Lecture 20: Causal article on risk assessment

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

Motivation

- Motivation for class: Review an article that uses causal inference to study the use of an algorithm for sentencing.

Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment*

Kosuke Imai† Zhichao Jiang‡ D. James Greiner§ Ryan Halen¶ Sooahn Shin‖

First Draft: July 9, 2020
This Draft: October 2, 2021

For Review

Abstract

Despite an increasing reliance on fully-automated algorithmic decision-making in our day-to-day lives, human beings still make highly consequential decisions. As frequently seen in business, healthcare, and public policy, recommendations produced by algorithms are provided to human decision-makers to guide their decisions. While there exists a fast-growing literature evaluating

Motivation

- Algorithms appeal: A growing body of literature has suggested the potential superiority of algorithmic decision-making over purely human choices across a variety of tasks.
- Human appeal: Many are resistant to have algorithms take over decision-making tasks, especially those that are high-stakes.
- Hybrid systems appeal: The desire for a human decision-maker as well as the precision and efficiency of algorithms have led to the adoption of hybrid systems involving both. By far the most popular system uses algorithmic recommendations to inform human decision-making. This has affected us in medicine, hiring, credit lending, investment decisions, and online shopping.
- Risk assessment in criminal justice: Of particular interest, algorithmic recommendations are increasingly used in risk assessment in the criminal justice system, with the goal of improving incarceration rulings and other decisions made by judges.

Contribution of this paper

- Bias/fairness: There is a fast-growing literature in computer science/statistics/ML that studies the bias and fairness of algorithms.
- An overlooked question is: Whether those algorithms actually help humans make better decisions. In this paper, the authors develop a methodological framework for experimentally evaluating the impacts of algorithmic recommendations on human decision-making.
- Method: They conducted the first-ever real-world field experiment by providing, for a randomly selected set of cases, information from a system consisting of Public Safety Assessment (PSA) risk scores, and a recommendation from a Decision Making Framework (DMF) to a judge who makes an initial release decision.

Outcomes

- What does it actually test?: They evaluate whether the PSA-DMF system (A) helps judges achieve their goal of preventing arrestees from committing a new crime or failing to appear in court while avoiding an unnecessarily harsh decision (Y).

Algorithms in court

To date, algorithmic outputs have appeared in

- i) At the "first appearance" or arraignment hearing, during which a judge decides whether to release an arrestee pending disposition of any criminal charges, (in response to an arraignment the accused is expected to enter a plea) and
- ii) At sentencing, in which the judge imposes a punishment on a defendant found guilty.
- This paper deals with point (i).

Setting

- What happens at hearing? We describe a typical first appearance hearing. The key decision the judge must make at a first appearance hearing is whether to release the arrestee pending disposition of any criminal charges, and if the arrestee is to be released, what conditions to impose.
- Seek to avoid incarceration: Because arrestees have not yet been adjudicated guilty of any charge at the time of a pretrial hearing, there exists a consensus that pretrial incarceration is to be avoided unless the risks associated with release are sufficiently high.

Judges deciding whether to release arrestees ordinarily consider risk factors among a variety of other concerns:

- 1) FTA: Failure to appear - the risk that the arrestee will fall to appear (FTA) at subsequent court dates.
- 2) NCA: New criminal activity - the risk that the arrestee will engage in new criminal activity (NCA) before the case is resolved.
- 3) NVCA: New violent criminal activity.

Risk assessment algorithm

- Goal of Public Safety Assessment (PSA) algorithm is to classify arrestees according to FTA and NCA risks.
- They are generally constructed by fitting a statistical model to a training dataset based on past first appearance hearings and the subsequent incidences (or lack thereof) of FTA and NCA.
- The hope is that providing such instrments will improve the assessment of FTA and NCA risks and thereby lead to better decisions.
- The goal of this paper is to develop a general methodological framework for evaluating the impact of providing the PSA to judges at first appearance hearings using an RCT.

Experiment

- Dane County, Wisconsin.
- PSA scores are based on the weighted indices of nine factors drawn from criminal history information, primarily prior convictions and FTA, and a single demographic factor, age.
- Notably, gender and race are not used to compute the PSA.
- The weights are calculated using past data.
- More information here: https://advancingpretrial.org/psa/factors/

Data

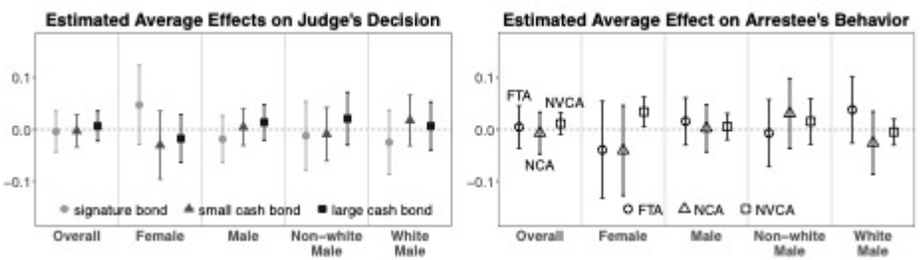
- 30 month treatment assignment period (2017-2019), followed by the collection of data on FTA, NCA, NCVA and other outcomes
- N = 1891 cases, 40% are white male arrestees and 13% are white female arrestees (non-white male is 39% and non-white female is 8%).

	no PSA (Control Group)			PSA (Treatment Group)			
	Signature bond	Cash bond ≤$1000	Cash bond >$1000	Signature bond	Cash bond ≤$1000	Cash bond >$1000	Total (%)
Non-white Female	64	11	6	67	6	0	154
	(3.4)	(0.6)	(0.3)	(3.5)	(0.3)	(0.0)	(8.1)
White Female	91	17	7	104	17	10	246
	(4.8)	(0.9)	(0.4)	(5.5)	(0.9)	(0.5)	(13.0)
Non-white Male	261	56	49	258	53	57	734
	(13.8)	(3.0)	(2.6)	(13.6)	(2.8)	(3.0)	(38.8)
White Male	289	48	44	276	54	46	757
	(15.3)	(2.5)	(2.3)	(14.6)	(2.9)	(2.4)	(40.0)
FTA committed	218	42	16	221	45	16	558
	(11.5)	(2.2)	(0.8)	(11.7)	(2.4)	(0.8)	(29.4)
not committed	487	90	90	484	85	97	1333
	(25.8)	(4.8)	(4.8)	(25.6)	(4.5)	(5.1)	(70.6)
NCA committed	211	39	14	202	40	17	523
	(11.2)	(2.1)	(0.7)	(10.7)	(2.1)	(0.9)	(27.7)
not committed	494	93	92	503	90	96	1368
	(26.1)	(4.9)	(4.9)	(26.6)	(4.8)	(5.1)	(72.4)
NVCA committed	36	10	3	44	10	6	109
	(1.9)	(0.5)	(0.2)	(2.3)	(0.5)	(0.3)	(5.7)
not committed	669	122	103	661	120	107	1782
	(35.4)	(6.5)	(5.4)	(35.0)	(6.3)	(5.7)	(94.3)
Total	705	132	106	705	130	113	1891
	(37.3)	(7.0)	(5.6)	(37.3)	(6.9)	(6.0)	(100)

Not Review

Findings

- In general, we observe a positive association between the PSA scores and judge's decisions, implying that a higher PSA score is associated with a harsher decision.
- This figure presents the estimated average causal effect of PSA provisions on the judge's decisions (left plot) and three outcomes of interest (right plot).



- The vertical bars are the 95% confidence intervals.
- PSA provision appears to have little overall effect on the judge's decision and arrestee's behavior, on average, although it may slightly increase NVCA among female arrestees.
- Results: The results imply that PSA provision, on average, has little effect on the judge's decisions.

Causal analysis about subgroups

Z_i: Binary treatment variable indicating whether the PSA is presented to the judge of case $i = 1, 2, \ldots, n$.

D_i: Binary detention decision made by the judge to either detain ($D_i=1$) or release ($D_i=0$) the arrestee prior to the trial. ($D_i(z)$ is potential outcome under treatment z).

Y_i: Binary outcome - they code all their outcomes (NCA, NVCA, FTA) as binary variables. For example, $Y_i = 1$ implies that the arrestee of case i commits an NCA.

$\mathbf{X_i}$: A vector of observed pre-treatment covariates for case i. They include age, gender, race, and prior criminal history.

$$\text{Preventable cases: } APCE_p = E(D_i(1) - D_i(0) | Y_i(1) = 0, Y_i(0) = 1)$$

$$\text{Risky cases: } APCE_r = E(D_i(1) - D_i(0) | Y_i(1) = 1, Y_i(0) = 1)$$

$$\text{Safe cases: } APCE_s = E(D_i(1) - D_i(0) | Y_i(1) = 0, Y_i(0) = 0)$$

Assumptions

- Randomization of the treatment assignment
- Exclusion restriction.
- Monotonicity.
- They also use nonparametric estimation.

Paper conclusions

- First, we find that PSA provision has little overall impact on the judge's decisions.
- Second, we find potentially suggestive evidence PSA provision may encourage the judge to make more lenient decisions for female arrestees regardless of their risk levels while leading to more stringent decisions for males who are classified as risky.
- Third, PSA provisio appears to widen the existing gender difference of the judge's decision against male arrestees whereas it does not seem to alter decision-making across race among male arrestees.
- Our results suggest that the PSA's recommendations may be harsher than necessary.