When the Oracle Misleads: Modeling the Consequences of Using Observable Rather than Potential Outcomes in Risk Assessment Instruments

Carnegie Mellon University

Alan Mishler, Niccolò Dalmasso
Department of Statistics & Data Science

"Do the right thing": machine learning and causal inference for improved decision making · workshop at NeurIPS 2019

Risk Assessment Instruments (RAIs)

- Used in medicine, criminal justice, child welfare, etc. [1, 2, 3]
- Predict risk of negative outcome (death, recidivism, neglect)
- Typically predict observable outcome (what *will* happen)
- Should predict potential outcomes (what *would* happen under available decisions) [5, 4]

Research Question:

What's the consequence of using RAIs that predict observable outcomes?

Findings:

RAIs based on observable outcomes can **make things worse.**

True even with the oracle predictor and no unmeasured confounding.

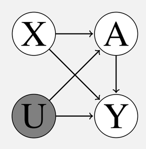
1. Setup

Example: Which patients need to be hospitalized to reduce mortality risk?

X	Observed covariates (features)
U	Unobserved confounders
A	Binary treatment (1 = hospitalization)
Y	Binary outcome (1 = death)
Y^0, Y^1	Potential outcomes under A = 0, 1

Time 0: training the predictor

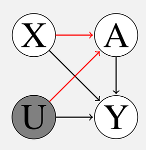
Training data: $(X, A, Y) \overset{iid}{\sim} \mathbb{P}_0$.
Train a predictor $s(X) = \hat{\mathbb{E}}_0[Y|X]$.



$s(X)$ is made available to decision makers to inform treatment A.

Time 1: the predictor in action

New distribution: $(X, A, Y) \sim \mathbb{P}_1$.
New average outcome: $\mathbb{E}_1[Y]$



When is $\mathbb{E}_1[Y] < \mathbb{E}_0[Y]$, as desired? (the predictor reduced mortality)
How far is $\mathbb{E}_1[Y]$ from $\mathbb{E}_1[Y^{d^{\text{opt}}}]$, the mortality rate under the optimal treatment rule?

2. RAIs can make things worse

Difference in mortality rates:

$$\Delta := \mathbb{E}_1[Y] - \mathbb{E}_0[Y] = \mathbb{E}\left\{\Gamma(X,U)(\mu^1(X,U) - \mu^0(X,U))\right\}$$

with

$$\Gamma(X,U) = \mathbb{P}_1(A=1|X,U) - \mathbb{P}_0(A=1|X,U) \quad \text{(difference in treatment propensities)}$$
$$\mu^a(X,U) = \mathbb{E}[Y|X,U,A=a] \quad \text{(outcome regression functions)}$$

Clearly, Δ can be positive! (more patients die at time 1)
Even if $\Delta > 0$, could have greater mortality in some strata (x, u).

2.1. Simulated example

$X \sim \text{Unif}(0,1)$	Marker of disease severity
$U = \emptyset$	No unobserved confounders
$\mathbb{P}_0(A=1|X) = X$	Treatment propensity at time 0
$\mathbb{E}_0[Y|X] = X$	Risk of non-treatment
$\mathbb{E}_1[Y|X] = (0.7 - X)^2$	Risk under hospitalization

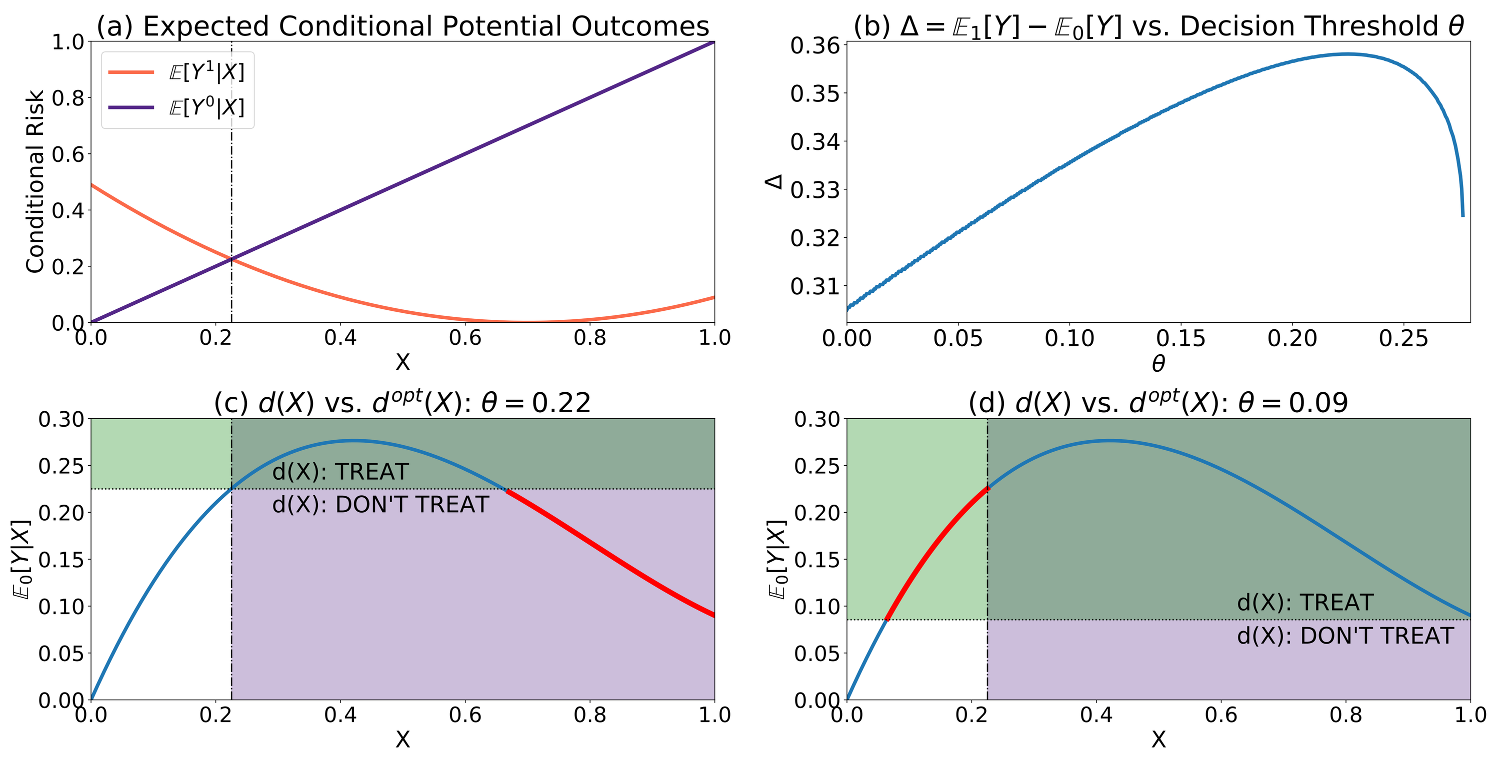
Optimal treatment rule:

$$d^{\text{opt}}(X) = \mathbb{I}\{X \geq 0.22\} \quad \text{(treat if disease severity above a certain level)}$$

Treatment rule implemented at time 1:

$$d(X) = \mathbb{I}\{s(X) \geq \theta\} \text{ for some } \theta \in [0,1] \quad \text{(treat if predicted risk above a certain level)}$$

Let $s(X) = \mathbb{E}_0[Y|X]$, the oracle predictor.



Simulated Example Results:

(a) Optimal treatment rule: treat if X above 0.22.

(b) $\Delta \approx 0.30$: More patients die at time 1 under $d(X)$, regardless of threshold θ.

(c)-(d) Groups treated (green) or not (purple) under $d(X)$. Red: patients harmed by $d(X)$.

2.2. Other undesirable properties of $s(X) = \mathbb{E}_0[Y|X]$

1. Expertise can make things worse.

Assume two medical systems, $\mathbb{P}_0^*, \mathbb{P}_0$.
Doctors in $\mathbb{P}_0^*$ are better at identifying who needs to be hospitalized:

$$\mathbb{P}_0^*(A=1|d^{\text{opt}}(X)=1) > \mathbb{P}_0(A=1|d^{\text{opt}}(X)=1)$$

Then, under a threshold rule:

Time 0: $\mathbb{E}_0^*[Y] < \mathbb{E}_0[Y]$

Time 1: $\mathbb{E}_1^*[Y] > \mathbb{E}_1[Y]$

$\mathbb{P}^*$ is **better** than $\mathbb{P}$ at time 0 and **worse** at time 1.

2. Procedure is unstable under iteration.

Suppose:

For time $t = 1, 2, \ldots$ we have $d(X) = \mathbb{I}\{\mathbb{E}_{t-1}[Y|X] > \theta\}$.

Suppose for some X we have $\mathbb{E}_0[Y^1|X] < \theta, \mathbb{E}_0[Y^0|X] > \theta$ and $\mathbb{E}_0[Y|X] > \theta$.

Then, treatment rule alternates between optimal and non-optimal:

Time t	Treatment decision	$\mathbb{E}[Y_t|X]$	$\mathbb{E}_t[Y|X]$ relative to θ
0	Treat with probability $\pi_0(X)$	$\mathbb{E}_0[Y|X]$	$> \theta$
1	Treat all	$\mathbb{E}[Y^1|X]$	$< \theta$
2	Treat none	$\mathbb{E}[Y^0|X]$	$> \theta$
3	Treat all	$\mathbb{E}[Y^1|X]$	$< \theta$
4	Treat none	$\mathbb{E}[Y^0|X]$	$> \theta$
…			

3. $s(X)$ doesn't map to a quantity of interest like $\mathbb{E}[Y^0|X], \mathbb{E}[Y^1|X]$, or $d^{\text{opt}}(X)$.

It's not clear how $s(X)$ could help decision makers get closer to optimal.

3. Conclusion

Risk Assessment Instruments based on observable outcomes **can make things worse**.

Solutions:

Estimate potential outcomes instead: $\mathbb{E}[Y^d|X]$.

Estimate optimal treatment regime $d^{\text{opt}}(X)$.

References

[1] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior*, 36(1):21–40, January 2009.

[2] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA, 2015. ACM.

[3] Alexandra Chouldechova, Emily Putnam-Hornstein, Suzanne Dworak-Peck, Diana Benavides-Prado, Oleksandr Fialko, Rhema Vaithianathan, Sorelle A Friedler, and Christo Wilson. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research*, 81:1–15, 2018.

[4] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness, 2019.

[5] Peter Schulam and Suchi Saria. Reliable Decision Support using Counterfactual Models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 1697–1708. Curran Associates, Inc., 2017.