A Causal Framework for Observational Studies of Discrimination*

Johann Gaebler
Stanford University

William Cai
Stanford University

Guillaume Basse
Stanford University

Ravi Shroff
New York University

Sharad Goel
Harvard University

Jennifer Hill
New York University

Abstract

In studies of discrimination, researchers often seek to estimate a causal effect of race or gender on outcomes. For example, in the criminal justice context, one might ask whether arrested individuals would have been subsequently charged or convicted had they been a different race. It has long been known that such counterfactual questions face measurement challenges related to omitted-variable bias, and conceptual challenges related to the definition of causal estimands for largely immutable characteristics. Another concern, which has been the subject of recent debates, is post-treatment bias: many studies of discrimination condition on apparently intermediate outcomes, like being arrested, that themselves may be the product of discrimination, potentially corrupting statistical estimates. There is, however, reason to be optimistic. By carefully defining the estimand—and by considering the precise timing of events—we show that a primary causal quantity of interest in discrimination studies can be estimated under an ignorability condition that may hold approximately in some observational settings. We illustrate these ideas by analyzing both simulated data and the charging decisions of a prosecutor's office in a large county in the United States.

*We thank Alex Chohlas-Wood, Avi Feller, Andrew Gelman, Zhiyuan "Jerry" Lin, Julian Nyarko, Brendan O'Flaherty, Elizabeth Ogburn, José Luis Montiel Olea, Steven Raphael, James Robins, Rajiv Sethi, Amy Shoemaker, and Ilya Shpitser for helpful conversations. Code to replicate our analysis is available online at: `https://github.com/stanford-policylab/gcbsgh-rep`.

To assess the role of race or gender in decision making, researchers often examine disparities between groups after adjusting for relevant factors. For example, to measure racial discrimination in lending decisions, one might estimate race-specific approval rates after adjusting for creditworthiness, typically via a regression model. This simple statistical strategy—sometimes called benchmark analysis—has been used to study discrimination in a wide variety of domains, including banking [Munnell et al., 1996], employment [Berg and Lien, 2002], education [Baum and Goodstein, 2005], healthcare [Balsa et al., 2005], housing [Edelman and Luca, 2014, Greenberg et al., 2016], and criminal justice [Ayres, 2002, Fryer Jr, 2019, Gelman et al., 2007, MacDonald and Raphael, 2021, Rehavi and Starr, 2014].

The results of benchmark analyses are often framed in causal terms (e.g., as an effect of race on outcomes), but it is well understood that such an approach suffers from at least three significant statistical challenges when used to estimate causal quantities. First, at a conceptual level, it is unclear how best to rigorously define causal estimands of interest when the treatment is race, gender, or another largely immutable trait. Second, estimates can be plagued by omitted-variable bias if one does not appropriately adjust for all relevant covariates. Third—and the focus of our paper—there are worries that estimates are corrupted by post-treatment bias when one adjusts for covariates or restricts to samples of individuals determined downstream from race, gender, or another such treatment variable. This concern, in particular, has raised doubts about the reliability of the literature on police discrimination, where many studies rely on administrative stop records, and hence implicitly condition on officers stopping an individual, an action that itself is likely discriminatory [Heckman and Durlauf, 2020, Knox et al., 2020].

Here we present a causal framework for conceptualizing and estimating a measure of discrimination that is suitable for many applied problems. Our framing specifically addresses concerns about post-treatment bias. To do so, we first define a causal quantity—the second-stage sample average treatment effect, or SATE_M—which closely maps to the legal notion of disparate treatment. For this estimand, by carefully considering the timing of events, we show that treatment assignment conceptually occurs after selection into the sample of interest. We then introduce the notion of subset ignorability, show that this condition formally justifies the use of benchmark analysis to estimate the SATE_M, and discuss settings in which it is likely to hold approximately. We illustrate these ideas by analyzing synthetic data, as well as a detailed dataset of prosecutorial charging decisions for approximately 20,000 felony arrests in a major U.S. county. By developing this statistical foundation, we hope to place discrimination studies on more solid theoretical footing.

1 A Motivating Example

Consider the problem of measuring racial discrimination in prosecutorial charging decisions. After an individual has been arrested, prosecutors in the district attorney's office read the arresting officer's incident report and then decide whether or not to press charges. For simplicity, suppose prosecutors only have access to the incident report—and to no other information—when making their decisions. We allow for the possibility that the arrest decision that preceded the charging decision may have suffered from racial discrimination in complex ways that cannot be inferred from the incident reports themselves. Finally, suppose that a researcher has access to these incident reports for arrested individuals, but, importantly, not to any data on individuals that officers considered but ultimately decided against arresting. What, if anything, might one hope to discover about racial discrimination in charging decisions in light of the fact that the people about whom the prosecutor makes charging decisions have been selected—that is, arrested—not randomly, but rather in ways that likely depended on their race?

The first challenge is to rigorously define causal estimands of interest. The inherent difficulty is captured by the statistical refrain "no causation without manipulation" [Holland, 1986], since it is often unclear what it means to alter attributes like race and gender [Sekhon, 2008]. One common maneuver is to instead consider the causal effect of *perceived* attributes (e.g., perceived race or perceived gender), which ostensibly can be manipulated—for example, by changing the name listed on an employment application [Bertrand and Mullainathan, 2004], or by masking an individual's appearance [Goldin and Rouse, 2000, Grogger and Ridgeway, 2006, Pierson et al., 2020]. In our case, one might imagine a hypothetical experiment in which explicit mentions of race in the incident report are altered (e.g., replacing "white" with "Black"). The causal effect is then, by definition, the difference in charging rates between those cases in which arrested individuals were randomly described (and hence may be perceived) as "Black" and those in which they were randomly described as "white." This conceptualization of discrimination conforms to one common causal understanding of discrimination used, for example, in audit studies. This framing also maps closely to the legal notion of disparate treatment, a form of discrimination in which actions are motivated by animus or otherwise discriminatory intent [Goel et al., 2017].

While researchers have carried out such audit studies—including in the case of prosecutorial charging decisions [Chohlas-Wood et al., 2021, Robertson et al., 2019][1]—it is often infeasible to study important policy questions through randomized experiments. In the absence of a controlled experiment, one can in theory identify this type of causal estimand from purely observational data by comparing charging rates across pairs of cases that are identical in all aspects other than the stated race of the arrested individual.[2] That strategy, which mimics the key features of the hypothetical randomized experiment described above, is formally justified when treatment assignment (i.e., description of race on the incident report, and subsequent perception by the prosecutor) is *ignorable* given the observed covariates (i.e., features of the incident report) [Imbens and Rubin, 2015]. In practice, though, this approach may suffer from omitted-variable bias when the full incident report is not available to researchers, and may suffer from lack of overlap when suitable matches cannot be found for each case—limitations common to many observational studies of causal effects. To address these issues, one can restrict attention to the overlap region and gauge the robustness of estimates to varying forms and degrees of unmeasured confounding [Cinelli and Hazlett, 2020, Cornfield et al., 1959, Rosenbaum and Rubin, 1983b], an approach we demonstrate below.

Finally, there is the issue of post-treatment bias, especially due to sample selection. Knox et al. [2020] argue that researchers often inadvertently introduce post-treatment bias in observational studies of discrimination by subsetting on apparently intermediate outcomes—such as, in our charging example, being arrested—that themselves may be the product of discrimination. As a result, the authors caution that causal quantities of interest cannot be identified by the data in the absence of implausible assumptions, such as lack of discrimination in the initial arrest decision. In making their argument, Knox et al. focus on the use of force by police officers in civilian encounters, but they suggest their formal critique applies more broadly, casting doubt on a wide range of observational studies of discrimination.

Here we show that such customary subsetting does not pose an insurmountable threat to

[1] There are some differences between the idealized audit study described above and these two experiments. Chohlas-Wood et al. conduct a quasi-random field trial in which they mask—but do not switch—the stated race of individuals in police narratives used to make actual charging decisions. Robertson et al. survey prosecutors in a randomized lab experiment and ask them, hypothetically, what their charging decision would be based on fact patterns in which the race of the suspect is manipulated. Although neither of these studies maps exactly to the hypothetical experiment motivating our estimand, both demonstrate the feasibility of conducting such an experiment.

[2] It suffices to compare groups of cases that have the same distribution of potential outcomes—even if the cases themselves are not identical—a property we formalize in Definition 2 below.

discrimination research. To understand why, one must precisely define the causal estimand, and carefully consider the timing of events. For instance, in our charging example, there are two relevant treatments, the officer's perception of race, affecting the officer's arrest decision, and the prosecutor's perception of race, affecting the prosecutor's charging decision. The arrest decision is post-treatment relative to the officer's perception of race but, importantly, it is pre-treatment relative to the prosecutor's perception of race. Similarly, the features of the incident report—which we must adjust for in this type of benchmark analysis—are post-treatment relative to the officer's perception of race but pre-treatment relative to the prosecutor's perception of race. In such a two-decider situation, as Greiner and Rubin [2011] suggest, it is possible to recover estimates of discrimination by the second decider (e.g., in the charging decision) even if there is discrimination by the first decider (e.g., in the arrest decision).

2 A Measure of Discrimination

We present a simple two-stage model to characterize discriminatory decision making in a variety of real-world situations and define our main causal quantity of interest—the second-stage sample average treatment effect, or SATE_M—within this general framework. In the context of our motivating example, the SATE_M corresponds to the quantity that would be measured in the hypothetical audit study of prosecutorial decisions described in Section 1. A central aim of this paper is to formalize technical assumptions that allow one to statistically identify discrimination—more precisely, disparate treatment—in the second stage (e.g., in prosecutorial charging decisions) when data are only available for individuals who made it past the first stage (e.g., those who were arrested). Importantly, our formalization accommodates scenarios in which first-stage decisions may themselves be discriminatory.

In the first stage, we assume each individual in some population is subject to a binary decision M, such as an offer of employment, admission to college, or law enforcement action. Those who receive a "positive" first-stage decision (e.g., those who are arrested) proceed to a second stage, where another binary decision Y is made. In our running example, the case of each arrested individual is reviewed in the second stage by a prosecutor who may or may not choose to press charges. Those who are not arrested do not have a case that requires review by a prosecutor and, indeed, there may be no administrative record of those individuals.

When considering racial discrimination in decisions involving Black and white individuals, our primary quantity of interest is the second-stage sample average treatment effect, $\mathbb{E}[Y(b) - Y(w)]$, where $Y(z)$ indicates the potential second-stage decision and the expectation is taken over individuals reaching the second stage. Here, we imagine that the perception of race is counterfactually determined after the first-stage decision but before the second-stage decision (e.g., after arrest but before charging, perhaps by altering the description of race on the incident report viewed by a prosecutor). The second-stage sample average treatment effect thus captures discrimination in the second-stage decision among those who made it past the first stage (e.g., discrimination in charging decisions among those who were arrested). This estimand maps onto a common understanding of disparate treatment in second-stage decisions, including in our charging example.

2.1 A formal model of discrimination

We now formalize the above discussion to explicitly include decisions made at both the first and second stages. For ease of interpretation, we follow Greiner and Rubin [2011] and motivate our statistical model by considering settings where there are two deciders (e.g., an officer and a prosecutor) whose perceptions of race—or gender, or another trait—can in theory be independently altered prior

to their decisions. There are, however, examples in which one can plausibly intervene twice even when a single decider makes both decisions. For instance, an officer may decide to stop a motorist based in part on a brief impression of the motorist's skin tone as they drive past [Grogger and Ridgeway, 2006, Pierson et al., 2020]. This visual impression of race could subsequently be altered if the motorist presents a driver's license bearing a name characteristic of another race group, or speaks a dialect of English at odds with the officer's expectation. It thus may be possible to apply our framework whether one imagines there are two deciders or a single one.

We begin by denoting the race of an individual as perceived by the first decider at the first stage by $D \in \{w, b\}$, where, for simplicity, we consider a population consisting of only white and Black individuals. We focus on racial discrimination for concreteness, but similar considerations apply to discrimination based on other attributes, such as gender. Assuming that there is no interference between units [Imbens and Rubin, 2015], we let the binary variables $M(w)$ and $M(b)$ denote the potential first-stage decisions for an individual (e.g., whether they were arrested), and write $M = M(D)$ for the observed first-stage decision. To avoid triviality, we assume throughout that $\Pr(M = 1) > 0$.

Next, we let $Z \in \{w, b\}$ denote the race of an individual as perceived by the second decider, at the second stage. In our running example, Z denotes race as perceived by the prosecutor reviewing that person's file, while D denotes race as perceived by the police officer during the encounter. Finally, we define the second-stage potential outcomes as a function of both the first-stage outcome M (e.g., the arrest decision) and the second decider's perception of race Z. Thus, assuming once again that there is no interference, the observed second-stage outcome for an individual can be denoted $Y = Y(Z, M)$, where we consider four potential second-stage outcomes for each individual: $Y(z, m)$, where $z \in \{w, b\}$ and $m \in \{0, 1\}$. In our example, only those who were arrested can be charged, and so $Y(b, 0) = Y(w, 0) = 0$ for all individuals.[3]

We further allow each individual to have an associated vector of (non-race) covariates X, representing, for example, their behavior during a police encounter, their recorded criminal history, or both. We imagine these covariates are fixed prior to the second-stage treatment (e.g., prior to the prosecutor's perception of race), since otherwise the key ignorability assumption in Definition 2 below is unlikely to hold. In practice, X is only observed for a subset of the population (e.g., those who were arrested and hence in the dataset), but we nonetheless define the covariate vector for all individuals in our population of interest. These covariates are not necessary to define our causal estimands of interest, but they play an important role in constructing our statistical estimators.

In this model of discrimination, we have taken care to distinguish between the (realized) first- and second-stage perceptions of race, D and Z, because this helps to clarify the timing of events and the meaning of causal quantities. Importantly, this makes it clear that we can conceive of D and Z as separately manipulable. At the same time, our focus is observational settings, in which disagreement between Z and D may be realized only rarely, if at all, in the data we observe. For instance, barring manipulation of the incident report, it seems unlikely that an arresting officer's perception of race will frequently differ from a prosecutor's perception. Our simulation in Section 3 thus imposes the further constraint that perceived race is the same at each stage, though this restriction is not necessary in general.

With this framing, we now formally describe the primary causal estimand we consider. This quantity, which we call the second-stage sample average treatment effect (SATE_M) reflects discrimination in the second stage of the decision-making process outlined above, such as discrimination in

[3]To avoid imagining values of Z for individuals not arrested, one could also make them "missing" by setting $Z = Z(D, M)$, $Z(d, 0) = \text{NA}$, $Z(d, 1) = d$, $Y(\text{NA}, 1) = \text{NA}$, and $Y(z, 0) = 0$ for $z \in \{w, b, \text{NA}\}$, as we do in the simulation in Section 3 below. This does not affect any of the mathematical details in what follows.

the prosecutor's charging decision.[4]

Definition 1 (SATE_M)**.** The *second-stage sample average treatment effect*, denoted SATE_M, is:

$$\text{SATE}_M = \mathbb{E}[Y(b,1) - Y(w,1) \mid M = 1]. \tag{1}$$

The estimand in Eq. (1) compares the potential second-stage decisions under two race perception scenarios. For example, it compares the potential charging decisions when the prosecutor perceives the individual to be either Black or white; importantly, though, the estimand does not explicitly consider the arresting officer's perception of race. Moreover, this estimand restricts to the subset of individuals who had a "positive" first-stage decision (e.g., those who were in reality arrested).

Because we condition on $M = 1$ in the definition of the SATE_M, we may equivalently write Eq. (1) as

$$\text{SATE}_M = \mathbb{E}[Y(b,M) - Y(w,M) \mid M = 1]. \tag{2}$$

We can further write

$$\text{SATE}_M = \mathbb{E}[Y(b) - Y(w) \mid M = 1], \tag{3}$$

where we define $Y(z) = Y(z, M)$. Among those who reach the second stage (i.e., individuals with $M = 1$), $Y(z) = Y(z, 1)$ denotes the outcome of intervening *only* on the second decider's perception of race. Among those who do not reach the second stage (i.e., individuals with $M = 0$), $Y(z) = Y(z, 0) = 0$. Eqs. (1), (2), and (3), as well as the informal estimand introduced at the beginning of Section 2, are equivalent ways of capturing the same underlying quantity, varying only in the degree to which they are explicit about the staged nature of the process.

2.2 Estimating the $\texttt{sate}_M$

Having defined the SATE_M, our goal is now to estimate it using only second-stage data. That is, we aim to estimate the SATE_M only using observations for those individuals who received a "positive"—and potentially discriminatory—decision in the first stage. For example, we seek to estimate discrimination in charging decisions based only on data describing those who were arrested. As we show now, an ignorability assumption, together with an overlap condition, is sufficient to guarantee the SATE_M is nonparametrically identified by data on the second-stage decisions.

Definition 2 (Subset ignorability)**.** We say that $Y(z,1)$, Z, M, and X satisfy *subset ignorability* if

$$Y(z,1) \perp\!\!\!\perp Z \mid X, M = 1 \tag{4}$$

for $z \in \{w, b\}$.

In our recurring example, subset ignorability means that among arrested individuals, after conditioning on available covariates, race (as perceived by the prosecutor) is independent of the potential outcomes for the charging decision. As above, we can equivalently write Eq. (4) as

$$Y(z) \perp\!\!\!\perp Z \mid X, M = 1. \tag{5}$$

This latter expression makes clear that subset ignorability is closely related to the traditional ignorability assumption in causal inference, but where we have explicitly referenced the first-stage outcomes to accommodate a staged model of decision making.

[4]The SATE_M is notationally equivalent to the $\text{CDE}_{M=1}$ defined in Knox et al. [2020]. In our case, however, we have taken care to specify that the first parameter in the quantity $Y(z,m)$ denotes intervening on the *second-stage* perception of race. Moreover, the SATE_M is distinct from what Knox et al. call the $\text{ATE}_{M=1}$.

In our prosecutorial setting, subset ignorability would fail if, for example, there were a factor that prosecutors used to make their charging decisions but which was not accounted for in the analysis (e.g., if prosecutors reviewed witness statements that were not in the case files provided to the analyst), and, further, that factor were unbalanced between groups (e.g., if all else equal, witness statements were more commonly available in the cases of white individuals). See Sections 3 and 4 for further discussion of such unobserved confounders and their statistical consequences.

Almost all causal analyses implicitly rely on a version of subset ignorability, since researchers rarely make inferences about the full population of interest. For example, analyses are typically limited to the individuals who agreed to participate in the study. Even randomized experiments, while ideal for internal validity, frequently lack external validity because the study participants do not resemble a larger population of interest. Whenever ascribing causal interpretations to non-experimental data, it is important to carefully consider the plausability of ignorability and other assumptions, as we discuss in detail in Sections 3 and 4 below. We note, though, that the assumptions we rely on are similar to those invoked in nearly every observational study of causal effects.

Ignorability assumptions typically require a corresponding overlap condition to guarantee consistent estimation.[5]

Definition 3 (Overlap)**.** We say that *overlap* holds when $\Pr(Z = z \mid X = x, M = 1) > 0$ for all z and x such that $\Pr(X = x, M = 1) > 0$.

Overlap states that there are no covariate levels for which the probability of receiving one of the treatments is zero within the population of interest. In our prosecution example, overlap ensures that every case has a "twin", identical in all aspects other than the stated race of the arrested individual, against which it can be compared. Overlap would fail, therefore, in the prosecutorial setting, if, for instance, there were alleged offenses for which only Black individuals were arrested. We note that, unlike ignorability, overlap can be assessed directly from the data; see Section 4. In cases where overlap fails to hold, one can still elicit valid causal estimates by restricting to the subset of the population where overlap holds. For example, in assessing discrimination in prosecutorial charging decisions, one might only consider those alleged offenses for which both Black and white individuals have a non-zero probability of being arrested. But this restriction comes at the cost of inferential validity for the original population. In such cases, one is estimating the causal effect *only* for the restricted population; the causal effect for the original population may differ, sometimes substantially.

In the traditional, single-stage setting, ignorability and overlap are sufficient to obtain consistent estimates of the average treatment effect. Likewise, we now show that in our two-stage model of discrimination, subset ignorability and overlap are sufficient to guarantee consistent estimates of the SATE_M. In practice, if one can adjust for (nearly) all relevant factors affecting second-stage decisions, one can (approximately) satisfy subset ignorability, and in particular, one can estimate the SATE_M only using data available at the second stage. In the Appendix, we compare subset ignorability to several alternatives, and show that those variants tend either to be too weak to guarantee identifiability, or unnecessarily demanding for real-world applications. We emphasize that since the first-stage decision, M, and the covariates, X, can be viewed as pre-treatment relative to the second-stage intervention, concerns about post-treatment bias corrupting estimates of the SATE_M are more naturally thought of as familiar concerns about omitted-variable bias.

[5]In the following, we assume that X is discrete for simplicity of exposition; for continuous analogues of these results, see Appendix C.

Theorem 4. *Suppose $Y(z,1)$, Z, M, and X satisfy subset ignorability and overlap. Then, the* SATE_M *equals*

$$\sum_x \mathbb{E}[Y \mid Z=b, X=x, M=1] \cdot \Pr(X=x \mid M=1) \\ - \sum_x \mathbb{E}[Y \mid Z=w, X=x, M=1] \cdot \Pr(X=x \mid M=1).$$

Proof. Conditioning on X in Eq. (1), we have

$$\text{SATE}_M = \sum_x \mathbb{E}[Y(b,1) \mid X=x, M=1] \cdot \Pr(X=x \mid M=1) \\ - \sum_x \mathbb{E}[Y(w,1) \mid X=x, M=1] \cdot \Pr(X=x \mid M=1). \tag{6}$$

By subset ignorability and overlap, we can condition the summands in Eq. (6) on $Z=b$ and $Z=w$, respectively, without changing their values, yielding

$$\text{SATE}_M = \sum_x \mathbb{E}[Y(b,1) \mid Z=b, X=x, M=1] \cdot \Pr(X=x \mid M=1) \\ - \sum_x \mathbb{E}[Y(w,1) \mid Z=w, X=x, M=1] \cdot \Pr(X=x \mid M=1) \tag{7}$$

$$= \sum_x \mathbb{E}[Y(Z,M) \mid Z=b, X=x, M=1] \cdot \Pr(X=x \mid M=1) \\ - \sum_x \mathbb{E}[Y(Z,M) \mid Z=w, X=x, M=1] \cdot \Pr(X=x \mid M=1). \tag{8}$$

Finally, the statement of the proposition follows by consistency, as $Y = Y(Z,M)$. □

Corollary 5. *Suppose subset ignorability and overlap hold, and that we have n i.i.d. observations $(X_i, Z_i, Y_i)_{i=1}^n$ with $M_i = 1$. Let $S_x^{(n)} = \{1 \le i \le n : X_i = x\}$ represent the set of observations with $X = x$, and let $S_{zx}^{(n)} = \{1 \le i \le n : Z_i = z \wedge X_i = x\}$ represent the set of observations with $X = x$ and $Z = z$. Then the stratified difference-in-means estimator,*

$$\Delta_n = \sum_x \left[\frac{1}{\left|S_{bx}^{(n)}\right|} \sum_{i \in S_{bx}^{(n)}} Y_i \right] \frac{\left|S_x^{(n)}\right|}{n} - \sum_x \left[\frac{1}{\left|S_{wx}^{(n)}\right|} \sum_{i \in S_{wx}^{(n)}} Y_i \right] \frac{\left|S_x^{(n)}\right|}{n}, \tag{9}$$

yields a consistent estimate of the SATE_M.

Proof. Note that by the strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{\left|S_{zx}^{(n)}\right|} \sum_{i \in S_{zx}^{(n)}} Y_i \overset{\text{a.s.}}{=} \mathbb{E}[Y \mid Z=z, X=x, M=1], \text{ and}$$

$$\lim_{n\to\infty} \frac{\left|S_x^{(n)}\right|}{n} \overset{\text{a.s.}}{=} \Pr(X=x \mid M=1).$$

Consequently,

$$\lim_{n\to\infty} \Delta_n \overset{\text{a.s.}}{=} \sum_x \mathbb{E}[Y \mid Z=b, X=x, M=1] \cdot \Pr(X=x \mid M=1) \\ - \sum_x \mathbb{E}[Y \mid Z=w, X=x, M=1] \cdot \Pr(X=x \mid M=1),$$

which is the SATE_M, by Theorem 4. □

A straightforward calculation further shows that the following expression yields a consistent estimate of the standard error of Δ_n:

$$\widehat{\text{SE}}(\Delta_n) = \sqrt{\sum_x \left(\frac{\left|S_x^{(n)}\right|}{n}\right)^2 \left[\frac{c_{bx}(1-c_{bx})}{\left|S_{bx}^{(n)}\right|} + \frac{c_{wx}(1-c_{wx})}{\left|S_{wx}^{(n)}\right|}\right]}, \tag{10}$$

where

$$c_{zx} = \frac{1}{\left|S_{zx}^{(n)}\right|} \sum_{i \in S_{zx}^{(n)}} Y_i.$$

Eq. (10) accordingly allows us to form confidence intervals for Δ_n.

The nonparametric stratified difference-in-means estimator Δ_n is the basis for nearly all applications of benchmark analysis in discrimination studies. In practice, as we discuss further in Section 3, it is common to approximate Δ_n via a parametric regression model—but the two estimators share the same theoretical underpinnings. As such, our analysis above simply grounds traditional benchmark analysis within a specific causal framework, and demonstrates that a particular ignorability assumption, together with overlap, is sufficient to yield valid estimates.

2.3 An alternative measure of discrimination

To better understand the SATE_M, we now contrast it with the total effect (TE) [Imai et al., 2010a], a second estimand considered by discrimination researchers [Heckman and Durlauf, 2020, Knox et al., 2020, Zhao et al., 2021]. The total effect and the SATE_M differ in our setting in two ways: (1) the population of individuals about which we make inferences; and (2) the potential outcomes being contrasted. The total effect is not restricted to individuals who had a "positive" first-stage decision (e.g., it is not restricted to those who were arrested). Additionally, we imagine a causal variable that reflects a situation where the perception of race is counterfactually determined *before* the first-stage decision (instead of *after* the first-stage decision, as with the SATE_M), and is the same at both stages.

We note that, in general—as discussed in Section 1 and below—there is no fully coherent notion of a "total effect" of race, since one cannot intervene on race, *per se*. In our running example, the two treatments (i.e., the officer's perception of race and the prosecutor's perception of race) represent distinct, situation-specific notions of intervening on race. In this restricted context, then, there is a natural estimand that captures the spirit of a "total effect": comparing an individual's potential outcomes had they been perceived as white or Black when *both* the first- and second-stage decisions were made. We formalize this as follows:

Definition 6 (TE)**.** The *total effect*, denoted TE, is given by:

$$\text{TE} = \mathbb{E}[Y(b, M(b)) - Y(w, M(w))]. \tag{11}$$

Unlike the SATE_M, which only measures discrimination in the second decision, the total effect measures cumulative discrimination across *both* decisions. In our recurring example, the total effect captures the effect of race at the time of arrest on the subsequent charging decision. In particular, if a charged Black individual had instead been perceived as white by an officer, they might never have been arrested, and hence never been at risk of being charged, a possibility encompassed by the definition of the total effect, but not by the SATE_M.

We stress, however, that in studies of discrimination—particularly racial discrimination—there is often no clear intervention point, and the difference between the TE and the SATE_M is largely

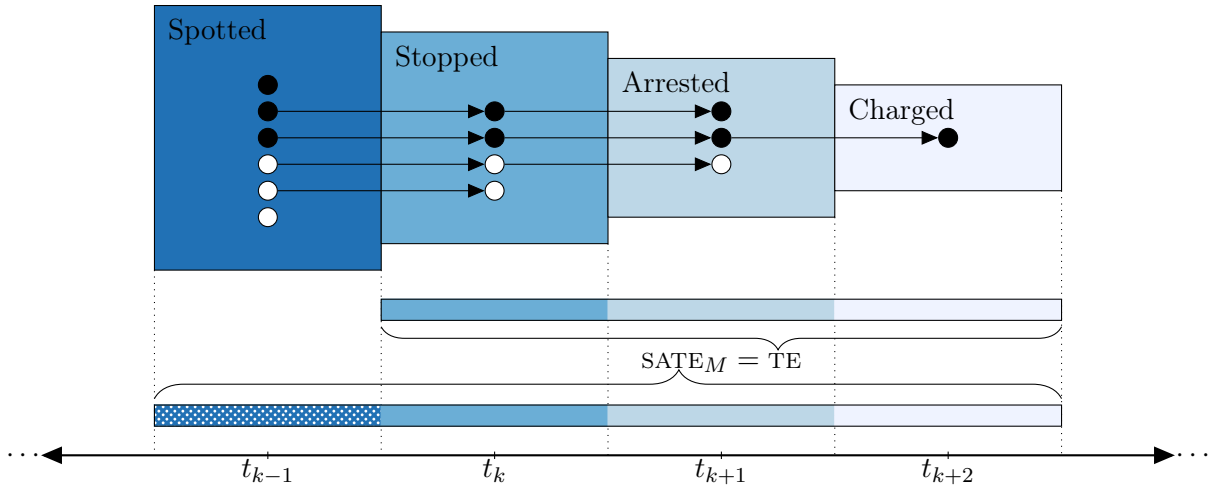


Figure 1: *This figure illustrates estimands one could consider, and the populations they concern, as individuals move through one segment of the criminal justice system. For instance, one can measure combined discrimination in arrest and charging decisions either via the* TE *or the* SATE_M*. In studies of discrimination, there is no clear point at which race is "assigned" and so both the* TE *and the* SATE_M *can be used interchangeably to express the same underlying causal effect, the* TE *with respect to the population of stopped individuals, and the* SATE_M *with respect to the population of spotted individuals. More generally, the diagram illustrates a multistage process, where one seeks to measure discrimination culminating at stage* t_{k+2} *(e.g., charging decisions) among those who make it to stage* t_k *(e.g., those who were stopped by the police). This quantity can be viewed as the* TE*, where one imagines the process starting at time* t_k*. Alternatively, it can be viewed as the* SATE_M*, where one views the process as starting earlier (at, say,* t_{k-1}*, indicating that an officer spotted an individual), and then conditioning on those who made it to stage* t_k*. Note that the quantities themselves are formally defined—and equivalent in the manner just described—even absent any considerations of estimation and randomization, which are not illustrated here.*

an artifact of how one defines both the population of interest and the start of the decision-making process. What is the TE in one description of events may be the SATE_M in another, equally valid description of the same events, as we describe next.

In our running example, the implicit population of interest consists of those individuals stopped by the police, and the TE reflects a description of events in which the decision-making process starts—and perception of race is counterfactually determined—when the arrest decision is made. We can, however, imagine moving back the clock and starting the process when the stop decision is made, with the population of interest now comprising those individuals spotted by an officer. In this case, the original TE is equivalent to the SATE_M on this newly defined population, where the first-stage decision indicates whether an individual was *stopped*. Both the original TE and the new SATE_M capture combined discrimination in the arrest and charging decisions, among the subset of individuals who were stopped.[6]

[6]To be explicit, our point is that the original TE and the new SATE_M are the same quantities, and hence are estimable using the same data. However, the new SATE_M (which subsets on individuals who are *stopped* among those who are *spotted*) and the original SATE_M (which subsets on individuals who are *arrested* among those who are *stopped*), are, in contrast, not equal in general, and not necessarily estimable using the same data. In particular, if one wants to estimate either the original TE or, equivalently, the new SATE_M, the arrest decision can be viewed as an intermediate variable, and, accordingly, subsetting to arrested individuals would in general introduce post-treatment bias.

But the moment when an individual is spotted is no more statistically privileged as a starting point than the moment when an officer makes a stop decision. One could similarly measure cumulative discrimination that includes the stop decision itself, either in terms of the TE or the SATE_M. For the TE, as above, we imagine time starting immediately after a potential police encounter, with the first-stage decision indicating whether an individual was stopped (among a population of individuals spotted by the officer). For the SATE_M, we back up the clock once again and imagine the first-stage decision indicating whether an individual was spotted by an officer, among an even larger population of people walking through the neighborhood where the officer patrols. Figure 1 provides a graphical depiction of this interchangeability.[7]

Although the TE may appear to avoid conditioning on intermediate outcomes, it simply masks a complex chain of events that came before the nominal start of the process, a chain that itself was likely influenced by discriminatory decisions. For instance, the officer spotting and stopping motorists in our running example could be patrolling the neighborhood in question because of its racial composition.[8] The very idea of "intermediate outcomes"—a concept central to concerns about post-treatment bias—is a slippery notion in the context of discrimination studies, where there is no clear point in time where one can imagine that race is "assigned." Even birth cannot be considered the ultimate starting point since, in theory, one might include, at the least, the race of a child's parents, determined at an earlier stage, when assessing discrimination.[9] Indeed, such generational counterfactuals may be critical for understanding systemic, institutional discrimination.

Our discussion of discrimination in multi-staged, multi-decider scenarios applies widely, but it is not universal. In particular, measuring discrimination in a single-decider case—and, specifically, in officer use of force—is challenging. In many of these single-decider scenarios, it is hard to imagine intervening on race after the decision-making process begins, making it difficult to isolate discrimination in later stages.

3 Assessing Second-Stage Discrimination in a Stylized Scenario

Subset ignorability, in theory, is sufficient to ensure nonparametrically identified estimates of the SATE_M, even when the first-stage decisions are discriminatory. We illustrate that idea by investigating in detail a hypothetical scenario involving discriminatory arrest decisions in the first stage and discriminatory charging decisions in the second stage. We explore the properties of simple estimators in this setting through a simulation study. We demonstrate that failing to adjust for a factor that directly influences charging decisions can result in biased estimates of discrimination in those decisions, but by accounting for all factors that directly influence charging decisions—and hence satisfying subset ignorability—one can accurately estimate the SATE_M, even when there is unmeasured confounding in arrest decisions. This example further clarifies the conceptual importance of distinguishing between an officer's perception of race and a prosecutor's perception of race when defining and estimating our quantities of interest.

[7] The formalism above shows a certain statistical equivalence between estimands having different starting points of the decision-making process. Nonetheless, the choice of starting point corresponds to measuring discrimination across different parts of the process, and so different estimands are relevant in different contexts. As such, we do not assert any normative ordering among them.

[8] Importantly, even if the population of individuals spotted by police at a street corner is a (near) random sample of people living or working in the neighborhood, we still cannot think of race as being randomly assigned in that subset. In particular, spotted individuals may still differ on a variety of dimensions (e.g., socio-economic status) across race groups. As such, one would need to statistically account for these differences in any analysis that seeks to measure disparate treatment.

[9] In the case of biological sex, one might consider assignment to occur at conception, though that is typically not the primary moment of interest in studies of sex discrimination.

We consider a hypothetical jurisdiction in which police officers observe the behavior and race of individuals who are potentially engaged in specific criminal activity (e.g., a drug transaction) and then decide whether or not to make an arrest. Subsequently, the case files of arrested individuals—consisting of a written copy of the officer's description of the encounter and the arrested individual's criminal history—are brought to a prosecutor who decides whether or not to press charges. We assume the prosecutor only observes the documented race and criminal record of the arrestee, and the arresting officer's written description of the encounter; accordingly, by construction, the charging decision depends only on these three factors. For example, the prosecutor may choose only to charge individuals who have several previous drug convictions and who were reported to be engaging in a drug transaction. Importantly, while the prosecutor has access to an officer's written report, the prosecutor does not directly observe the individual's behavior leading up to the arrest.

Our goal is to estimate discrimination in charging decisions, formalized in terms of the SATE_M. Intuitively, if we observe every arrested individual's criminal history, race, and officer report, then subset ignorability would hold because the prosecutor's charging decision depends only on these factors. Thus, with these three covariates, we could generate valid estimates of discrimination in prosecutorial decisions, even without knowing all of the factors that led to an arrest, a decision that may itself have been discriminatory. However, if any of these three covariates—criminal history, race, or officer report—are unobserved, we will, in general, be unable to accurately assess discrimination in prosecutorial decisions. In both scenarios, with and without unmeasured confounding, our analysis is based on the subpopulation of arrested individuals, where we note that the subsetting (i.e., arrest) is not influenced by the prosecutor's perception of race. In this setting, the primary concern is thus omitted-variable bias, not post-treatment bias.

We emphasize that we seek only to estimate discrimination in the second-stage charging decision, not cumulative discrimination stemming from both the arrest and charging decisions. In particular, while officer reports may represent an inaccurate—and discriminatory—account of events, such discrimination is distinct from that in the charging decision itself. Similarly, criminal histories reflect a form of complex, long-term discrimination that we do not aim to measure here. Alternative, and more expansive, notions of discrimination are important to understand, but here we focus on assessing the prosecutor's narrow contribution to inequities at a specific point in the process, a common statistical objective closely tied to policy decisions and legal theories of disparate treatment [Jung et al., 2018].

3.1 The data-generating process

We now formally describe the data-generating process for our stylized example. Under the structural causal model we consider, we can both compute the true SATE_M and compute estimates based only on select information available to the prosecutor. In defining the generative process, we closely follow the terminology and conventions of Pearl [2009] and Pearl et al. [2016].[10]

Our model is defined in terms of the causal directed acyclic graph (DAG) depicted in Figure 2. In this model, $S \in \{w, b\}$ indicates one's self-identified race, and D and Z indicate, respectively, an officer's and a prosecutor's perception of race. Further, $M \in \{0, 1\}$ indicates the arrest decision, and $Y \in \{0, 1\}$ indicates the charging decision. Finally, A corresponds to an individual's behavior, as

[10] In particular, we follow Pearl [2009] in representing unobserved confounding by bidirectional dashed arrows; see Section 1.2.1. We do deviate from Pearl in one aspect of our notation: we write counterfactuals as $Y(z, m)$ instead of $Y_{z,m}(u)$, suppressing the notational dependence on u. The former notation aligns with the popular Rubin-Neyman potential outcome notation that we use when defining the SATE_M. We further note that this SEM is included primarily for illustrative purposes, and consequently contains some simplifications, such as strictly binary covariates. In practice, we recommend reasoning about subset ignorability and its relevant potential outcomes directly.

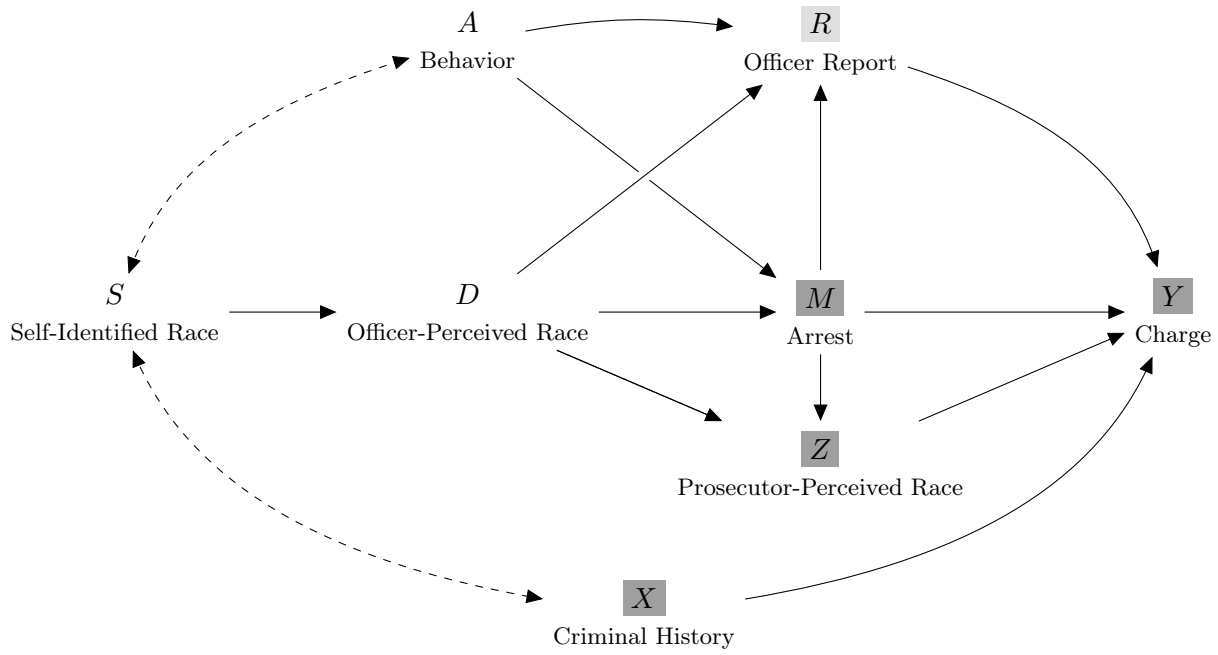


Figure 2: *A causal DAG depicting our stylized example of arrest and charging decisions, where D represents the officer's perception of race, and Z represents the prosecutor's perception of race. Officer arrest decisions (M) are directly influenced by observed criminal behavior (A) and officer-perceived race (D); the officer reports of the encounters (R) are directly influenced by A and D. Prosecutorial charging decisions are made for all arrested individuals, and are directly influenced by officer reports (R), criminal history (X), and prosecutor-perceived race (Z). Finally, an individual's self-identified race (S) influences the officer's perception of race (D), and is confounded with criminal history (X) and behavior (A). We consider two scenarios. The variables highlighted in dark gray (i.e., M, Z, X, and Y) are always observed. In one scenario, the analyst also observes the officer report R, highlighted in light gray, obtaining the full set of information available to the prosecutor; in the other, the analyst does not observe the officer report R (i.e., only M, Z, X, and Y are observed), leading to omitted-variable bias.*

observed by an officer, and X and R correspond, respectively, to criminal history and an officer's description of an encounter, as included in the arrest report. For simplicity, in our example these latter three variables are operationalized as being binary—for example, one can imagine that X indicates whether an individual had at least one previous drug conviction, A indicates whether they were seen actively engaging in a drug transaction, and R indicates whether they were reported by the officer to be actively engaging in a drug transaction. Officers observe D, A, and R for all individuals; prosecutors observe Z, X, and R only for the subset of arrested individuals. Note that we also allow for Z and R to be missing (i.e., to take the value NA) in cases where an individual is not arrested.

Structural causal models are defined by a set of exogenous random variables and deterministic structural equations specifying the values of all other variables in the DAG. In our example, the independent exogenous variables are:

$$U_L \sim \textsc{Bern}(\mu_L),$$
$$U_A,\ U_M,\ U_X,\ U_R,\ U_Y \sim \textsc{Unif}(0,1),$$

where μ_L is an appropriately defined constant.

We define self-identified race (S), behavior (A), and criminal history (X) in terms of U_L, which captures latent confounding. For constants μ_A, γ, μ_X, and δ, the structural equations for these three variables are given by:

$$\begin{aligned} f_S(u_L) &= \begin{cases} w & u_L = 0, \\ b & u_L = 1, \end{cases} \\ f_A(u_A, u_L) &= \mathbb{1}(u_A \leq \mu_A + \gamma \cdot u_L), \\ f_X(u_X, u_L) &= \mathbb{1}(u_X \leq \mu_X + \delta \cdot u_L). \end{aligned}$$

This specification allows for the distributions of criminal history and behavior to vary by race due to exogenous factors like disparate police deployment and historical discrimination. For example, stopped Black individuals may be less likely to be engaged in criminal activity than stopped white individuals, corresponding to $\gamma < 0$.

In line with our discussion in Section 2.1, we set the prosecutor's perception of race (Z) equal to the officer's perception of race (D), and, for simplicity, we set both equal to one's self-identified race (S). This choice yields the following structural equations:

$$\begin{aligned} f_D(s) &= s, \\ f_Z(m, d) &= \begin{cases} d & m = 1, \\ \text{NA} & m = 0. \end{cases} \end{aligned}$$

Note that, when someone is not arrested, we represent the prosecutor's perception of race as an explicit missing value. The arrest report, R, is treated similarly below.

Finally, for constants α_0, α_A, α_{black}, λ_0, λ_A, λ_{black}, β_0, β_X, β_R, and β_{black}, the structural equations for arrest decisions (M), police reports (R), and charging decisions (Y) are given by:

$$\begin{aligned} &f_M(d, a, u_M) \\ &\qquad = \mathbb{1}(u_M \leq \alpha_0 + \alpha_A \cdot a + \alpha_{\text{black}} \cdot \mathbb{1}(d = b)), \\ &f_R(d, a, m, u_R) \\ &\qquad = \begin{cases} \mathbb{1}(u_R \leq \lambda_0 + \lambda_A \cdot a + \lambda_{\text{black}} \cdot \mathbb{1}(d = b)) & m = 1, \\ \text{NA} & m = 0, \end{cases} \\ &f_Y(z, m, r, x, u_Y) \\ &\qquad = \begin{cases} \mathbb{1}(u_Y \leq \beta_0 + \beta_X \cdot x + \beta_R \cdot r + \beta_{\text{black}} \cdot \mathbb{1}(z = b)) & m = 1 \wedge\ z \neq \text{NA} \wedge r \neq \text{NA}, \\ \text{NA} & m = 1 \wedge\ (z = \text{NA} \vee r = \text{NA}), \\ 0 & m = 0. \end{cases} \end{aligned}$$

In particular, arrest decisions and police reports depend on an officer's perception of race, whereas charging decisions depend on a prosecutor's perception of race. This model incorporates both discrimination in arrest decisions, via α_{black}, and discrimination in police reports—e.g., by omitting potentially exculpatory details or by falsifying information—via λ_{black}. Discrimination in charging decisions is encoded by β_{black}.

The above structural equations, together with the distributions on the exogenous variables, fully

Table 1: *A sample of potential and realized outcomes for individuals in our hypothetical example. The data-generating process produces the full set of entries, but the prosecutor only observes the realized outcomes for those who were arrested, indicated by the shaded cells. In the first scenario we consider, the analyst also observes all the information in the shaded cells; in the second scenario, the analyst only observes the information in the dark gray cells (i.e., the analyst does not observe the officer report R), leading to omitted variable bias.*

S	D	A	X	$M(b)$	$M(w)$	M	Z	R	$R(0)$	$R(1)$	$Y(b,1)$	$Y(w,1)$	Y
b	b	0	0	0	0	0	NA	NA	NA	0	0	0	0
b	b	0	1	0	0	0	NA	NA	NA	1	1	0	0
b	b	1	1	1	0	1	b	0	NA	0	1	1	1
w	w	0	0	1	1	1	w	1	NA	1	0	0	0
w	w	0	1	0	0	0	NA	NA	NA	0	0	0	0

define the joint distribution of realized and potential outcomes. In particular,

$$\begin{aligned} S &= f_S(U_L), & D &= f_D(S), \\ Z &= f_Z(M, D), & A &= f_A(U_A, U_L), \\ X &= f_X(U_X, U_L), & M &= f_M(D, A, U_M), \\ R &= f_R(D, A, M, U_R), & Y &= f_Y(Z, M, R, X, U_Y). \end{aligned}$$

The primary causal quantity we seek to estimate—the SATE_M—is defined in terms of counterfactuals $Y(z, m)$. As discussed in Pearl [2009] and Pearl et al. [2016], such counterfactuals require some care to define, as one must appropriately account for the exogenous variables U. In particular, for the causal DAG in Figure 2, the bivariate charge potential outcomes, for counterfactual versions of prosecutor-perceived race, are given by $Y(z, m) = f_Y(z, m, R(m), X, U_Y)$, where $R(m) = f_R(D, A, m, U_R)$ are the counterfactual versions of the officer report. Further, the arrest potential outcomes—where we consider counterfactual versions of officer-perceived race—are given by $M(d) = f_M(d, A, U_M)$. In general, counterfactuals defined in this way obey the consistency rule, meaning that $M = M(D)$ and $Y = Y(Z, M)$.

When $\alpha_{\text{black}} \geq 0$, anyone who would be arrested if white would also be arrested if Black (i.e., $M(b) \geq M(w)$). When $\alpha_{\text{black}} > 0$, we say arrest decisions are discriminatory since, all else being equal, an individual is more likely to be arrested if they were Black than if they were white. Likewise, $Y(b, 1) \geq Y(w, 1)$ when $\beta_{\text{black}} \geq 0$, meaning that an individual who would be charged if arrested and white would also be charged if arrested and Black. We say the charging decision is discriminatory when $\beta_{\text{black}} > 0$.

Features of our data-generating process. Table 1 displays a sample of five rows of data generated from our model. From the full set of potential outcomes, we can compute the true SATE_M by directly applying Definition 1 to the generated data, taking the average difference between $Y(b, 1)$ and $Y(w, 1)$ among arrested individuals.[11] However, given the simple linear form of our structural equations, a straightforward calculation also shows that the SATE_M is exactly equal to β_{black}.

[11] Because Z and D are separately manipulable in our framing, this quantity—obtained by first subsetting on arrested individuals, and then computing the average difference between potential outcomes—can also be expressed in the *do*-calculus: $\text{SATE}_M = \mathbb{E}[Y \mid do(Z = b), M = 1] - \mathbb{E}[Y \mid do(Z = w), M = 1]$. However, as is common in

Our hypothetical example captures three key features of real-world discrimination studies. First, prosecutorial records do not contain all information that influenced officers' first-stage arrest decisions (i.e., prosecutors only observe R, not A). Second, our set-up allows for situations where the arrest decisions are themselves discriminatory—those where $\alpha_{\text{black}} > 0$—or the officer's report is discriminatory, e.g., because of omission of exculpatory information or deliberate falsification—those where $\lambda_{\text{black}} > 0$. Third, the prosecutor's records include the full set of information on which charging decisions are based (i.e., Z, X, and R).

Among those who were arrested, the charging potential outcomes depend only on one's criminal history (X) and the arrest report (R). In particular, they do not depend on one's realized, prosecutor-perceived race (Z). Consequently, $Y(z,1) \perp\!\!\!\perp Z \mid X, R, M = 1$, meaning that the model satisfies subset ignorability relative to X and R. As a result, access to X and R, along with overlap, guarantees the stratified difference-in-means is a consistent estimator of the SATE_M, even if one does not have access to A.[12] However, in general, $Y(z,1) \not\perp\!\!\!\perp Z \mid X, M = 1$ (and, likewise, $Y(z,1) \not\perp\!\!\!\perp Z \mid R, M = 1$), and so if one only has partial information on charging decisions there is no guarantee the SATE_M can be consistently estimated.[13] Indeed, when there is such unmeasured confounding in the prosecutor's decisions, one should expect biased estimates of the SATE_M.

3.2 Estimating the $\texttt{sate}_M$

Although the data-generating procedure produces the full set of potential outcomes for each individual, the prosecutor only observes a subset of the cells—realized outcomes for arrested individuals, highlighted in gray in Table 1. While this circumscribes the causal effects one can estimate—e.g., discrimination by police will no longer be identifiable in the reduced dataset—one can still learn about the SATE_M. We explore the performance of two statistical methods for estimating the SATE_M based on data observed by the prosecutor: the stratified difference-in-means estimator described in Eq. (9), and a regression-based estimator. We apply each of these methods to two types of data: the full set of information available to prosecutors (i.e., Y, Z, X and R), and an incomplete dataset comprised only of Y, Z, and X (highlighted in dark gray in Table 1), in which case we view R as an unmeasured confounder.

One can compute the stratified difference-in-means estimate in three steps. First, partition arrested individuals into subsets that have the same value of the available control variables (i.e., X and R in the complete data setting, and X alone in the partial data setting). Second, on each resulting subset, compute the average difference in charging rates between Black and white individuals. Third, take a weighted average of these differences, where the weights reflect the proportion of arrested individuals in each subset. In addition, one can apply Eq. (10) to estimate the standard error of this point estimate to generate confidence intervals.

The stratified difference-in-means estimator is theoretically appealing in that it is guaranteed to yield consistent estimates of the SATE_M when subset ignorability and overlap hold. But the estimator can have high variance when the dimension of the covariate space is high and the sample

causal mediation analysis, if there were only one indecomposable treatment (e.g., if one instead imagined directly manipulating S) then the corresponding estimand could no longer be expressed using *do*-operations alone [Pearl, 2009, 2015].

[12] In general, first-stage discrimination such as discriminatory arrest decisions or fabrication of evidence in arrest reports does not affect the consistency of the stratified difference-in-means estimator, since subset ignorability will continue to hold. Consistency may fail if discrimination is so extreme that overlap fails, e.g., if no white people are arrested.

[13] In the prosecutorial context, sufficiently diligent data gathering can mitigate this possibility; many offices maintain detailed case files, and we make use of such records in our empirical analysis in Section 4. In general studies of discrimination, it is important to ensure that decision factors are accurately captured and made available to analysts.

size is small. Thus, in practice, it is common to model potential outcomes as a function of observed covariates—also known as response surface modeling [Hill, 2011]. In particular, on the subset of arrested individuals, one can estimate the SATE_M via a parametric model that estimates observed charging decisions as a function of the available information.

To demonstrate this latter approach, we use a linear probability model. In the complete data setting, we have:

$$\mathbb{E}[Y \mid Z, X, R] = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 R, \tag{12}$$

where the model is fit on the full set of arrests seen by the prosecutor. Under this model, the SATE_M is approximated by the fitted coefficient $\hat{\beta}_1$, since that term captures the difference in charging potential outcomes after adjusting for the observed covariates. For our specific stylized example, the linear regression model in Eq. (12) is in fact perfectly specified—exactly mirroring the prosecutor's charging decisions—and so we are guaranteed to obtain statistically consistent estimates. In the partial data setting, where an analyst only has access to X, one must fit a reduced model that excludes R:

$$\mathbb{E}[Y \mid Z, X] = \beta_0 + \beta_1 Z + \beta_2 X. \tag{13}$$

In this case, $\hat{\beta}_1$ in general yields a biased estimate of the SATE_M, because of the omitted variable R. The stratified difference-in-means estimator will in general similarly yield a biased estimate of the SATE_M in this omitted-variable setting.

3.3 Simulation results

We perform a simulation study to understand the properties of the above estimators, varying our assumptions about discrimination and confounding. We simulate 10,000 datasets of size 100,000 for each of 25 different parameter settings. Each setting is defined as a combination of our two key discrimination parameters, α_{black} and β_{black}, where each parameter is allowed to take one of five values: 0.20, 0.25, 0.30, 0.35, and 0.40. Across all simulation settings, we assume the population of individuals encountered by police is 30% Black (i.e., $\mu_L = 0.3$); that 30% of white individuals and 40% of Black individuals have a past drug conviction, indicated by X; and that 30% of white individuals and 20% of Black individuals are seen engaging in a drug transaction, indicated by A.[14] These settings allow for a substantial amount of overlap across race groups with regard to the key covariates.

On each synthetic dataset, we estimate the SATE_M using both the stratified difference-in-means estimator and the regression-based estimator, and compare the results to the true population-level SATE_M in two scenarios. To illustrate the impact of omitted variable bias, in the first scenario, we assume the officer's report R is unavailable—meaning there is unmeasured confounding—and therefore only stratify based on X in the difference-in-means estimator, and fit the model in Eq. (13) for the regression-based estimator. In the second scenario, we assume that R is available, and stratify on both X and R in the difference-in-means estimator, and fit the model in Eq. (12) for the regression-based estimator. For each combination of α_{black} and β_{black}, the estimates on the 10,000 synthetic datasets yield the approximate sampling distributions for the difference-in-means and regression-based estimators. In Figure 3, we summarize each sampling distribution by its mean, 2.5th percentile, and 97.5th percentile. The solid points correspond to the difference-in-means estimator,

[14] More specifically, the full set of parameters in our simulation was set as follows: $\mu_L = 0.3, \mu_X = 0.3$, $\mu_A = 0.3$, $\delta = 0.1, \gamma = -0.1$, $\alpha_0 = 0.1$, $\alpha_A = 0.3$, $\alpha_{\text{black}} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$, $\lambda_0 = 0.2$, $\lambda_A = 0.6$, $\lambda_{\text{black}} = 0.1$, $\beta_0 = 0.2$, $\beta_X = 0.4$, $\beta_R = 0.2$, and $\beta_{\text{black}} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$.

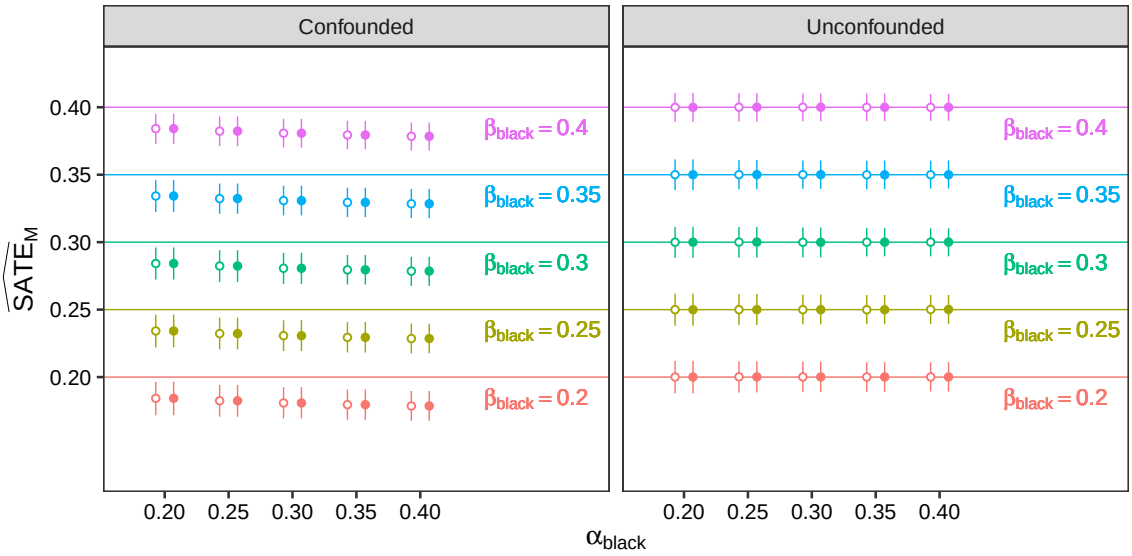


Figure 3: *In our hypothetical example of officer and prosecutor behavior, estimates of discrimination in charging decisions are biased when information directly influencing those decisions—in this case, an officer's report—is omitted (left). However, one can obtain accurate estimates of discrimination when accounting for all information directly influencing charging decisions (right). Each plot shows the results of 10,000 simulations for each of 25 different combinations of discrimination in officer and prosecutor decisions, given by α_{black} and β_{black}, respectively. The true value of the* SATE_M*, indicated by the horizontal colored lines, is computed based on the full set of potential outcomes for each individual, and does not depend on the degree of discrimination in the first stage, as seen by the constant value of the* SATE_M *across different values of α_{black}. For each parameter choice, we display the mean of the sampling distribution for the stratified difference-in-means estimator (solid circle) and the regression-based estimator (hollow circle), along with the interval spanned by the 2.5th and 97.5th percentiles of the sampling distribution. In the right plot ("unconfounded"), estimates are based on all three factors that directly influence charging decisions: race, criminal history, and officer report; in the left plot ("confounded"), we omit the report. When all variables directly influencing charging decisions are available, both estimators recover the true value of the* SATE_M*, even when there is an unknown degree of discrimination in arrest decisions.*

and the hollow points to the regression-based estimator. The horizontal lines indicate the true population-level SATE_M.

In the left panel ("confounded") of Figure 3, the points lie below the horizontal lines in all cases, meaning we underestimate discrimination in charging decisions. In this setting, estimates do not account for the officer reports R, and so there is unmeasured confounding in the charging decisions. We set $\gamma < 0$ in our simulations, and thus stopped and arrested Black individuals are less likely to be engaging in criminal activity, a pattern (noisily) reflected in the officer reports. Because we assume these arrest reports are not available for analysis, we cannot fully adjust for their direct influence on prosecutor decisions. As a result, by adjusting for X alone, we miss an important, unmeasured difference between arrested white and Black individuals, leading us to underestimate discrimination in prosecutorial decisions.

In the right panel ("unconfounded") of Figure 3, the points lie on the horizontal lines in all cases,

meaning the estimators are unbiased, and the range between the 2.5th and 97.5th percentiles is relatively narrow, indicating estimates are typically close to the true value. These results hold even when one is unable to assess the degree of discrimination α_{black} in the arrest decisions. As implied by Theorem 4, to accurately estimate the SATE_M, it is sufficient to measure all covariates that directly influence the prosecutor's decisions. In practice, it is nearly always impossible to do so perfectly; for instance, decision factors such as forensic evidence may not be readily available, or non-obvious factors, such as the time of day, may play a role in the prosecutor's charging decision. Thus it is important to gauge the sensitivity of estimates to unmeasured confounding in those decisions, as we demonstrate with real-world data in Section 4 below. The key point is that it is sufficient to adjust for unmeasured confounding in the charging decisions alone; to estimate discrimination in these charging decisions—formalized by the SATE_M—one need not account for unmeasured confounding in either the documents generated by police, such as arrest reports, or the arrest decisions themselves.

Finally, in addition to examining the sampling distributions, we assessed the coverage of our 95% confidence intervals. For the difference-in-means estimator, confidence intervals were constructed via the estimated standard error given by Eq. (10); and for the regression-based estimator, we used the conventional OLS estimate of standard error. For each parameter setting, we computed the proportion of confidence intervals for the 10,000 datasets that contained the true value of the SATE_M. In the no-confounding scenario, we found the true coverage was in line with the nominal coverage, ranging from 94% to 96% across parameter specifications. In the confounding scenario, the intervals rarely covered the true values, as expected, with coverage ranging from 1% to 30% across parameters.

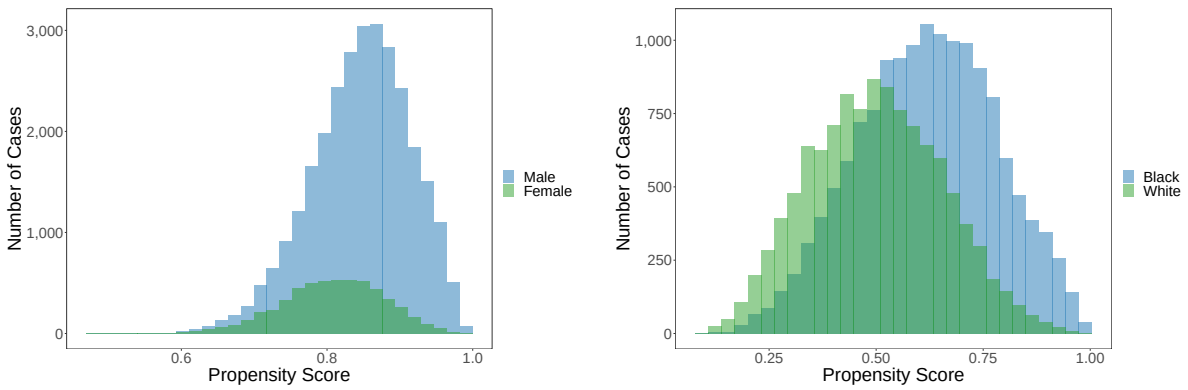
4 An Empirical Analysis of Prosecutorial Charging Decisions

We now apply the statistical framework developed above to assess possible race and gender discrimination in real-world prosecutorial charging decisions. We start with the set of individuals in a major U.S. county who were arrested for a felony offense between 2013 and 2019. For our race-based analysis, we then limit to the 25,918 instances in which the race of the arrested individual was identified as either Black (14,686) or non-Hispanic white (11,232), and for our gender-based analysis we limit to the 34,871 instances in which the gender of the arrested individual was recorded as either male (29,283) or female (5,588).[15]

Our dataset includes a variety of information about each case, including the criminal history of the arrested individual; the alleged offenses (e.g., burglary); the location, date, and time of the incident; whether there is body-worn camera footage; whether a weapon was involved; whether an elderly victim was involved; and whether there was gang involvement. We also know the ultimate charging decision for each case. Disaggregating by gender, 51% of cases involving a male arrestee were charged, compared to 45% of cases involving a female arrestee; and disaggregating by race, 51% of cases involving a Black arrestee were charged, compared to 50% of cases involving a white arrestee.

To gauge the extent to which charging decisions may suffer from disparate treatment by race or gender, we estimate the SATE_M. We start by checking that overlap is satisfied for both our race-based and our gender-based analyses. Recall that overlap means $\Pr(Z = z \mid X = x, M = 1) > 0$, where $Z = 1$ indicates an individual's "treatment" status (i.e., whether an individual is male in our analysis of gender discrimination, or Black in our analysis of racial discrimination), X is a vector of observed case features, and $M = 1$ means we restrict to those individuals who were arrested. In contrast to

[15] Both Hispanic and non-Hispanic white individuals in our dataset appear to have been recorded simply as "white". To disentangle these two categories, we followed past work and imputed Hispanic ethnicity from surnames [Pierson et al., 2020, Word and Perkins, 1996, Word et al., 2008].



(a) Gender-based analysis

(b) Race-based analysis

Figure 4: *We plot, for both our gender-based (left) and race-based (right) analyses, the distribution of propensity scores, disaggregated by observed treatment status. We find that the propensity scores are concentrated away from the interval endpoints, satisfying overlap.*

ignorability, overlap can be assessed directly by examining the data. To do so, we estimate propensity scores [Rosenbaum and Rubin, 1983a], $\Pr(Z = z \mid X = x, M = 1)$, via an L^1-regularized (lasso) logistic regression model. In Figure 4, we plot the distribution of the estimated propensity scores. In the left panel we disaggregate by gender, and in the right panel we disaggregate by race (Black and white). In situations where overlap does not hold, it is common to restrict one's analysis to a region of the covariate space where it does hold. In our case, however, the vast majority of the data are already far from the endpoints of the unit interval, so we work with the dataset in its entirety.

As discussed in Section 3, regression-based estimators can be viewed as a parametric variant of the stratified difference-in-means estimator Δ_n. Thus, to help account for the high dimensionality of our feature set, we now estimate the SATE_M via linear regression. In particular, for ease of interpretation, we use a linear probability model:

$$\mathbb{E}[Y \mid Z, X] = \beta_0 + \beta_1 Z + \beta_2^T X, \tag{14}$$

where Y indicates whether an arrested individual was charged, and X denotes the vector of covariates.

In the gender model, we find that the $\widehat{\text{SATE}}_M$—as given by $\hat{\beta}_1$—is 0.025 (95% CI: [0.014, 0.037]); and in the race model, we have $\widehat{\text{SATE}}_M$ is -0.008 (95% CI: $[-0.018, 0.002]$). These results indicate that the charging rate for men is slightly higher than the rate for similar women, and that the charging rate for Black individuals is on par with that of similar white individuals, mirroring the patterns we saw with the raw, unadjusted charging rates. If there are no unmeasured confounders (i.e., if subset ignorability holds) and our parametric model is appropriate, these results suggest race and gender have a relatively modest impact on charging decisions in the jurisdiction we consider.

To help contextualize these results, we note that past studies have found mixed evidence of disparate treatment in prosecutorial charging decisions, likely due in part to differences in the jurisdictions and time periods analyzed, and the methods employed. In one of the most comprehensive investigations to date, Rehavi and Starr [2014] examined nearly 40,000 individuals in the federal criminal justice system from initial arrest to final sentencing. The authors found that disparate treatment in prosecutorial charging decisions—specifically for charges with statutory mandatory minimum sentences—was a primary driver for sentencing disparities between Black and white individuals. In contrast, in a recent experimental study, Robertson et al. [2019] found no evidence of

racial bias in charging decisions when they presented prosecutors with vignettes in which the race of the suspect was randomly varied. Similarly, in an observational analysis of prosecutors at the San Francisco District Attorney's Office, MacDonald and Raphael [2021] found little evidence of discrimination in charging decisions—in fact, the authors found that white individuals were charged slightly more often than similarly situated Black individuals. Finally, in a recent quasi-random study of charging decisions at a large metropolitan district attorney's office, Chohlas-Wood et al. [2021] similarly found little evidence of disparate treatment.

The AUC of our outcome model in Eq. (14) above—fit with all available covariates, including race and gender—is 86%, indicating that it can predict charging decisions well. Our model, however, cannot capture all aspects of prosecutorial decision making, as at least some information used by prosecutors (e.g., forensic evidence) is not recorded in our dataset, meaning that subset ignorability likely does not hold exactly. To check the robustness of our causal estimates to such unmeasured confounding, one may use a variety of statistical methods for sensitivity analysis [Carnegie et al., 2016, Dorie et al., 2016, Franks et al., 2019, Imbens, 2003, Jung et al., 2020, McCandless and Gustafson, 2017, McCandless et al., 2007, Rosenbaum and Rubin, 1983b]. At a high level, these methods posit relationships between the unmeasured confounder and both the treatment variable (e.g., race or gender) and the outcome (e.g., the charging decision), and then examine the sensitivity of estimates under the model of confounding.

We apply a technique for sensitivity analysis recently introduced by Cinelli and Hazlett [2020]. In brief, their approach bounds the extent to which a coefficient estimate in a linear model—like $\hat{\beta}_1$ in Eq. (14)—might change if one were to refit the model including an unmeasured confounder U. More specifically, under the extended model

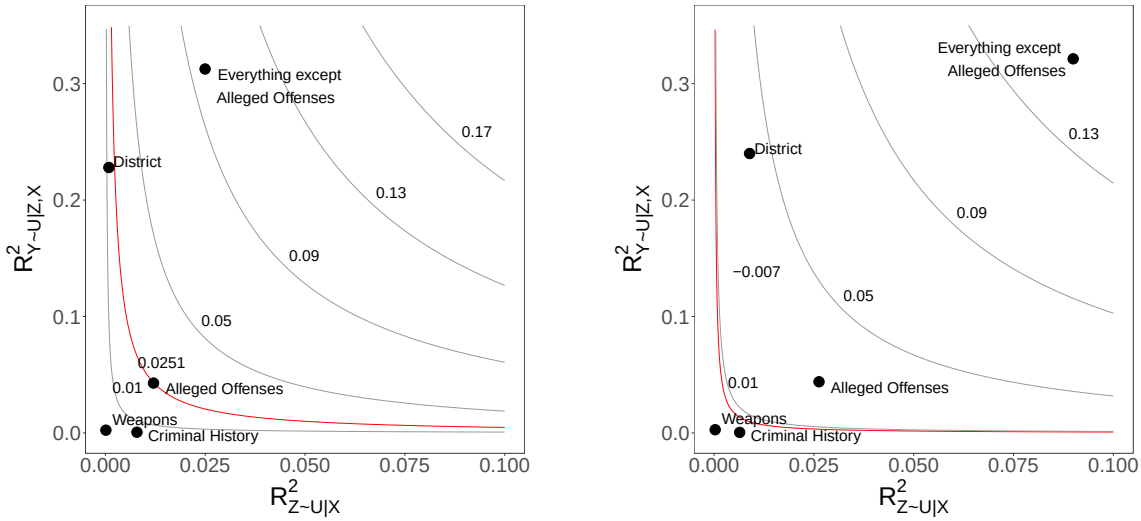
$$\mathbb{E}[Y \mid Z, X, U] = \beta_0 + \beta_1 Z + \beta_2^T X + \gamma U,$$

Cinelli and Hazlett bound the change in $\hat{\beta}_1$ in terms of two partial R^2 values: $R^2_{Y\sim U|Z,X}$ and $R^2_{Z\sim U|X}$. These two values respectively quantify how much residual variance in the outcome Y and treatment Z is explained by U. Formally, $R^2_{Y\sim U|Z,X}$ is defined in terms of the R^2 of two linear regressions: one using all the covariates X, Z, and U to estimate Y (R^2_{full}), and one excluding U (R^2_{red}). Then, $R^2_{Y\sim U|Z,X} = (R^2_{\text{full}} - R^2_{\text{red}})/(1 - R^2_{\text{red}})$. The quantity $R^2_{Z\sim U|X}$ is defined analogously. As these partial R^2 values increase, so does the amount by which $\hat{\beta}_2$ could change.

The contour plots in Figure 5 show the maximum amount by which the $\widehat{\text{SATE}}_M$ may change as a function of $R^2_{Y\sim U|Z,X}$ and $R^2_{Z\sim U|X}$ for our analysis of gender and race—with that change potentially increasing or decreasing the estimate. The red lines trace out values for which the maximum change equals our empirical point estimates of the $\widehat{\text{SATE}}_M$. In particular, an unmeasured confounder lying above the red line could be sufficient to change the sign of our estimate.

A key hurdle in sensitivity analysis is positing a reasonable range for the strength of a possible unmeasured confounder. To aid interpretation, we compute the partial R^2 values for various subsets of observed covariates, as recommended by Cinelli and Hazlett. For each such subset, we fit the regression model in Eq. (14) both with and without that subset, which in turn yields a pair of partial R^2 values for that subset of covariates.

The contour plots in Figure 5 contain these reference points for five different subsets of covariates: (1) the subset describing criminal history (e.g., number of prior convictions and number of prior arrests); (2) the alleged offenses (e.g., burglary); (3) the subset of all covariates except for the alleged offenses; (4) the district in which the alleged incident took place; and (5) whether a weapon was alleged to have been used. We find that the partial R^2 values associated with criminal history and whether a weapon was used are below the red curves for both our analysis of gender and race, indicating that a confounder with comparable marginal explanatory power to these covariates would



(a) Gender-based analysis

(b) Race-based analysis

Figure 5: *Contour plots describing the sensitivity of the* $\widehat{\text{SATE}}_M$ *to unmeasured confounding, for our analysis of gender (left) and race (right). The plots indicate the maximum amount the* $\widehat{\text{SATE}}_M$ *may change under the Cinelli and Hazlett [2020] model of confounding, parameterized by two partial* R^2 *values. The red curves correspond to a change equalling the magnitude of the* $\widehat{\text{SATE}}_M$ *estimated from the available data. Thus, an unobserved confounder corresponding to a point above the red curve would be capable of changing the sign of our estimate. To aid interpretation, both plots display the partial* R^2 *values associated with several observed subsets of covariates.*

not be sufficient to change the sign of our estimates. However, the partial R^2 values corresponding to the alleged offenses and the district in which the charges were filed are near the red curve for our gender-based analysis and far above the curve for our race-based analysis, meaning that omitting a covariate with similar explanatory power could qualitatively change our conclusions. Furthermore, the partial R^2 values corresponding to everything except the alleged offenses are far above the red curve in both cases, suggesting that an unobserved confounder of similar strength could again substantially alter our results. For instance, in this extreme scenario, inclusion of a currently omitted confounder with similar characteristics in the race-based analysis could yield an estimated treatment effect of more than 13%.

One cannot know the exact nature and impact of unmeasured confounding. Thus, as in many applied statistical problems, we must rely in large part on domain expertise and intuition to form reasonable conclusions. In this case, we interpret our results as providing evidence that perceived gender and race have limited effects on prosecutorial charging decisions in the jurisdiction we consider. As with the SATE_M, our sensitivity analysis is solely focused on discrimination in the charging decision, and, in particular, is not designed to capture the cumulative effects of discrimination stemming from arrests and other earlier decision points.

5 Discussion

We have outlined a formal causal framework to ground observational studies of discrimination. We specifically showed that subset ignorability, together with overlap, is sufficient to guarantee that one important causal measure of discrimination (the SATE_M) is nonparametrically identified in a canonical two-stage decision-making setting. In this context, potential issues of post-treatment bias are more appropriately thought of as concerns about omitted variables. We demonstrated that a traditional regression-based analysis can be used to assess discrimination in real-world prosecutorial charging decisions, even though the underlying arrests may have been discriminatory in unknown ways. In that example—as in many applied settings—subset ignorability may only hold approximately, and our empirical analysis illustrates the importance of sensitivity analysis for robust inference.

Stepping back, there are at least two broad notions of discrimination, which approximately map to the legal concepts of disparate treatment and disparate impact. Both involve causal interpretations, though with key differences in the definition of the estimand. Disparate treatment concerns the causal effect of race on outcomes, with behavior often driven by animus or explicit racial categorization. Disparate impact, on the other hand, concerns the causal effect of policies or practices on unjustified racial disparities, regardless of intent. Disparate treatment and disparate impact both play important roles in legal and policy discussions, and the perspective one adopts in any given situation affects the choice of statistical estimation strategy and the interpretation of results [Jung et al., 2018].

We have throughout focused on the statistical foundations and measurement of disparate treatment. In our primary example, we estimate—assuming subset ignorability holds—that perceived race and gender have relatively small effects on prosecutorial charging decisions in the jurisdiction we examine. We further demonstrate that these estimates are moderately robust to potential omitted-variable bias. However, that finding, in and of itself, does not mean charging decisions are equitable in a broader sense. Consider, for example, the 1,637 cases in our data involving alleged possession of controlled substances by Black or non-Hispanic white individuals. Of these, 748 cases (46%) were ultimately charged, and charging rates by race were nearly identical across race groups, offering little prima facie evidence of disparate treatment. However, among the 748 charged cases, 464 (62%) involved a Black individual—far exceeding the proportion of Black residents in the county we study. Charging decisions for these cases thus impose a heavy burden on Black individuals, even if those decisions were not tainted by animus. To the extent that prosecution of drug crimes is misaligned with community goals, these decisions create an unjustified, and discriminatory, disparate impact.

Rigorously estimating discrimination is a daunting task that requires careful consideration. At an empirical level, it is often difficult to obtain detailed data on individual decisions, in which case benchmark analysis may be inadequate—even if coupled with sensitivity analysis. At a theoretical level, we have a limited statistical language to make precise concepts such as animus and implicit bias that are central to discrimination research. Further, as we note above, past work has often framed discrimination as the causal effect of race on behavior, but other conceptions of discrimination, such as disparate impact, are equally important for assessing and reforming practices. Finally, the conclusions of discrimination studies are generally limited to specific decisions that happen within a long chain of potentially discriminatory actions. Quantifying discrimination at any one point (e.g., in charging decisions) does not yield estimates of specific or cumulative discrimination at other points (e.g., in arrest decisions). Despite these important considerations, we hope our work helps place discrimination research on more solid statistical footing, and provokes further interest in the subtle conceptual and methodological issues at the heart of discrimination studies.

References

I. Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2): 131–142, 2002.

A. I. Balsa, T. G. McGuire, and L. S. Meredith. Testing for statistical discrimination in health care. *Health Services Research*, 40(1):227–252, 2005.

S. Baum and E. Goodstein. Gender imbalance in college applications: Does it lead to a preference for men in the admissions process? *Economics of Education Review*, 24(6):665–675, 2005.

N. Berg and D. Lien. Measuring the effect of sexual orientation on income: Evidence of discrimination? *Contemporary Economic Policy*, 20(4):394–414, 2002.

M. Bertrand and S. Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

P. Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

N. B. Carnegie, M. Harada, and J. L. Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

A. Chohlas-Wood, J. Nudell, K. Yao, Z. Lin, J. Nyarko, and S. Goel. Blind justice: Algorithmically masking race in charging decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 35–45, 2021.

C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *J. R. Statist. Soc. B*, 2020.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.

V. Dorie, M. Harada, N. B. Carnegie, and J. Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, 35(20):3453–3470, 2016.

B. G. Edelman and M. Luca. Digital discrimination: The case of airbnb.com. *Harvard Business School NOM Unit Working Paper*, 14(054), 2014.

A. Franks, A. D'Amour, and A. Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, pages 1–33, 2019.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics, New York, 2001.

R. G. Fryer Jr. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261, 2019.

A. Gelman, J. Fagan, and A. Kiss. An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC press, 2013.

S. Goel, M. Perelman, R. Shroff, and D. A. Sklansky. Combatting police discrimination in the age of big data. *New Criminal Law Review: An International and Interdisciplinary Journal*, 20(2): 181–232, 2017.

C. Goldin and C. Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, 2000.

D. Greenberg, C. Gershenson, and M. Desmond. Discrimination in evictions: Empirical evidence and legal challenges. *Harv. CR-CLL Rev.*, 51:115, 2016.

D. J. Greiner and D. B. Rubin. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785, 2011.

J. Grogger and G. Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475):878–887, 2006.

J. J. Heckman and S. N. Durlauf. Comment on "An empirical analysis of racial differences in police use of force" by Roland G. Fryer Jr. *Journal of Political Economy*, 2020.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986.

K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychol. Methods*, 15(4):309–334, 2010a.

K. Imai, L. Keele, and T. Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.*, 25(1):51–71, 2010b.

G. W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, 93(2):126–132, 2003.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

J. Jung, S. Corbett-Davies, R. Shroff, and S. Goel. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*, 2018.

J. Jung, R. Shroff, A. Feller, and S. Goel. Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 64–70, 2020.

D. Knox, W. Lowe, and J. Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 2020.

J. MacDonald and S. Raphael. Effect of scaling back punishment on racial and ethnic disparities in criminal case outcomes. *Criminology & Public Policy*, 2021.

L. C. McCandless and P. Gustafson. A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Statistics in Medicine*, 2017.

L. C. McCandless, P. Gustafson, and A. Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26(11):2331–2347, 2007.

A. H. Munnell, G. M. B. Tootell, L. E. Browne, and J. McEneaney. Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, 86(1):25–53, 1996.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. Conditioning on post-treatment variables. *Journal of Causal Inference*, 3(1):131–137, 2015.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, R. Shroff, C. Phillips, and S. Goel. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(5), 2020.

M. M. Rehavi and S. B. Starr. Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6):1320–1354, 2014.

T. Richardson and J. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical report, Technical Report 128, Center for Statistics and the Social Sciences, Univ. Washington, Seattle, WA., 2013.

C. Robertson, S. B. Baughman, and M. S. Wright. Race and Class: A Randomized Experiment with Prosecutors. *Journal of Empirical Legal Studies*, 16(4):807–847, 2019.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393 – 1512, 1986.

J. M. Robins, T. S. Richardson, and I. Shpitser. An interventionist approach to mediation analysis, 2020.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983a.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*, 45(2):212–218, 1983b.

J. Sekhon. The Neyman–Rubin model of causal inference and estimation via matching methods. In *The Oxford Handbook of Political Methodology*. Oxford University Press, 2008.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

D. Word and C. Perkins. *Building a Spanish Surname List for the 1990's: A New Approach to an Old Problem*. Population Division, US Bureau of the Census Washington, DC, 1996.

D. Word, C. Coleman, R. Nunziata, and R. Kominski. Demographic aspects of surnames from Census 2000. Technical report, U.S. Census Bureau Population Division, 2008. URL `http://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf`.

Q. Zhao, L. J. Keele, D. S. Small, and M. M. Joffe. A note on posttreatment selection in studying racial discrimination in policing. *American Political Science Review*, page 1–14, 2021.

A A Comparison to Alternative Ignorability Conditions

To better understand subset ignorability, we compare it to alternative conditions that recently have been proposed in the context of discrimination studies. In particular, we compare subset ignorability to a set of assumptions introduced by Knox et al. [2020], which they call treatment ignorability, mediator ignorability, and mediator monotonicity. We show that this set of assumptions, like subset ignorability, is sufficient—but not necessary—to ensure the SATE_M is nonparametrically identified by data on second-stage decisions. Importantly, however, the Knox et al. conditions are unlikely to be satisfied in important examples of potentially discriminatory decision making where subset ignorability holds (either exactly or approximately) and the SATE_M accordingly can be estimated, like those situations presented in Sections 3 and 4.

Aside from the Knox et al. conditions, it is instructive to compare subset ignorability to sequential ignorability [Imai et al., 2010a,b], a popular and often useful concept that was introduced to formalize causal mediation analysis, and one that is closely related to the Knox et al. conditions. Sequential ignorability is strictly stronger than subset ignorability, meaning that the former implies the latter but that the converse does not hold. In the setting of discrimination studies, there is little reason to believe sequential ignorability—or reasonable approximations of it—would be satisfied, and we primarily discuss the idea to clarify its distinction from subset ignorability.

The alternative ignorability conditions considered here were developed in the context of a single treatment. Therefore, to facilitate a direct comparison between subset ignorability and these alternatives, we adopt this single-treatment perspective throughout the Appendix. As discussed in the main text, there are substantive issues with positing a single manipulation of (perceived) race, gender, or other immutable characteristics in many multi-stage settings. Formally, however, it is straightforward to collapse Z and D to a single treatment—which we call Z—that affects both the first-stage and the second-stage decisions. In particular, we now assume the potential outcomes $M(z)$ and $Y(z, m)$ satisfy the consistency relations $M = M(Z)$ and $Y = Y(Z, M)$. We emphasize that in this new framing, the definition of subset ignorability in Eq. (4) remains the same and that Theorem 4 likewise holds unaltered—since neither explicitly references the first-stage potential outcomes.[16]

We start by formally considering sequential ignorability, following Imai et al. [2010a,b].

Definition A.1 (Sequential ignorability)**.** We say that *sequential ignorability* is satisfied when the following two conditional independence criteria hold:

$$\{Y(z', m), M(z)\} \perp\!\!\!\perp Z \mid X, \tag{A.15}$$

$$Y(z', m) \perp\!\!\!\perp M \mid Z, X, \tag{A.16}$$

for $z, z' \in \{w, b\}$ and $m \in \{0, 1\}$.

[16] As noted in Footnote 11 above, since we have restricted to the context of a single treatment Z, many of the quantities we consider are not expressible via the *do*-calculus, though they are still expressible in terms of potential outcomes. We emphasize that these potential outcomes should be understood in the conventional sense [Pearl, 2009]: $Y(z, 1)$ represents what would have resulted for an individual if, counterfactually, one had intervened on M so that $M = 1$ and Z so that $Z = z$. Although directly manipulating the first-stage decision so that $M = 1$ may be implausible in some situations—for instance, it may be challenging in practice to intervene on an arresting officer's decision—no issue arises in our setting as we are only concerned with the outcomes $Y(z, 1)$ for individuals who would be arrested in the absence of such an intervention (i.e., where it is already the case that $M = 1$). Moreover, while the FFRCISTG framework [Richardson and Robins, 2013, Robins, 1986] may consider these to be "cross-world" counterfactual quantities, we note that recent extensions of these frameworks discussed in Robins et al. [2020] could accommodate our estimand and identifying assumptions by allowing for the race variable to be split into race variables that are time- and context-specific, as we did in the main body of the paper.

The two key conditional independence assumptions we list are the same as in the definition of sequential ignorability given by Imai et al. [2010a,b], but to facilitate direct comparison with other ignorability criteria, we omit from our definition the accompanying overlap conditions. Also, for ease of exposition, we present the definition in the setting of binary treatment and mediator variables, though the original was more general. In the context of our running example, sequential ignorability means that: (1) conditional on the observed covariates X, the potential outcomes for charging $Y(z, m)$ and arrest $M(z)$ are jointly independent of an individual's actual race Z; and (2) conditional on the observed covariates X and an individual's race Z, the arrest decision M is independent of the potential charging outcomes $Y(z, m)$.

Theorem A.5, below, shows that sequential ignorability implies subset ignorability, but also, importantly, that sequential ignorability is a strictly stronger condition. To understand why, consider the stylized model of Section 3.1, in which one has all of the information that drives a prosecutor's charging decision—satisfying subset ignorability—but not all of the information that drives an officer's arrest decision. For example, suppose the prosecutor has access to the officer's report, but not the arrested individual's actual behavior. In this case, one would in general expect the first condition of sequential ignorability—in Eq. (A.15)—to be violated. In particular, without detailed data on what an officer observes, there is little reason to think the arrest potential outcomes, $M(z)$, would be independent of an individual's race, even controlling for factors available to the prosecutor.

We next formally present the definitions of treatment ignorability, mediator ignorability, and mediator monotonicity proposed by Knox et al., starting with treatment ignorability.

Definition A.2 (Treatment ignorability)**.** *Treatment ignorability* is the combination of the following two conditional independence criteria: for $z, z' \in \{w, b\}$ and $m \in \{0, 1\}$,

$$M(z) \perp\!\!\!\perp Z \mid X, \tag{A.17}$$

$$Y(z', m) \perp\!\!\!\perp Z \mid M(w), M(b), X. \tag{A.18}$$

In the context of arrest and charging decisions, treatment ignorability means that: (1) the potential outcomes for the arrest decision $M(z)$ are independent of race Z, after conditioning on the observed covariates X; and (2) the potential outcomes for the charging decision $Y(z', m)$ are independent of race Z after conditioning on both the covariates X and the arrest potential outcomes $M(w)$ and $M(b)$.

The first condition of treatment ignorability is similar to the first condition of sequential ignorability, and it is unlikely to hold in our setting for the same reason. In general, given only information about what motivates the second-stage decision (e.g., charging, in our case) one cannot say much about what occurs in the first stage (e.g., arrest). But, critically, such information about the first stage is not necessary to estimate the SATE_M, which only quantifies discrimination in the second-stage decision. Theorem 4 makes that statement precise, showing that subset ignorability—which does not consider first-stage potential outcomes—is sufficient to ensure the SATE_M is nonparametrically identified by the second-stage data.

The second criterion of treatment ignorability appears similar in spirit to subset ignorability, but it conditions on the potential outcomes $M(w)$ and $M(b)$ rather than on the actual outcome M. In practice, that distinction may not be too significant; in theory, however, the difference between the two is large. As we show in Theorem A.5 below, treatment ignorability alone—even with its strong assumption on the first stage—is not sufficient to ensure the SATE_M is identified by the second-stage data.

Finally, we consider mediator ignorability and the related mediator monotonicity condition.

Definition A.3 (Mediator ignorability)**.** For $z \in \{w, b\}$ and $m \in \{0, 1\}$, *mediator ignorability* is satisfied when

$$Y(z, m) \perp\!\!\!\perp M(w) \mid Z = z, M(b) = 1, X. \tag{A.19}$$

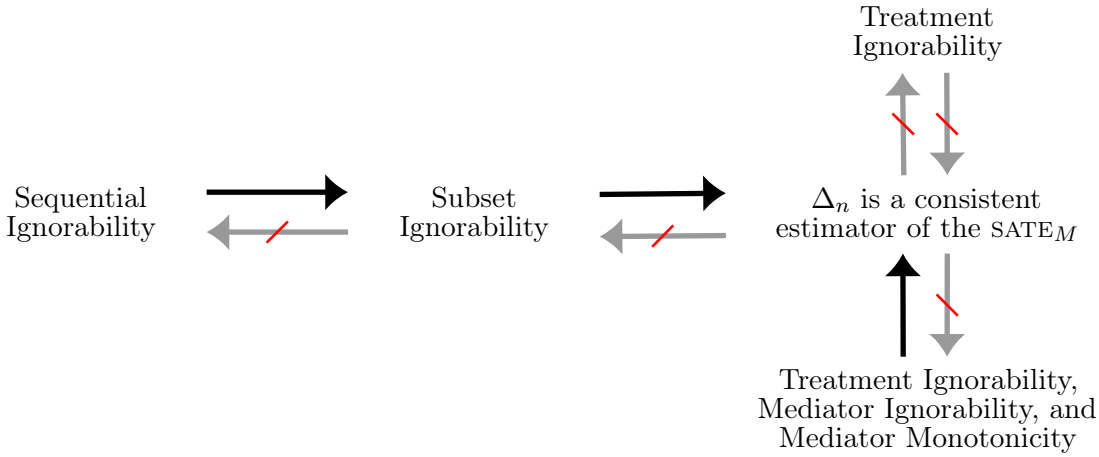
Definition A.4 (Mediator monotonicity)**.** *Mediator monotonicity* is satisfied when

$$M(b) \geq M(w). \tag{A.20}$$

In our running example, mediator ignorability means that the charging potential outcomes $Y(z, m)$ are independent of *one* of the arrest potential outcomes—$M(w)$, the arrest decision for (counterfactually) white individuals—conditional on the observed covariates X, and among individuals of race $Z = z$, who would be arrested if they were Black. The asymmetry in this condition stems from the additional mediator monotonicity constraint considered by Knox et al.: $M(b) \geq M(w)$, meaning that an individual who would be arrested if white would also be arrested if Black. The monotonicity condition is perhaps intuitively plausible given our understanding of racial discrimination, but the conditional independence assumption of mediator ignorability appears harder to interpret.

Having introduced the key definitions, we now present our main analytic result, Theorem A.5, which summarizes and formalizes our discussion of the various ignorability assumptions and their connections to estimating discrimination. In particular, we show that sequential ignorability is a strictly stronger assumption than subset ignorability, and recapitulate (from Theorem 4) that subset ignorability is a sufficient condition for the difference-in-means estimator Δ_n to yield consistent estimates of the SATE_M. Further, we show that treatment ignorability is not a necessary condition for Δ_n to yield consistent estimates. We show this by explicitly constructing examples for which $\Delta_n \xrightarrow{\text{a.s.}} \text{SATE}_M$, but which violate the treatment ignorability condition. We additionally show that treatment ignorability is not a sufficient condition to guarantee consistency, despite its formal resemblance to the (sufficient) subset ignorability condition. To do so, we construct a family of observationally equivalent examples that satisfy treatment ignorability but which have different values of the SATE_M. Accordingly, no estimator, including Δ_n, can yield a consistent estimate of the SATE_M for every instance in the family. Importantly, the more conventional assumption of subset ignorability is sufficient to ensure the SATE_M can be identified from data on the second-stage decisions.

Theorem A.5. *Assume overlap holds, meaning that* $\Pr(Z = z \mid X = x, M = 1) > 0$ *for all* x *and* z*. Then we have the following collection of implications and non-implications:*



Proof. Theorem 4 shows that subset ignorability implies that Δ_n is a consistent estimator of the SATE_M. We show the remaining seven implications and non-implications in turn, starting with the claim that sequential ignorability implies subset ignorability. In particular, we prove that the conjunction of treatment ignorability, mediator ignorability, and mediator monotonicity implies that Δ_n is a consistent estimator of the SATE_M—a fact initially suggested by Knox et al.

Case 1 (Sequential ignorability implies subset ignorability)**.** The first condition of sequential ignorability, in Eq. (A.15), states that $Y(z,m)$ and $M(z')$ are jointly independent of Z given X: $\{Y(z,m), M(z')\} \perp\!\!\!\perp Z \mid X$. From this, it immediately follows that $Y(z,m)$ alone is independent of Z given X: $Y(z,m) \perp\!\!\!\perp Z \mid X$. Now, because $Y(z,m) \perp\!\!\!\perp M \mid Z, X$—which is the second condition of sequential ignorability, in Eq. (A.16)—we have that $Y(z,m) \perp\!\!\!\perp \{Z, M\} \mid X$, by the contraction property of conditional independence. Therefore, by the weak-union property,

$$Y(z,m) \perp\!\!\!\perp Z \mid M, X. \tag{A.21}$$

Subset ignorability now follows, as it is the special case in which $M = 1$ in Eq. (A.21).

Case 2 (Subset ignorability does not imply sequential ignorability)**.** Sequential ignorability is an intuitively stronger condition than subset ignorability, as the former requires that Z is independent of the mediator potential outcomes $M(z)$ given X. Indeed, the synthetic example given in Section 3 satisfies subset ignorability but violates sequential ignorability.

To formally establish our claim, we construct an even simpler example that satisfies subset ignorability but not sequential ignorability. First, suppose that $Y(z,1) = 1$ and $Y(z,0) = 0$, deterministically for $z \in \{b, w\}$. In particular, using the language of our policing and prosecution application, everyone who is arrested is charged, regardless of race. We further set $X = 1$, which effectively means that there are no contextual variables. Finally, we set

$$\begin{aligned} &\Pr(Z = z, M(b) = m_b, M(w) = m_w) \\ &\quad = \Pr(Z = z) \cdot \Pr(M(b) = m_b \mid Z = z) \cdot \Pr(M(w) = m_w \mid Z = z), \end{aligned} \tag{A.22}$$

where $\Pr(Z = z) = \frac{1}{2}$, $\Pr(M(z) = 1 \mid Z = w) = \frac{1}{2}$, and $\Pr(M(z) = 1 \mid Z = b) = 1$. Note that $M = M(Z)$ and $Y = Y(Z, M)$, and so the above description fully defines the joint distribution on all the relevant variables.

Now, because $Y(z,1) = 1$, we trivially have that $Y(z,1) \perp\!\!\!\perp Z \mid M$, meaning that subset ignorability is satisfied. But, because $M(z) \not\perp\!\!\!\perp Z$, sequential ignorability is violated.

Case 3 (Consistency of Δ_n does not imply subset ignorability holds)**.** At a high level, even if the potential outcomes $Y(z,1)$ are not independant of Z—violating subset ignorability—Δ_n can still be a consistent estimator when there is appropriate cancellation. For a concrete illustration of this in the context of our two-stage arrest and charging application, consider a simple example in which: (1) there are no contextual variables (i.e., $X = 1$); (2) the population is evenly split across race groups (i.e., $\Pr(Z = z) = \frac{1}{2}$); (3) everyone in the population is arrested (i.e., $M = 1$); and (4) the prosecutor's *potential* decisions depend on an arrestee's *actual* race. Specifically, we set $Y(z,0) = 0$ and $Y(z,1)$ to be a Bernoulli random variable distributed as follows:

$$\Pr(Y(z,1) = 1 \mid Z) = \begin{cases} 1 & z = b \wedge Z = b, \\ 0 & z = w \wedge Z = b, \\ \frac{1}{2} & Z = w. \end{cases} \tag{A.23}$$

Because $Y = Y(Z, M)$, the above relationships completely specify the joint distribution of Y, Z, M, and X.

Subset ignorability is violated in this example since, by Eq. (A.23), $Y(z,1) \not\perp\!\!\!\perp Z$. (Because X and M are constant, we need not condition on them when considering the subset ignorability criterion.) We further have,

$$\begin{aligned}
\textsc{sate}_M &= \mathbb{E}[Y(b,1) \mid M=1] - \mathbb{E}[Y(w,1) \mid M=1] \\
&= (\mathbb{E}[Y(b,1) \mid Z=b] - \mathbb{E}[Y(w,1) \mid Z=b]) \cdot \Pr(Z=b) \\
&\qquad + (\mathbb{E}[Y(b,1) \mid Z=w] - \mathbb{E}[Y(w,1) \mid Z=w]) \cdot \Pr(Z=w) \\
&= (1-0) \cdot \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{2}\right) \cdot \frac{1}{2} \\
&= \frac{1}{2}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\lim_{n\to\infty} \Delta_n &\overset{\text{a.s.}}{=} \mathbb{E}[Y \mid Z=b, M=1] - \mathbb{E}[Y \mid Z=w, M=1] \\
&= 1 - \frac{1}{2} \\
&= \textsc{sate}_M \,.
\end{aligned}$$

Thus, even though subset ignorability is violated in this example, Δ_n yields a consistent estimate of the $\textsc{sate}_M$.

Case 4 (Consistency of Δ_n does not imply treatment ignorability holds)**.** Consider the example described in Case 2. As discussed there, subset ignorability is satisfied in that example and so, by Theorem 4, Δ_n is a consistent estimator of the $\textsc{sate}_M$. However, that example does not satisfy treatment ignorability, as $M(z) \not\perp\!\!\!\perp Z$, contrary to Eq. (A.17). (Because X is constant, we need not condition on it when evaluating the treatment ignorability criterion.)

Case 5 (Consistency of Δ_n does not imply that treatment ignorability, mediator ignorability, and mediator monotonicity hold)**.** This is directly implied by Case 4.

Case 6 (Treatment ignorability does not imply Δ_n is a consistent estimator of the $\textsc{sate}_M$)**.** We show, more generally, that the $\textsc{sate}_M$ is not identifiable under treatment ignorability alone. To do so, we construct a family of observationally equivalent examples that satisfy treatment ignorability but which have different values of $\textsc{sate}_M$. As a result, no estimator—including Δ_n—can consistently estimate the $\textsc{sate}_M$ for every example in this family.

We construct the family of examples as follows. First, as in the other cases, we set $X = 1$, so that there are effectively no contextual variables, and we set $Y(z,0) = 0$, meaning that if an individual were not arrested, that individual could not be charged. Second, we set $M(b) = 1$, meaning that everyone in the population would be arrested if they were Black. Finally, we set

$$\begin{aligned}
&\Pr(Y(z,1) = y_{zm}, M(w) = m_w, Z = z) \\
&\quad = \Pr(Y(z,1) = y_{zm} \mid M(w) = m_w) \cdot \Pr(M(w) = m_w) \cdot \Pr(Z = z),
\end{aligned} \tag{A.24}$$

where $\Pr(Z=z) = \frac{1}{2}$, $\Pr(M(w) = m_w) = \frac{1}{2}$, and, for $\alpha \in [0,1]$,

$$\Pr(Y(z,1) = 1 \mid M(w)) = \begin{cases} \alpha & M(w) = 0 \wedge z = w, \\ 1 & \text{otherwise.} \end{cases} \tag{A.25}$$

The examples we construct thus differ only in the choice of α.

Now, regardless of α, these examples all satisfy treatment ignorability. To see this, note that $M(w) \perp\!\!\!\perp Z$ by Eq. (A.24) and $M(b) \perp\!\!\!\perp Z$ since $M(b)$ is constant. Consequently, the first condition of treatment ignorability is satisfied. Eq. (A.24) further implies that $Y(z,1) \perp\!\!\!\perp Z \mid M(w)$ and, since $Y(z,0)$ is constant, $Y(z,0) \perp\!\!\!\perp Z \mid M(w)$, establishing the second condition of treatment ignorability. (Because $M(b)$ and X are constant, we need not condition on them when considering the two treatment ignorability conditions.)

We next show that all these examples are observationally equivalent. Intuitively, observational equivalence stems from the fact that the only difference between the examples is in the distribution of $Y(w,1)$ for those individuals with $M(w) = 0$. But for those with $M(w) = 0$, who would not be arrested if they were white, we never observe $Y(w,1)$.

Now, to rigorously establish observational equivalence, we must show that $\Pr(X = x, Y = y, Z = z \mid M = 1)$ does not depend on the value of α. Because X is constant, we need only consider $\Pr(Y = y, Z = z \mid M = 1)$. First, observe that

$$\begin{aligned}\Pr(M = 1) &= \Pr(M(w) = 1, Z = w) + \Pr(M(b) = 1, Z = b) \\ &= \Pr(M(w) = 1) \cdot \Pr(Z = w) + \Pr(Z = b) \\ &= \tfrac{3}{4}.\end{aligned}$$

Further, note that

$$\Pr(Y = y, Z = z, M = 1) = \Pr(Y(z,1) = y, Z = z, M(z) = 1),$$

and consider the case $z = b$. Then, because $Y(b,1) = 1$ and $M(b) = 1$,

$$\Pr(Y = y, Z = b, M = 1) = \begin{cases} 0 & y = 0, \\ \frac{1}{2} & y = 1. \end{cases} \tag{A.26}$$

Now consider the case $z = w$. By Eq. (A.24),

$$\begin{aligned}&\Pr(Y(w,1) = y, Z = w, M(w) = 1) \\ &\quad = \Pr(Y(w,1) = y \mid M(w) = 1) \cdot \Pr(M(w) = 1) \cdot \Pr(Z = w) \\ &\quad = \Pr(Y(w,1) = y \mid M(w) = 1) \cdot \tfrac{1}{4}.\end{aligned}$$

By Eq. (A.25), $\Pr(Y(w,1) = 1 \mid M(w) = 1) = 1$, and so,

$$\Pr(Y = y, Z = w, M = 1) = \begin{cases} 0 & y = 0, \\ \frac{1}{4} & y = 1. \end{cases} \tag{A.27}$$

Finally, combining Eqs. (A.26) and (A.27) with the fact that $\Pr(M = 1) = \frac{3}{4}$, we have

$$\Pr(Y = y, Z = z \mid M = 1) = \begin{cases} 0 & y = 0, \\ \frac{2}{3} & y = 1 \wedge z = b, \\ \frac{1}{3} & y = 1 \wedge z = w. \end{cases}$$

In particular, $\Pr(Y = y, Z = z \mid M = 1)$ does not depend on α, and so the examples are all observationally equivalent.

We conclude the proof by showing that the SATE_M differs across these examples. First, it remains to calculate $\Pr(M(w) = m_w \mid M = 1)$. To do so, note that

$$\begin{aligned}
\Pr(M(w) = 1, M = 1) &= \Pr(M(w) = 1, Z = w) + \Pr(M(w) = 1, M(b) = 1, Z = b) \\
&= \Pr(M(w) = 1) \cdot \Pr(Z = w) + \Pr(M(w) = 1) \cdot \Pr(M(b) = 1) \cdot \Pr(Z = b) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 1 \cdot \frac{1}{2} \\
&= \frac{1}{2},
\end{aligned}$$

and so, since $\Pr(M = 1) = \frac{3}{4}$, it follows that $\Pr(M(w) = 1 \mid M = 1) = \frac{2}{3}$.

Consequently, we have

$$\begin{aligned}
\text{SATE}_M &= \mathbb{E}[Y(b,1) \mid M = 1] - \mathbb{E}[Y(w,1) \mid M = 1] \\
&= \Pr(M(w) = 1 \mid M = 1) \\
&\qquad \cdot (\mathbb{E}[Y(b,1) \mid M(w) = 1, M = 1] - \mathbb{E}[Y(w,1) \mid M(w) = 1, M = 1]) \\
&\quad + \Pr(M(w) = 0 \mid M = 1) \\
&\qquad \cdot (\mathbb{E}[Y(b,1) \mid M(w) = 0, M = 1] - \mathbb{E}[Y(w,1) \mid M(w) = 0, M = 1]) \\
&= \frac{2}{3} \cdot (1 - 1) + \frac{1}{3} \cdot (1 - \mathbb{E}[Y(w,1) \mid M(w) = 0, Z = b]) \\
&= \frac{1 - \alpha}{3},
\end{aligned}$$

where second to last equality follows from Eq. (A.25) and the fact that the event $\{M(w) = 0 \wedge M = 1\}$ equals $\{M(w) = 0 \wedge Z = b\}$; the final equality also follows from Eq. (A.25), as well as the fact that $Y(z,1) \perp\!\!\!\perp Z \mid M(w)$. We have thus constructed a family of observationally equivalent examples that satisfy treatment ignorability but which have different SATE_M, implying that the SATE_M is not in general identifiable under treatment ignorability alone.

Case 7 (Treatment, mediator ignorability, and mediator monotonicity jointly imply Δ_n is a consistent estimator of the SATE_M)**.** The proof is in two pieces. First, we derive an expression for the SATE_M holding X constant, and then prove the general claim.

Supposing $X = x$ is constant, recall that by definition $M = 1$ if and only if $M(z) = 1$ where $Z = z$. By mediator monotonicity, $M(b) \geq M(w)$. Therefore, the event $\{M = 1\}$ can be partitioned into the following two events:

- $E_1 = \{M(b) = 1 \wedge M(w) = 1\}$,
- $E_2 = \{Z = b \wedge M(b) = 1 \wedge M(w) = 0\}$.

Recall the definition of the SATE_M in Definition 1. It follows from the law of total expectation that:

$$\begin{aligned}
\text{SATE}_M &= \mathbb{E}[Y(b,1) - Y(z,1) \mid M = 1] \\
&= \mathbb{E}[Y(b,1) - Y(z,1) \mid E_1] \cdot \Pr(E_1 \mid M = 1) \\
&\qquad + \mathbb{E}[Y(b,1) - Y(z,1) \mid E_2] \cdot \Pr(E_2 \mid M = 1)
\end{aligned} \tag{A.28}$$

Now, we examine each of these summands in turn. First, consider the E_1 term:

$$\mathbb{E}[Y(b,1) - Y(w,1) \mid E_1] = \mathbb{E}[Y(b,1) \mid E_1] - \mathbb{E}[Y(w,1) \mid E_1]$$

By the definition of $E_1 = \{M(b) = 1 \wedge M(w) = 1\}$ and the second treatment ignorability condition, Eq. (A.18), we are free to condition both terms on the right hand side by levels of Z, yielding

$$\mathbb{E}[Y(b,1) \mid Z = b, E_1] - \mathbb{E}[Y(w,1) \mid Z = w, E_1] = \mathbb{E}[Y \mid Z = b, E_1] - \mathbb{E}[Y \mid Z = w, E_1], \quad \text{(A.29)}$$

where equality follows from replacing potential outcomes by their realized values according to the definition of $Y = Y(M, Z)$.

Next, consider the E_2 term. Again,

$$\mathbb{E}[Y(b,1) - Y(w,1) \mid E_2] = \mathbb{E}[Y(b,1) \mid E_2] - \mathbb{E}[Y(w,1) \mid E_2].$$

It follows from mediator ignorability, Eq. (A.19), and the definition of E_2 that

$$\begin{aligned}
\mathbb{E}[Y(w,1) \mid E_2] &= \mathbb{E}[Y(w,1) \mid Z = b, M(b) = 1, M(w) = 0] \\
&= \mathbb{E}[Y(w,1) \mid Z = b, M(b) = 1, M(w) = 1] \\
&= \mathbb{E}[Y(w,1) \mid Z = w, M(b) = 1, M(w) = 1],
\end{aligned}$$

where the last equality follows from treatment ignorability, Eq. (A.18). Replacing potential outcomes with their realizations, it follows that

$$\mathbb{E}[Y(b,1) - Y(w,1) \mid E_2] = \mathbb{E}[Y \mid Z = b, E_2] - \mathbb{E}[Y \mid Z = w, E_1]. \quad \text{(A.30)}$$

Now, we substitute Eqs. (A.29) and (A.30) into Eq. (A.28).

$$\begin{aligned}
\textsc{sate}_M &= (\mathbb{E}[Y \mid Z = b, E_1] - \mathbb{E}[Y \mid Z = w, E_1]) \cdot \Pr(E_1 \mid M = 1) \\
&\quad + (\mathbb{E}[Y \mid Z = b, E_2] - \mathbb{E}[Y \mid Z = w, E_1]) \cdot \Pr(E_2 \mid M = 1) \\
&= (\mathbb{E}[Y \mid Z = b, E_1] - \mathbb{E}[Y \mid Z = w, M = 1]) \cdot \Pr(E_1 \mid M = 1) \\
&\quad + (\mathbb{E}[Y \mid Z = b, E_2] - \mathbb{E}[Y \mid Z = w, M = 1]) \cdot \Pr(E_2 \mid M = 1) \\
&= \big(\mathbb{E}[Y \mid Z = b, E_1] \cdot \Pr(E_1 \mid M = 1) + \mathbb{E}[Y \mid Z = b, E_2] \cdot \Pr(E_2 \mid M = 1)\big) \\
&\quad - \big(\mathbb{E}[Y \mid Z = w, M = 1] \cdot (\Pr(E_1 \mid M = 1) + \Pr(E_2 \mid M = 1))\big) \\
&= \mathbb{E}[Y \mid Z = b, M = 1] - \mathbb{E}[Y \mid Z = w, M = 1], \quad \text{(A.31)}
\end{aligned}$$

where the second equality follows from the fact that $\{M = 1 \wedge Z = w\} = \{E_1 \wedge Z = w\}$ by mediator monotonicity, and the last equality follows from the facts that $\{M = 1 \wedge Z = b \wedge E_1\} = \{Z = b \wedge E_1\}$, $\{M = 1 \wedge Z = b \wedge E_2\} = \{Z = b \wedge E_2\}$, and $\Pr(E_1 \mid M = 1) + \Pr(E_2 \mid M = 1) = 1$.

Now, suppose that X is not constant. Conditioning Y, Z, and M on $X = x$, it follows from the law of total expectation that

$$\begin{aligned}
\mathbb{E}[Y(b,1) - Y(w,1) \mid M = 1] &= \sum_x \mathbb{E}[Y(b,1) - Y(w,1) \mid M = 1, X = x] \cdot \Pr(X = x \mid M = 1) \\
&= \sum_x \mathbb{E}[Y \mid Z = b, M = 1, X = x] \cdot \Pr(X = x \mid M = 1) \\
&\quad - \mathbb{E}[Y \mid Z = w, M = 1, X = x] \cdot \Pr(X = x \mid M = 1), \quad \text{(A.32)}
\end{aligned}$$

where the second equality follows from Eq. (A.31), using the fact that X is constant on each of the events $\{X = x\}$. Eq. (A.32) is identical to the expression in the statement of Theorem 4, and so the estimator Δ_n converges almost surely to the quantity on the right-hand side of Eq. (A.32) by precisely the same argument as there.

□

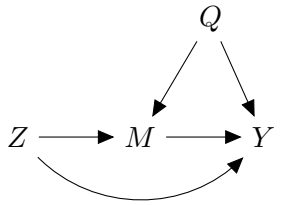


Figure B.6: *A causal DAG considered by Knox et al. In the context of our charging example, Z indicates race, M indicates arrest decisions, Y indicates charging decisions, and Q is an unobserved confounder. Even when one restricts to distributions compatible with this DAG, the Knox et al. conditions are not necessary to non-parametrically identify the* SATE_M *from data on second-stage decisions.*

B Analysis of a Restricted Family of Distributions

Theorem A.5 shows that treatment ignorability, mediator ignorability, and mediator monotonicity are jointly sufficient but not necessary to identify the SATE_M from data on second-stage decisions. We show that this non-necessity holds even if we restrict to distributions compatible with a particular causal DAG considered by Knox et al., shown in Figure B.6, where an unobserved confounder Q directly influences the first-stage decisions M (e.g., arrests) and the second-stage decisions Y (e.g., charging). To do so, we explicitly construct a counterexample in which: (1) the joint distribution of random variables is compatible with this causal DAG; (2) mediator ignorability is violated; and (3) subset ignorability is satisfied, which in turn implies that the stratified difference-in-means Δ_n is a consistent estimator of the SATE_M, by Theorem 4.

Proposition B.1. *There exists a structural causal model (SCM) compatible with the causal DAG in Figure B.6 which violates mediator ignorability but satisfies subset ignorability.*

Proof. We start by explicitly constructing an SCM that is (faithfully) compatible with the DAG in Figure B.6. Our SCM has the following independent exogenous variables:

$$\begin{aligned} U_Z &\sim \text{Unif}(\{w, b\}), \\ U_Q &\sim \text{Unif}(\{1, 2, 3, 4\}), \\ U_M &\sim \text{Unif}((0, 1)), \\ U_Y &\sim \text{Unif}((0, 1)), \end{aligned}$$

where U_Z and U_Q are uniformly distributed over the specified discrete sets, and U_M and U_Y are uniform over the unit interval. Now, the structural equations are given by:

$$\begin{aligned} f_Z(u_z) &= u_z, \\ f_Q(u_q) &= u_q, \\ f_M(z, q, u_m) &= \mathbb{1}\left(u_m \leq (1 + \mathbb{1}(z = b)) \cdot \frac{\mathbb{1}(q = 1) + \mathbb{1}(z = b \wedge q = 3) + \mathbb{1}(z = w \wedge q = 2)}{2}\right), \\ f_Y(z, m, q, u_y) &= m \cdot \mathbb{1}\left(u_y \leq (1 + \mathbb{1}(z = b)) \cdot \frac{\mathbb{1}(q = 1)}{2}\right), \end{aligned}$$

where $\mathbb{1}$ denotes the indicator function and $\wedge$ denotes conjunction (i.e., the `and` operator). For avoidance of doubt, $Z = f_Z(U_Z)$, $Q = f_Q(U_Q)$, $M = f_M(Z, Q, U_M)$, and $Y = f_Y(Z, M, Q, U_Y)$.

Further, the potential arrest outcomes are given by $M(z) = f_M(z, Q, U_M)$, and the bivariate potential charge outcomes are given by $Y(z, m) = f_Y(z, m, Q, U_Y)$.

Mediator ignorability is violated. First, note that $Z \perp\!\!\!\perp \{Y(b,1), M(w), M(b)\}$, because Z is a function of U_Z, and $\{Y(b,1), M(w), M(b)\}$ are functions of U_Y, U_Q, and U_M, which are jointly independent of U_Z. Now, applying this fact and conditioning on Q, we have that,

$$\begin{aligned}
&\Pr(Y(b,1) = 1 \mid M(w) = m_w, M(b) = 1, Z = z) \\
&\quad = \Pr(Y(b,1) = 1 \mid M(w) = m_w, M(b) = 1) \\
&\quad = \sum_{q=1}^{4} \Pr(Y(b,1) = 1 \mid M(w) = m_w, M(b) = 1, Q = q) \\
&\qquad\qquad \cdot \Pr(Q = q \mid M(w) = m_w, M(b) = 1) \\
&\quad = \sum_{q=1}^{4} \Pr(f_Y(b,1,q,U_Y) = 1 \mid M(w) = m_w, M(b) = 1, Q = q) \\
&\qquad\qquad \cdot \Pr(Q = q \mid M(w) = m_w, M(b) = 1).
\end{aligned}$$

Next, observe that $f_Y(b, 1, q, U_Y) = \mathbb{1}(q = 1)$, and so

$$\begin{aligned}
&\Pr(Y(b,1) = 1 \mid M(w) = m_w, M(b) = 1, Z = z) \\
&\quad = \Pr(Q = 1 \mid M(w) = m_w, M(b) = 1) \\
&\quad = \frac{\Pr(M(w) = m_w, M(b) = 1 \mid Q = 1) \cdot \Pr(Q = 1)}{\sum_{q=1}^{4} \Pr(M(w) = m_w, M(b) = 1 \mid Q = q) \cdot \Pr(Q = q)} \\
&\quad = \frac{\Pr(M(w) = m_w, M(b) = 1 \mid Q = 1)}{\sum_{q=1}^{4} \Pr(M(w) = m_w, M(b) = 1 \mid Q = q)}.
\end{aligned} \tag{B.33}$$

The second equality above follows from Bayes' rule, and the third follows from the fact that $\Pr(Q = q) = 1/4$.

Finally, we compute $\Pr(M(w) = m_w, M(b) = 1 \mid Q = q)$. Note that

$$\begin{aligned}
M(w) &= f_M(w, Q, U_M) \\
&= \mathbb{1}\left(U_M \leq \frac{\mathbb{1}(Q = 1) + \mathbb{1}(Q = 2)}{2}\right) \\
&= \begin{cases} \mathbb{1}(U_M \leq 1/2) & Q \in \{1, 2\} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{B.34}$$

Likewise,

$$\begin{aligned}
M(b) &= f_M(b, Q, U_M) \\
&= \mathbb{1}(U_M \leq (\mathbb{1}(Q = 1) + \mathbb{1}(Q = 3))) \\
&= \begin{cases} 1 & Q \in \{1, 3\} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{B.35}$$

As a result,

$$\Pr(M(w) = m_w, M(b) = 1 \mid Q = q) = \begin{cases} 1/2 & q = 1 \\ 1 & q = 3 \wedge m_w = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, by Eq. (B.33),

$$\Pr(Y(b,1)=1 \mid M(w)=1, M(b)=1, Z=z)=1,$$

while

$$\Pr(Y(b,1)=1 \mid M(w)=0, M(b)=1, Z=z)=\frac{1}{3}.$$

Therefore, $Y(b,1) \not\perp\!\!\!\perp M(w) \mid M(b)=1, Z=z$, meaning that mediator ignorability does not hold.

Subset ignorability holds. Similar to the above, we have that

$$\begin{aligned} Y(b,1) &= f_Y(b,1,Q,U_Y) \\ &= \mathbb{1}(U_Y \leq \mathbb{1}(Q=1)) \\ &= \begin{cases} 1 & Q=1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{B.36}$$

and

$$\begin{aligned} Y(w,1) &= f_Y(w,1,Q,U_Y) \\ &= \mathbb{1}(U_Y \leq \mathbb{1}(Q=1)/2) \\ &= \begin{cases} \mathbb{1}(U_Y \leq 1/2) & Q=1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{B.37}$$

Now, as before, $Z \perp\!\!\!\perp \{Q, M(w), M(b)\}$, since Z is a function of U_Z, and $\{Q, M(w), M(b)\}$ are functions of U_Q and U_M, which are jointly independent of U_Z. Consequently,

$$\begin{aligned} \Pr(Q=1 \mid M=1, Z=z) &= \Pr(Q=1 \mid M(z)=1, Z=z) \\ &= \Pr(Q=1 \mid M(z)=1) \\ &= \frac{\Pr(M(z)=1 \mid Q=1)\cdot \Pr(Q=1)}{\sum_{q=1}^{4} \Pr(M(z)=1 \mid Q=q)\cdot \Pr(Q=q)} \\ &= \frac{1}{2}, \end{aligned} \tag{B.38}$$

where the last equality follows from Eqs. (B.34) and (B.35), together with the fact that $\Pr(Q=q)=1/4$, and that $U_M \perp\!\!\!\perp Q$.

Finally, conditioning on Q, we have

$$\begin{aligned} &\Pr(Y(b,1)=1 \mid M=1, Z=z) \\ &\quad= \sum_{q=1}^{4} \Pr(Y(b,1)=1 \mid M=1, Z=z, Q=q)\cdot \Pr(Q=q \mid M=1, Z=z) \\ &\quad= \Pr(Q=1 \mid M=1, Z=z) \\ &\quad= \frac{1}{2}, \end{aligned}$$

where the second equality follows from Eq. (B.36), and the third from Eq. (B.38). Similarly,

$$\begin{aligned}
&\Pr(Y(w,1)=1 \mid M=1, Z=z) \\
&\quad= \sum_{q=1}^{4} \Pr(Y(w,1)=1 \mid M=1, Z=z, Q=q) \cdot \Pr(Q=q \mid M=1, Z=z) \\
&\quad= \Pr(U_Y \leq 1/2 \mid M=1, Z=z, Q=1) \cdot \Pr(Q=1 \mid M=1, Z=z) \\
&\quad= \Pr(U_Y \leq 1/2) \cdot \Pr(Q=1 \mid M=1, Z=z) \\
&\quad= \frac{1}{4},
\end{aligned}$$

where the second equality follows from Eq. (B.37), the third from the fact that $U_Y \perp\!\!\!\perp \{M, Z, Q\}$, and the fourth from Eq. (B.38). Therefore, $\Pr(Y(b,1)=y \mid M=1, Z=b) = \Pr(Y(b,1) \mid M=1, Z=w)$ and similarly for $Y(w,1)$. In particular, this means that $Y(z,1) \perp\!\!\!\perp Z \mid M=1$, and so subset ignorability holds. □

C Extending Theorem 4 to Allow for Continuous Covariates

Theorem 4 in the main text shows that subset ignorability—together with overlap—implies the SATE_M is nonparametrically identified, where, for simplicity, we proved the result for discrete covariates X. We now extend that result to allow for continuous covariates. At a conceptual level, the extension is straightforward: we first condition on X, then appeal to subset ignorability to condition on Z, and, finally, use consistency to replace potential outcomes by their observed values. In the general case, however, typically $\Pr(X=x)=0$, and so one must take care to define expressions that nominally condition on these probability-zero events.

Recall that in the discrete case, the primary conditional expectations, treated as functions of z and x, are of the form

$$\begin{aligned}
\mathbb{E}[Y|Z=z, X=x, M=1] &= \sum_y y \frac{\Pr(Y=y, Z=z, X=x \mid M=1)}{\Pr(Z=z, X=x \mid M=1)} \\
&= \sum_y y \frac{\Pr(Y=y, Z=z, X=x \mid M=1)}{\Pr(Z=z \mid X=x, M=1)\Pr(X=x \mid M=1)}. \qquad \text{(C.39)}
\end{aligned}$$

Overlap ensures that the denominator in (C.39) is non-zero, and, accordingly, that the conditional expectation is well-defined. In the continuous case, to address conditioning on probability-zero events, conditional probabilities are defined as random variables rather than simple numeric quantities (cf. Billingsley [2008]). Further, if the random variables $\Pr(Z=z \mid X, M=1) > 0$ a.s. for $z \in \{w, b\}$—a condition that we call generalized overlap—then the expression $\mathbb{E}[Y \mid Z=z, X=x, M=1]$ is a well-defined function of z and x, as in the discrete case, up to a set of measure zero with respect to the pushforward measure $\mu_{X|M=1}$ for each fixed z.[17,18]

We now state and prove the extension of Theorem 4, with the understanding that the conditional probabilities and expectations below are defined according to the usual measure-theoretic conventions.

[17]The pushforward measure $\mu_{X|M=1}$ is the measure on $\mathcal{X}$—the range of X—given by $\mu_{X|M=1}[A] = \Pr(X \in A \mid M=1)$ for measurable $A \subseteq \mathcal{X}$.

[18]To see this, first note that, in general, $\mathbb{E}[Y \mid Z=z, X=x, M=1]$ is uniquely defined up to a set of measure zero with respect to the pushforward measure $\mu_{Z,X|M=1}$. Now, for fixed z, suppose, toward a contradiction, that $f_1(x)$ and $f_2(x)$ are two versions of $\mathbb{E}[Y \mid Z=z, X=x, M=1]$ that differ on a set A such that $\Pr(X \in A \mid M=1) > 0$. Then, by the generalized overlap condition, $\Pr(Z=z, X \in A \mid M=1) = \int_A \Pr(Z=z \mid X=x, M=1)\, \mathrm{d}F_{X|M=1} > 0$, contradicting the fact that $f_1(x) \neq f_2(x)$ only on a null set with respect to the pushforward measure $\mu_{Z,X|M=1}$.

Theorem C.1. *Suppose $Y(z,1)$, Z, M, and X satisfy subset ignorability, and that generalized overlap holds—i.e., for $z \in \{b,w\}$, $\Pr(Z = z \mid X, M = 1) > 0$ a.s. Then, the* $\textsc{sate}_M$ *equals*

$$\begin{aligned}\int_{\mathcal{X}} \mathbb{E}[Y \mid Z = b, X = x, M = 1] \, \mathrm{d}F_{X|M=1} \\ - \int_{\mathcal{X}} \mathbb{E}[Y \mid Z = w, X = x, M = 1] \, \mathrm{d}F_{X|M=1},\end{aligned} \tag{C.40}$$

where $\mathcal{X}$ denotes the range of X and $\mathrm{d}F_{X|M=1}$ denotes integration over $\mathcal{X}$ with respect to the pushforward measure $\mu_{X|M=1}$.

Proof. By conditioning on X, we have,

$$\begin{aligned}\textsc{sate}_M &= \mathbb{E}[Y(b,1) - Y(w,1) \mid M = 1] \\ &= \int_{\mathcal{X}} \mathbb{E}[Y(b,1) - Y(w,1) \mid X = x, M = 1] \, \mathrm{d}F_{X|M=1} \\ &= \int_{\mathcal{X}} \mathbb{E}[Y(b,1) \mid X = x, M = 1] - \mathbb{E}[Y(w,1) \mid X = x, M = 1] \, \mathrm{d}F_{X|M=1}. \end{aligned} \tag{C.41}$$

Now, subset ignorability gives that

$$\mathbb{E}[Y(z,1) \mid X = x, M = 1] = \mathbb{E}[Y(z,1) \mid X = x, Z = z, M = 1] \text{ a.s.}, \tag{C.42}$$

where generalized overlap ensures that the right-hand side of Eq. (C.42) is well-defined up to a set of measure zero with respect to $\mathrm{d}F_{X|M=1}$. Substituting this expression into Eq. (C.41), and then appealing to consistency to replace potential outcomes with their observed values, we have

$$\begin{aligned}\textsc{sate}_M &= \int_{\mathcal{X}} \mathbb{E}[Y(b,1) \mid X = x, Z = b, M = 1] - \mathbb{E}[Y(w,1) \mid X = x, Z = w, M = 1] \, \mathrm{d}F_{X|M=1} \\ &= \int_{\mathcal{X}} \mathbb{E}[Y(Z,M) \mid X = x, Z = b, M = 1] - \mathbb{E}[Y(Z,M) \mid X = x, Z = w, M = 1] \, \mathrm{d}F_{X|M=1} \\ &= \int_{\mathcal{X}} \mathbb{E}[Y \mid X = x, Z = b, M = 1] - \mathbb{E}[Y \mid X = x, Z = w, M = 1] \, \mathrm{d}F_{X|M=1}.\end{aligned}$$

□

All of the quantities in Eq. (C.40) (i.e., the distribution of X and the conditional expectations) are functions of observables, establishing that the $\textsc{sate}_M$ is identified by data on second-stage decisions. One may adopt a variety of approaches to estimate the terms in Eq. (C.40), including model-based strategies, as we do in Section 4. One may also adopt non-parametric estimation strategies, wherein continuous covariates are appropriately binned into discrete sets. For further treatment of these issues, see, for example, Gelman et al. [2013], Friedman et al. [2001], and Tsybakov [2008].