Experimental Evaluation of Computer-Assisted Human Decision Making

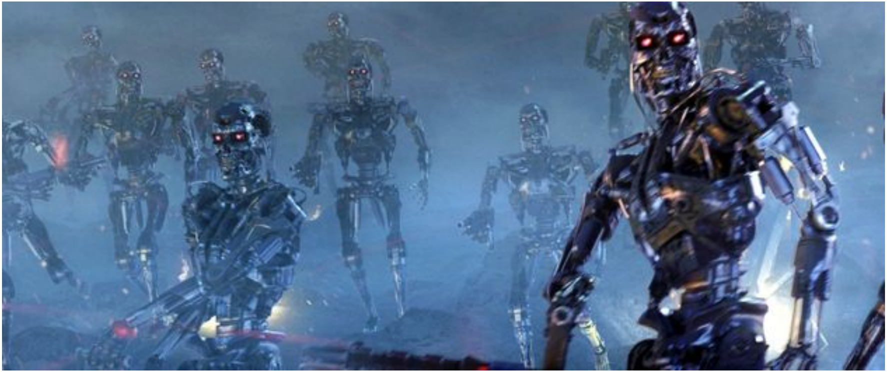
Kosuke Imai

Harvard University

Virtual Workshop on Missing Data Challenges in Computation, Statistics and Applications
Institute for Advanced Studies, Princeton New Jersey
September 8, 2020

Joint work with Zhichao Jiang (UMass. Amherst)
Jim Greiner, Ryan Halen (Harvard Law School)
and Sooahn Shin (Harvard)

Rise of the Machines



- Statistics, machine learning, artificial intelligence in our daily lives
- Nothing new but accelerated due to technological advances
- Examples: factory assembly lines, ATM, home appliances, autonomous cars and drones, games (Chess, Go, Shogi), ...

Motivation

- But, humans still make many consequential decisions
  - this is true even when human decisions can be suboptimal
  - we may want to hold *someone*, rather than *something*, accountable
- Computer-assisted human decision making
  - humans make decisions with the aid of machine recommendations
  - routine decisions made by individuals in daily lives
  - consequential decisions made by judges, doctors, etc.
- How do machine recommendations influence human decisions?
  - Do they help human decision-makers achieve a goal?
  - Do they help humans improve the fairness of their decisions?
- Many have studied the accuracy and fairness of machine recommendations rather than their impacts on human decisions
- We develop a set of statistical methodology for experimentally evaluating computer-assisted human decision making

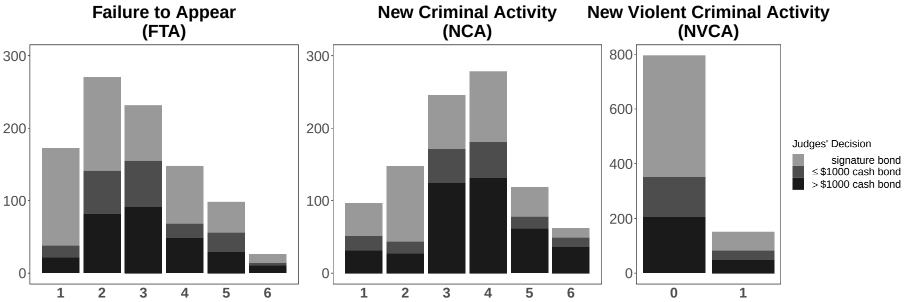
Application: Pretrial Risk Assessment Instrument (PRAI)

- Machine recommendations often used in US criminal justice system
- At the first appearance hearing, judges primarily make two decisions
  1. whether to release an arrestee pending disposition of criminal charges
  2. what conditions (e.g., bail and monitoring) to impose if released
- Goal: avoid predispositional incarceration while preserving public safety
- Judges are required to consider three risk factors along with others
  1. arrestee may fail to appear in court (FTA)
  2. arrestee may engage in new criminal activity (NCA)
  3. arrestee may engage in new violent criminal activity (NVCA)
- PRAI as a machine recommendation to judges
  - classifying arrestees according to FTA and NCA/NVCA risks
  - derived from an application of a machine learning algorithm or a statistical model to a training data set based on past observations
- Controversy over the potential racial bias of COMPAS score
  - Propublica's analysis and Northpointe's rebuttal
  - Almost all existing work focus on the accuracy and fairness of PRAI

A Field Experiment for Evaluating a PRAI

- An anonymous (for now) county
- PRAI
  1. based on criminal history (prior convictions and FTA) and age
  2. two separate ordinal risk scores for FTA and NCA
  3. one binary risk score for new violent criminal activity (NVCA)
- Judges have other information about an arrestee
  - affidavit by a police officer about the arrest
  - defense attorney may inform about the arrestee's connections to the community (e.g., family, employment)
  - assistant district attorney may provide additional information
- Field experiment
  - clerk assigns case numbers sequentially as cases enter the system
  - PRAI is calculated for each case using a computer system
  - if the first digit of case number is even, PRAI is given to the judge

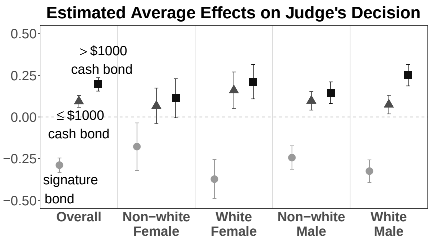
(Somewhat Empirically Informed) Synthetic Data Set



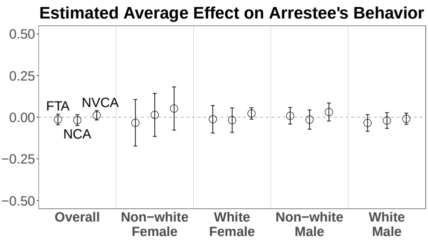
- PRAI
  1. 6-point scale for FTA and NCA
  2. binary flag for NVCA
- Trichotomized ordinal decisions
  1. signature bond
  2. $\leq$ \$1,000 cash bond
  3. $>$ \$1,000 cash bond

Intention-to-Treat Analysis of PRAI Provision

(a) Estimated effects on judges' decisions



(b) Estimated effects on outcomes



- Large effects on judges' decisions
- But little effects on outcomes
  - Do judges' decisions have no effect on outcomes? $\rightsquigarrow$ unlikely
  - Are the heterogeneous effects being masked?

The Setup of the Proposed Methodology

- Notation:
  - $i = 1, 2, \ldots, n$: cases
  - Z_i: whether PRAI is presented to the judge ($Z_i = 1$) or not ($Z_i = 0$)
  - D_i: judge's binary decision to release ($D_i = 1$) or detain ($D_i = 0$)
  - Y_i: binary outcome (NCA, FTA, or NVCA)
  - X_i: observed (by researchers) pre-treatment covariates

- Potential outcomes:
  - $D_i(z)$: potential value of the release decision when $Z_i = z$
  - $Y_i(z, d)$: potential outcome when $Z_i = z$ and $D_i = d$
  - Relationship to observed data: $D_i = D_i(Z_i)$ and $Y_i = Y_i(Z_i, D_i(Z_i))$
  - No interference across cases: we analyze the first arrest cases only

- Assumptions maintained throughout our analysis:
  1. Randomized treatment assignment: $\{D_i(z), Y_i(z, d), \mathrm{X}_i\} \perp\!\!\!\perp Z_i$
  2. Exclusion restriction: $Y_i(z, d) \ = \ Y_i(d)$
  3. Monotonicity: $Y_i(0) \leq Y_i(1)$

Causal Quantities of Interest

- Principal stratification (Frangakis and Rubin 2002)
  - $(Y_i(1), Y_i(0)) = (1, 0)$: preventable cases
  - $(Y_i(1), Y_i(0)) = (1, 1)$: risky cases
  - $(Y_i(1), Y_i(0)) = (0, 0)$: safe cases
  - $\cancel{(Y_i(1), Y_i(0)) = (0, 1)}$: eliminated by monotonicity
- Average principal causal effects of PRAI on judge's decisions:

$$\begin{aligned}
\text{APCEp} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 0\}, \\
\text{APCEr} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 1\}, \\
\text{APCEs} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 0\}.
\end{aligned}$$

- If PRAI is helpful, we should have APCEp < 0 and APCEs > 0
- The desirable sign of APCEr depends on various factors
- Partial identification (e.g., the signs of APCE) is possible under the assumptions of randomization, exclusion restriction, and monotonicity

Point Identification under Unconfoundedness

- Unconfoundedness:

$$Y_i(d) \perp\!\!\!\perp D_i \mid \mathrm{X}_i, Z_i = z$$

  for $z = 0, 1$ and all d.
- Violated if judges base their decision on additional information they have about arrestees $\rightsquigarrow$ sensitivity analysis

- Principal scores (Ding and Lu 2017)

$$\begin{aligned} e_P(\mathrm{x}) &= \Pr\{Y_i(1) = 1, Y_i(0) = 0 \mid \mathrm{X}_i = \mathrm{x}\} \\ e_R(\mathrm{x}) &= \Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid \mathrm{X}_i = \mathrm{x}\} \\ e_S(\mathrm{x}) &= \Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid \mathrm{X}_i = \mathrm{x}\} \end{aligned}$$

Identification Results

Under the assumptions of randomization, monotonicity, exclusion restriction, and unconfoundedness, we can identify causal effects as

$$\begin{aligned}
\text{APCEp} &= \mathbb{E}\{w_P(\mathrm{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_P(\mathrm{X}_i)D_i \mid Z_i = 0\}, \\
\text{APCEr} &= \mathbb{E}\{w_R(\mathrm{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_R(\mathrm{X}_i)D_i \mid Z_i = 0\}, \\
\text{APCEs} &= \mathbb{E}\{w_S(\mathrm{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_S(\mathrm{X}_i)D_i \mid Z_i = 0\},
\end{aligned}$$

where

$$w_P(\mathrm{x}) = \frac{e_P(\mathrm{x})}{\mathbb{E}\{e_P(\mathrm{X}_i)\}}, \quad w_R(\mathrm{x}) = \frac{e_R(\mathrm{x})}{\mathbb{E}\{e_R(\mathrm{X}_i)\}}, \quad w_S(\mathrm{x}) = \frac{e_S(\mathrm{x})}{\mathbb{E}\{e_S(\mathrm{X}_i)\}}.$$

and

$$\begin{aligned}
e_P(\mathrm{x}) &= \Pr\{Y_i = 1 \mid D_i = 1, \mathrm{X}_i = \mathrm{x}\} - \Pr\{Y_i = 1 \mid D_i = 0, \mathrm{X}_i = \mathrm{x}\}, \\
e_R(\mathrm{x}) &= \Pr\{Y_i = 1 \mid D_i = 0, \mathrm{X}_i = \mathrm{x}\}, \\
e_S(\mathrm{x}) &= \Pr\{Y_i = 0 \mid D_i = 1, \mathrm{X}_i = \mathrm{x}\}.
\end{aligned}$$

Extension to Ordinal Decision

- Judge's decision is typically ordinal (e.g., bail amount)
  - $D_i = 0, 1, \ldots, k$: a bail of increasing amount
  - Monotonicity: $Y_i(d_1) \ \geq \ Y_i(d_2) \quad \text{for} \quad d_1 \leq d_2$
- Principal strata based on an ordinal measure of risk

$$R_i \ = \ \begin{cases} \min\{d : Y_i(d) = 0\} & \text{if } \ Y_i(k) = 0 \\ k + 1 & \text{if } \ Y_i(k) = 1 \end{cases}$$

  - Least amount of bail that keeps an arrestee from committing NCA
  - Example with $k = 2$: risky cases ($R_i = 3$), preventable cases ($R_i = 2$ and $R_i = 1$), safe cases ($R_i = 0$)
- Causal quantities of interest: reduction in the proportion of NCA attributable to the PRAI within each principal strata $r = 1, \ldots, k$

$$\text{APCEp}(r) \ = \ \Pr\{D_i(1) \geq r \mid R_i = r\} - \Pr\{D_i(0) \geq r \mid R_i = r\}$$

- Nonparametric identification under unconfoundedness

Principal Fairness (Imai and Jiang, 2020)

- Literature focuses on the fairness of machine-recommendations/PRAI
- We focus on the fairness of human decision
- Existing statistical fairness definitions do not take into account how a decision affects individuals

- Principal fairness: decision should not (statistically) depend on a protected attribute A_i (e.g., race and gender) within a principal strata

$$D_i \perp\!\!\!\perp A_i \mid R_i = r \quad \text{for all } r \in \{-1, 0, 1, \ldots, k\}$$

Measuring and Estimating the Degree of Fairness

- How fair are the judges' decisions?

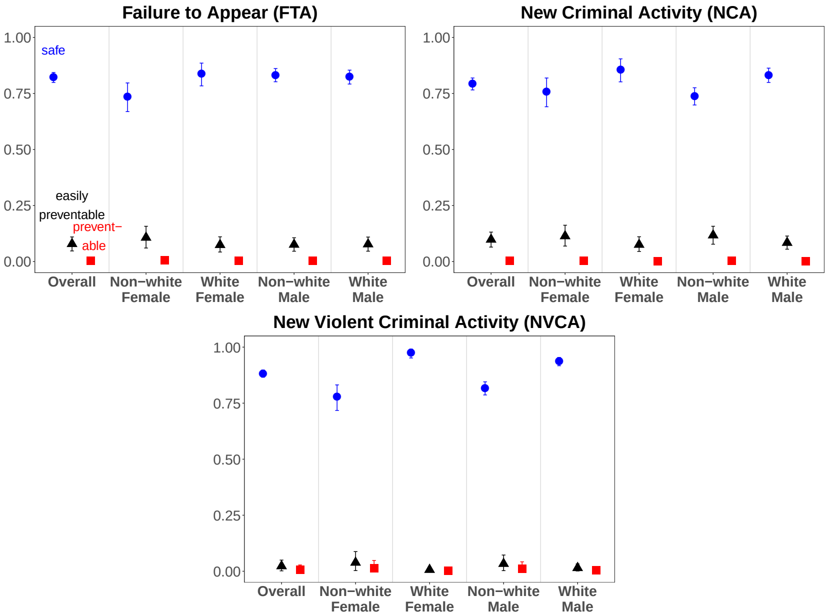
$$\begin{aligned}\Delta_r(z) \;=\; \max_{a,a',d} \big| & \Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} \\ & - \; \Pr\{D_i(z) \geq d \mid A_i = a' R_i = r\}\big|\end{aligned}$$

for $1 \leq d \leq k$ and $0 \leq r \leq k$

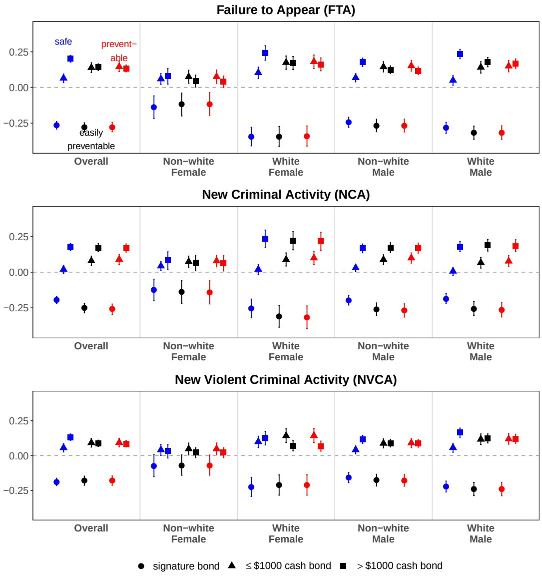
- Does the provision of PRAI improve the fairness of judges' decision?

$$\Delta_r(1) - \Delta_r(0)$$

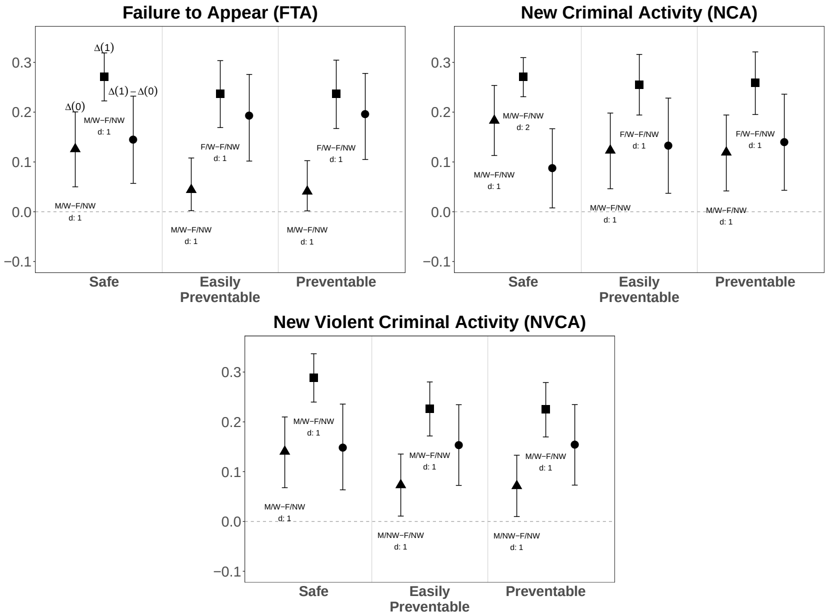
Estimated Proportion of Principal Strata



Estimated Average Principal Causal Effects



Principal Fairness



Concluding Remarks

- We offer a set of statistical methods for experimentally evaluating computer-assisted human decision making
- Application to pretrial risk assessment instrument
  - first field experiment since the 1981–82 Philadelphia experiment
  - actual empirical results will be made public in the future
- Future research
  - extension to multi-dimensional decision
  - optimal PRAI provision vs. optimal PRAI
  - effects of PRAI on judges and arrestees over time
- Papers available at
  https://imai.fas.harvard.edu/research/PRAI.html