Experimental Evaluation of Computer-Assisted Human Decision-Making: Application to Pretrial Risk Assessment Instrument

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

July 9, 2020

Abstract

Despite an increasing reliance on computerized 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 statistical models and machine learning algorithms are provided to human decision-makers in order to guide their decisions. The prevalence of such computer-assisted human decision making calls for the development of a methodological framework to evaluate its impact. Using the concept of principal stratification from the causal inference literature, we develop a statistical methodology for experimentally evaluating the causal impacts of machine recommendations on human decisions. We also show how to examine whether machine recommendations improve the fairness of human decisions. We apply the proposed methodology to the randomized evaluation of a pretrial risk assessment instrument (PRAI) in the criminal justice system. Judges use the PRAI when deciding which arrested individuals should be released and, for those ordered released, the corresponding bail amounts and release conditions. We analyze how the PRAI influences judges' decisions and impacts their gender and racial fairness.

Keywords: algorithmic fairness, causal inference, machine-recommendation, principal stratification, randomized experiments, sensitivity analysis

Note to Readers: We are not yet authorized to make our empirical results public. Thus, all of the results presented in this paper are based on a synthetic data set. Once we obtain the permission, we will post the new version of the paper with the empirical results based on the real data set.

[*] Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu URL: https://imai.fas.harvard.edu

[†] Assistant Professor, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst MA 01003. Email: zhichaojiang@umass.edu

[‡] Honorable S. William Green Professor of Public Law, Harvard Law School, 1525 Massachusetts Avenue, Griswold 504, Cambridge, MA 02138.

[§] Data Analyst, Access to Justice Lab at Harvard Law School, 1607 Massachusetts Avenue, Third Floor, Cambridge, MA 02138.

[¶] Ph.D. student, Department of Government, Harvard University, Cambridge, MA 02138. Email: sooahnshin@g.harvard.edu URL: https://sooahnshin.com

1 Introduction

A growing body of literature has suggested the potential superiority of algorithmic or machine-based decision making over purely human choices across a variety of tasks (e.g., Hansen and Hasan, 2015; He et al., 2015). Although some of this evidence is decades old (e.g., Dawes et al., 1989), it has recently gained a significant public attention by the spectacular defeats of humanity's best in cerebral games (e.g., Silver et al., 2018). Yet, even in contexts where research has warned of human frailties, we humans still make many important decisions to give ourselves agency and be held accountable for highly consequential choices.

The desire for a human decision maker as well as the precision and efficiency of machines has led to the adoption of hybrid systems involving both. By far the most popular system uses machine recommendations to inform human decision making. Such computer-assisted human decision making has been deployed in many aspects of our daily lives, including medicine, hiring, credit lending, investment decisions, and online shopping to name a few. And of particular interest, machine recommendations are increasingly of use in the realm of evidence-based public policy making. A prominent example, studied in this paper, is the use of risk assessment instruments in the criminal justice system that are designed to improve incarceration and other decisions made by judges.

In this paper, we develop a methodological framework for experimentally evaluating the impacts of machine recommendations on human decision making. Our primary goal is to assess whether machine recommendations improve human decision making. We conduct a field experiment by providing a pretrial risk assessment instrument (PRAI) to a judge who makes an initial release decision for randomly selected cases. We evaluate whether or not the PRAI helps judges make better decisions so that arrestees do not commit a new crime or fail to appear in court. This requires researchers to infer how an arrestee, who is ordered by a judge to remain in custody, would behave if released.

Using the concept of principal stratification (Frangakis and Rubin, 2002) from the causal inference literature, we propose the evaluation quantities of interest, identification assumptions, and estimation strategies. We also develop a sensitivity analysis to assess the robustness of empirical findings to the potential violation of the key identifying assumption. In addition, we consider how the data from experimental evaluation can be used to inform an optimal decision rule and how machine recommendations can be used to help humans make the optimal decision. Finally, we examine whether machine recommendations improve the fairness of human decisions, using the concept of principal fairness (Imai and Jiang, 2020). Although the proposed methodology is described and applied in the context of evaluating PRAIs, it is directly applicable or at least extendable to many other settings of computer-assisted human decision making.

PRAIs, which serve as the main application of the current paper, have played a prominent role in the literature on algorithmic fairness ever since the controversy over the potential racial bias of COMPAS risk assessment score (see Angwin et al., 2016; Dieterich et al., 2016; Flores et al., 2016; Dressel and Farid, 2018). However, with few exceptions, much of this debate focused upon the accuracy and fairness properties of the PRAI itself rather than how the PRAI affects the decisions by judges (see e.g., Berk et al., 2018; Kleinberg et al., 2018, and references therein). Even the studies that directly estimate the impacts of PRAIs on human decisions are based on either observational data or hypothetical survey questions (e.g., Miller and Maloney, 2013; Berk, 2017; Stevenson, 2018; Green and Chen, 2019). We contribute to this literature by demonstrating how to evaluate the PRAIs using a real-world field experiment. An experimental study relevant to this paper is the 1981–82 Philadelphia Bail Experiment, in which randomly selected 8 out of 16 participating judges are asked to use a new guideline to set bail amounts (Goldkamp and Gottfredson, 1984, 1985). Although this study did not evaluate the PRAI, its experimental design is somewhat similar to the one used for our study. As noted above, the methodology proposed in this paper can be applied or extended to the evaluation of other recommendation systems.

2 Experimental Evaluation of Pretrial Risk Assessment Instruments

2.1 Background

The United States criminal justice apparatus consists of thousands of diverse systems. Some are similar in the decision points they feature as an individual suspected of a crime travels from investigation to sentencing. Common decision points include whether to stop and frisk an individual in a public place, whether to arrest or issue a citation to an individual suspect of committing a crime, whether to release the arrestee while they await the disposition of any charges against them (the subject of this paper), the charge(s) to be filed against the individual, whether to find the defendant guilty of those charges, and what sentence to impose on a defendant found guilty.

At present, human beings make all of these decisions. In theory, algorithms could inform any of them, and could even make some of these decisions without human involvement. To date, algorithmic outputs have appeared most frequently in two settings: (i) at the “first appearance” hearing, during which a judge decides whether to release an arrestee pending disposition of any criminal charges, and

(ii) at sentencing, in which the judge imposes a punishment on a defendant found guilty. The first of these two motivates the present paper, but the proposed methodology is applicable or extendable to other criminal justice, legal, and non-legal settings.

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. Almost all jurisdictions allow the judge to release the arrestee with only a promise to reappear at subsequent court dates. In addition, most jurisdictions let the judge impose a money bail or bond. Since the arrestee has not yet been adjudicated guilty of any charge at the time of a predisposition hearing, there exists a consensus that predisposition incarceration is to be avoided unless the cost is sufficiently high.

Judges deciding whether to release an arrestee ordinarily consider two risk factors among a variety of other concerns; the risk that the arrestee if released will fail to appear (FTA) at subsequent court dates, and the risk that the arrestee, if released, will engage in new criminal activity (NCA) before the case is resolved (e.g., 18 U.S.C. § 3142(e)(1)). Jurisdiction laws vary regarding how these two risks are to be weighed. Some jurisdictions direct judges to consider both simultaneously along with other factors (e.g., Ariz. Const. art. II, § 22, Iowa Code § 811.2(1)(a)), while others focus on only FTA risk (e.g., N.Y. Crim. Proc. Law § 510.30(2)(a)). Despite these variations, NCA or FTA are constant and prominent in the debate over the first appearance decisions.

The concerns about the consequential nature of the first appearance decision have led to the development of PRAIs, which are ordinarily offered as inputs to first appearance judges. PRAIs can take various forms, but most focus on classifying arrestees according to FTA and NCA risks. PRAIs are generally derived by fitting a statistical model to a training dataset based on the past observations from first appearance hearings and the subsequent incidences (or lack thereof) of FTA and NCA. The hope is that providing a PRAI will improve assessment of FTA and NCA risks and thereby lead to better decisions. The goal of this paper is to develop a methodological framework for evaluating the impact of PRAIs on judges' decisions using an RCT, to which we now turn.

2.2 The Experiment

We conducted a field RCT in a county, which remains anonymous in this paper, to evaluate the impacts of a PRAI on judges' decisions. The PRAI used in our RCT depends on criminal history information, primarily prior convictions and FTA, as well as a single demographic factor, age. This PRAI purports to band arrestees based on risk of NCA/NVCA or FTA while it also provides a binary risk score for new violent criminal activity. The field operation was straightforward. In this county, a court employee assigned each matter a case number sequentially as matters entered the system. No one but this clerk was aware of the pending matter numbers, so manipulation of the number by charging assistant district attorneys was not possible. Employees of the Clerk's office scanned online record systems to calculate the PRAI for all cases. If the last digit of case number was even, these employees made the PRAI available to the judge at the first appearance hearing. Otherwise, no PRAI was made available. Thus, the provision of PRAI to judges was essentially randomized. Indeed, comparisons of distributions of observed covariates (not shown) suggests that this scheme produced groups comparable on background variables.

The judge presiding over the first appearance hearing by law was to consider risk of FTA and NCA, along with other factors as prescribed by statute. The judge could order the arrestee released with or without a bail of varying amount. The judge could also condition release on compliance with certain conditions such as certain levels of monitoring, but for the sake of simplicity we focus on bail decisions and ignore other conditions in this paper. When making decisions, the judge also had information other than the PRAI and its inputs. In all cases, the judge has a copy of an affidavit sworn to by a police officer recounting the circumstances of the incident that led to the arrest. The defense attorney sometimes informs the judge of the following regarding the arrestee's connections to the community: length of time lived there, employment there, and family living there. When available, this information ordinarily stems from an arrestee interview conducted earlier by a paralegal. The assistant district attorney sometimes provides additional information regarding circumstances of the arrest or criminal history.

The field operation design calls for approximately a 30-month treatment assignment period followed by collection of data on FTA, NCA, and other outcomes for a period of two years after randomization. As of the time of this writing, the outcome data have been collected for a period of 24 months for the arrestees who were involved in arrest events during the first 12 months. We focus on the first arrest cases in order to avoid potential spillover effects across cases. This leads to a total of 1890 cases for our analysis, of which 40.1% (38.8%) involve White (non-White) male arrestees and 13.0% (8.1%) are White (non-White) female arrestees.

The heatmaps in Figure 1 represent the distributions of three PRAIs given judges' decision among the cases in the treatment group, to which the PRAIs are provided. The PRAIs for FTA and NCA are ordinal, ranging from 1 (safest) to 6 (riskiest), whereas the PRAI for New Violent Criminal Activity (NVCA) is binary, 0 (safe) and 1 (risky). Judges' decision is an ordinal variable with three categories, based on the bail amount. Specifically, we code the cases into three categories:

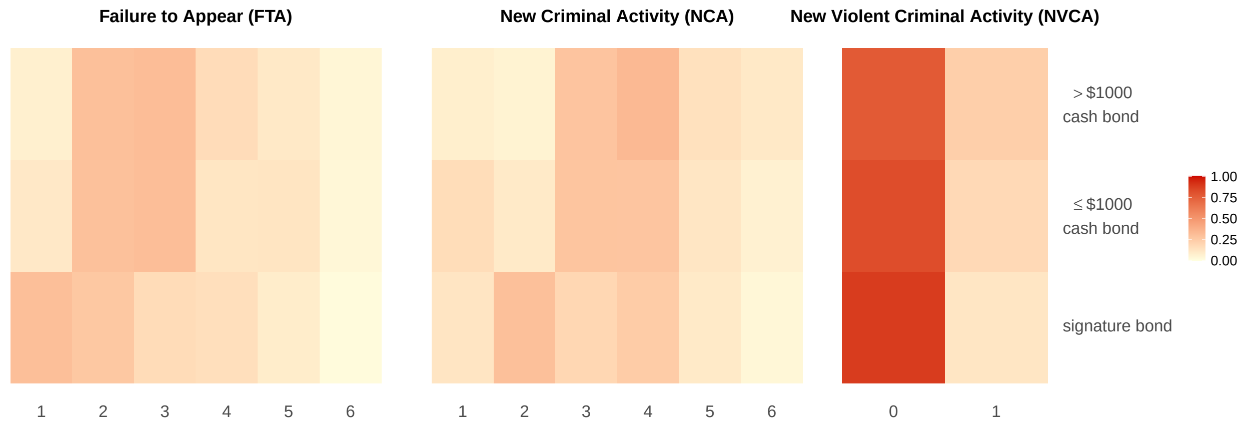

Figure 1: The Distributions of Pretrial Risk Assessment Instruments (PRAIs) given Judges' Decision among the Cases in the Treatment Group. There are three PRAIs, two of which are ordinal, ranging from 1 to 6, — Failure to Appear (FTA) and New Criminal Activity (NCA) — while the other is dichotomous — New Violent Criminal Activity (NVCA). Judges' decision is coded as a three-category ordinal variable based on the type and the amount of bail. The darker colors represent higher proportions with all proportions sum to 1 for each row.

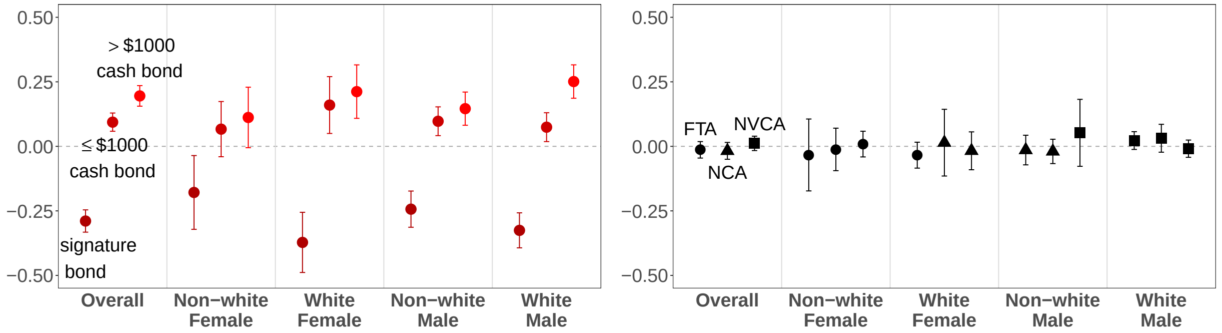

Figure 2: Estimated Average Causal Effects of the PRAI Provision on Judges' Decision and Outcome Variables (FTA, NCA, and NVCA). The results are based on the difference-in-means estimator. The vertical bars represent the 95% confidence intervals.

the signature bond,[1] the bail amount of 1,000 dollars or less, and the bail amount of greater than 1,000 dollars. The darker colors represent higher proportions, which are normalized such that they sum to 1 in each row. In general, we observe a positive association between the PRAIs and judges' decisions, implying that the cases with less amount of bail tend to have lower scores of the PRAIs. One exception is FTA, where a majority of cases with the greatest amount of cash bond receive lower risk scores.

Figure 2 presents the estimated average causal effect of the PRAI provision on judge's decision (left plot) and three outcomes of interest (right plot). We use the difference-in-means estimator and display the 95% confidence intervals as well as the point estimates. The results imply that the

[1] An arrestee is not required to pay the signature bond to be released.

PRAI provision, on average, makes it more likely for judges to impose a cash bail amount of greater than 1,000 dollars while reducing the likelihood of imposing a signature bond. These effects are consistently observed for all subgroups, but are particularly pronounced for white arrestees. The effects of the PRAI provision on FTA, NCA, and NVCA are estimated to much closer to zero with none being statistically distinguishable from zero.

Although these results are informative, we can learn much more from the data as to how the PRAI affects judges' decisions. For example, we wish to find out why the PRAI affects judges' decision and yet has no discernable effects on the arrestees' behavior. It is also of interest to know how the PRAI affects the gender and racial fairness of judges' decision. In the next section, we develop a suite of statistical methods that directly address these and other questions.

3 The Proposed Methodology

In this section, we describe the proposed methodology for experimentally evaluating the impacts of machine recommendation on human decision-making. Although we refer to our specific application throughout, the proposed methodology can be applied to other settings, in which humans make decisions using machine recommendations as an input.

3.1 The Setup

Let Z_i be a binary treatment variable indicating whether the PRAI is presented to the judge of case $i = 1, 2, \ldots, n$. We use D_i to denote the binary detention decision made by the judge to either detain ($D_i = 1$) or release ($D_i = 0$) the arrestee prior to the trial. Section 3.4 considers an extension to an ordinal decision based on bail amount. In addition, let Y_i represent the binary outcome. All of the outcomes in our application — NCA, NVCA, and FTA — are binary variables. For example, $Y_i = 1$ ($Y_i = 0$) implies that the arrestee of case i commits (does not commit) an NCA. Finally, we use $\mathbf{X}_i$ to denote a K-dimensional vector of observed pre-treatment covariates for case i. They include age, gender, race, and prior criminal history.

We adopt the potential outcomes framework of causal inference and assume the stable unit treatment value assumption (SUTVA) (Rubin, 1990). In particular, we assume no interference among cases, implying that the treatment assignment for one case does not influence the judge's decision and the outcome variable in another case. Let $D_i(z)$ be the potential value of the pretrial detention decision if case i is assigned to the treatment condition $z \in \{0, 1\}$. Furthermore, $Y_i(z, d)$ represents the potential outcome under the scenario, in which case i is assigned to the treatment condition z and the judge makes the detention decision $d \in \{0, 1\}$. Then, the observed decision is given by $D_i = D(Z_i)$ whereas the observed outcome is denoted by $Y_i = Y_i(Z_i, D_i(Z_i))$.

Throughout this paper, we maintain the following three assumptions, which we believe are reasonable in our application. First, because the treatment assignment is essentially randomized, the following conditional independence assumption is automatically satisfied.

Assumption 1 (Randomization of the Treatment Assignment)

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

for $z, d \in \{0, 1\}$ and all i.

Second, we assume that the provision of the PRAI influences the outcome only through the judge's detention decision. Because an arrestee would not care (or, perhaps, even know) whether the judge is presented with the PRAI at their first appearance, it is reasonable to assume that their behavior, be it NCA, NVCA, or FTA, is not affected directly by the treatment assignment.

Assumption 2 (Exclusion Restriction)

$$Y_i(z,d) \ = \ Y_i(z',d)$$

for $z, z', d \in \{0, 1\}$ and all i.

Under Assumption 2, we can simplify our notation by writing $Y_i(z,d)$ as $Y_i(d)$. A potential violation of this assumption is that the PRAI may directly influence the judge's decision about release conditions, which can in turn affect the outcome. The extension of the proposed methodology to multi-dimensional decisions is left for future research.

Finally, we assume that the judge's decision monotonically affects the outcome. Thus, for NCA (NVCA), the assumption implies that each arrestee is no less likely to commit a new (violent) crime if released. If FTA is the outcome of interest, this assumption implies that an arrestee is no more likely to appear in court if released. The assumption is reasonable since being held in custody of a court makes it difficult, if not impossible, to engage in NCA, NVCA, and FTA.

Assumption 3 (Monotonicity)

$$Y_i(1) \ \leq \ Y_i(0)$$

for all i.

3.2 Causal Quantities of Interest

We define causal quantities of interest using principal strata that are determined by the joint values of potential outcomes, i.e., $(Y_i(1), Y_i(0)) = (y_1, y_0)$ where $y_1, y_0 \in \{0, 1\}$ (Frangakis and Rubin, 2002). Since Assumption 3 eliminates one principal stratum, $(Y_i(1), Y_i(0)) = (1, 0)$, there are three remaining principal strata. The stratum $(Y_i(1), Y_i(0)) = (0, 1)$ consists of those who would engage in NCA (NVCA or FTA) only if they are released. We call members of this stratum as "preventable cases" because keeping the arrestee in custody would prevent the negative outcome. The stratum $(Y_i(1), Y_i(0)) = (1, 1)$ is called "risky cases," and corresponds to those who always engage in NCA (or FTA) regardless of the judge's decision. In contrast, the stratum $(Y_i(1), Y_i(0)) = (0, 0)$ represent "safe cases," in which the arrestees would never engage in NCA (NVCA or FTA) regardless of the detention decision.

We are interested in examining how the PRAI influences judges' detention decisions across different types of cases. Specifically, we focus on the following three principal causal effects,

$$\begin{aligned} \mathsf{APCEp} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 1\}, && (1) \\ \mathsf{APCEr} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 1\}, && (2) \\ \mathsf{APCEs} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 0\}. && (3) \end{aligned}$$

If the PRAI is helpful, it should make judges more likely to detain the arrestees of the preventable cases. That is, the principal causal effect on the detention decision for the preventable cases (APCEp) should be positive. In addition, the PRAI should encourage judges to release the arrestees of the safe cases, implying that the principal causal effect for the safe cases (APCEs) should be negative. The desirable direction of the principal causal effect for risky cases (APCEr) depends on various factors including the societal costs of holding the arrestees of this category in custody.

3.3 Nonparametric Identification

We consider the nonparametric identification of the principal strata effects defined above. The following theorem shows that under the aforementioned assumptions, these effects can be identified up to the marginal distributions of $Y_i(d)$ for $d = 0, 1$.

Theorem 1 (Identification) *Under Assumptions 1, 2, and 3,*

$$\begin{aligned} \mathit{APCEp} &= \frac{\Pr(Y_i = 1 \mid Z_i = 0) - \Pr(Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\} - \Pr\{Y_i(1) = 1\}}, \\ \mathit{APCEr} &= \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 1) - \Pr(D_i = 1, Y_i = 1 \mid Z_i = 0)}{\Pr\{Y_i(1) = 1\}}, \\ \mathit{APCEs} &= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{1 - \Pr\{Y_i(0) = 1\}}. \end{aligned}$$

Proof is given in Appendix S1.2. Because $\Pr\{Y_i(d)\}$ is not identifiable without additional assumptions, we cannot estimate the causal effects based on Theorem 1. However, the denominators of the expressions on the right-hand side of Theorem 1 are positive under the required assumptions.

As a result, the signs of the causal effects are identified from Theorem 1, which allows us to draw qualitative conclusions.

In addition, the theorem implies, for example, that the sign of APCEp is the opposite of the sign of the ITT effect. This is intuitive because if the provision of the PRAI increases the probability of NCA (NVCA or FTA), then the judges must release more arrestees for preventable cases. Furthermore, we can obtain the nonparametric bounds on these causal quantities by bounding $\Pr\{Y_i(d) = y\}$ that appears in the denominators. By the law of total probability,

$$\begin{aligned}\Pr\{Y_i(d) = 1\} &= \Pr\{Y_i(d) = 1 \mid D_i = d\} \Pr(D_i = d) + \Pr\{Y_i(d) = 1 \mid D_i = 1 - d\} \Pr(D_i = 1 - d) \\ &= \Pr\{Y_i = 1 \mid D_i = d\} \Pr(D_i = d) + \Pr\{Y_i(d) = 1 \mid D_i = 1 - d\} \Pr(D_i = 1 - d)\end{aligned}$$

for $d = 0, 1$. Then, the bounds on $\Pr\{Y(d) = 1\}$ are obtained by replacing $\Pr\{Y_i(d) = 1 \mid D_i = 1-d\}$ with 0 or 1.

For point identification, we consider the following unconfoundedness assumption, which states that conditional on a set of observed pre-treatment covariates $\mathbf{X}_i$ and the PRAI, the judge's detention decision is independent of the potential outcomes.

Assumption 4 (Unconfoundedness)

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

for $z = 0, 1$ and all d.

Assumption 4 holds if $\mathbf{X}_i$ contains all the information a judge has access to when making the detention decision under each treatment condition. On the other hand, if researchers do not have some of the information used by judges for their decision making, then the assumption is unlikely to be satisfied. In particular, as noted in Section 2.2, a judge may receive and use additional information regarding whether the arrestee has a job in, or family living in, the jurisdiction, or perhaps regarding the length of time the arrestee has lived in the jurisdiction. Later, we address this issue by developing a sensitivity analysis for the potential violation of Assumption 4 (see Section 3.5).

To derive the identification result, consider the following principal scores (Ding and Lu, 2017), which represent in our application the population proportion (conditional on $\mathbf{X}_i$) of preventable, risky, and safe cases, respectively,

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

Under Assumptions 2, 3, and 4, we can identify the principal scores as,

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

Thus, the identification result is given as follows,

Theorem 2 (Identification under Unconfoundedness) *Under Assumptions 1, 2, 3 and 4, APCEp, APCEr and APCEs are identified as,*

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

where

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

Proof is given in Appendix S1.2.

In some situations, we might consider the following strong monotonicity assumption instead of Assumption 3.

Assumption 5 (Strong Monotonicity)

$$Y_i(1) \;=\; 0$$

for all i.

The assumption implies, for example, that an arrestee do not commit a new crime unless released. For FTA, this assumption is plausible. This assumption may not hold for NCA or NVCA in some cases, but NCA/NVCA among incarcerated arrestees may be sufficiently rare that the violation of the assumption may not be consequential.

Under Assumption 5, the risky cases do not exist and hence the APCEr is not defined, whereas the APCEp simplifies to $\mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(0) = 1\}$. This leads to the following identification result.

Theorem 3 (Identification under Strong Monotonicity) *Under Assumptions 1, 2, and 5,*

$$\begin{aligned} \mathit{APCEp} &= \frac{\Pr(D_i = 0, Y_i = 1 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\}}, \\ \mathit{APCEs} &= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{\Pr\{Y_i(0) = 0\}}. \end{aligned}$$

Proof is given in Appendix S1.4. As in Theorem 1, the APCEp and APCEs depend on the distribution of $Y_i(0)$, which is not identifiable. However, as before, the sign of each effect is identifiable.

For point identification, we invoke the unconfoundedness assumption. Note that under the strong monotonicity assumption, Assumption 4 is equivalent to a weaker conditional independence relation concerning only one of the two potential outcomes,

$$Y_i(1) \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$$

for $z = 0, 1$. We now present the identification result.

Theorem 4 (Identification under Unconfoundedness and Strong Monotonicity) *Under Assumptions 1, 2, 4 and 5,*

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

where

$$w_P(\mathbf{x}) = \frac{1 - e_S(\mathbf{x})}{\mathbb{E}\{1 - e_S(\mathbf{X}_i)\}}, \quad w_S(\mathbf{x}) = \frac{e_S(\mathbf{x})}{\mathbb{E}\{e_S(\mathbf{X}_i)\}}.$$

Proof is straightforward and hence omitted. The identification formulas are identical to those in Theorem 2. However, with Assumption 5, we can simply compute the principal score as $e_S(\mathbf{x}) = \Pr\{Y_i = 0 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}\}$.

3.4 Ordinal Decision

We generalize the above identification results to an ordinal decision. In our application, this extension is important as the judge's release decision often is based on different amounts of money bail or varying levels of supervision of an arrestee. We first generalize the monotonicity assumption (Assumption 3) by requiring that a decision with a greater amount of bail (or a stricter level of supervision) is no less likely to make an arrestee engage in NCA (NVCA or FTA). The assumption may be reasonable, for example, because a greater amount of bail is expected to imply a greater probability of being held in custody. The assumption could be violated if arrestees experience financial strain in an effort to post bail, causing them to commit NCA (NVCA or FTA).

Formally, let D_i be an ordinal decision variable where $D_i = 0$ is the least amount of bail (or most permissive release supervision condition), and $D_i = 1, \ldots, k$ represents a bail of increasing amount (or increasingly restrictive supervision conditions), i.e., $D_i = k$ is the largest bail amount (or detention without a bail). For simplicity, we refer to D_i as the increasing amount of bail for the remainder of this paper. Then, the monotonicity assumption for an ordinal decision is given by, Assumption 6 (Monotonicity with Ordinal Decision)

$$Y_i(d_1) \ \leq \ Y_i(d_2)$$

for $d_1 \geq d_2$.

To generalize the principal strata introduced in the binary decision case, we define the decision with the least amount of bail that prevents an arrestee from committing NCA (NVCA or FTA) as follows,

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

We may view R_i as an ordinal measure of risk with a greater value indicating a higher degree of risk. Note that when D_i is binary, R_i takes one of the three values, $\{0, 1, 2\}$, representing safe, preventable, and risky cases, respectively. Thus, R_i generalizes the principal strata to the ordinal case under the monotonicity assumption.

Now, we can define the principal causal effects in the ordinal decision case. Specifically, for $r = 1, \ldots, k$ (excluding the cases with $r = 0$ and $r = k + 1$), we define the average principal causal effect of the PRAI on the judges' decision as a function of this risk measure,

$$\mathsf{APCEp}(r) \ = \ \Pr\{D_i(1) \geq r \mid R_i = r\} - \Pr\{D_i(0) \geq r \mid R_i = r\}. \tag{4}$$

Since the arrestees with $R_i = r$ would not commit NCA (NVCA or FTA) under the decision with $D_i \geq r$, the $\mathsf{APCEp}(r)$ represents a reduction in the proportion of NCA (NVCA or FTA) that is attributable to the PRAI. Thus, the expected proportion of NCA (NVCA or FTA) that would be reduced by the PRAI is given by,

$$\sum_{r=1}^{k} \mathsf{APCEp}(r) \cdot \Pr(R_i = r).$$

This quantity equals the overall ITT effect of the PRAI provision.

Furthermore, the arrestees with $R_i = 0$ would never commit a new crime regardless of the judges' decisions. Therefore, we may be interested in estimating the increase in the proportion of the lightest decision for these safest cases. This generalizes the APCEs to the ordinal decision case to the following quantity,

$$\mathsf{APCEs} \ = \ \Pr\{D_i(1) = 0 \mid R_i = 0\} - \Pr\{D_i(0) = 0 \mid R_i = 0\}.$$

For the cases with $R_i = k+1$ that would always result in a new criminal activity, a desirable decision may depend on a number of factors. Note that if we assume the strict monotonicity, i.e., $Y_i(k) = 0$ for all i, then such cases do not exist.

The identification of these principal causal effects requires the knowledge of the distribution of R_i. Fortunately, under the monotonicity and unconfoundedness assumptions (Assumptions 4 and 6), this distribution is identifiable conditional on $\mathbf{X}_i$,

$$\begin{aligned}
e_r(\mathbf{x}) &= \Pr(R_i = r \mid \mathbf{X}_i = \mathbf{x}) \\
&= \Pr(R_i \geq r \mid \mathbf{X}_i = \mathbf{x}) - \Pr(R_i \geq r+1 \mid \mathbf{X}_i = \mathbf{x}) \\
&= \Pr\{Y_i(r-1) = 1 \mid \mathbf{X}_i = \mathbf{x}\} - \Pr\{Y_i(r) = 1 \mid \mathbf{X}_i = \mathbf{x}\} \\
&= \Pr\{Y_i = 1 \mid D_i = r-1, \mathbf{X}_i = \mathbf{x}\} - \Pr\{Y_i = 1 \mid D_i = r, \mathbf{X}_i = \mathbf{x}\}, \text{ for } \ r = 1, \ldots, k, \quad (5) \\
e_0(\mathbf{x}) &= \Pr\{Y_i(0) = 0 \mid \mathbf{X}_i = \mathbf{x}\} = \Pr\{Y_i = 0 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}\}.
\end{aligned}$$

Since $e_r(\mathbf{x})$ cannot be negative for each r, this yields a set of testable conditions for Assumptions 1, 4 and 6.

Finally, we formally present the identification result for the ordinal decision case,

Theorem 5 (Identification with Ordinal Decision) *Under Assumptions 1, 2, 4 and 6, $\mathsf{APCEp}(r)$ is identified by*

$$\begin{aligned}
\mathsf{APCEp}(r) &= \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 1\} - \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 0\}, \\
\mathsf{APCEs} &= \mathbb{E}\{w_0(\mathbf{X}_i)\mathbf{1}(D_i = 0) \mid Z_i = 1\} - \mathbb{E}\{w_0(\mathbf{X}_i)\mathbf{1}(D_i = 0) \mid Z_i = 0\},
\end{aligned}$$

where $w_r(\mathbf{x}) = e_r(\mathbf{x})/\mathbb{E}\{e_r(\mathbf{X}_i)\}$.

Proof is given in Appendix S1.5.

3.5 Sensitivity Analysis

The unconfoundedness assumption, which enables the nonparametric identification of causal effects, may be violated when researchers do not observe all the information the judges have when making the detain decision. As noted in Section 2.2, many in the criminal justice community believe that judges consider variables not included in some PRAIs, such as the length of time the arrestee has lived in the community, and whether the arrestee has family and/or a job in the community. Therefore, it is important to develop a sensitivity analysis for the potential violation of Assumption 4.

We begin by proposing a sensitivity analysis for the binary decision under the strong monotonicity assumption, i.e., Assumption 5. We introduce the following sensitivity parameter $\xi(\mathbf{x})$ to characterize the deviation from the unconfoundedness assumption,

$$\xi(\mathbf{x}) = \frac{\Pr\{D_i(1) = 1 \mid Y_i(0) = 1, \mathbf{X}_i = \mathbf{x}\}}{\Pr\{D_i(1) = 1 \mid Y_i(0) = 0, \mathbf{X}_i = \mathbf{x}\}},$$

which is equal to 1 for all $\mathbf{x}$ when the unconfoundedness assumption holds.

For a given value of $\xi(\mathbf{x})$, we have

$$\begin{aligned}
\Pr\{D_i(1)=1, Y_i(0)=0 \mid \mathbf{X}_i=\mathbf{x}\} &= \Pr\{D_i(1)=1 \mid Y_i(0)=0, \mathbf{X}_i=\mathbf{x}\}\Pr\{Y_i(0)=0 \mid \mathbf{X}_i=\mathbf{x}\}, \\
\Pr\{D_i(1)=1, Y_i(0)=1 \mid \mathbf{X}_i=\mathbf{x}\} &= \Pr\{D_i(1)=1 \mid Y_i(0)=1, \mathbf{X}_i=\mathbf{x}\}\Pr\{Y_i(0)=1 \mid \mathbf{X}_i=\mathbf{x}\} \\
&= \xi(\mathbf{x}) \cdot \Pr\{D_i(1)=1 \mid Y_i(0)=0, \mathbf{X}_i=\mathbf{x}\}\Pr\{Y_i(0)=1 \mid \mathbf{X}_i=\mathbf{x}\}.
\end{aligned}$$

Solving these equations yields,

$$\Pr\{Y_i(0)=1 \mid \mathbf{X}_i=\mathbf{x}\} = \frac{\Pr(Y_i=1, D_i=1 \mid Z_i=1, \mathbf{X}_i=\mathbf{x})}{\xi(\mathbf{x}) \cdot \Pr(Y_i=0, D_i=1 \mid Z_i=1, \mathbf{X}_i=\mathbf{x}) + \Pr(Y_i=1, D_i=1 \mid Z_i=1, \mathbf{X}_i=\mathbf{x})}.$$

By using this result and the expressions in Theorem 3, we can identify the APCEp and APCEs with a given value of the sensitivity parameter $\xi(\mathbf{x})$. Similarly, We can conduct a sensitivity analysis for the binary decision under the monotonicity assumption, i.e., Assumption 3, which requires more sensitivity parameters.

Next, we consider the ordinal decision case. A nonparametric sensitivity analysis is difficult to develop in this case, and hence we propose a parametric sensitivity analysis. We consider the following bivariate ordinal probit model for the observed judge's decision D and the latent risk measure R_i,

$$\begin{aligned}
D_i^*(z) &= \beta_Z z + \mathbf{X}_i^\top \beta_X + Z_i \mathbf{X}_i^\top \beta_{ZX} + \epsilon_{i1}, && (6) \\
R_i^* &= \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2}, && (7)
\end{aligned}$$

where

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and

$$D_i(z) = \begin{cases} 0 & D^*(z) \leq \theta_{z1} \\ 1 & \theta_{z1} < D_i^*(z) \leq \theta_{z2} \\ \vdots & \vdots \\ k-1 & \theta_{z,k-1} < D_i^*(z) \leq \theta_{zk} \\ k & \theta_{zk} < D_i^*(z) \end{cases}, \quad R_i = \begin{cases} 0 & R_i^* \leq \delta_1 \\ 1 & \delta_1 < R_i^* \leq \delta_2 \\ \vdots & \vdots \\ k & \delta_k < R_i^* \leq \delta_{k+1} \\ k+1 & \delta_{k+1} < R_i^* \end{cases}.$$

Because $(\epsilon_{i1}, \epsilon_{i2})$ follow bivariate normal distributions, (D_i, R_i) follow a bivariate ordinal probit models. In the literature, Frangakis et al. (2002), Barnard et al. (2003), and Forastiere et al. (2016) also model the distribution of principal strata using the ordinal probit model.

Under this model, ρ represents a sensitivity parameter since Assumption 4 implies $\rho = 0$. If the value of ρ is known, then the other coefficients, i.e., β_X, α_X and β_Z, can be estimated, which in turn leads to the identification of the $\mathsf{APCEp}(r)$ and APCEs. Because R_i is a latent variable, the estimation of this model is not straightforward. In our empirical application, we conduct a Bayesian analysis to estimate the causal effects (see e.g., Hirano et al., 2000; Schwartz et al., 2011; Mattei et al., 2013; Jiang et al., 2016, for other applications of Bayesian sensitivity analysis). Appendix S2 presents the details of the Bayesian estimation.

3.6 Optimal Decision Rule

The discussion so far has focused on estimating the impacts of machine recommendations on human decisions. However, the experimental evaluation considered here can also shed light on how humans may make an optimal decision given a certain objective. In our application, suppose that the goal is to prevent as many NCAs (NVCAs or FTAs) as possible with the minimal amount of bail (or minimally restrictive monitoring conditions). We show how the experiment (or unconfoundedness assumption) can help identify the optimal decisions to best achieve this goal.

Formally, let δ be the judge's decision based on $\mathbf{X}_i$, which may include the PRAI. Thus, $\delta(\mathbf{x}) = d$ if $\mathbf{x} \in \mathcal{X}_d$ where $\mathcal{X}_d$ is a non-overlapping partition of the covariate space $\mathcal{X}$ with $\mathcal{X} = \bigcup_{r=0}^{k} \mathcal{X}_r$ and $\mathcal{X}_r \cap \mathcal{X}_{r'} = \emptyset$. We consider the 0–1 utility function, i.e., $\mathbf{1}\{\delta(\mathbf{X}_i) = R_i\}$. This implies that the optimal decision for a safe case is $\delta(\mathbf{X}_i) = k$, i.e., the least amount of bail (or the most lenient monitoring condition). We ignore the cases with $R_i = k + 1$ by assuming that their utilities are zero because those cases result in the negative outcome regardless of the judges' decisions. It is possible to use other utility functions. For example, one alternative is $\mathbf{1}\{\delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i) \geq R_i\}$, which implies that we do not require the decision to be the least stringent so long as it can prevent NCA (NVCA or FTA).

We derive the optimal decision rule δ^* that maximizes the expected utility,

$$\delta^* \quad = \quad \underset{\delta}{\operatorname{argmax}}\, U(\delta) \quad \text{where} \quad U(\delta) \; = \; \mathbb{E}[\mathbf{1}\{\delta(\mathbf{X}_i) = R_i\}].$$

For $r = 0, \ldots, k$, we can write

$$\mathbb{E}[\mathbf{1}\{\delta(\mathbf{X}_i) = R_i = r\}] \quad = \quad \mathbb{E}[\mathbf{1}\{\mathbf{X}_i \in \mathcal{X}_r, R_i = r\}] \; = \; \mathbb{E}\left[\mathbf{1}\{\mathbf{X}_i \in \mathcal{X}_r\} \cdot e_r(\mathbf{X}_i)\right].$$

Therefore,

$$U(\delta) \quad = \quad \sum_{r=0}^{k} \mathbb{E}[\mathbf{1}\{\delta(\mathbf{X}_i) = r, R_i = r\}] \; = \; \sum_{r=0}^{k} \mathbb{E}\left\{\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_r) \cdot e_r(\mathbf{X}_i)\right\}.$$

This yields the following optimal decision,

$$\delta^*(\mathbf{x}) = \underset{r\in\{0,1,\ldots,k\}}{\operatorname{argmax}} \ e_r(\mathbf{x}),$$

where the optimal expected utility is given by,

$$\mathbb{E}\left[\max\left\{e_0(\mathbf{X}_i), \ldots, e_k(\mathbf{X}_i)\right\}\right].$$

Therefore, we can use the experimental estimate $e_r(\mathbf{x})$ to inform the judge's decision.

3.7 Optimal PRAI Provision Rule

Policy makers could, for example, use the decision rule derived above to encourage the judge to make an optimal decision. However, this may not be useful if the judge decides not to follow the recommendation or only partially do so for some cases. We next consider the optimal provision of the PRAI given the same goal considered above (i.e., prevent as many NCAs (NVCAs or FTAs) as possible with the minimal amount of bail). That is, we investigate the settings, in which policy makers can decide whether to provide the judge with the PRAI, depending on the case characteristics. Unfortunately, we cannot derive an optimal PRAI itself (as opposed to its optimal provision) unless the PRAI can be directly randomized.

Let ξ be a PRAI provision rule, i.e., $\xi(\mathbf{x}) = 1$ (the PRAI is provided) if $\mathbf{x} \in \mathcal{B}_1$ and $\xi(\mathbf{x}) = 0$ (the PRAI is not provided) if $\mathbf{x} \in \mathcal{B}_0$, where $\mathcal{X} = \mathcal{B}_0 \bigcup \mathcal{B}_1$ and $\mathcal{B}_0 \cap \mathcal{B}_1 = \emptyset$. The judges will make their decisions based on the PRAI and other available information included in $\mathbf{X}_i = \mathbf{x}$. To consider the influence of the PRAI on judges' decision, we define δ_{i1} the potential decision rule of case i if the judge received the PRAI and δ_{i0} if not. Thus, $\delta_{iz}(\mathbf{x}) = d$ if $\mathbf{x} \in \mathcal{X}_{i,zd}$ where $\mathcal{X}_{i,zd}$ is a partition of the covariate space with $\mathcal{X} = \bigcup_{d=0}^{k} \mathcal{X}_{i,zd}$ and $\mathcal{X}_{i,zd} \cap \mathcal{X}_{i,zd'} = \emptyset$. Although we allow the judge to make a different decision even if the observed case characteristics $\mathbf{X}_i$ are identical, we assume that the judges' decisions are identically distributed given the observed case characteristics and the PRAI provision. That is, we assume $\Pr\{\delta_{iz}(\mathbf{x}) = d\} = \Pr\{\delta_{i'z}(\mathbf{x}) = d\}$ for fixed $\mathbf{x}$, z and $i \neq i'$, where the probability is taken with respect to the super population of all cases.

Given this setup, we can derive the optimal PRAI provision rule. As before, we consider the 0–1 utility $\mathbf{1}\{\delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i) = R_i\}$. This utility equals one, if the judge makes the most lenient decision to prevent an arrestee from engaging in NCA (NVCA or FTA), and equals zero otherwise. As before, we begin by rewriting the utility in the following manner,

$$U(\xi) \quad = \quad \mathbb{E}\left[\mathbf{1}\{R_i = \delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i)\}\right]$$

$$\begin{aligned} &= \sum_{r=0}^{k} \mathbb{E}\left[\mathbf{1}\{R_i = r, \delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i) = r\}\right] \\ &= \sum_{r=0}^{k}\sum_{z=0}^{1} \mathbb{E}[\mathbf{1}\{R_i = r, \delta_{iz}(\mathbf{X}_i) = r, \mathbf{X}_i \in \mathcal{B}_z\}]. \end{aligned}$$

We can write,

$$\begin{aligned} \mathbb{E}[\mathbf{1}\{R_i = r, \delta_{iz}(\mathbf{X}_i) = r, \mathbf{X}_i \in \mathcal{B}_z\}] &= \mathbb{E}[\Pr(R_i = r \mid \mathbf{X}_i) \cdot \Pr\{\delta_{iz}(\mathbf{X}_i) = r \mid \mathbf{X}_i\} \cdot \mathbf{1}\{\mathbf{X}_i \in \mathcal{B}_z\}] \\ &= \mathbb{E}[e_r(\mathbf{X}_i) \cdot \Pr\{\delta_{iz}(\mathbf{X}_i) = r\} \cdot \mathbf{1}\{\mathbf{X}_i \in \mathcal{B}_z\}]. \end{aligned}$$

Because in the experiment, the provision of PRAI is randomized, we can estimate $\Pr\{\delta_{iz}(\mathbf{X}_i) = r\} = \Pr(D_i = r \mid Z_i = z, \mathbf{X}_i)$ from the data. Therefore, we obtain

$$U(\xi) = \sum_{z=0,1} \mathbb{E}\left(\left[\sum_{r=0}^{k} e_r(\mathbf{X}_i) \cdot \Pr(D_i = r \mid Z_i = z, \mathbf{X}_i)\right] \cdot \mathbf{1}\{\mathbf{X}_i \in \mathcal{B}_z\}\right).$$

Then, the optimal PRAI provision rule is,

$$\xi(\mathbf{x}) = \operatorname*{argmax}_{z} h_z(\mathbf{x}) \quad \text{where} \quad h_z(\mathbf{x}) = \sum_{r=0}^{k} e_r(\mathbf{x}) \cdot \Pr(D_i = r \mid Z_i = z, \mathbf{X}_i). \tag{8}$$

Thus, we can use the experimental data to derive the optimal PRAI provision rule.

3.8 Principal Fairness

Finally, we discuss how the causal effects discussed above relate to the fairness of decision. In particular, Imai and Jiang (2020) introduce the concept of "principal fairness." The basic idea is that within each principal strata a fair decision should not depend on protected attributes (race, gender, etc.). Imai and Jiang (2020) provide a detailed discussion about how principal fairness differs from the existing definitions of fairness, which are based on the predictive accuracy (see also Corbett-Davies et al., 2017; Chouldechova and Roth, 2018, and references therein).

Formally, let $A_i \in \mathcal{A}$ be a protected attribute such as race and gender. We first consider a binary decision. We say that decisions are fair on average with respect to A_i if it does not depend on the attribute within each principal strata, i.e.,

$$\Pr(D_i = 1 \mid A_i, Y_i(1) = y_1, Y_i(0) = y_0) = \Pr(D_i = 1 \mid Y_i(1) = y_1, Y_i(0) = y_0) \tag{9}$$

for all $y_1, y_0 \in \{0, 1\}$. We can generalize this definition to the ordinal case as,

$$\Pr(D_i \geq d \mid A_i, R_i = r) = \Pr(D_i \geq d \mid R_i = r)$$

for $1 \leq d \leq k$ and $0 \leq r \leq k$. The degree of fairness for principal stratum $R_i = r$ can be measured as,

$$\Delta_r(z) \; = \; \max_{a,a',d} \left| \Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} \; - \; \Pr\{D_i(z) \geq d \mid A_i = a', R_i = r\} \right| \qquad (10)$$

for $z = 0, 1$.

By estimating $\Delta_r(z)$, we can use the experimental data to examine whether or not the provision of the PRAI improves the fairness of judge's decision. Specifically, the PRAI improves the fairness of judge's decision for the principal stratum r if $\Delta_r(1) \leq \Delta_r(0)$.

4 Empirical Results

We apply the proposed methodology to the synthetic data based on the field RCT described in Section 2. We use the ordinal decision variable with three categories; the signature bond ($D_i = 0$), the bail amount of \$1,000 or less ($D_i = 1$), and the bail amount of greater than \$1,000 ($D_i = 2$). Given this ordinal decision, we call the principal strata as safe ($R_i = 0$), easily preventable ($R_i = 1$), preventable ($R_i = 2$), and risky cases ($R_i = 3$). None of the estimated proportions of principal strata is negative (see equation (5)), suggesting that the monotonicity assumption (Assumption 6) may be appropriate in this application.

We fit the Bayesian model defined in equations (6) and (7) with a diffuse prior,[2] separately for each of three binary outcome variables — FTA, NCA, and NVCA. The model incorporates following pre-treatment covariates: sex (male or female), race (white or non-white), age, age at current arrest, and several indicator variables regarding the current and past charges.[3] We use the Gibbs sampler described in Appendix S2 and run five Markov chains of 50,000 iterations each with random starting values independently drawn from the prior distribution. Based on the Gelman-Rubin statistic for convergence diagnostics, we retain the second half of each chain and combine them to be used for our analysis.

4.1 Estimated Average Principal Causal Effects of the PRAI Provision

We begin by presenting the estimated population proportion of each principal stratum (see equation (5)). Figure 3 shows that the overall proportion of safe cases (blue circles) is estimated to be 82%, whereas those of easily preventable (black triangles) and preventable (red squares) cases are less than 10% and 1%, respectively. This finding is consistent across all outcomes and all subgroups.

[2] See Appendix S2 for the prior specification.

[3] They include the presence of current violent offense, pending charge at time of offense, prior misdemeanor conviction, prior violent conviction, and prior sentence to incarceration.

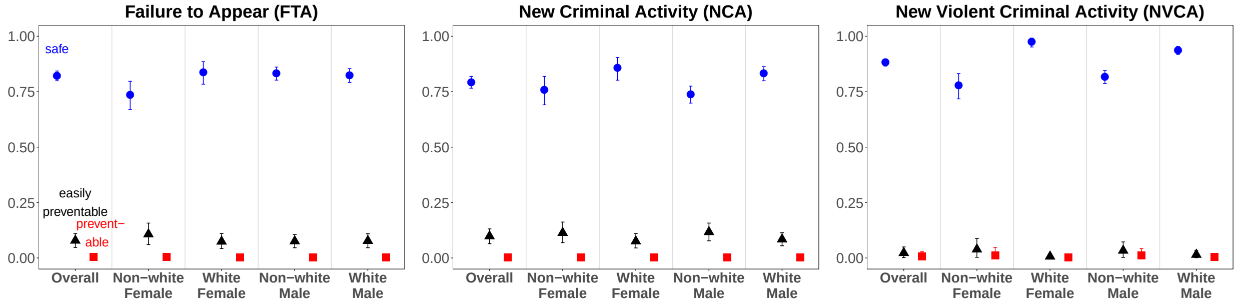

Figure 3: Estimated Population Proportion of Each Principal Stratum. Each panel represents the result using three different outcome variables (FTA, NCA, and NVCA). In each column, the blue circle, black triangle, and red square represent the estimates for safe (blue circles), easily preventable (black triangles), and preventable (red squares) cases, respectively. These three estimates do not sum to one because there is an additional principal stratum of risky cases that represents a group of arrestees who will commit a new FTA (or NCA/NVCA) regardless of judges' decisions. The solid vertical lines represent 95% Bayesian credible intervals.

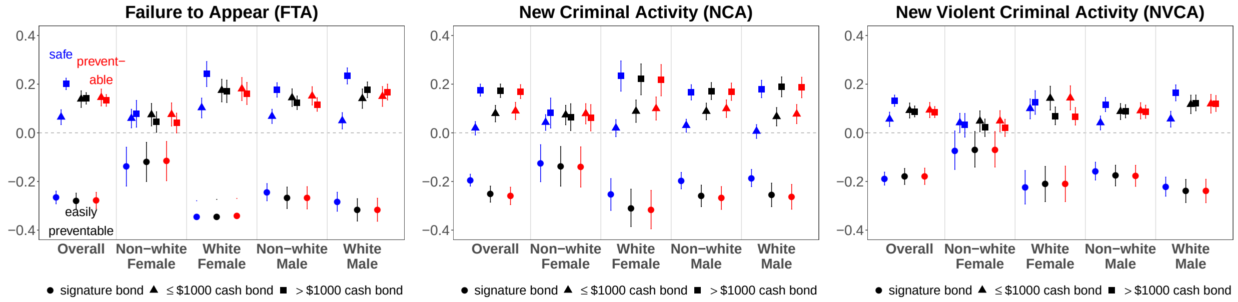

Figure 4: Estimated Average Principal Causal Effects (APCE) of the PRAI Provision on Judges' Decision. Each plot presents the overall and subgroup-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PRAI provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge the signature bond (circles), the cash bail amount of 1,000 dollars or less (triangles), and the cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the Bayesian 95% credible interval.

In most cases, the estimated proportion of safe cases is smaller among non-white arrestees than white arrestees. Given its rarity, it is not surprising that NVCA has the smallest estimated proportion of safe cases.

Next, Figure 4 presents the estimated APCE of the PRAI provision on the three ordinal decision categories, separately for each of the three outcomes (see equation (4)). The overall and subgroup-specific results are given for each of the three principal strata — safe (blue), easily preventable (black), and preventable (red) cases. The figure shows that the provision of the PRAI tends to make judges' decisions stricter across all principal strata and subgroups. This is indicated by the fact that the estimated APCE is negative for signature bond (circles) whereas it is positive and of the greatest

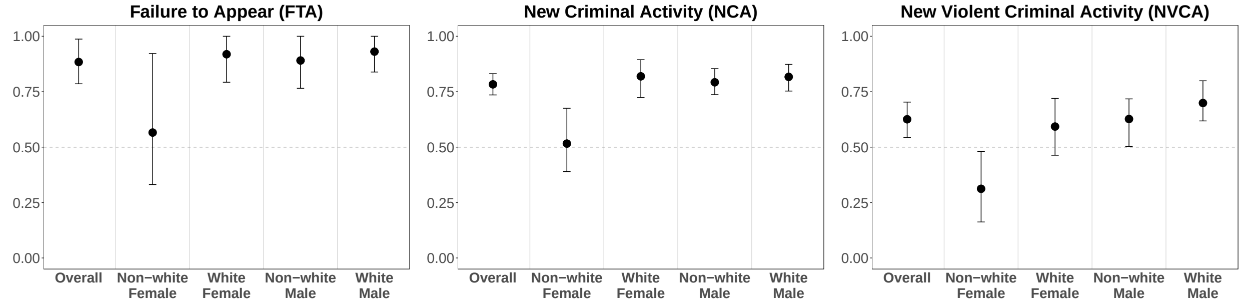

Figure 5: Estimated Optimal PRAI Provision Rule. For each different outcome variable, we present the estimated population proportion of cases, for which providing the PRAI is optimal. The overall and subgroup-specific estimates are reported. For example, according to the overall estimate for FTA, for of 88% of cases the PRAI provision would encourage judges to give the most lenient decision that prevents an arrestee from engaging in FTA. The vertical lines represent 95% confidence interval of the estimates.

magnitude for the cash bound of over \$1,000 (squares). The effect sizes vary across subgroups and outcomes, but they tend to be the largest for white females and the smallest for non-white females.

4.2 Optimal PRAI Provision Rule

We estimate the optimal PRAI provision rule derived in Section 3.7. Figure 5 presents the estimated population proportion of the cases where the PRAI provision is optimal (see equation (8)). Overall, the results show that the PRAI provision is optimal for a majority of cases. In other words, for about 88% (78% or 63%) of cases, the PRAI provision would encourage judges to make the most lenient decision that prevents an arrestee from engaging in FTA (NCA or NVCA). All the subgroups show similar results except for non-white females. For these arrestees, the PRAI provision is optimal only for about 57% (52% or 31%) of cases. This is consistent with the finding in Section 4.1 that the PRAI provision decreases only a small proportion of FTAs (NCA/NVCA) engaged by non-white females.

4.3 Principal Fairness of the PRAI Provision

Finally, we evaluate the principal fairness of the PRAI provision as discussed in Section 3.8. We use race and gender as protected attributes, and analyze the four subgroups defined by those variables. Figure 6 presents the results for two principal strata (easily preventable and preventable cases) and separately for each of the three outcomes. Each column presents $\Delta_r(z)$ defined in equation (10), representing the maximal subgroup difference in the judges' decision probability within the same principal stratum $R_i = r$ under the provision of PRAI $z = 1$ (no provision $z = 0$). We also present the estimated difference of the two $\Delta_r(1) - \Delta_r(0)$. If this difference is estimated to be positive, then

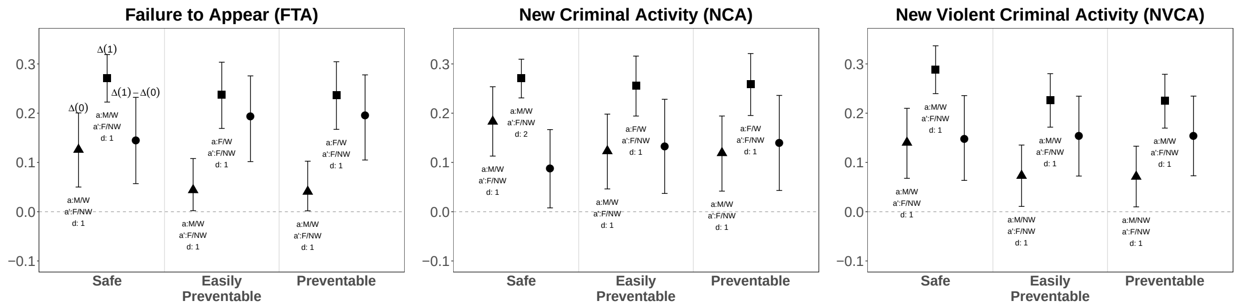

Figure 6: The PRAI Provision and Principal Fairness when Gender and Race are Protected Attributes. Each panel presents the results for a different outcome; males and females are represented by M and F whereas whites and non-whites are indicated by W and NW. Within each plot, we show three estimates separately for each principal stratum r (easily preventable and preventable cases) — $\Delta_r(z)$ representing the maximal subgroup difference in the judges' decision probability with ($z = 1$; squares) and without ($z = 0$; triangles) the PRAI provision. A positive value of the difference (circles), i.e., $\Delta_r(1) - \Delta_r(0) > 0$, implies that the PRAI reduces the fairness of judges' decision. The vertical solid lines represent 95% Bayesian credible intervals.

the PRAI provision reduces the fairness by increasing the maximal subgroup difference.

We find that the PRAI provision indeed has a negative impact on the principal fairness of judges' decisions for all outcomes and across allq principal strata. For example, in the principal stratum of safe cases, we find that when the PRAI is provided, the maximal difference in the judges' decision probability is greater than that without the PRAI provision. For safe cases, for example, the difference between white males and non-white females is the greatest across all outcomes with white males receiving less lenient decisions than non-white females, and the PRAI provision further widens this gap. Thus, unfortunately, the PRAI appears to reduce the fairness of judges' decisions with respect to gender and race in our synthetic data set.

5 Concluding Remarks

In today's data-rich society, many human decisions are guided by machine-recommendations. While some of these computer-assisted human decisions may be trivial and routine (e.g., online shopping and movie suggestions), others that are much more consequential include judicial and medical decision-making. As algorithmic recommendation systems play increasingly important roles in our lives, we believe that it is important to empirically evaluate the impacts of such systems on human decisions. In this paper, we present a set of new statistical methods that can be used for the experimental evaluation of computer-assisted human decision making. We applied these methods to a synthetic data based on an original randomized experiment for evaluating the impacts of the PRAI on judges' pretrial decisions. Our findings suggest that the PRAI only marginally reduces failures to appear or new crimes while the judges' decisions are fairer without the PRAI than with it.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the American Statistical Association 98*(462), 299–323.

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology 13*, 193–216.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*. http://doi.org/10.1177/0049124118782533.

Chouldechova, A. and A. Roth (2018). The frontiers of fairness in machine learning. Technical report, arXiv:1810.08810.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. In *KDD'17*, August 13–17, 2017, Halifax, NS, Canada.

Dawes, R. M., D. Faust, and P. E. Meehl (1989). Clinical versus actuarial judgment. *Science 243*(4899), 1668–1674.

Dieterich, W., C. Mendoza, and T. Brennan (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf. Northpointe Inc. Research Department.

Ding, P. and J. Lu (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 79*(3), 757–777.

Dressel, J. and H. Farid (2018, January). The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*(1), eaao5580.

Flores, A. W., K. Bechtel, and C. Lowenkamp (2016, September). False positives, false negatives, and false analyses: A rejoinder to ”machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks.”. *Federal Probation Journal 80*(2), 28–46.

Forastiere, L., F. Mealli, and T. J. VanderWeele (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using bayesian principal stratification. *Journal of the American Statistical Association 111*(514), 510–525.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Frangakis, C. E., D. B. Rubin, and X.-H. Zhou (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics 3*(2), 147–164.

Goldkamp, J. S. and M. R. Gottfredson (1984). *Judicial Guidelines for Bail: The Philadelphia Experiment.* Washington D.C.: U.S. Department of Justice, National Institute of Justice.

Goldkamp, J. S. and M. R. Gottfredson (1985). *Policy Guidelines for Bail: An Experiment in Court Reform.* Temple University Press.

Green, B. and Y. Chen (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *FAT* ’19: Conference on Fairness, Accountability, and Transparency (FAT* ’19)*, January 29–31, 2019, Atlanta, GA, USA, pp. 90–99.

Hansen, J. H. L. and T. Hasan (2015, November). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine 32*(6), 74–99.

He, K., X. Zhang, S. Ren, and J. Sun (2015, December). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics 1*(1), 69–88.

Imai, K. and Z. Jiang (2020). Principal fairness for human and algorithmic decision-making. *Working paper available at https://imai.fas.harvard.edu/research/fairness.html.*

Jiang, Z., P. Ding, and Z. Geng (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(4), 829–848.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018, January). Human decisions and machine predictions. *Quarterly Journal of Economics 133*(1), 237–293.

Mattei, A., F. Li, F. Mealli, et al. (2013). Exploiting multiple outcomes in bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics 7*(4), 2336–2360.

Miller, J. and C. Maloney (2013, July). Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Criminal Justice and Behavior 40*(7), 716–736.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science 5*, 472–480.

Schwartz, S. L., F. Li, and F. Mealli (2011). A bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association 106*(496), 1331–1344.

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabi (2018, December). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science 362*(6419), 1140–1144.

Stevenson, M. (2018). Assessing risk assessment in action. *Minnesota Law Review*, 303–384.

Supplementary Appendix

S1 Proofs of the Theorems

S1.1 Lemmas

To prove the theorems, we need some lemmas.

Lemma S1 *Consider two random variables X and Y with finite moments. Let $f_1(x)$ and $f_2(y)$ be their density functions. Then, any function $g(\cdot)$*

$$\mathbb{E}\{g(X)\} = \mathbb{E}\left\{\frac{f_1(Y)}{f_2(Y)} g(Y)\right\}.$$

Proof is straightforward and hence omitted.

Lemma S2 *For a binary decision, Assumption 4 implies $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$ under Assumption 3. For a discrete decision, Assumption 4 implies $R_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$ under Assumption 6.*

Proof of Lemma S2. For a binary decision, we have

$$\begin{aligned}
\Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid D_i, \mathbf{X}_i, Z_i = z\} &= \Pr\{Y_i(1) = 1 \mid D_i, \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(1) = 1 \mid \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid \mathbf{X}_i, Z_i = z\},
\end{aligned}$$

where the first and third equality follow from Assumption 3 and the second equality follows from Assumption 4. Similarly, we have

$$\begin{aligned}
\Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid D_i, \mathbf{X}_i, Z_i = z\} &= \Pr\{Y_i(0) = 0 \mid D_i, \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(0) = 0 \mid \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid \mathbf{X}_i, Z_i = z\},
\end{aligned}$$

where the first and third equality follow from Assumption 3 and the second equality follows from Assumption 4. As a result, $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$ because $\{Y_i(1), Y_i(0)\}$ takes only three values.

For a discrete decision D_i taking values in $\{0, \ldots, k\}$, we have

$$\begin{aligned}
\Pr(R_i = r \mid D_i, \mathbf{X}_i, Z_i = z) &= \Pr(R_i \geq r \mid D_i, \mathbf{X}_i, Z_i = z) - \Pr(R_i \geq r + 1 \mid D_i, \mathbf{X}_i, Z_i = z) \\
&= \Pr(Y_i(r-1) = 1 \mid D_i, \mathbf{X}_i, Z_i = z) - \Pr(Y_i(r) = 1 \mid D_i, \mathbf{X}_i, Z_i = z) \\
&= \Pr(Y_i(r-1) = 1 \mid \mathbf{X}_i, Z_i = z) - \Pr(Y_i(r) = 1 \mid \mathbf{X}_i, Z_i = z) \\
&= \Pr(R_i \geq r \mid \mathbf{X}_i, Z_i = z) - \Pr(R_i \geq r + 1 \mid \mathbf{X}_i, Z_i = z) \\
&= \Pr(R_i = r \mid D_i, \mathbf{X}_i, Z_i = z),
\end{aligned}$$

where the second and the fourth equality follow from the definition of R_i and the third equality follows from Assumption 4. As a result, $R_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$. □

S1.2 Proof of Theorem 1

First, Assumption 3 implies,

$$\begin{aligned}\Pr\{Y_i(0)=0,Y_i(1)=0\} &= \Pr\{Y_i(0)=0\}, \quad \Pr\{Y_i(0)=1,Y_i(1)=1\}=\Pr\{Y_i(1)=1\},\\ \Pr\{Y_i(0)=1,Y_i(1)=0\} &= 1-\Pr\{Y_i(0)=0\}-\Pr\{Y_i(1)=1\}.\end{aligned}$$

Second, we have

$$\begin{aligned}&\Pr\{D_i(z)=1,Y_i(0)=0,Y_i(1)=0\}\\ =\;&\Pr\{Y_i(0)=0,Y_i(1)=0\}-\Pr\{D_i(z)=0,Y_i(0)=0,Y_i(1)=0\}\\ =\;&\Pr\{Y_i(0)=0\}-\Pr\{D_i(z)=0,Y_i(0)=0\}\\ =\;&\Pr\{Y_i(0)=0\}-\Pr\{D_i(z)=0,Y_i(D_i(z))=0\mid Z_i=z\}\\ =\;&\Pr\{Y_i(0)=0\}-\Pr(D_i=0,Y_i=0\mid Z_i=z),\end{aligned}$$

where the second equality follows from Assumption 3 and the third equality follows from Assumption 1. Similarly, we can obtain

$$\begin{aligned}\Pr\{D_i(z)=1,Y_i(0)=1,Y_i(1)=1\} &= \Pr\{D_i(z)=1,Y_i(1)=1\}\\ &= \Pr\{D_i(z)=1,Y_i(D_i(z))=1\mid Z_i=z\}\\ &= \Pr(D_i=1,Y_i=1\mid Z_i=z).\end{aligned}$$

Therefore,

$$\begin{aligned}&\Pr\{D_i(z)=1,Y_i(0)=1,Y_i(1)=0\}\\ =\;&\mathrm{pr}\{D_i(z)=1\}-\Pr\{D_i(z)=1,Y_i(0)=0,Y_i(1)=0\}-\Pr\{D_i(z)=1,Y_i(0)=1,Y_i(1)=1\}\\ =\;&\mathrm{pr}\{D_i=1\mid Z_i=z\}-\Pr\{Y_i(0)=0\}+\Pr(D_i=0,Y_i=0\mid Z_i=z)-\Pr(D_i=1,Y_i=1\mid Z_i=z)\\ =\;&\Pr(Y_i=0\mid Z_i=z)-\Pr\{Y_i(0)=0\}.\end{aligned}$$

Finally, we have,

$$\begin{aligned}\mathsf{APCEp} &= \frac{\Pr\{D_i(1)=1,Y_i(0)=1,Y_i(1)=0\}-\Pr\{D_i(0)=1,Y_i(0)=1,Y_i(1)=0\}}{\Pr\{Y_i(0)=1,Y_i(1)=0\}}\\ &= \frac{\Pr(Y_i=1\mid Z_i=0)-\Pr(Y_i=1\mid Z_i=1)}{\Pr\{Y_i(0)=1\}-\Pr\{Y_i(1)=1\}},\end{aligned}$$

$$\begin{aligned}\mathsf{APCEr} &= \frac{\Pr\{D_i(1)=1,Y_i(0)=1,Y_i(1)=1\}-\Pr\{D_i(0)=1,Y_i(0)=1,Y_i(1)=0\}}{\Pr\{Y_i(0)=1,Y_i(1)=1\}}\\ &= \frac{\Pr(D_i=1,Y_i=1\mid Z_i=1)-\Pr(D_i=1,Y_i=1\mid Z_i=0)}{\Pr\{Y_i(1)=1\}},\end{aligned}$$

and

$$\begin{aligned}\mathsf{APCEs} &= \frac{\Pr\{D_i(1)=1,Y_i(0)=0,Y_i(1)=0\}-\Pr\{D_i(0)=1,Y_i(0)=0,Y_i(1)=0\}}{\Pr\{Y_i(0)=0\}}\\ &= \frac{\Pr(D_i=0,Y_i=0\mid Z_i=0)-\Pr(D_i=0,Y_i=0\mid Z_i=1)}{\Pr\{Y_i(0)=0\}}.\end{aligned}$$

□

S1.3 Proof of Theorem 2

Assumption 4 and Lemma S2 imply,

$$\begin{aligned}\mathbb{E}\{D_i(z) \mid Y_i(1)=y_1, Y_i(0)=y_0\} &= \mathbb{E}\left[\mathbb{E}\{D_i(z) \mid \mathbf{X}_i, Y_i(1)=y_1, Y_i(0)=y_0\} \mid Y_i(1)=y_1, Y_i(0)=y_0\right] \\ &= \mathbb{E}\left[\mathbb{E}\{D_i(z) \mid \mathbf{X}_i\} \mid Y_i(1)=y_1, Y_i(0)=y_0\right].\end{aligned}$$

Based on Lemma S1,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\{D_i(z) \mid \mathbf{X}_i\} \mid Y_i(1)=y_1, Y_i(0)=y_0\right] \\ =\;& \mathbb{E}\left[\frac{\Pr\{\mathbf{X}_i \mid Y_i(1)=y_1, Y_i(0)=y_0\}}{\Pr(\mathbf{X}_i)}\mathbb{E}\{D_i(z) \mid \mathbf{X}_i\}\right] \\ =\;& \mathbb{E}\left(\mathbb{E}\left[\frac{\Pr\{\mathbf{X}_i \mid Y_i(1)=y_1, Y_i(0)=y_0\}}{\Pr(\mathbf{X}_i)}D_i(z)\middle|\mathbf{X}_i\right]\right) \\ =\;& \mathbb{E}\left(\mathbb{E}\left[\frac{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0 \mid \mathbf{X}_i\}}{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0\}}D_i(z)\middle|\mathbf{X}_i\right]\right) \\ =\;& \mathbb{E}\left[\frac{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0 \mid \mathbf{X}_i\}}{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0\}}D_i(z)\right] \\ =\;& \mathbb{E}\left[\frac{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0 \mid \mathbf{X}_i\}}{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0\}}D_i\middle|Z_i=1\right], \qquad \text{(S1)}\end{aligned}$$

where the last equality follows from Assumption 1. We can then obtain the expressions for APCEp, APCEr and APCEs by choosing different values of y_1 and y_0 in (S1). □

S1.4 Proof of Theorem 3

Assumption 1 implies,

$$\Pr\{D_i(z)=d, Y_i(d)=y\} = \Pr\{D_i(z)=d, Y_i(D_i(z))=y \mid Z_i=z\} = \Pr(D_i=d, Y_i=y \mid Z_i=z).$$

Therefore,

$$\begin{aligned}\Pr\{D_i(z)=1 \mid Y_i(0)=y\} &= \frac{\Pr\{D_i(z)=1, Y_i(0)=y\}}{\Pr\{Y_i(0)=y\}} \\ &= \frac{\Pr\{Y_i(0)=y\} - \Pr\{D_i(z)=0, Y_i(0)=y\}}{\Pr\{Y_i(0)=y\}} \\ &= \frac{\Pr\{Y_i(0)=y\} - \Pr(D_i=0, Y_i=y \mid Z_i=z)}{\Pr\{Y_i(0)=y\}}\end{aligned}$$

As a result, we have

$$\begin{aligned}\mathsf{APCEp} &= \frac{\Pr(D_i=0, Y_i=1 \mid Z_i=0) - \Pr(D_i=0, Y_i=1 \mid Z_i=1)}{\Pr\{Y_i(0)=1\}}, \\ \mathsf{APCEs} &= \frac{\Pr(D_i=0, Y_i=0 \mid Z_i=0) - \Pr(D_i=0, Y_i=0 \mid Z_i=1)}{\Pr\{Y_i(0)=0\}}.\end{aligned}$$

□

S1.5 Proof of Theorem 5

Using the law of total expectation, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid R_i = r] &= \mathbb{E}(\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i, R_i = r] \mid R_i = r) \\
&= \mathbb{E}(\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i] \mid R_i = r) \\
&= \mathbb{E}\left(\frac{\Pr(\mathbf{X}_i \mid R_i = r)}{\Pr(\mathbf{X}_i)} \mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i]\right) \\
&= \mathbb{E}\left(\frac{\Pr(R_i = r \mid \mathbf{X}_i)}{\Pr(R_i = r)} \mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i]\right) \\
&= \mathbb{E}\left[\frac{\Pr(R_i = r \mid \mathbf{X}_i)}{\Pr(R_i = r)} \mathbf{1}\{D_i(z) \geq r\}\right] \\
&= \mathbb{E}\left[\frac{\Pr(R_i = r \mid \mathbf{X}_i)}{\Pr(R_i = r)} \mathbf{1}\{D_i \geq r\} \mid Z_i = z\right],
\end{aligned}$$

where the second equality follows from Assumption 4 and Lemma S2, and the last equality follows from Assumption 1. Thus,

$$\mathsf{APCEp}(r) \;=\; \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 1\} - \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 0\}.$$

We can prove the expression for APCEs similarly. □

S2 Details of the Bayesian Estimation

We only consider the algorithm for sensitivity analysis with ordinal decision since the computation of the original analysis is straightforward by setting the sensitivity parameters to zero. Consider the model given in equations (6) and (7). We can write equation (6) in terms of the observed data as,

$$D_i^* \;=\; \beta_Z Z_i + \mathbf{X}_i^\top \beta_X + Z_i \mathbf{X}_i^\top \beta_{ZX} + \epsilon_{i1}, \tag{S2}$$

where

$$D_i = \begin{cases} 0 & D^* \leq \theta_{Z_i,1} \\ 1 & \theta_{Z_i,1} < D_i^* \leq \theta_{Z_i,2} \\ \vdots & \vdots \\ k-1 & \theta_{Z_i,k-1} < D_i^* \leq \theta_{Z_i,k} \\ k & \theta_{Z_i,k} < D_i^* \end{cases}.$$

We then consider equation (7). For $r = 0, \ldots, k$, because $R_i \geq r + 1$ is equivalent to $Y_i(r) = 1$, we have

$$\Pr\{Y(r) = 1\} \;=\; \Pr\{R_i^* > \delta_r\} = \Pr(\boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2} > \delta_r) = \Pr(-\delta_r + \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2} > 0).$$

Therefore, we can introduce a latent variable $Y^*(r)$, and write

$$Y_i^*(r) = -\delta_r + \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2}, \tag{S3}$$

where $Y_i(r) = 1$ if $Y_i^*(r) > 0$ and $Y_i(r) = 0$ if $Y_i^*(r) \leq 0$. We can further write (S3) in terms of the observed data as

$$Y_i^* = -\sum_{r=0}^{k} \delta_r \mathbf{1}(D_i = r) + \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2}, \tag{S4}$$

where $Y_i = 1$ if $Y_i^* > 0$ and $Y_i = 0$ if $Y_i^* \leq 0$.

Combining (S2) and (S4), we have

$$\begin{aligned} D_i^* &= \beta_Z Z_i + \mathbf{X}_i^\top \beta_X + Z_i \mathbf{X}_i^\top \beta_{ZX} + \epsilon_{i1}, \\ Y_i^* &= -\sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) + \mathbf{X}_i^\top \alpha_X + \epsilon_{i2}, \end{aligned}$$

where

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and

$$D_i = \begin{cases} 0 & D^* \leq \theta_{Z_i,1} \\ 1 & \theta_{Z_i,1} < D_i^* \leq \theta_{Z_i,2} \\ \vdots & \vdots \\ k-1 & \theta_{Z_i,k-1} < D_i^* \leq \theta_{Z_i,k} \\ k & \theta_{Z_i,k} < D_i^* \end{cases}, \quad Y_i = \begin{cases} 0 & Y_i^* \leq 0 \\ 1 & Y_i^* > 0 \end{cases}$$

with $\delta_d \leq \delta_{d'}$ for $d \leq d'$.

We choose multivariate normal priors for the regression coefficients, $(\beta_Z, \beta_X^\top, \beta_{ZX}^\top) \sim \boldsymbol{N}_{2p+1}(\mathbf{0}, \mathbf{\Sigma}_D)$ and $\alpha_X \sim \boldsymbol{N}_p(\mathbf{0}, \mathbf{\Sigma}_R)$. We choose the priors for θ and δ in the following manner. We first choose a normal prior for θ_{z1} and α_1, $\theta_{z1} \sim N(0, \sigma_0^2)$ and $\delta_1 \sim N(0, \sigma_0^2)$ for $z = 0, 1$. We then choose truncated normal priors for other parameters, $\theta_{zj} \sim N(0, \sigma_0^2)\mathbf{1}(\theta_{zj} \geq \theta_{z,j-1})$ for $j = 2, \ldots, k$ and $\delta_l \sim N(0, \sigma_0^2)\mathbf{1}(\delta_l \geq \delta_{l-1})$ for $l = 1, \ldots, k$. In this way, we guarantee that θ's and δ's are increasing. In simulation studies, we choose $\mathbf{\Sigma}_D = 0.01 \cdot \mathbf{I}_{2p+1}$, $\mathbf{\Sigma}_D = 0.01 \cdot \mathbf{I}_p$, and $\sigma_0 = 10$.

Treating Y_i^* and D_i^* as missing data, we can write the complete-data likelihood as

$$\begin{aligned} & L(\theta, \beta, \delta, \alpha) \\ = & \prod_{i=1}^{n} L_i(\theta, \beta, \delta, \alpha) \\ \propto & \prod_{i=1}^{n} \exp\left(-\frac{1}{2(1-\rho^2)} \left[(D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})^2 + \left\{ Y_i^* + \sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\}^2 \right. \right. \\ & \left. \left. -2\rho(D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i) \left\{ Y_i^* + \sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\} \right] \right) \end{aligned}$$

Imputation Step. We first impute the missing data given the observed data and parameters. Using R package *tmvtnorm*, we can jointly sample Y_i^* and D_i^*. Given $(D_i, Y_i, Z_i, \mathbf{X}_i^\top, \theta, \beta, \alpha, \delta)$, (D_i^*, Y_i^*) follows a truncated bivariate normal distribution whose means are given by $\mathbf{X}_i^\top \beta_X + \beta_Z Z_i + Z_i \mathbf{X}_i^\top \beta_{ZX}$ and $-\sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) + \mathbf{X}_i^\top \alpha_X$, and whose covariance matrix has unit variances and correlation ρ where D^* is truncated within interval $[\theta_{zd}, \theta_{z,d+1}]$ if $Z_i = z$ and $D_i = d$ (we define $\theta_0 = -\infty$ and $\theta_{k+1} = \infty$) and Y_i^* is truncated within $(-\infty, 0)$ if $Y_i = 0$ and $[1, \infty)$ if $Y_i = 1$.

Posterior Sampling Step. The posterior distribution is proportional to

$$\begin{aligned}
&\prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)}\left[(D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})^2 + \left\{Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X\right\}^2\right.\right.\\
&\qquad\qquad \left.\left. -2\rho(D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})\left\{Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X\right\}\right]\right)\\
&\cdot \exp\left\{-\frac{(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)\mathbf{\Sigma}_D^{-1}(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top}{2}\right\} \cdot \exp\left(-\frac{\alpha_X^\top \mathbf{\Sigma}_R^{-1} \alpha_X}{2}\right)\\
&\cdot \exp\left(-\frac{\theta_{11}^2}{2\sigma_0^2}\right) \exp\left(-\frac{\delta_0^2}{2\sigma_0^2}\right) \prod_{j=2}^k \left\{\exp\left(-\frac{\theta_{1j}^2}{2\sigma_0^2}\right) \mathbf{1}(\theta_{1j} \geq \theta_{1,j-1})\right\} \prod_{l=1}^k \left\{\exp\left(-\frac{\delta_l^2}{2\sigma_0^2}\right) \mathbf{1}(\delta_l \geq \delta_{l-1})\right\}\\
&\cdot \exp\left(-\frac{\theta_{01}^2}{2\sigma_0^2}\right) \prod_{j=2}^k \left\{\exp\left(-\frac{\theta_{0j}^2}{2\sigma_0^2}\right) \mathbf{1}(\theta_{0j} \geq \theta_{0,j-1})\right\}.
\end{aligned}$$

We first sample $(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)$. From the posterior distribution, we have

$$\begin{aligned}
& f(\beta_Z, \beta_X^\top, \beta_{ZX}^\top \mid \cdot)\\
\propto\ & \prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)}\left[(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})^2\right.\right.\\
& \left.\left. -2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})\left\{Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X\right\}\right]\right) \cdot \exp\left\{-\frac{(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top \mathbf{\Sigma}_D^{-1}(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)}{2}\right\}\\
\propto\ & \prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)}\left[(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)(Z_i, \mathbf{X}_i^\top, Z_i \mathbf{X}_i^\top)^\top (Z_i, \mathbf{X}_i^\top, Z_i \mathbf{X}_i^\top)(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top - 2D_i^*(Z_i, \mathbf{X}_i^\top, Z_i \mathbf{X}_i^\top)(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top\right.\right.\\
& \left.\left. +2\rho\left\{Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X\right\}(Z_i, \mathbf{X}_i^\top, Z_i \mathbf{X}_i^\top)(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top\right]\right) \cdot \exp\left\{-\frac{(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top \mathbf{\Sigma}_D^{-1}(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)}{2}\right\}.
\end{aligned}$$

Therefore, we can sample

$$(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top \mid \cdot \sim \boldsymbol{N}_{p+1}(\widehat{\mu}_D, \widehat{\mathbf{\Sigma}}_D)$$

where

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_D &= \left\{\frac{1}{1-\rho^2}\sum_{i=1}^n (Z_i, \boldsymbol{X}_i^\top, Z_i \mathbf{X}_i^\top)^\top (Z_i, \boldsymbol{X}_i^\top, Z_i \mathbf{X}_i^\top) + \mathbf{\Sigma}_D^{-1}\right\}^{-1},\\
\widehat{\mu}_D &= \widehat{\mathbf{\Sigma}}_D \left(\frac{1}{1-\rho^2}\sum_{i=1}^n (Z_i, \boldsymbol{X}_i^\top, Z_i \mathbf{X}_i^\top)^\top \left[D_i^* - \rho\left\{Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X\right\}\right]\right).
\end{aligned}$$

We then consider sampling α_X. We have

$$\begin{aligned}
& f(\alpha_X \mid \cdot) \\
\propto \; & \prod_{i=1}^{n} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left\{ Y_i^* + \sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\}^2 \right.\right. \\
& \qquad \left.\left. -2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \left\{ Y_i^* + \sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\} \right] \right) \cdot \exp\left(-\frac{\alpha_X^\top \mathbf{\Sigma}_R^{-1} \alpha_X}{2} \right) \\
\propto \; & \prod_{i=1}^{n} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\alpha_X^\top \mathbf{X}_i^\top \mathbf{X}_i \alpha_X - 2 \left\{ Y_i^* + \sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) \right\} \mathbf{X}_i \alpha_X + 2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \mathbf{X}_i \alpha_X \right] \right) \\
& \quad \cdot \exp\left(-\frac{\alpha_X^\top \mathbf{\Sigma}_R^{-1} \alpha_X}{2} \right).
\end{aligned}$$

Therefore, we can sample

$$\alpha_X \mid \cdot \sim \boldsymbol{N}_p(\widehat{\mu}_R, \widehat{\mathbf{\Sigma}}_R),$$

where

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_R &= \left\{ \frac{1}{1-\rho^2} \sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{X}_i + \mathbf{\Sigma}_R^{-1} \right\}^{-1}, \\
\widehat{\mu}_R &= \widehat{\mathbf{\Sigma}}_D \left(\frac{1}{1-\rho^2} \sum_{i=1}^{n} \boldsymbol{X}_i \left[\left\{ Y_i^* + \sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) \right\} - \rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \right] \right).
\end{aligned}$$

To sample δ's, we write $\sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) = \delta_0 + \sum_{d=1}^{k} (\delta_d - \delta_{d-1}) \mathbf{1}(D_i \geq d)$ and denote $\boldsymbol{W}_i = (1, \mathbf{1}(D_i \geq 1), \ldots, \mathbf{1}(D_i \geq k))$ and $\delta = (\delta_0, \delta_1 - \delta_0, \ldots, \delta_k - \delta_{k-1})$. Thus, we have

$$\begin{aligned}
& f(\delta \mid \cdot) \\
\propto \; & \prod_{i=1}^{n} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left\{ Y_i^* + \boldsymbol{W}_i \delta - \mathbf{X}_i^\top \alpha_X \right\}^2 - 2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \left\{ Y_i^* + \boldsymbol{W}_i \delta - \mathbf{X}_i^\top \alpha_X \right\} \right] \right) \\
& \quad \cdot \exp\left(-\frac{\delta_0^2}{2\sigma_0^2} \right) \prod_{d=1}^{k} \left\{ \exp\left(-\frac{\delta_l^2}{2\sigma_0^2} \right) \mathbf{1}(\delta_d - \delta_{d-1} \geq 0) \right\} \\
\propto \; & \prod_{i=1}^{n} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\delta^\top \boldsymbol{W}_i^\top \boldsymbol{W}_i \delta + 2\left(Y_i^* - \mathbf{X}_i^\top \alpha_X \right) \boldsymbol{W}_i \delta - 2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \boldsymbol{W}_i \delta \right] \right) \\
& \quad \cdot \exp\left(-\frac{\delta_0^2}{2\sigma_0^2} \right) \prod_{d=1}^{k} \left\{ \exp\left(-\frac{\delta_l^2}{2\sigma_0^2} \right) \mathbf{1}(\delta_d - \delta_{d-1} \geq 0) \right\}.
\end{aligned}$$

Therefore, we can draw from a truncated normal distribution with mean and covariance matrix

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_\delta &= \left\{ \frac{1}{1-\rho^2} \sum_{i=1}^{n} \boldsymbol{W}_i^\top \boldsymbol{W}_i + \sigma_0^{-2} \right\}^{-1}, \\
\widehat{\mu}_\delta &= \widehat{\mathbf{\Sigma}}_D \left[\frac{1}{1-\rho^2} \sum_{i=1}^{n} \boldsymbol{W}_i^\top \left\{ \rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) - \left(Y_i^* - \mathbf{X}_i^\top \alpha_X \right) \right\} \right],
\end{aligned}$$

where the 2-th to $(k+1)$-th element is truncated within interval $[0, \infty)$. We can then transform δ to obtain $(\delta_0, \delta_1, \ldots, \delta_k)$.

Finally, we sample

$$\theta_{z1} \mid \cdot \sim TN(0, \sigma_0^2; \max_{i:Z_i=z, D_i=0} D_i^*, \min_{i:Z_i=z, D_i=1} (D_i^*, \theta_2)).$$

We then sample

$$\theta_{zj} \mid \cdot \sim TN(0, \sigma_0^2; \max_{i:Z_i=z, D_i=j-1} (D_i^*, \theta_{j-1}), \min_{i:Z_i=z, D_i=j} (D_i^*, \theta_{j+1}))$$

for $j = 2, \dots, k-1$, and

$$\theta_{zk} \mid \cdot \sim TN(0, \sigma_0^2; \max_{i:Z_i=z, D_i=k-1} (D_i^*, \theta_{k-1}), \min_{i:Z_i=z, D_i=k} D_i^*).$$

The MCMC gives the posterior distributions of the parameters and therefore we can obtain the posterior distributions of $\Pr(D_i \mid R_i, \mathbf{X}_i = \mathbf{x}, Z_i = z)$ and $\Pr(R_i \mid \mathbf{X}_i = \mathbf{x})$. As a result, for $r = 0, \dots, k-1$, we have

$$\begin{aligned} \mathsf{APCE}(r) &= \Pr\{D_i(1) \geq r \mid R_i = r\} - \Pr\{D_i(0) \geq r \mid R_i = r\} \\ &= \frac{\mathbb{E}\{\Pr(D_i(1) \geq r, R_i = r \mid \mathbf{X}_i)\}}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i)\}} - \frac{\mathbb{E}\{\Pr(D_i(0) \geq r, R_i = r \mid \mathbf{X}_i)\}}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i)\}}, \\ \mathsf{APCEs} &= \Pr\{D_i(1) = k \mid R_i = k\} - \Pr\{D_i(0) = k \mid R_i = k\} \\ &= \frac{\mathbb{E}\{\Pr(D_i(1) = 0, R_i = 0 \mid \mathbf{X}_i)\}}{\mathbb{E}\{\mathrm{pr}(R_i = 0 \mid \mathbf{X}_i)\}} - \frac{\mathbb{E}\{\Pr(D_i(0) = 0, R_i = 0 \mid \mathbf{X}_i)\}}{\mathbb{E}\{\Pr(R_i = 0 \mid \mathbf{X}_i)\}}. \end{aligned}$$

We can calculate the conditional probabilities $\Pr\{D_i(z), R_i \mid \mathbf{X}_i\}$ and $\Pr(R_i \mid \mathbf{X}_i)$ based on the posterior sample of the coefficients, and then replace the expectation with the empirical average to obtain the estimates.