12 - Prediction Policy Problems

ml4econ, HUJI 2025

Itamar Caspi

June 29, 2025 (updated: 2025-06-29)

Outline

- Prediction Policy Problems
- Prediction and Fairness

Prediction Policy Problems

Motivation

Consider the following toy example from Kleinberg, Ludwig, Mullainathan, and Obermeyer (AER 2015):

- $Y = \{\text{rain}, \text{no rain}\}$
- X atmospheric conditions
- D is a binary policy decision
- $\Pi(Y, D)$ payoff (utility)

The change in payoff resulting from a policy decision is given by

$$\underbrace{\frac{d\Pi(Y,D)}{dD}}_{\text{total effect of the decision}} = \underbrace{\left.\frac{\partial \Pi}{\partial D}\right|_{Y}}_{\substack{\text{direct / policy term} \\ \text{(needs a *prediction* of } Y\text{)}}} + \underbrace{\frac{\partial \Pi}{\partial Y}\frac{\partial Y}{\partial D}}_{\substack{\text{indirect term} \\ \text{(needs the *causal* effect of } D \text{ on } Y\text{)}}}$$

Prediction-Policy Problems

$$\frac{d\Pi}{dD} = \underbrace{\left.\frac{\partial \Pi}{\partial D}\right|_{Y=\hat{Y}}}_{\text{direct (prediction)}} + \underbrace{\frac{\partial \Pi}{\partial Y}\frac{\partial Y}{\partial D}}_{\text{indirect (causal)}}$$

Interpretation

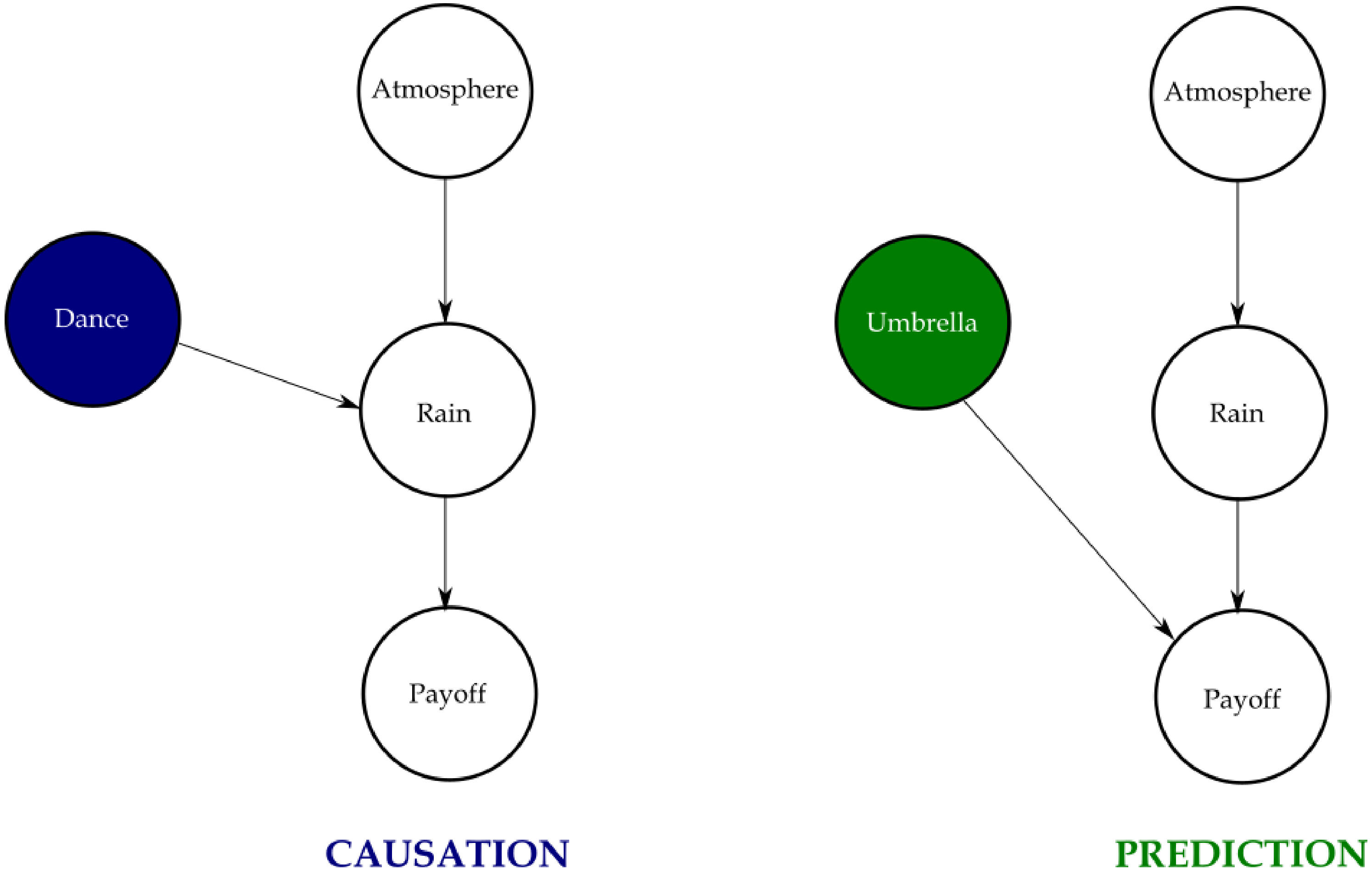
- The *direct* term asks: *If the outcome Y were frozen at its forecast $\hat{Y}$, what is the marginal payoff of changing D?*
  – needs accurate prediction, no causal ID.
- The *indirect* term multiplies how much an extra unit of Y is worth $(\partial\Pi/\partial Y)$ by how D *causes* Y to move $(\partial Y/\partial D)$.
  – needs causal evidence.

Edge Cases vs. Reality

$$\frac{d\Pi}{dD} = \underbrace{\left.\frac{\partial\Pi}{\partial D}\right|_{Y=\hat{Y}}}_{\text{direct (prediction)}} + \underbrace{\frac{\partial\Pi}{\partial Y}\,\frac{\partial Y}{\partial D}}_{\text{indirect (causal)}}$$

- Pure prediction $\partial Y/\partial D = 0$ – decision cannot move the outcome – *umbrella choice* (your umbrella won't change the weather)
- Pure causation $(\partial\Pi/\partial D)|_Y = 0$ – action matters *only* via its effect on Y – *vaccine dose* (ignoring side-effects)
- Most real decisions: both terms ≠ 0 → need good forecasts *and* causal evidence

Rain dance vs. umbrella



Prediction-Policy Problems

In the past two lectures we focused on assessing policy with causal inference and treatment effects.

Some decisions, however, depend *only* on prediction (Kleinberg, Ludwig, Mullainathan & Obermeyer 2015):

- Which applicants will be **most effective teachers?** (hiring & promotion)
- **How long will a worker stay unemployed?** (setting optimal savings)
- **Which restaurants are likeliest to violate health codes?** (targeting inspections)
- **Which youths face the highest risk of re-offending?** (allocating interventions)
- **How credit-worthy is a loan applicant?** (approval decisions)

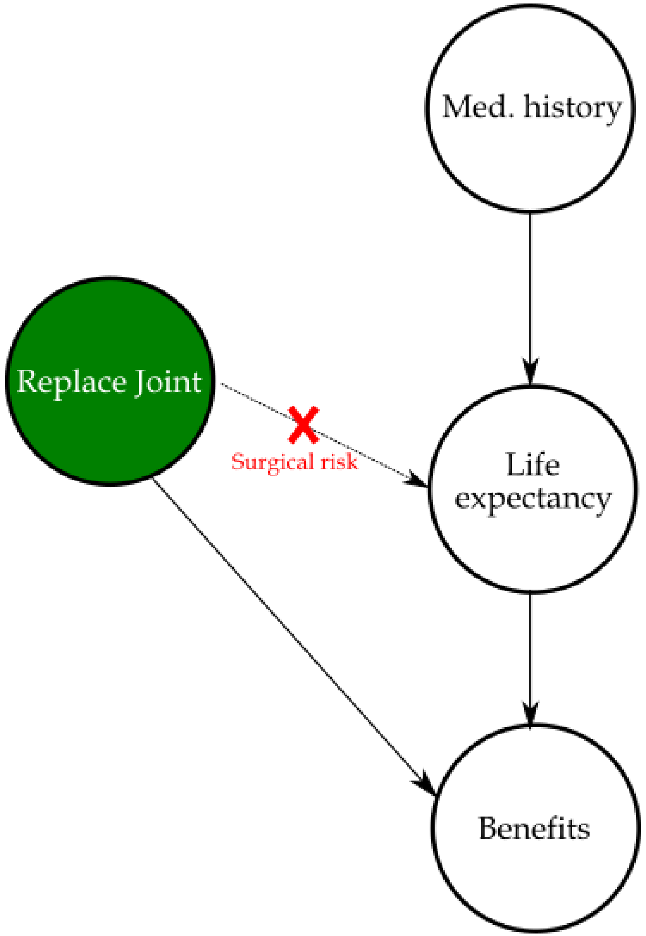
When the action cannot influence the outcome, causal methods add no value—the planner's job is to forecast as accurately as possible.

Real-world prediction-policy problem: Joint replacement

- 750,000 hip or knee replacements are performed in the United States each year.
- Benefits: substantial gains in mobility and pain relief.
- Costs: about $15,000 per procedure plus a painful, months-long recovery.

> **Working assumption:** the payoff, Π, from surgery rises with postoperative longevity, Y.
>
> **Key Question:** Using only information available before the operation, can we forecast which surgeries will be futile and redirect those resources?

Joint replacement DAG



Note: Kleinberg et al. (2015) abstract from surgical risk.

Data

- A 20% sample of 7.4 million Medicare beneficiaries was analyzed; 98,090 individuals (1.3%) had a claim for joint replacement in 2010.
- Among these patients, 1.4% died within one month of surgery—likely due to complications—and an additional 4.2% died within the subsequent 1–12 months.
- The average mortality rate is approximately 5%. On average, surgeries are not futile.
- However, this average may be misleading. A more relevant question is whether surgeries performed on the predictably highest-risk patients were futile.

Predicting mortality risk

Setup:

- **Outcome:** Mortality within 1–12 months post-surgery
- **Features:** Medicare claims data prior to joint replacement, including patient demographics (age, sex, geography); comorbidities, symptoms, injuries, acute conditions, and their progression over time; and healthcare utilization
- **Sample:** Training set: 65,000 observations; Test set: 33,000 observations
- **ML algorithm:** Lasso regression

The playbook:

- Rank test-set beneficiaries by model-predicted mortality risk percentiles
- Assign to each percentile its corresponding share of surgeries
- Demonstrate that the algorithm outperforms physician decision-making

The riskiest people receiving joint replacement

Predicted Mortality Percentile	Observed Mortality Rate	Total Number Annually
1	0.562 (.027)	4905
2	0.530 (.02)	9810
5	0.456 (.012)	24525
10	0.345 (.008)	49045
20	0.228 (.005)	98090
30	0.165 (.004)	147135
100	0.057 (.001)	490450

Source: Kleinberg et al. (2015).

How to read the table

- Col 1: Model-predicted mortality percentile (1 = riskiest 1 %).
- Col 2: Actual 12-month mortality for that percentile (standard errors in parentheses).
- Col 3: Number of joint-replacement surgeries performed each year in that bin.

Example: the top 1 % risk group received 4,905 surgeries, yet 56.2 % died within a year—so most of those operations were likely futile.

Can an algorithm beat physicians?

Econometric hurdle — selective-labels bias

We only observe post-operative mortality for people *who actually received a joint replacement*; the counterfactual outcome for untreated, yet eligible, patients is missing.

Counterfactual construction

1. Pull the pool of patients who satisfied Medicare eligibility but did not undergo surgery.
2. Working assumption: surgeons schedule operations roughly in order of *increasing* risk (lowest-risk first).
3. Reallocate a fixed number of surgeries from the highest-risk treated patients to the lowest-risk untreated eligibles and compare predicted mortality.

If this simulated swap lowers the expected death rate, the algorithm outperforms physician judgment.

So, can the lasso beat physicians?

Predicted Mortality Percentile	Observed Mortality Rate	Total Number Annually	Substitute with 50th percentile Eligibles: Futile Procedures Averted	Annual Savings (in millions)
1	0.562 (.027)	4905	2403	36
2	0.530 (.02)	9810	4485	67
5	0.456 (.012)	24525	9398	141
10	0.345 (.008)	49045	13350	200
20	0.228 (.005)	98090	15219	228
30	0.165 (.004)	147135	13548	203
100	0.057 (.001)	490450		

Source: Kleinberg et al. (2015).

Simulation logic: For each risk percentile, swap surgeries from the highest-risk treated patients with untreated, median-risk eligibles (50th percentile), holding total surgeries fixed.

The table reports:

- Futile procedures averted (col 4) – operations reallocated away from patients with ≥50 % predicted 12-month mortality.
- Annual savings (col 5) – hospital costs avoided (≈ $15 k per surgery).

Key line: replacing the top 10 % risk group prevents 13,350 likely futile operations and saves $200m per year—evidence that the algorithm screens better than current physician judgement.

What can still go wrong?

Econometric Concern #2 — Omitted Payoff Bias Physicians might observe benefits—such as pain relief—that are not captured in our data.

Testing the Hypothesis
Use proxies for post-operative pain relief:

- Physical therapy (PT) and joint injection claims
- Follow-up visits for osteoarthritis

Empirically, these proxies are flat across predicted risk percentiles. High-mortality patients do not appear to experience greater pain relief ⇒ Reallocating their surgeries is unlikely to reduce welfare.

Predicted Mortality Percentile	Observed Mortality Rate	Total Number Annually	PT + Joint Injections	Physician Visits for Osteo.
1	0.562 (.027)	4905	4.4 (.356)	1.4 (.173)
2	0.530 (.02)	9810	4.0 (.316)	1.8 (.13)
5	0.456 (.012)	24525	3.9 (.208)	2.0 (.092)
10	0.345 (.008)	49045	3.8 (.143)	2.1 (.066)
20	0.228 (.005)	98090	3.9 (.091)	1.8 (.042)
30	0.165 (.004)	147135	3.8 (.076)	1.9 (.035)
100	0.057 (.001)	490450	3.9 (.046)	2.1 (.023)

Leveling the Playing Field

To fairly compare physicians (or any human experts) with a machine learning model, both must operate under identical inputs and incentives:

- Same information – access to the full pre-decision dataset
- Same objective – a shared payoff or loss function
- Same constraints – identical budgetary, temporal, and policy limitations

> Only after aligning these conditions can we meaningfully ask: Who allocates resources more effectively?

Key take-aways

- A more accurate model does not automatically lead to better decisions. Alignment with payoffs and real-world constraints is essential.
- Selective-labels and omitted-payoff bias can obscure algorithmic gains—or generate illusory ones.
- Robust policy design requires marrying machine learning with social science: incentives, fairness, and welfare metrics must guide implementation.

Prediction and Fairness

Blind Algorithms

Algorithmic Fairness (Kleinberg, Ludwig, Mullainathan, and Rambachan, AER 2018):

Can we increase algorithmic fairness by ignoring variables that induce such bias such as race, age, sex, etc.?

Short answer: Not necessarily.

The basic setup

The context: Student admission to college.

Data: $\{Y_i, X_i, R_i\}_{i=1}^N$, where

- Y_i is performance
- X_i is a set of features
- R_i is a binary race indicator where $R_i = 1$ for individuals that belong to the minority group and $R_i = 0$ otherwise.

Predictors:

- "Aware": $\hat{f}\left(X_i, R_i\right)$

- "Blind": $\hat{f}\left(X_i\right)$

- "Orthogonality": $\hat{f}\left(\widetilde{X}_i\right)$, where $\widetilde{X}_i \perp R_i$.

Definitions

Let S denote the set of admitted students and $\phi(S)$ denote a function that depends only on the predicted performance, measured by $\hat{f}$, of the students in S.

Compatibility condition: If S and S' are two sets of students of the same size, sorted in descending order of predicted performance $\widehat{f}(X, R)$, and the predicted performance of the i^{th} student in S is at least as large as the predicted performance of the i^{th} in S' for all i, then $\phi(S) \geq \phi(S')$.

Intuition: if every member of class S is no worse on paper than the counterpart in S', any planner who claims to care only about student performance shouldn't prefer S'.

- The *efficient* planner maximizes $\phi(S)$ where $\phi(S)$ is compatible with $\hat{f}$.

- The *equitable* planner seeks to maximize $\phi(S) + \gamma(S)$, where $\phi(S)$ is compatible with $\hat{f}$, and $\gamma(S)$ is monotonically increasing in the number of students in S who have $R = 1$.

Main result: Keep R in

Kleinberg et al. (2018) main result:

> THEOREM 1: *For some choice of K_0 and K_1 with $K_0 + K_1 = K$, the equitable planner's problem can be optimized by choosing the K_0 applicants in the $R = 0$ group with the highest $\widehat{f}(X, R)$, and the K_1 applicants in the $R = 1$ group with the highest $\widehat{f}(X, R)$.*

(See Kleinberg et al., 2018 for a sketch of the proof.)

In words: If you want both quality and a fair share of minority students, first decide how many seats each group should get, then simply admit the highest-scoring people within each group—using the race-aware score.

Intuition

- Good ranking of applicants is desired for both types of planners.
- Equitable planners still care about ranking *within* groups.
- Achieving a more balanced acceptance rate is a *post* prediction step. Can be adjusted by changing the group-wise threshold.

Illustration of the result

Say that we have 10 open slots, 100 admissions from the majority group $(R = 0)$ and 20 form the minority group $(R = 1)$. In addition, assume that the acceptance rate for the minority group is set to 30%.

An equitable planner should:

1. Rank candidates within each group according to $\widehat{f}(X_i, R_i)$.
2. Accept the top 7 from the $R = 0$ group, and top 3 from the $R = 1$ group.

Empirical application

DATA: Panel data on This representative sample of students who entered eighth grade in the fall of 1988, and who were then followed up in 1990, 1992, 1994, and mid-20s).

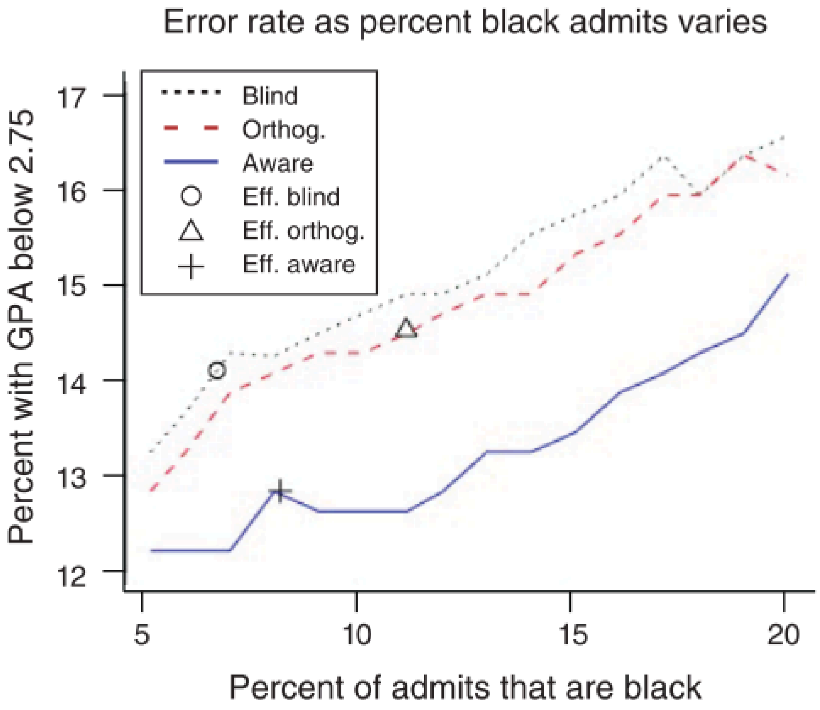
OUTCOME: GPA ≥ 2.75.

FEATURES: High school grades, course taking patterns, extracurricular activities, standardized test scores, etc.

RACE: White ($N_0 = 4,274$) and black ($N_1 = 469$).

PREDICTORS: OLS (random forest for robustness)

RESULT: The "aware" predictor dominates for both efficient planner and equitable planner.



Sources of disagreement

- On the right: The distribution of black students in the sample across predicted-outcome deciles according to the race-blind or race-aware predictors.
- How to read this: In the case of agreement between race-blind and race-aware, the values would be aligned on the main diagonal. By contrast, disagreement is characterized by off-diagonal non-zero values.
- Bottom line (again): Adding race to the equation improves within group ranking.

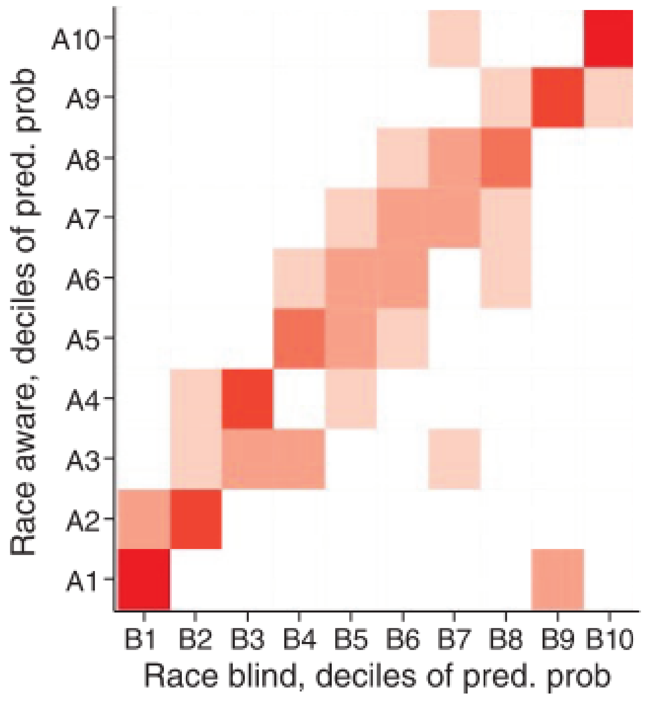


FIGURE 2. HEATMAP OF RANKINGS OF BLACK APPLICANTS BY PREDICTED PROBABILITY OF GPA < 2.75, USING RACE-AWARE VERSUS RACE-BLIND ALGORITHMS

Main takeaways

- Turning algorithms blind might actually do harm.
- What actually matters is the rankings within groups.
- Caveat: This is a very specific setup and source of bias.

```
slides |> end()
```

Source code

Selected references

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Athey, S. (2018). The Impact of Machine Learning on Economics.

Athey, S., & Wager, S. (2018). Efficient Policy Learning.

Kleinberg, B. J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review: Papers & Proceedings*, 105(5), 491–495.

Kleinberg, B. J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic Fairness. *American Economic Review: Papers & Proceedings*, 108, 22–27.

Selected references

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Descisions and Machine Predictions. *Quarterly Journal of Economics*, 133(1), 237–293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of the 8th Conference on Innovation in Theoretical Computer Science*, 43, 1–23.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87–106.