting a variable to a value in a debugger. Sometimes called an intervention.
- Differentiate between what variable does normally (A in the graph) and the hypothetical value under an intervention (a in the graph).

Introducing: Causal Graphs

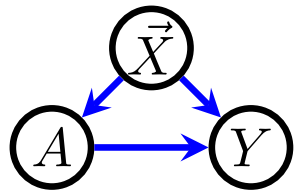
- Note: no path from A to $Y(a)$, which means $A \perp\!\!\!\perp Y(a)$.



- General method of constructing graphs like this, and reading independences is:
  - Split every variable we want to intervene on into a random and a fixed part.
  - Random parts inherit incoming edges. Fixed parts inherit outgoing edges.
  - Relabel any variable with a directed path from a fixed part by its value. Such a variable is now counterfactual.
- These are called Single World Intervention Graphs (SWIGs). Can read off independences in corresponding distribution by looking at paths.
- General method (d-separation criterion) for reading off independences via paths:
  - A d-separated from $Y(a)$ if all paths are "blocked."
  - A path is blocked if it has a blocking triplet.
  - Blocking triplets:
    - $V_1 \to V_2 \to V_3$ (V_2 conditioned on).
    - $V_1 \leftarrow V_2 \to V_3$ (V_2 conditioned on).
    - $V_1 \to V_2 \leftarrow V_3$ (neither V_2 nor any descendant of V_2 is conditioned on).

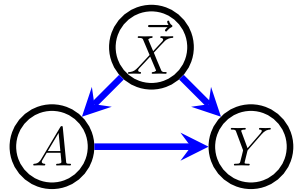
Dealing with Observed Confounders

- Observed situation (intuition – $\vec{X}$ are common causes of A and Y):

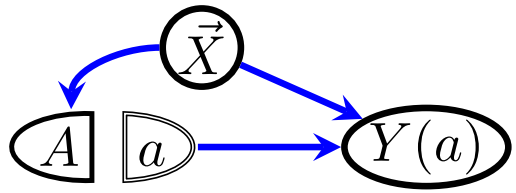


Dealing with Observed Confounders

- Observed situation (intuition – $\vec{X}$ are common causes of A and Y):

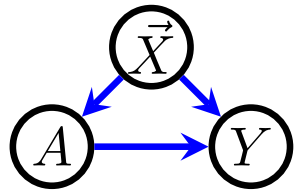


- Representing hypothetical $A = a$ as before:

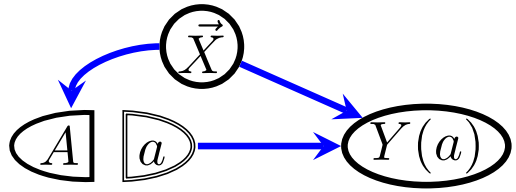


Dealing with Observed Confounders

- Observed situation (intuition – $\vec{X}$ are common causes of A and Y):



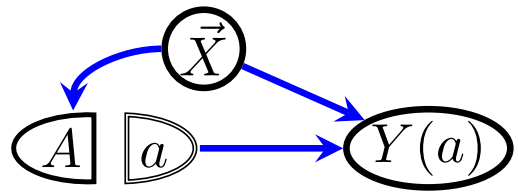
- Representing hypothetical $A = a$ as before:



- Note: A and $Y(a)$ connected via $\vec{X}$.
- Implies that $A \not\perp\!\!\!\perp Y(a)$, since A depends on $\vec{X}$ and $\vec{X}$ depends on $Y(a)$ (and A and $Y(a)$ are not common causes of $\vec{X}$).
- Association is not causation: $p(Y \mid A = a) \neq p(Y(a))$!
- Can we get $p(Y(a))$ in some other way?

Ignorability Conditional On A Confounder

- If the distribution $p(Y(a), A, \vec{X})$ is represented by the graph below, we can conclude that $A \perp\!\!\!\perp Y(a) \mid \vec{X}$ ($A \leftarrow \vec{X} \rightarrow Y(a)$ is a blocking triplet).
- In other words, if we know the value of $\vec{X}$, learning about $Y(a)$ gives us no additional information on how likely values of A are.

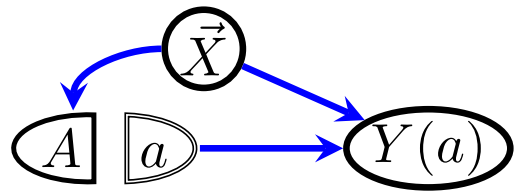


- This is called conditional ignorability.
- So, conditionally on $\vec{X}$, can repeat earlier reasoning:

$$p(Y(a) \mid \vec{X} = \vec{x}) = p(Y(a) \mid \vec{X} = \vec{x}, A = a) = p(Y \mid \vec{X}, A = a)$$

Ignorability Conditional On A Confounder

- If the distribution $p(Y(a), A, \vec{X})$ is represented by the graph below, we can conclude that $A \perp\!\!\!\perp Y(a) \mid \vec{X}$ ($A \leftarrow \vec{X} \rightarrow Y(a)$ is a blocking triplet).
- In other words, if we know the value of $\vec{X}$, learning about $Y(a)$ gives us no additional information on how likely values of A are.



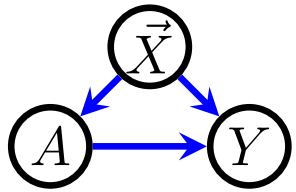
- This is called conditional ignorability.
- So, conditionally on $\vec{X}$, can repeat earlier reasoning:

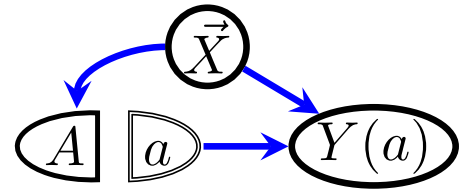
$$p(Y(a) \mid \vec{X} = \vec{x}) = p(Y(a) \mid \vec{X} = \vec{x}, A = a) = p(Y \mid \vec{X}, A = a)$$

- But what if we don't know value of $\vec{X}$?
- Can average across possible levels of $\vec{X}$ using prior probability $p(\vec{X} = \vec{x})$ of observing a particular $\vec{x}$.

Conditional Ignorability (Summary)

- Treatment A (usually binary, but not necessary).
- Outcome Y (discrete or continuous).
- A vector of baseline factors $\vec{X}$. Picture (observed and counterfactual):





- Predicting what will happen to Y if A were intervened on to value a:

$$p(Y(a)) = \sum_{\vec{x}} p(Y \mid A = a, \vec{X} = \vec{x}) p(\vec{X} = \vec{x}).$$

- Average causal effect (ACE):

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

- This is called stratification or adjustment formula.
- Old idea: 'adjusting for," "controlling for" refers to this.

Estimation of Identified Functionals

- Causal inference parameters such as the ACE are typically finite dimensional, but depend on potentially infinite dimensional *nuisance models*.
- The ACE (in the conditionally ignorable model) is a single number $\mathbb{E}[\mathbb{E}[Y|A=1,\vec{X}]-\mathbb{E}[Y|A=0,\vec{X}]]$ that depends on the nuisance function $\mathbb{E}[Y|A,\vec{X}]$ (the outcome model) or, potentially $p(A \mid \vec{X})$ (the treatment assignment model).
- Estimators for causal parameters (in the frequentist framework) often aim for parametric rates while using flexible nuisance models.
- This is done via the semi-parametric theory of influence functions, which yields semi-parametric efficient estimators in a wide class of estimators.
- For $\mathbb{E}[\mathbb{E}[Y|A=1,\vec{X}]-\mathbb{E}[Y|A=0,\vec{X}]]$ such an estimator is given by the celebrated *augmented inverse probability weighted (AIPW)* estimator.

Augmented IPW (Step By Step Guide)

Given n data points on $A, Y, \vec{X}$, and assuming conditional ignorablity and consistency:

1. Posit statistical model for $p(A \mid \vec{X}; \alpha_1), \mathbb{E}[Y \mid A, \vec{X}; \alpha_2]$.
2. Fit models in some way (MLE, minimization of some loss), yielding $\hat{\alpha}_1, \hat{\alpha}_2$.
3. Estimate ACE by:

$$\frac{1}{n}\sum_i \{Y_i - \mathbb{E}[Y \mid a_i = 1, \vec{x}_i; \hat{\alpha}_2]\}\frac{\mathbb{I}(a_i = 1)}{p(a = 1 \mid \vec{x}_i; \hat{\alpha}_1)} + \mathbb{E}[Y \mid a_i = 1, \vec{x}_i; \hat{\alpha}_2] -$$
$$\frac{1}{n}\sum_i \{Y_i - \mathbb{E}[Y \mid a_i = 0, \vec{x}_i; \hat{\alpha}_2]\}\frac{\mathbb{I}(A_i = 0)}{p(a = 0 \mid \vec{x}_i; \hat{\alpha}_1)} - \mathbb{E}[Y \mid a_i = 0, \vec{x}_i; \hat{\alpha}_2]$$

4. Report confidence intervals.

This estimator is doubly robust (remains consistent if one of the two nuisance models is misspecified) and has many other nice properties.

Mediation Analysis

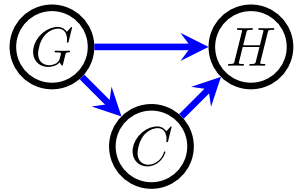
Motivating "Direct Effects": Discrimination

"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 (age, sex, religion, national origin etc.) and everything else had been the same."

In Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).

Motivating "Direct Effects": Discrimination

- G (gender), C (characteristics), H (hiring).



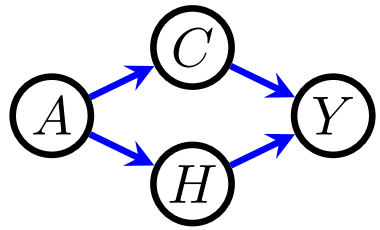
- Compare resumes of men:

$$H(G = male, C(G = male)) = H(G = male)$$

and *same* resumes with names switched to female ones:

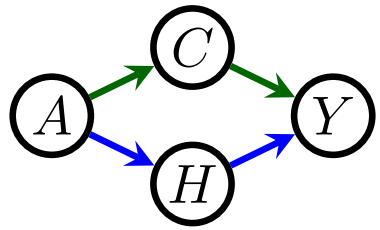
$$H(G = female, C(G = male)).$$

Motivating "Path-Specific Effects": Etiology by Pathway



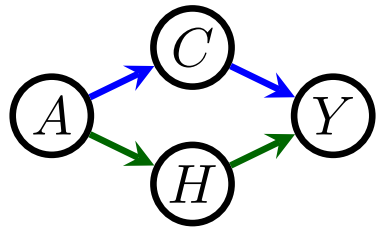
- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- Compare Y among non-smokers (or $Y(a')$) to:

Motivating "Path-Specific Effects": Etiology by Pathway



- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- Compare Y among non-smokers (or $Y(a')$) to:
- "Y if given nicotine-free cigarettes:" $Y(C(a), H(a'))$.

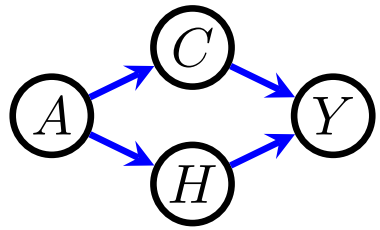
Motivating "Path-Specific Effects": Etiology by Pathway



- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- Compare Y among non-smokers (or $Y(a')$) to:

- "Y if given nicotine patches:" $Y(C(a'), H(a))$.

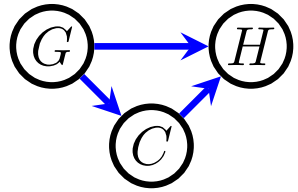
Motivating "Path-Specific Effects": Etiology by Pathway



- A (smoking), C (cancer), H (heart disease), Y (outcome).
- A affects Y via smoke (C pathway), and via nicotine (H pathway).
- Compare Y among non-smokers (or $Y(a')$) to:
- "Y if given nicotine-free cigarettes:" $Y(C(a), H(a'))$.
- "Y if given nicotine patches:" $Y(C(a'), H(a))$.

Defining Direct Effects

- Total effect: $\mathbb{E}[H(g=1)] - \mathbb{E}[H(g=0)]$:



- Say G : male (0) vs female (1) name on resume, H is hiring decision (1 is yes, 0 is no).
- What is a sensible question for discrimination?
- Compare hiring based on male resumes with same resumes but with names switched to female:

$$\mathbb{E}[H(g=0)] - \mathbb{E}[H(g=1, C(g=0))]$$

- Or compare female resumes with same resumes but with names switched to male:

$$\mathbb{E}[H(g=1)] - \mathbb{E}[H(g=0, C(g=1))]$$

- These are called natural direct effects.

Defining Indirect Effects

- Non-zero direct effect corresponds to discrimination in this setting.
- Can also define indirect effect similarly.
- Compare hiring based on a woman's resume with a man's name vs hiring based on a man's resume and a man's name:

$$\mathbb{E}[H(g=1, C(g=0))] - \mathbb{E}[H(g=1, C(g=1))].$$

- Or (with genders switched):

$$\mathbb{E}[H(g=0, C(g=1))] - \mathbb{E}[H(g=0, C(g=0))].$$

Defining Indirect Effects

- Non-zero direct effect corresponds to discrimination in this setting.
- Can also define indirect effect similarly.
- Compare hiring based on a woman's resume with a man's name vs hiring based on a man's resume and a man's name:

$$\mathbb{E}[H(g=1, C(g=0))] - \mathbb{E}[H(g=1, C(g=1))].$$

- Or (with genders switched):

$$\mathbb{E}[H(g=0, C(g=1))] - \mathbb{E}[H(g=0, C(g=0))].$$

- We get the following decomposition (same with flipped genders):

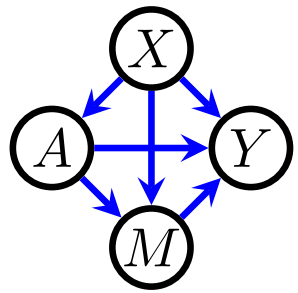
$$\underbrace{\mathbb{E}[H(1)] - \mathbb{E}[H(0)]}_{\text{ACE}} = \underbrace{(\mathbb{E}[H(1)] - \mathbb{E}[H(0, C(1))])}_{\text{Direct}} + \underbrace{(\mathbb{E}[H(0, C(1))] - \mathbb{E}[H(0)])}_{\text{Indirect}}$$

Nonparametric Effect Decomposition

- Defined direct and indirect effects and obtained a decomposition of the overall (total) causal effect.
- Did not mention statistical models (e.g. linear regressions) at all.
- Used potential outcomes directly.
- Can use any statistical model.
- But first, must make sure we are identified from observed data.

Simplest Interesting Mediation Setting

- $\vec{X}$ a vector of baseline factors/confounders (as before).
- A a treatment we are decomposing, M a mediator, Y an outcome.
- As before $\rightarrow$ means "directly causes."



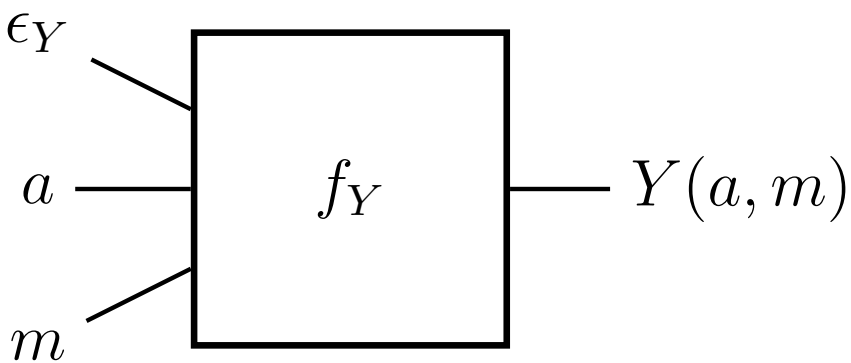
- To get direct and indirect effects, need to identify the following three distributions: $p(Y(1)), p(Y(0)), p(Y(1, M(0)))$.
- Need assumptions.

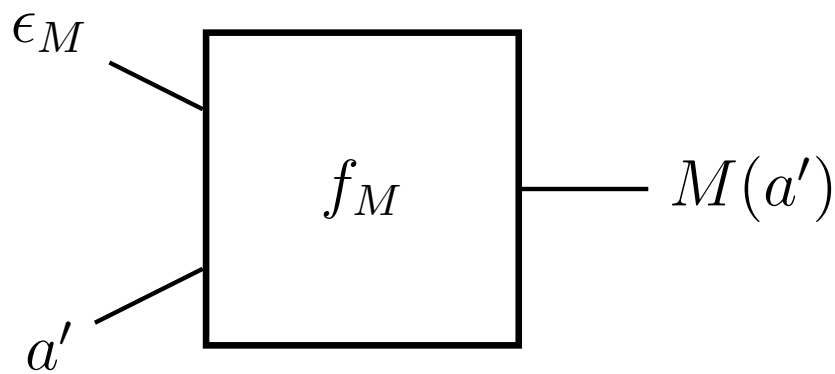
Identifying Assumptions (Causal Model)

- Identifying Assumptions:
  1. Conditional ignorability: $M(a) \perp\!\!\!\perp A \mid \vec{X}$ and $Y(a,m) \perp\!\!\!\perp \{A, M\} \mid \vec{X}$ for all a, m. This implies $Y(a) \perp\!\!\!\perp A \mid \vec{X}$. (meaning: conditioning on $\vec{X}$ suffices to deal with confounding between A, M and Y, and between A and M.)
  2. $M(a') \perp\!\!\!\perp Y(m,a) \mid \vec{X}$ for all a, a'. (meaning: within levels of $\vec{X}$, causal mechanisms for M and Y have independent sources of noise, even if treatments "mismatch.")

Identifying Assumptions (Causal Model)

- Identifying Assumptions:
  1. Conditional ignorability: $M(a) \perp\!\!\!\perp A \mid \vec{X}$ and $Y(a,m) \perp\!\!\!\perp \{A, M\} \mid \vec{X}$ for all a, m. This implies $Y(a) \perp\!\!\!\perp A \mid \vec{X}$. (meaning: conditioning on $\vec{X}$ suffices to deal with confounding between A, M and Y, and between A and M.)
  2. $M(a') \perp\!\!\!\perp Y(m,a) \mid \vec{X}$ for all a, a'. (meaning: within levels of $\vec{X}$, causal mechanisms for M and Y have independent sources of noise, even if treatments "mismatch.")
- Circuit analogy: $Y(a,m)$ and $M(a')$ are determined by causal mechanisms called *structural equations*, with observed directed causes and an exogenous disturbance input ϵ. These disturbance inputs are mutually independent regardless of how other inputs are set.





Identifying Functionals

- Since we have conditional ignorability for Y and A,

$$p(Y(a)) = \sum_{\vec{X}} p(Y \mid A = a, \vec{X}) p(\vec{X}).$$

- Tricky case is $p(Y(a, M(a')))$:

$$\begin{aligned}
p(Y(a, M(a'))) &= \sum_{m} p(Y(a, m), M(a') = m) \\
&= \sum_{m, \vec{X}} p(Y(a, m), M(a') = m \mid \vec{X}) p(\vec{X}) \\
&= \sum_{m, \vec{X}} p(Y(a, m) \mid \vec{X}) p(M(a') = m \mid \vec{X}) p(\vec{X}) \\
&= \sum_{m, \vec{X}} p(Y(a, m) \mid a, m, \vec{X}) p(M(a') = m \mid a', \vec{X}) p(\vec{X}) \\
&= \sum_{m, \vec{X}} p(Y \mid a, m, \vec{X}) p(M = m \mid a', \vec{X}) p(\vec{X})
\end{aligned}$$

- As before, can derive estimators based on semi-parametric theory.

A Triply Robust Estimator

- Model $p(A \mid \vec{X}; \eta_3)$, model $p(M \mid A, \vec{X}; \eta_2)$, model $\mathbb{E}[Y \mid A, M, \vec{X}; \eta_1]$.
- Fit by some reasonable method (MLE, minimizing some loss, etc.)
- Use empirical approximation for $p(\vec{X})$.
- Cleverly combine all three models, as follows:

$$\frac{1}{n}\sum_i \frac{\mathbb{I}(a_i = 1)p(m_i|A = 0, \vec{X}; \hat{\eta}_2)}{p(a_i|\vec{x}_i; \hat{\eta}_3)p(m_i|A = 1, \vec{x}_i; \hat{\eta}_2)}\{y_i - \mathbb{E}[Y|A = 1, m_i, \vec{x}_i; \hat{\eta}_1]\} +$$
$$\frac{\mathbb{I}(a_i = 0)}{p(a_i|\vec{x}_i; \hat{\eta}_3)}\{\mathbb{E}[Y|A = 1, m_i, \vec{x}_i; \hat{\eta}_1] - \int \mathbb{E}[Y|A = 1, m, \vec{x}_i; \hat{\eta}_1]p(m|A = 0, \vec{x}_i; \hat{\eta}_2)\}dm +$$
$$\int \mathbb{E}[Y|A = 1, m, \vec{x}_i; \hat{\eta}_1]p(m|A = 0, \vec{x}_i; \hat{\eta}_2)\}dm$$

Can use the bootstrap for confidence intervals under mild assumptions.

Causal Fairness Criteria

Why Causal Fairness Criteria? Past Convictions and Hiring

- Some US states forbid hiring discrimination based on prior conviction status.
- If given data on people with features $\vec{X}$, prior conviction A, hiring decision Y, how would we check for discrimination wrt A?
- Intuition: randomly assign people to (fake) serious records ($A = 1$), or light records ($A = 0$).
- Check average hiring rates in two groups in this hypothetical study: $\mathbb{E}^*[Y \mid A]$.
- In practice $p(A \mid \vec{X}) \neq p(A)$. Therefore, $\mathbb{E}[Y \mid A]$ (based on observed data) is very different from $\mathbb{E}^*[Y \mid A]$.
- Thus: fairness criteria based on associations of A and Y do not work.

Causal Versions of Associate Criteria

informal name	criterion
statistical parity	$\hat{Y} \perp\!\!\!\perp A$ in $p(\hat{Y}, A)$
equal opportunity	$\hat{Y} \perp\!\!\!\perp A \vert Y = 1$ in $p(\hat{Y}, Y, A)$
equalized odds	$\hat{Y} \perp\!\!\!\perp A \vert Y$ in $p(\hat{Y}, Y, A)$
equal accuracy	$p(\hat{Y} = Y \vert A) = p(\hat{Y} = Y)$
false positive rate (FPR) balance	$p(\hat{Y} = 1 \vert Y = 0, A) = p(\hat{Y} = 1 \vert Y = 0)$
false negative rate (FNR) balance	$p(\hat{Y} = 0 \vert Y = 1, A) = p(\hat{Y} = 0 \vert Y = 1)$
predictive parity	$p(Y = 1 \vert \hat{Y} = 1, A) = p(Y = 1 \vert \hat{Y} = 1)$

- Statistical criteria are formulated on $p(\hat{Y}, Y, A, C)$.
- May be re-defined in a straightforward way on a counterfactual distribution where A is manipulated to be drawn from some distribution $p^*(A)$ with no dependence on any covariate: $p(\hat{Y}, Y, C \vert \text{do}(a \sim p^*(A)))$.
- This addresses confounding if A is influenced by variables also influencing Y.

Work on Causal Versions of Fairness Criteria

A very large and active literature. A small sample of the work out there:

- FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes. Mishler, A. and Kennedy, E. `arxiv.org/abs/2109.00173`.
  An ensemble and semi-parametric based approach for navigating fairness/accuracy tradeoffs.
- Equality of Opportunity in Classification: A Causal Approach. Zhang, J. and Barenboim, E.

`https://proceedings.neurips.cc/paper/2018/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf`.

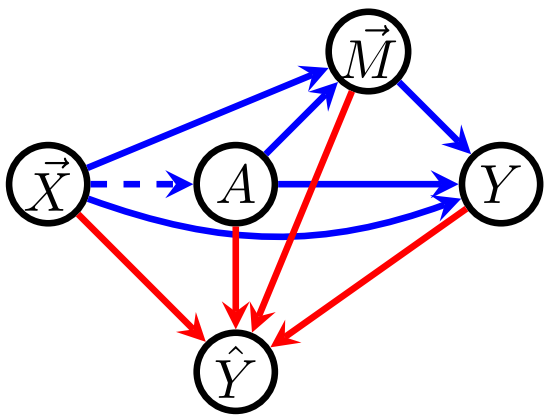
  An exploration of causal versions of equalized odds.
- Equal Opportunity and Affirmative Action via Counterfactual Predictions. Wang, Y., Sridhar, D. and Blei, D. `arxiv.org/pdf/1905.10870.pdf`.
  Training machine learning methods to satisfy causal versions of equalized odds (and related criteria).

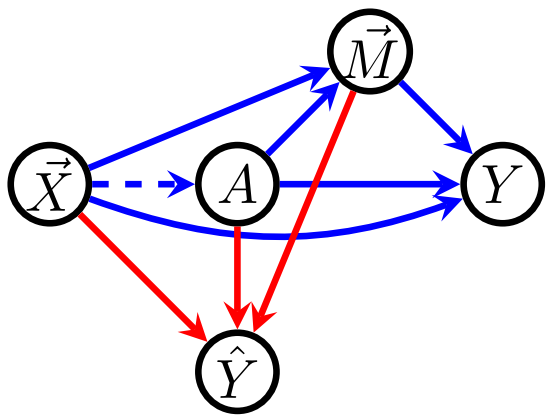
Common Features Of Discussed Causal Approaches

- A, a protected attribute or sensitive feature (race, sexual orientation, religion, etc.)
- $\vec{C}$, a set of covariates.
- Y, an outcome.
- Training a predictor $\hat{Y} \leftarrow f(\vec{C}; \beta)$ from data drawn i.i.d. from an observed data distribution $p(\vec{C}, A, Y)$.
- Either the distribution p, or the predictor f, or some combination is restricted in some way.
- Restrictions may be *population level*, or *individual level*.
- To formulate restrictions, a causal model is needed.
- Methods for creating predictors that satisfy criteria: "post-processing" (take a trained predictor, and modify), and "in-processing" (change the predictor training algorithm).

Formulating A Causal Model

- Further sub-divide covariates $\vec{C}$ into those causally prior to A ($\vec{X}$), and causal consequents of A ($\vec{M}$).
- Blue edges represent causal mechanisms in nature.
- Red edges represent a causal mechanism induced by training $\hat{Y} \leftarrow f(\vec{X}, A, \vec{M})$ trained on data drawn i.i.d. from $p(\vec{X}, A, \vec{M}, Y)$.
- Dashed edge from $\vec{X}$ to A may not exist: A may be "randomized by nature". Examples: biological sex.
- Graph on the left: $\hat{Y}$ is obtained via $f(\vec{X}, A, \vec{M})$ trained using observed data from the causal model represented by the graph.
- Graph on the right: $\hat{Y}$ is trained using some other data and is *held fixed*.





(Actual Cause) Counterfactual Fairness

(Kusner et al, 2017)

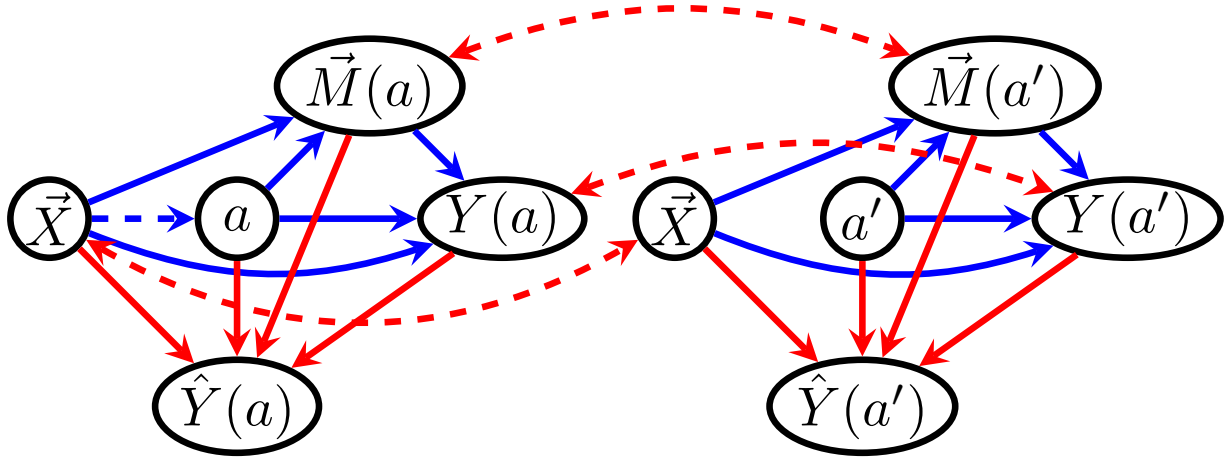
- A predictor for $\hat{Y}$ given a set of features C is (actual cause) counterfactually fair if, under any context $\vec{X} = \vec{x}$ and $A = a$:

$$p(\hat{Y}(a) = y | \vec{X} = \vec{x}, A = a) = p(\hat{Y}(a') | \vec{X} = \vec{x}, A = a).$$

- Note: the authors claims this to be an individual level criterion, but conditioning on $\vec{X}, A$ yields a population level criterion for the population with values $\vec{x}, a$. More on this issue later.
- An individual level criterion would equalize effects for every individual, e.g. $\hat{Y}_i(a) = \hat{Y}_i(a')$ for all i (and their values $\vec{x}_i, a_i$).
- Related to (conditional) effects of treatment on the treated in public health (with the protected feature serving as treatment).
- Justification based on the actual cause literature.
- See the Halpern/Pearl actual cause criterion (no time to discuss further).

(Actual Cause) Counterfactual Fairness

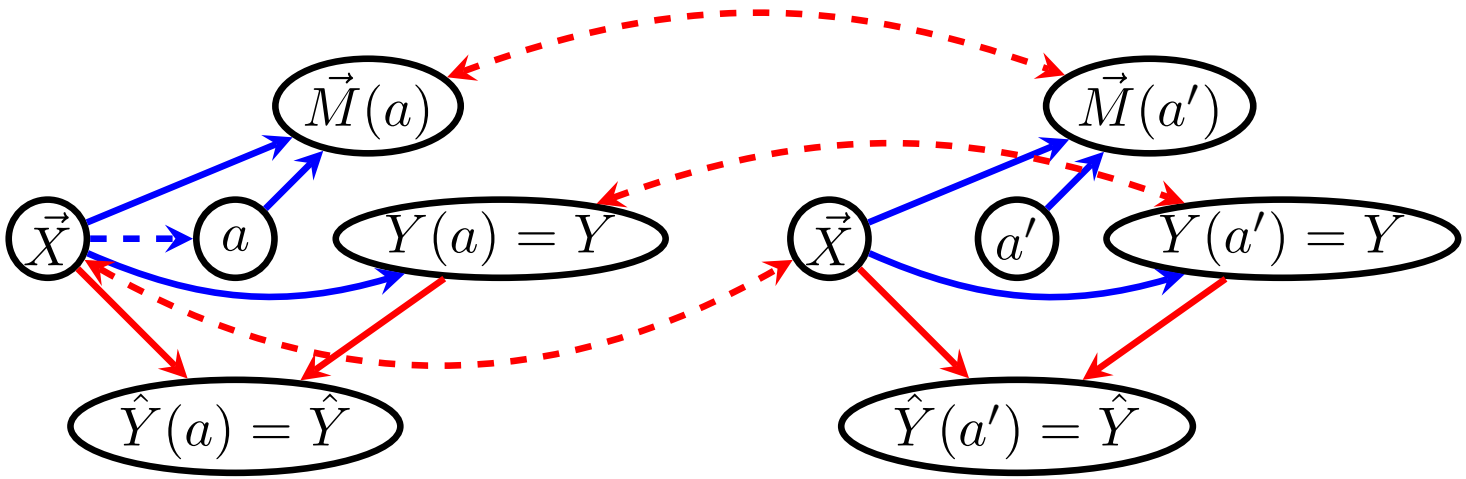
- Criterion involves two counterfactual situations simultaneously.
- Representable by *twin networks* (Balke and Pearl).



- Dashed edges represent association between different versions of counterfactuals, e.g. $\vec{M}(a)$ and $\vec{M}(a')$.
- Dependence because $M(a) \leftarrow f_M(a, \vec{X}, \epsilon_M)$ and $M(a') \leftarrow f_M(a', \vec{X}, \epsilon_M)$ are correlated since they share ϵ_M!
- Can read dependence and independence of cross-world counterfactuals via such graphs (using d-separation).

(Actual Cause) Counterfactual Fairness

- Simple sufficient criterion: a predictor that is (actual cause) "counterfactually-fair" is if it does not use causal descendants of the sensitive feature.
- Example:



$$p(\hat{Y}(a) = y | \vec{X} = \vec{x}, \vec{M}(a) = \vec{m}, a) = p(\hat{Y}(a') | \vec{X} = \vec{x}, \vec{M}(a') = \vec{m}, a) = p(\hat{Y} = y | \vec{X} = \vec{x}).$$

Issues With (Actual Cause) Counterfactual Fairness

- For many protected attributes (race, biological sex, sexual orientation) "causal non-descendants" probably excludes most variables predictive for Y.
- Can use causal consequents of A, provided above equality holds, due to "cancellation." Hard to verify.
- If individual level criterion fails to hold, difficult to see how to ensure it does.

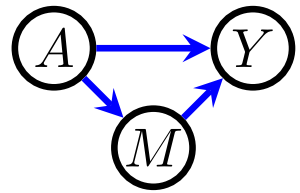
Fairness Based On Paths (Resolving Variables)

(Kilbertus et al, 2017)

- A variable V in a causal graph exhibits *unresolved discrimination* with respect to a protected attribute A and if:
  - There is a causal path from A to V.
  - V itself is non-resolving.
- "Resolving" roughly means "explaining a seemingly problematic relationship of the protected feature and a decision/outcome in a satisfactory way."
- In other words, all discrimination is "resolved" if decision/outcome is invariant to values of protected feature A via paths not through resolving variables.

Fairness Based On Paths (Resolving Variables)

- Example (Berkeley discrimination case): A is gender, Y is college admission decision, M is choice of department.



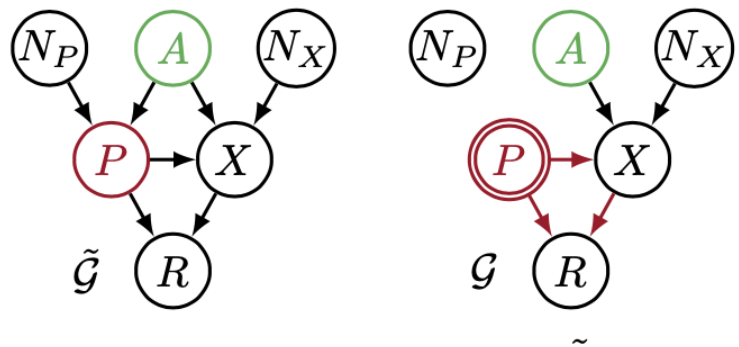
- No discrimination (?): $p(Y(a,m)) = p(Y(a',m))$ (or $p(Y|do(a,m)) = p(Y|do(a',m))$).
- Pessimistic view: by default we expect causal influence of a protected feature on outcome/decision to be a problem, unless there's a mitigating explanation.

Fairness Based On Paths (Proxies)

- A variable V in a causal graph exhibits *potential proxy discrimination* with respect to a protected attribute A and if:
  - There is a causal path from A to V blocked by a proxy variable.
  - V itself is not a proxy.
- Here "proxies" are possibly manipulable causal consequents of the protected feature which we may consider to be worrysome, if they cause the outcome/decision.
- Example: zip code in a segregated city like Baltimore is a proxy for race.
- This is an optimistic view: paths from the protected features to decision/outcome are fine, unless a worrysome proxy is involved.
- Discrimination with respect to a worrysome proxy may be checked by seeing of an overall effect on the outcome/decision exists.
- Thus, there is no proxy discrimination if $p(R(proxy)) = p(R(proxy'))$ (or $p(R|do(proxy)) = p(R|do(proxy'))$).

Algorithms Removing Proxy Discrimination

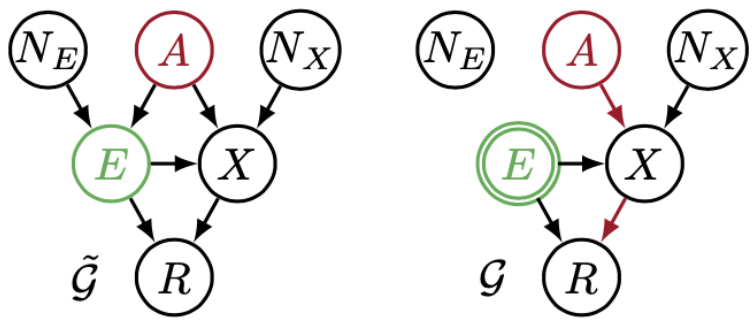
- Algorithm for satisfying proxy discrimination is as follows:
  - Intervene on the proxy directly (or identify $p(V \setminus P|do(p))$ from observed data).
  - Impose constraints on $p(V \setminus P|do(p))$ such that $p(R|do(p)) = p(R|do(p'))$.



- The authors implement this for linear models.

Algorithms Removing Resolving Discrimination

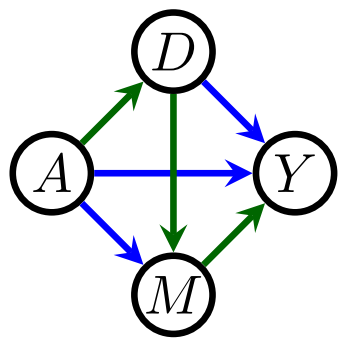
- Algorithm for satisfying resolving discrimination is as follows:
  - Intervene on the resolving E directly (or identify $p(V \setminus E|do(e))$ from observed data).
  - Impose constraints on $p(V \setminus E|do(e))$ such that $p(R|do(e,a)) = p(R|do(e,a'))$.



- The authors implement this for linear models.

Issue With Variable-Based Causal Path Definitions

- If we worry about the causal influence of a protected feature A on outcome/decision Y along paths, then we may end up in a situation where a worrysome path and a path we wish to allow both pass through the same variable.
- Example: A is ethnic group, D is growing up in a particular country, Y is hiring decision, M is a prior opportunity that looks good on a resume more likely to be available in the country in D.
- Discrimination via the path $A \to D \to Y$ is an issue as it is either about race or place of origin.
- Discrimination via the paths $A \to D \to M \to Y$ or $A \to M \to Y$ are possibly fine as they are about important lines on the resume.
- Can't think of D as a resolving variable or a proxy of A in the above sense.
- Path-specific effects provide a way of addressing this issue.



Fairness Via Path-Specific Effects

- Will consider a sensitive feature A (race, gender, etc.) and an outcome Y.
- Approach inspired by causal inference.
- Causal inference: move from factual to counterfactual worlds.

(1) Fair inference: move from "unfair" to "fair" worlds.

(2) Discrimination wrt A for Y is the presence of an effect of A on Y along "unfair causal paths."

(3) "Unfair causal paths" are a domain-specific issue.

- Fair world: counterfactual world closest to ours with no impermissible path-specific effect.

Our Approach (More Formally)

- Observed distribution $p(\vec{X}, A, \vec{M}, Y)$.
- $\vec{X}$ pre-treatment features, $\vec{M}$ post-treatment, pre-outcome features.
- Want to predict Y from $\vec{X}, A, \vec{M}$ in a fair way.
- Consider all causal paths from A to Y.
- Fix "bad ones." For now, only the direct path $A \to Y$ (as in hiring) is bad.
- If PDE is not zero, $p(\vec{X}, A, \vec{M}, Y)$ is "unfair" (in the hypothetical name-switching experiment sense).
- Want to find a "fair world" p^* close to p where PDE is zero.
- Natural choice: close in Kullback-Leibner (KL) divergence sense.

Our Approach (Parametric Finite Sample Version)

- Data from $p(\vec{X}, A, \vec{M}, Y)$, estimator $g(p)$ for PDE.
- Since p^* is KL-close to p, solve the following for $\hat{\alpha}$:

$$\begin{aligned} \arg\max_{\alpha} \quad & \mathcal{L}(p(\vec{X}, A, \vec{M}, Y); \alpha) \\ \text{such that} \quad & \epsilon_l \leq g(p) \leq \epsilon_u \end{aligned}$$

- In other words, maximize the likelihood subject to the fairness constraint.
- Parts of p we constrain depends on $g(p)$.
- Important: cannot classify new instances $(\vec{x}, a, \vec{m})$ using $\mathbb{E}[Y \mid \vec{x}, a, \vec{m}; \hat{\alpha}]$!
- Because new instances drawn from p, which is unfair.
- If $p^*(\vec{X}, A, \vec{M}, Y) = p(\vec{X})p^*(A, \vec{M}, Y \mid \vec{X})$, use $\mathbb{E}[Y \mid \vec{X}; \hat{\alpha}]$.
- Selectively "forget" some known info – entire problem must lie in p^*.
- Razieh Nabi will discuss the approach in more detail.

Bibliography (Fairness)

- "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`

Satisfying Causal Fairness Criteria: Algorithms and Open Problems

Part III. Removing unfair biases

Daniel Malinsky, Razieh Nabi, Ilya Shpitser

Joint Statistical Meeting

August 7, 2022

Two types of questions

- Let $O = (X, S, Y) \sim P \in \mathcal{M}$
- Assume $\Psi: \ P \in \mathcal{M} \mapsto \Psi(P) \in \mathbf{\Psi}$, 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:
    $$\Psi(P) = \arg\max_{f_a \in \mathcal{F}} \ \mathbb{E}[Y(f_a)], \text{ where } f_a : \mathcal{H} \mapsto a \in \mathcal{A}$$

Two types of questions

- Let $O = (X, S, Y) \sim P \in \mathcal{M}$
- Assume $\Psi: \ P \in \mathcal{M} \mapsto \Psi(P) \in \mathbf{\Psi}$, 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

     $$\Theta_\Psi : P \in \mathcal{M} \mapsto \Theta_\Psi(P) \in \mathbf{\Theta}$$

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

Outline

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)

1. Predictive decision support

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

Given a pre-specified $\Theta_\Psi(P)$, either

a. Pre-process the observed data O, or
b. Post-process the statistical output $\Psi(P)$, or
c. Re-train $\Psi(P)$ subject to fairness constraints,

in order to mitigate $\Theta_\Psi(P)$ from $\Psi(P)$.

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

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)

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

Find data representation Z by:
- Maximizing mutual information between $\{X, Z\}$, and
- Minimizing mutual information between $\{S, Z\}$.

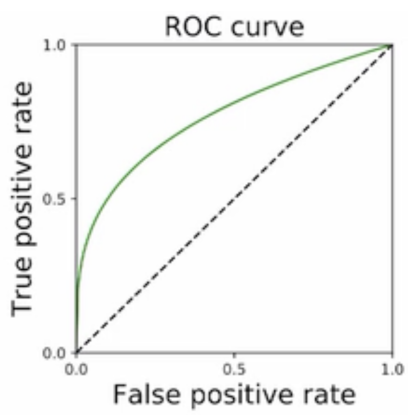
- Train $\Psi^*(P) = P(Y = 1 \mid Z)$

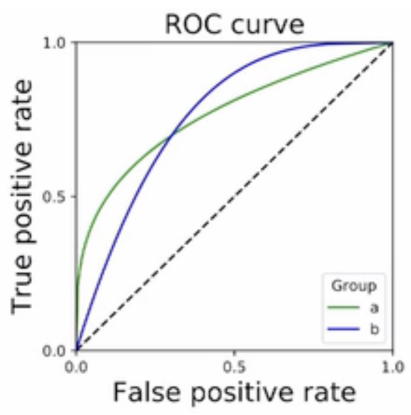
Post-process $\Psi(P)$: 1st example

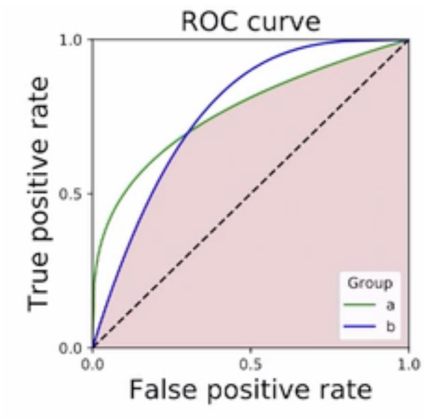
- Let $\Psi(P) = P(Y = 1 \mid X, S)$
- **Constraint** $\Theta_\Psi(P)$: $\widehat{Y} \perp\!\!\!\perp S \mid Y$
  - “Separation" (Hardt et al.; 2016)

Post-process $\Psi(P)$: 1st example

- Let $\Psi(P) = P(Y = 1 \mid X, S)$
- **Constraint** $\Theta_\Psi(P)$: $\widehat{Y} \perp\!\!\!\perp S \mid Y$
  - "Separation" (Hardt et al.; 2016)
- **Approach**: correction of $\widehat{Y}$ when $\Psi(P)$ is fixed (Hardt et al.; 2016)
  - Plot TPR and FPR for all possible thresholds that yield $\widehat{Y}$
  - Given a cost function between TPR and FPR: calculate the optimal threshold







Post-process $\Psi(P)$: 2nd example

- 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:

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

Post-process $\Psi(P)$: 2nd example

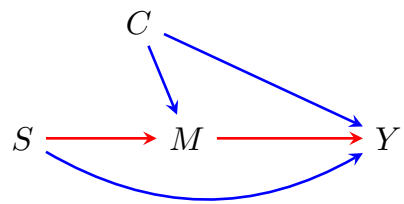
- Let $O = \{C, S, M, Y\}$ and $\Psi(P) = \mathbb{E}_P[Y \mid C, S, M]$
- **Constraint**: unfair path-specific effects (N, S; 2018)

  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
- The idea of constrained learning is very general
- Approaches on how to solve the optimization problem may vary
- We go over a few examples. Let
  - $\Psi(P) = P(Y = 1 \mid X, S)$,
  - $\widehat{Y} = 1$ if $\Psi(P) > \delta$, and 0 otherwise (dependent on Ψ, δ).

Constrained learning: 1st example

Constraint: separation (Hardt et al.; 2016)

$$\Theta_\Psi(P): \ \widehat{Y} \perp\!\!\!\perp S \mid Y$$

- Fix a function class $\boldsymbol{\Psi}$ and a loss function $\mathcal{L}$:

$$\Psi^*(P) = \underset{\Psi \in \boldsymbol{\Psi}}{\operatorname{argmin}} \ \mathbb{E}\left[\mathcal{L}(\Psi(P), Y)\right] \quad \text{s.t.} \ \ \widehat{Y}_\Psi \perp\!\!\!\perp S \mid Y$$

Highly intractable and thus a hard optimization problem

Constrained learning: 1st example

Constraint: separation (Hardt et al.; 2016)

$$\Theta_\Psi(P): \ \widehat{Y} \perp\!\!\!\perp S \mid Y$$

- Fix a function class $\boldsymbol{\Psi}$ and a loss function $\mathcal{L}$:

$$\Psi^*(P) = \underset{\Psi \in \boldsymbol{\Psi}}{\operatorname{argmin}} \ \mathbb{E}\left[\mathcal{L}(\Psi(P), Y)\right] \quad \text{s.t.} \ \ \widehat{Y}_\Psi \perp\!\!\!\perp S \mid Y$$

  Highly intractable and thus a hard optimization problem

Approach: Woodworth et al. (2017) assume data is Gaussian

$$\Theta_\Psi(P): \ \sigma_{\widehat{Y}S} \times \sigma_{YY} = \sigma_{\widehat{Y}Y} \times \sigma_{YS},$$

where $\sigma_{UV} = \mathbb{E}[(U - \mathbb{E}[U]) \times (V - \mathbb{E}[V])]$.

Constrained learning: 2nd example

Constraint: counterfactual fairness (Kusner et al.; 2017)

$$\Theta_\Psi(P) \coloneqq P\big[\widehat{Y}(s) = y \;\big|\; X, S = s\big] - P\big[\widehat{Y}(s') = y \;\big|\; X, S = s\big]$$

Constrained learning: 2nd example

Constraint: counterfactual fairness (Kusner et al.; 2017)

$$\Theta_\Psi(P) \coloneqq P\big[\widehat{Y}(s) = y \;\big|\; X, S = s\big] - P\big[\widehat{Y}(s') = y \;\big|\; X, S = s\big]$$

Approach: Kusner et al. (2017) suggest to learn $\Psi^*(P)$ as follows:

1. Define and fit a causal model over the triplet (U, O, F)
   - U: latent variables
   - O: observed variables
   - F: a set of functions such that for each $O_i \in O$, $O_i = f_i(\mathrm{pa}(O_i), U_{\mathrm{pa}(O_i)})$, a.k.a. structural equations
2. Obtain $\Psi^*(P)$ by solving the following optimization problem:

$$\Psi^*(P) = \underset{\Psi \in \mathbf{\Psi}}{\operatorname{argmin}} \; \mathbb{E}\left[\mathcal{L}(\Psi_{U, X \not\succ S}(P), Y) \mid X, S\right],$$

where $\Psi_{U, X \not\succ S}(P)$ denotes a classifier that only uses U and $X \not\succ S$ (i.e., features that are non-descendants of S)

Counterfactual fairness ctd.

- Kusner et al. (2017) suggest to estimate unobserved factors U
- Requiring assumptions on the latent variables and structural equations
  - Mainly that U factors are mutually independent
- Assumptions are hard and often impossible to verify in practice
- Follow up work on sensitivity analysis to the no-unmeasured confounding assumption
- 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^*

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 $\Psi(P_\alpha)$

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 $\widehat{\Theta}_n(P_\alpha)$

#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 $P(O)$:

#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: $P(M, Y \mid X, S)$

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

#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: $P(S, M, Y \mid X)$

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

#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

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

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

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

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.

“Optimal" fair world

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

“Optimal" fair world

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

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

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

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

“Optimal" fair world

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

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

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

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^*

#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:

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

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

#3 Complex optimization problem

- Typical learning problem of the form:

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

- 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

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)

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

2. Sequential Decision Making

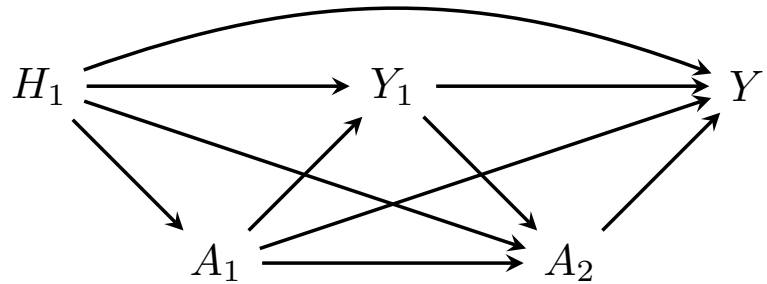
More complex statistical targets

Example: **Sequential decision making**

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

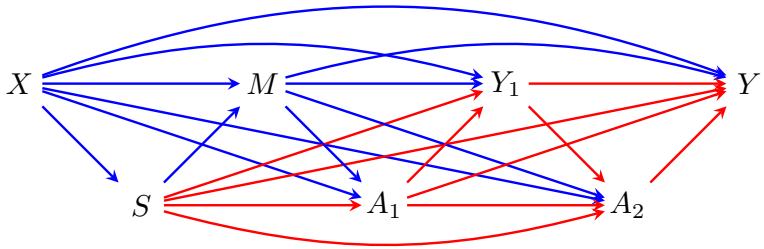
Policy: $f_A = \{f_{A_1}, f_{A_2}\}$

(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



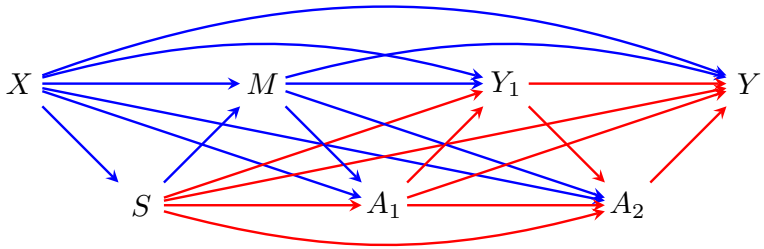
- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

  Example: unfair paths from S to Y:

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

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

Sources of bias in policy learning

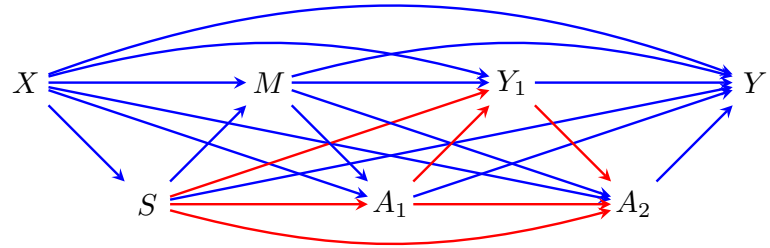


- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

  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:

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

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

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:

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

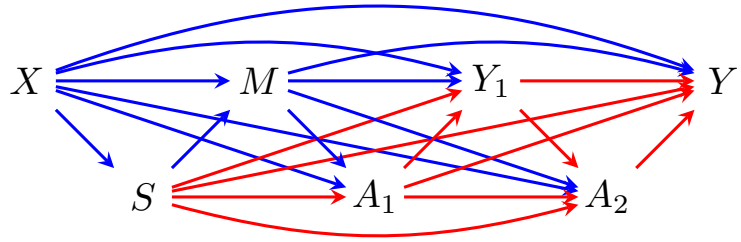
$$\text{such that} \quad \epsilon_j^- \leq \widehat{\Theta}_{j,n}(P_\alpha) \leq \epsilon_j^+, \; j = 1, \ldots, J.$$

Example: a two-stage decision point

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

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

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

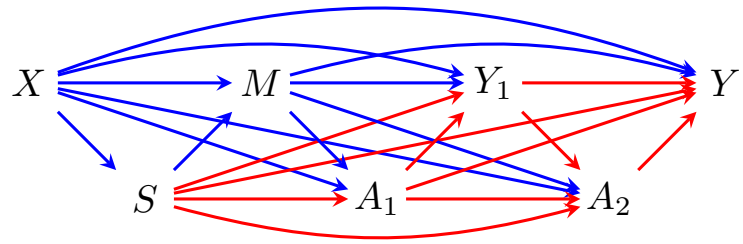


Example: a two-stage decision point

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}:

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

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

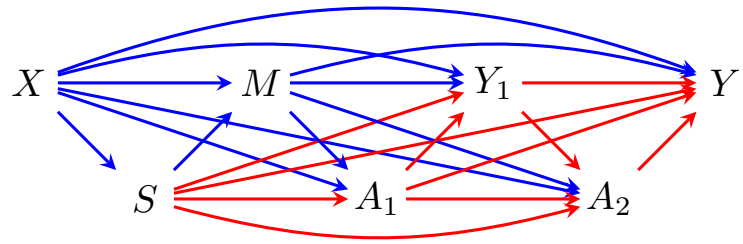
$$\widehat{\Theta}_{sa_2}(P) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{P(S_n|X_n)}\,\frac{P(M_n|s',X_n)}{P(M_n|s,X_n)} - \frac{\mathbb{I}(S_n = s')}{P(S_n|X_n)}\right\}A_{2n}.$$

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 $\text{PSE}^{sy} = \text{PSE}^{sa_1} = \text{PSE}^{sa_2} = 0$

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 $P(O)$

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

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

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

3. Data application

- https://github.com/raziehna/fair-inference-on-outcomes
- https://github.com/raziehna/learning-optimal-fair-policies

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

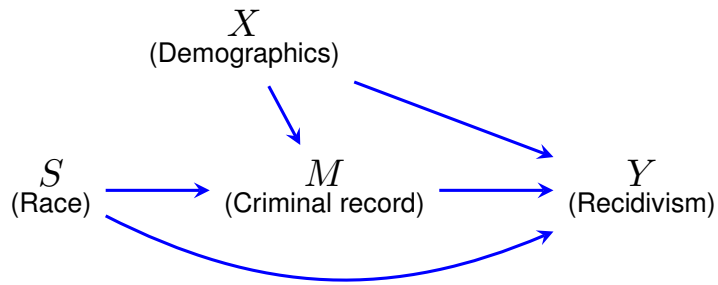
- Race: African Americans (60%) and Caucasians (40%)
- Demographics: sex and age
- Criminal record: binary indicator of crime counts $>$ one
- Recidivism: binary indicator
- COMPAS scores

Task: prediction

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

Unfair mechanisms:

$S \rightarrow Y$

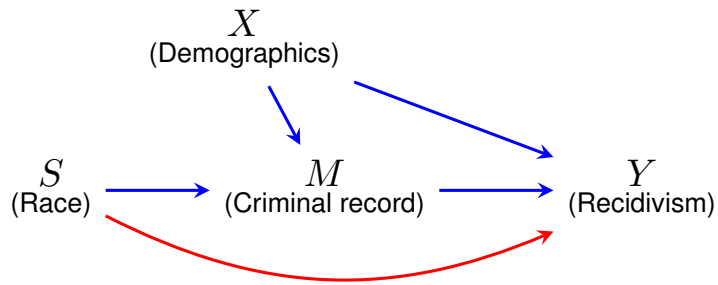


Task: prediction

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

Unfair mechanisms:

$S \rightarrow Y$



1. Fairness notion $\Theta(P)$

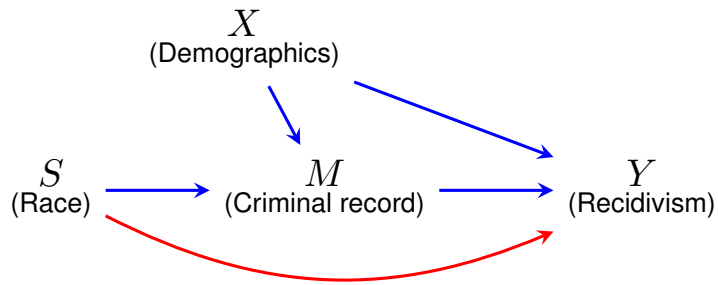
- $\Theta(P)$: direct effect of S on Y
- Let {Blacks: $S = 1$} and {Whites: $S = 0$}
  - $\mathbb{E}[Y(1, M(0))]$: risk of recidivism had individuals been Black and everything else had been as if they were White
  - $\mathbb{E}[Y(0)]$: risk of recidivism had individuals been White
- Let $\Theta(P)$ be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

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

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

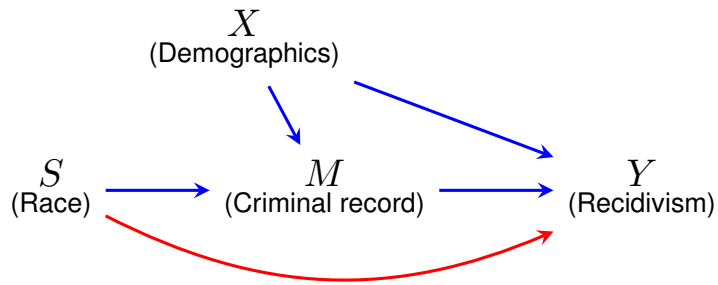
- Use BART to fit the outcome model $\mathbb{E}[Y \mid S, X, M]$
- $\mathbb{E}[Y(1, M(0))] = 0.47, \quad \mathbb{E}[Y(0)] = 0.40$
- $\Theta(P) = 1.3\ (1.01, 1.45)$ (odds ratio scale)

Task: prediction

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

Unfair mechanisms:

$$S \rightarrow Y$$

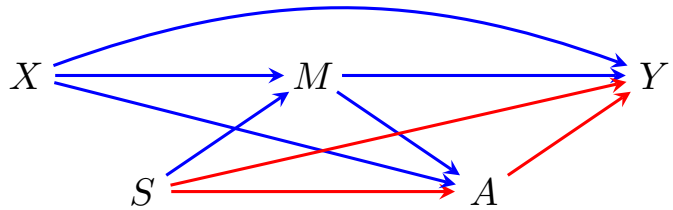


3. Remove $\Theta(P)$ from $\Psi(P) = P(Y = 1 \mid S, X, M)$

	$\Theta(P)$ (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

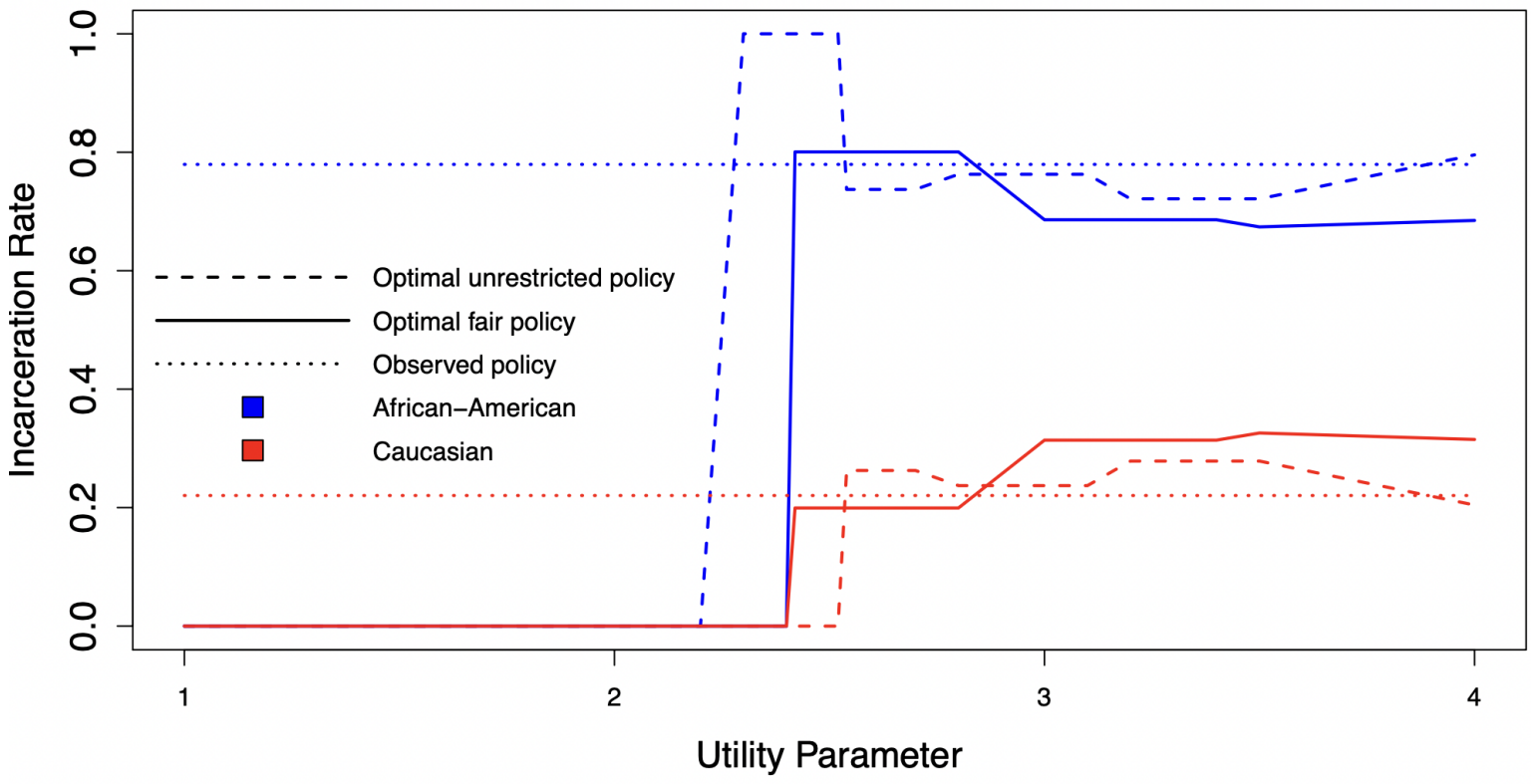
- BART and constrained MCMC to obtain the fair distribution
- Less than 2% relative change in out of sample performance

Task: automated decision making



- S: race, X: other demographics, M: prior convictions
- A: incarceration (based on risk of recidivism)
  - Assumed score ≥ 7 were incarcerated;
  - This accounts for 28.9% of individuals
- 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 $R = 0$)

Task: automated decision making



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.

Task: automated decision making

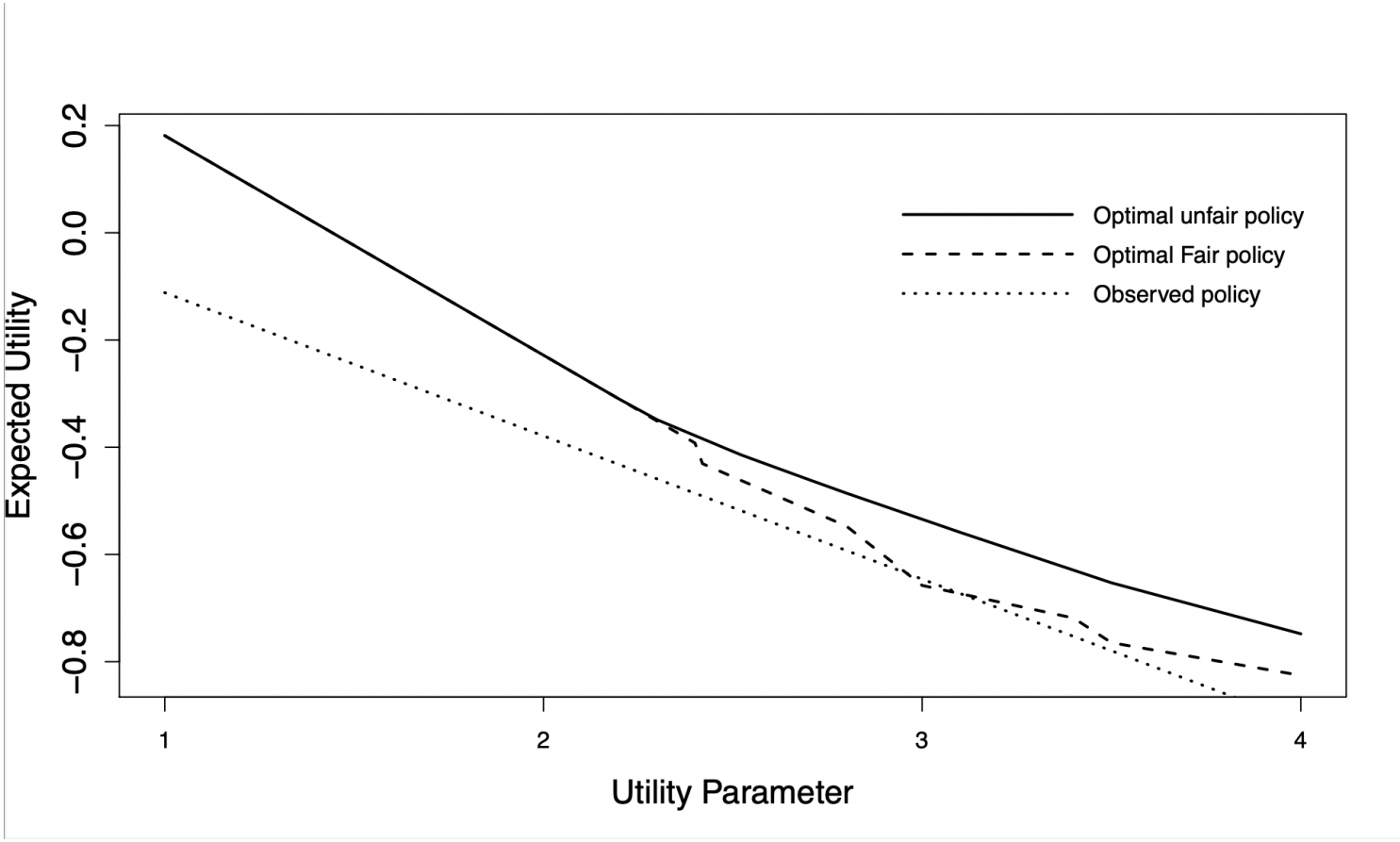


Figure: Relative expected utilities for policies as function of θ

Up next: **concluding remarks**

References

1. 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.
2. Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).
3. Woodworth, Blake, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. "Learning non-discriminatory predictors." In Conference on Learning Theory, pp. 1920-1953. PMLR, 2017.
4. 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.
5. 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.
6. Art Owen. Empirical Likelihood. Chapman & Hall, 2001.
7. Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." Advances in neural information processing systems 30 (2017).
8. Chiappa, Silvia. "Path-specific counterfactual fairness." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7801-7808. 2019.
9. Nabi, Razieh, and Ilya Shpitser. "Fair inference on outcomes." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1. 2018.
10. Nabi, R., Malinsky, D. and Shpitser, I., 2019, May. Learning optimal fair policies. In International Conference on Machine Learning (pp. 4674-4682). PMLR.
11. 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.

Satisfying Causal Fairness Criteria: Algorithms and Open Problems

Concluding Remarks.

Daniel Malinsky, Razieh Nabi, Ilya Shpitser

Joint Statistical Meeting

August 5, 2022

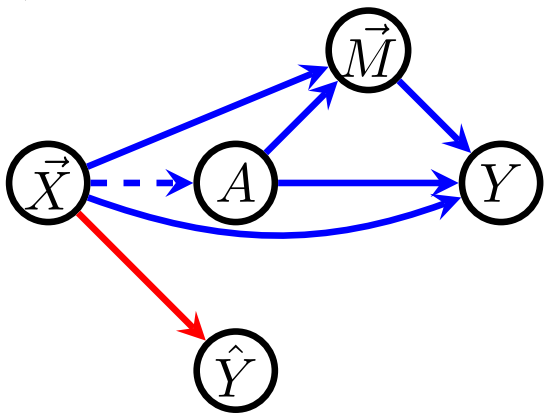
Overview

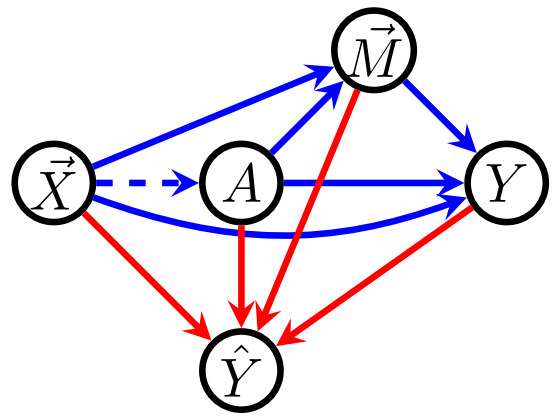
- True versus predicted outcomes.
- Individual versus group level criteria.
- Structural equations versus regression noise terms.
- Decision support: should actions be used as outcome or features?
- Should we use race as a predictor in diagnosis?
- The landscape of fairness/conclusions.

True Versus Predicted Outcomes

(I am grateful to Alan Mishler for first pointing this out to me.)

- Some discussed criteria mention the true outcome/decision Y, and others the predicted outcome/decision $\hat{Y}$.
- An important point: a fair criterion on $\hat{Y}$ will not yield a fair criterion on Y – and vice versa.
- Graph on the left shows a trivially fair (but poorly performing) classifier for $\hat{Y}$. This classifier remains fair regardless of what constraint is imposed on $p(X, A, M, Y)$. Thus: fair classifier $\not\Rightarrow$ fair world.
- Graph on the right shows the usual relationship of $\hat{Y}$ and X, A, M, Y. Constructing a criterion on Y (say that some $g(X, A, M, Y) = 0$) will not necessarily imply that $g(X, A, M, \hat{Y}) = 0$, as $\hat{Y}$ will not in general behave as Y. Thus: fair world $\not\Rightarrow$ fair classifier.



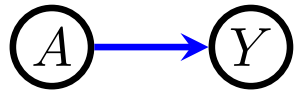


Individual Versus Group Level Fairness

- 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.

Structural Equations Versus Regressions

- Consider a simple graph:



- And the structural equations:

$$A \leftarrow f_A(\epsilon_A)$$
$$Y \leftarrow f_Y(a, \epsilon_Y).$$

- Have access to data from $p(Y, A) = p(Y|A)p(A)$.
- We may choose to model $p(Y|A)$ via $g_Y(A, \delta_Y)$.
- Important: there is no particular reason to suspect $g_Y = f_Y$ and $\delta_Y = \epsilon_Y$!
- In fact, f_Y and ϵ_Y are not identified from observed data!
- g_Y is an "epistemological object" (a statistical model we choose).
- f_Y is an "ontological object" (a part reality pertaining to the causal mechanism for Y).
- Be very careful when you see claims of algorithms that can learn f_Y or ϵ_Y: almost always this is talking about some particular choice g_Y and δ_Y.

Decision Support: Predictors Versus Policies

Two approaches to decision support (given covariates $\vec{X}$, action A, outcome Y).

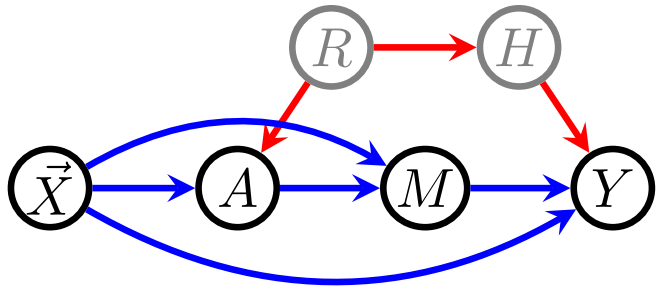
1. Learn a predictor $f(a, \vec{x}) = \mathbb{E}[Y|a, \vec{x}]$, choose A to maximize $f(., \vec{x})$ for a patient with features $\vec{x}$.
2. Learn a (counterfactual) predictor $\tilde{f}(a, \vec{x}) = \mathbb{E}[Y(a)|\vec{x}]$, choose A to maximize $\tilde{f}(., \vec{x})$ for a patient with features $\vec{x}$

- Not the same in general!
- Hypothetical example:

  Doctors in a particular hospital decide on one of two courses of treatment $A \in \{0, 1\}$ for patients with covariates $\vec{X}$, which exclude race R. Treatment effect on the outcome Y is mediated by a mechanism measurable by a biomarker M. Race is associated with a feature H which affects the outcomes Y. As it turns out, treatment decisions were dependent on the patient's race in inappropriate ways. How should retrospective data on a set of patients, past treatment decisions, and recorded outcomes be used to inform future decisions?

Decision Support: Predictors Versus Policies (cont.)

- Example as a causal diagram:



- Identification theory in causal inference (skipping details) tells us that:

$$\mathbb{E}[Y(a)|\vec{x}] = \sum_m \left(\sum_{a'} \mathbb{E}[Y \mid m, a', \vec{x}] p(a' \mid \vec{x}) \right) p(m \mid a, \vec{x}) \neq \mathbb{E}[Y|a, \vec{x}].$$

- Using $\mathbb{E}[Y|a, \vec{x}]$ captures an inappropriate association of A and Y via R and H, whereas treatment decisions should be made exclusively based on the causal relationship of A and Y.

Under Representation And Disparate Outcomes

- Lively debate in the literature on disentangling contributions to unequal health and life outcomes in minority groups (Geiger, 2003), many others.
- Determinants of health (not an exclusive list):
  - social, physical, economic environment
  - social norms, societal stratification, structural racism
  - lifestyle choices
  - health correlates
  - healthcare system bias
- "Race" typically has "simple" coding in biomedical datasets, but in reality corresponds to a very complicated vector (self-identity, perception, social context, ancestry, etc.)
- Very complicated statistical methodology issues are involved – looking at variable associations and predictive modeling is not sufficient!

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!.

Thank you for listening!

Contact info:

Razieh Nabi	razieh.nabi@emory.edu
Daniel Malinsky	dsm2128@cumc.columbia.edu
Ilya Shpitser	ilyas@cs.jhu.edu

Bibliography (Race and Diagnosis)

- H. Jack Geiger. "Racial and ethnic disparities in diagnosis and treatment: a review of the evidence and a consideration of causes." In Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care, Washington (DC): National Academies Press (US); 2003.
- JW Jackson, TJ VanderWeele. "Decomposition analysis to identify intervention targets for reducing disparities." Epidemiology (Cambridge, Mass.) 29 (6), 825.
- Lesley A. Inker, M.D., Nwamaka D. Eneanya, M.D., M.P.H., Josef Coresh, M.D., Ph.D., Hocine Tighiouart, M.S., Dan Wang, M.S., Yingying Sang, M.S., Deidra C. Crews, M.D., Alessandro Doria, M.D., Ph.D., M.P.H., Michelle M. Estrella, M.D., M.H.S., Marc Froissart, M.D., Ph.D., Morgan E. Grams, M.D., M.H.S., Ph.D., Tom Greene, Ph.D., et al. "New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race." N Engl J Med 2021; 385:1737-1749.
- James A. Diao, B.S., Lesley A. Inker, M.D., Andrew S. Levey, M.D., Hocine Tighiouart, M.S., Neil R. Powe, M.D., M.P.H., M.B.A., and Arjun K. Manrai, Ph.D. "In Search of a Better Equation – Performance and Equity in Estimates of Kidney Function." N Engl J Med 2021; 384:396-399.
- Chi-yuan Hsu, M.D., Wei Yang, Ph.D., Rishi V. Parikh, M.P.H., Amanda H. Anderson, Ph.D., Teresa K. Chen, M.D., Debbie L. Cohen, M.D., Jiang He, M.D., Ph.D., Madhumita J. Mohanty, M.D., James P. Lash, M.D., Katherine T. Mills, Ph.D., Anthony N. Muiru, M.D., Afshin Parsa, M.D., M.P.H., et al. "Race, Genetic Ancestry, and Estimating Kidney Function in CKD." N Engl J Med 2021; 385:1750-1760.