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

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 \( A = 0, 1 \)

2.2. Other undesirable properties of \( s(X) = E[Y|X] \)

1. Expertise can make things worse.
Assume two medical systems, \( P_0, P_1 \), Doctors in \( P_1 \) are better at identifying who needs to be hospitalized:

\[ P_1(A = 1 | \text{E(RAIs) > 0.5}) > P_0(A = 1 | \text{E(RAIs) > 0.5}) \]

Then, under a threshold rule:
Time 0: \( E[Y|X] < E[Y^0|X] \)
Time 1: \( E[Y|X] > E[Y^0|X] \)

\( P_1 \) is better than \( P_0 \) at time 0 and worse at time 1.

2. Procedure is unstable under iteration.
Suppose:
For time \( t = 1, 2, \ldots \) we have \( A_t = \{ E_{Y_t}[Y|X] > \theta \} \).
Suppose for some \( X \) we have \( E_{Y_t}[Y|X] > \theta \) and \( E_{Y^0_t}[Y|X] > \theta \).

Then, treatment rule alternates between optimal and non-optimal:
Time 1: Treatment decision \( Y_{t+1}|X \) relative to \( \theta \):
1. Treat all \( E[Y|X] < \theta \)
2. Treat none \( E[Y|X] > \theta \)
3. Treat all \( E[Y|X] > \theta \)
4. Treat none \( E[Y|X] < \theta \)

3. \( s(X) \) doesn’t map to a quantity of interest \( E[Y^0|X], E[Y^1|X] \), or \( d_{opt}(X) \).
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: \( E[Y^0|X] \).
Estimate optimal treatment regime \( d_{opt}(X) \).

References