Modeling Risk and Achieving Algorithmic Fairness Using Potential Outcomes

Alan Mishler, Edward H. Kennedy, Amanda Coston, Alexandra Chouldechova

Carnegie Mellon University

Predictive Algorithms

Widely used in
- Criminal Justice (pretrial release, sentencing, parole decisions)
- Healthcare (choosing among treatment options)
- Consumer finance (issuing loans)

Typically designed to predict *observable* outcomes:
- Recidivism
- Health outcomes
- Default on a loan

Fairness also often defined in terms of observables
- Error rates (False Positives, False Negatives)
- Calibration, predictive parity
- Equalized odds, equal opportunity

Problem: Observable outcomes confound *risk* and the *effect of interventions*.
$\Longrightarrow$ Limited use to decision-makers.
$\Longrightarrow$ Hard to evaluate performance or fairness.

Solution: Use **potential outcomes** under various intervention(s) instead.
$\Longrightarrow$ More useful information for decision-makers.
$\Longrightarrow$ More sensible definitions of fairness.

Observable and potential outcomes

Notation

Observable variables

A = Exposure (e.g. incarceration)
Y = Outcome (e.g. recidivism)
R = Race (b = black, w = white)
$\mathbf{X}$ = Other covariates
$S = \hat{Y}$ = Predictor of outcome Y

Potential outcomes

$Y^{A=a}$ = Outcome under treatment a

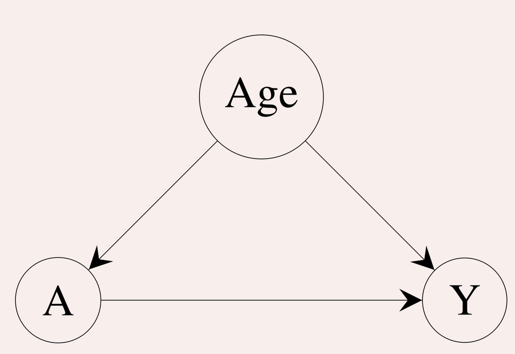
Assumption:
$Y = \sum_a Y^a \mathbb{1}\{A = a\}$.
Potential outcome Y^a is observed when treatment is set to $A = a$.

Confounding with observable outcomes

Example: predicting outcome for pneumonia patients.

Predicting observable outcomes:

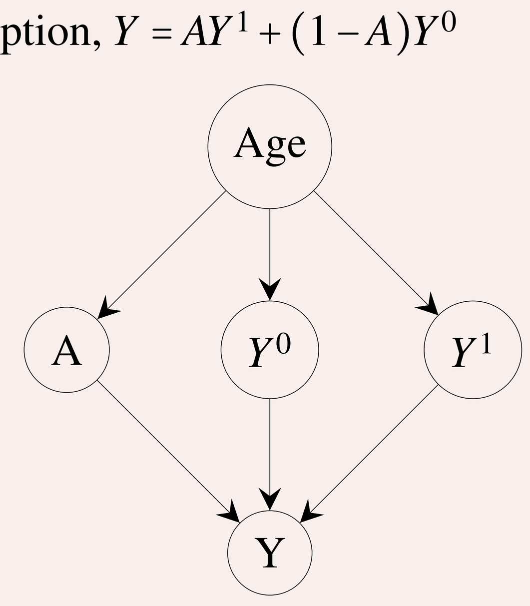
$A \in \{0,1\}$ = hospitalization indicator
$Y \in \{0,1\}$ = death indicator



Doctors treat older patients more aggressively.
Result: **Spurious negative correlation** between age and death.

Predicting potential outcomes:

Y^0 = death indicator under no hospitalization
Y^1 = death indicator under hospitalization
By assumption, $Y = AY^1 + (1-A)Y^0$



$A \perp\!\!\!\perp Y^0, Y^1 | \text{Age}$
Result: **Age positively correlated** with Y^0, Y^1.

COMPAS: Potential Outcomes Reanalysis

The COMPAS recidivism prediction tool (Northpointe, inc.) predicts rearrest for a crime within 2 years.

Data (ProPublica, 2016)
- 5,278 arrest cases from Broward County, FL
- 3,175 black; 2,103 white
- Jail durations, recidivism outcomes, covariates
- $S \in \{0,1\}$ = COMPAS score, 1 = "high risk"
- 2 scores: General (G) and recidivism risk (G)

  Violent recidivism risk (V)

Previous Analyses
- ProPublica (2016): Found different error rates and predicted score ratios based on race.
- Northpointe (2016): Found scores show predictive parity for white and black defendants.

Analysis 1: False Positive Rates

ProPublica:

$$\hat{P}(S=1|Y=0,R=r)$$

Reanalysis (doubly robust estimator):

$$\hat{P}(S=1|Y^{A=0}=0,R=r)$$

Assumes $A \perp\!\!\!\perp Y^{A=a} | S=1, R=r, X$

Results:

	(G)		(V)	
	White	Black	White	Black
ProPublica	0.23	0.45	0.18	0.38
Reanalysis	0.24	0.43	0.17	0.30

Counterfactual scores show similar bias.

Analysis 2: False Negative Rates

ProPublica:

$$\hat{P}(S=0|Y=1,R=r)$$

Reanalysis (doubly robust estimator):

$$\hat{P}(S=0|Y^{A=0}=1,R=r)$$

Assumes $A \perp\!\!\!\perp Y^{A=a} | S=0, R=r, X$.

Results:

	(G)		(V)	
	White	Black	White	Black
ProPublica	0.48	0.28	0.63	0.38
Reanalysis	0.51	0.29	0.71	0.45

Counterfactual scores show similar bias.

Analysis 3: Positive Predictive Values

Northpointe:

$$\hat{P}(Y=1|S=1,R=r)$$

Reanalysis (doubly robust estimator):

$$\hat{P}(Y^{A=0}=1|S=1,R=r)$$

Assumes $A \perp\!\!\!\perp Y^{A=a} | S=1, R=r, X$.

Results:

	(G)		(V)	
	White	Black	White	Black
Northpointe	0.59	0.63	0.17	0.21
Reanalysis	0.65	0.69	0.14	0.18

COMPAS Reanalysis Conclusions

- Error rates (FPR, FNR) similar to ProPublica results.
- Approximate predictive parity, similar to Northpointe results.
- Slightly higher PPVs for blacks than whites.
- General recidivism: Slightly higher PPVs than from observed outcomes.
- Bias in score ratios, but much less than in ProPublica results (not shown).

Predicting Recidivism in Pennsylvania

Background
- State of Pennsylvania currently developing a recidivism prediction instrument.
- Mandated by 2010 legislation.
- Goal: identify low- and high-risk defendants for further analysis.

Data
- Records from 131,076 criminal defendants from 2004–2006 (Pennsylvania Commission on Sentencing)
- $A \in \{0,1\}$: Indicator for minimum sentence served
- X: Covariates
- Y: Rearrest within 3 years of release

Analyses
- Compare "naive" modeling approach to potential outcomes-based approach

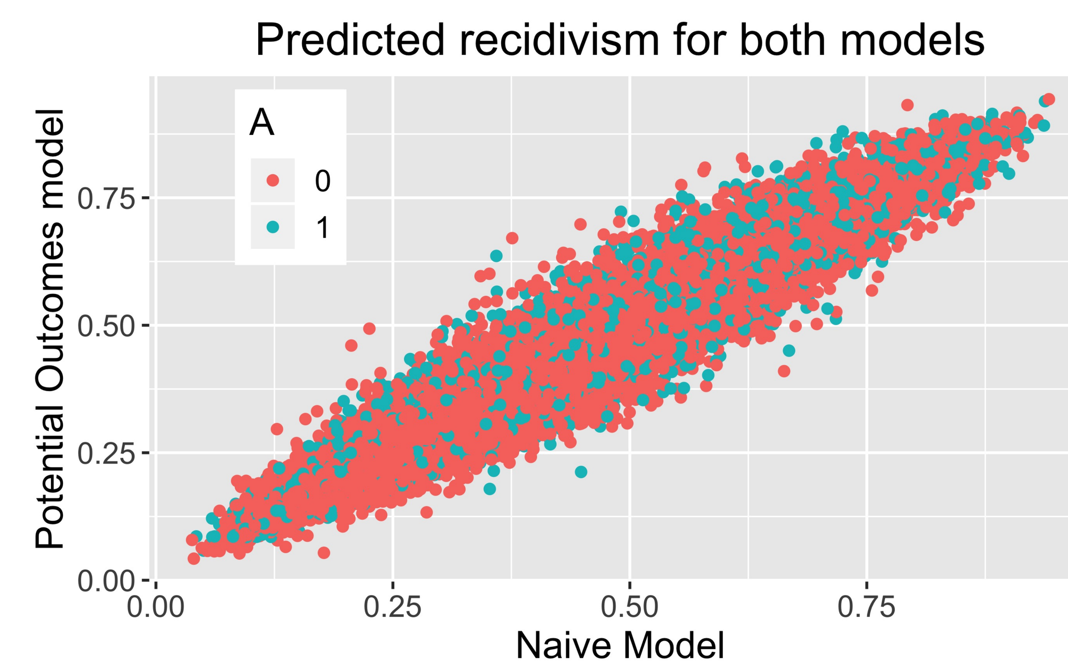
Naive model

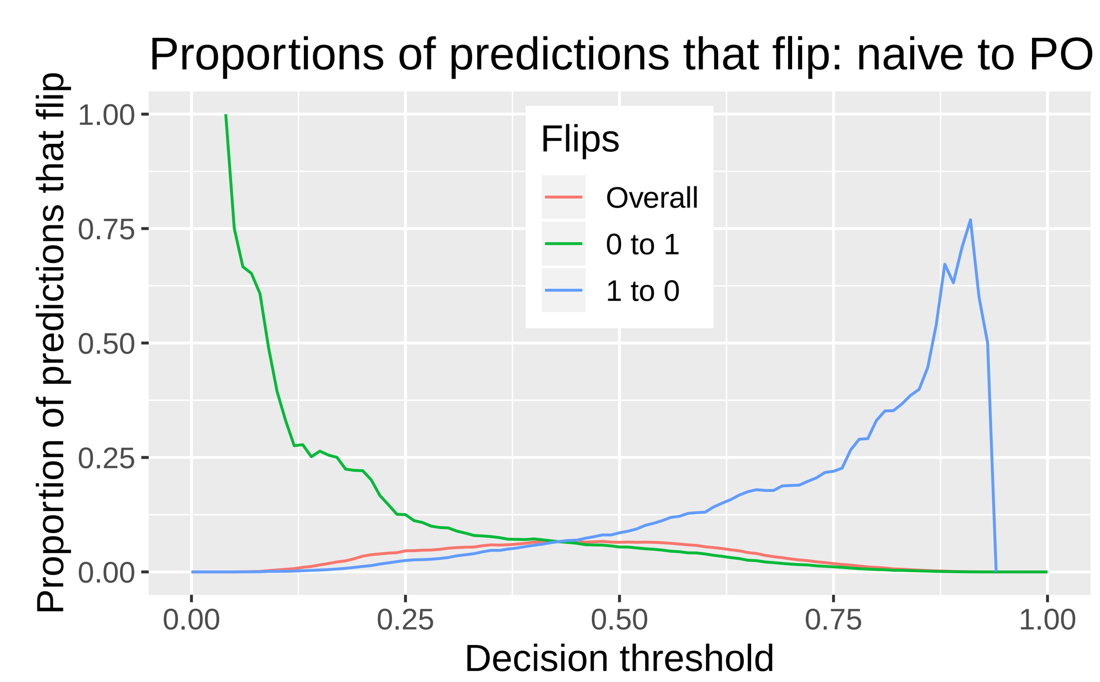
$$S := \hat{\mathbb{E}}[Y|X]$$

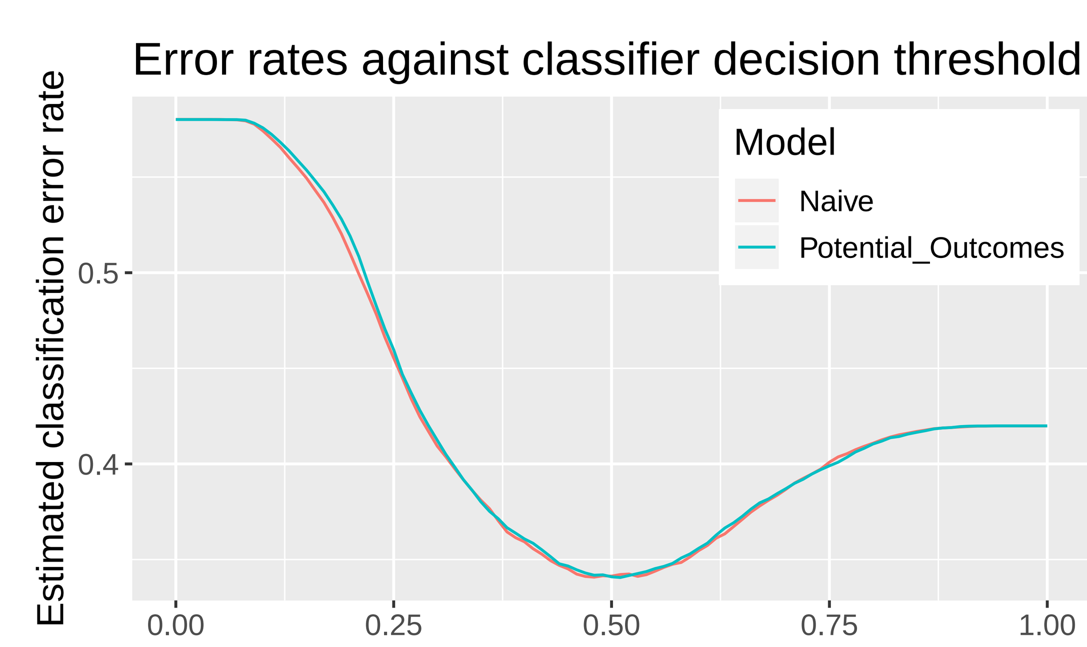
Potential outcomes model

$$S := \hat{\mathbb{E}}[Y^0|X]$$

Identifiable under assumptions of consistency, exchangeability, and positivity.







Pennsylvania Recidivism Conclusions

- Non-trivial proportions of changes in predicted outcome.
- Error rates nearly identical.
- Unclear if incarceration has an effect beyond aging.
- Positivity violation means we can't estimate Y^1.
- Further work:
  - Are there systematic differences in the two models for certain subpopulations?
  - Comparing the models on fairness criteria.