Deconstructing Claims of Post-Treatment Bias in Observational Studies of Discrimination*

Johann Gaebler
Stanford University

William Cai
Stanford University

Guillaume Basse
Stanford University

Ravi Shroff
New York University

Sharad Goel
Stanford University

Jennifer Hill
New York University

June 28, 2020

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, raised most recently in Knox et al. [2020], is post-treatment bias. The authors argue that many studies of discrimination condition on intermediate outcomes, like being arrested, which themselves may be the product of discrimination, corrupting statistical estimates. Here we show that the Knox et al. critique is itself flawed, suffering from a mathematical error. Within their causal framework, we prove that a primary quantity of interest in discrimination studies is nonparametrically identifiable under a standard ignorability condition that is common in causal inference with observational data. More generally, though, we argue that it is often problematic to conceptualize discrimination in terms of a causal effect of protected attributes on decisions. We present an alternative perspective that avoids the common statistical difficulties, and which closely relates to long-standing legal and economic theories of disparate impact. We illustrate these ideas both with synthetic data and by analyzing the charging decisions of a prosecutor's office in a large city in the United States.

*We thank Alex Chohlas-Wood, Avi Feller, Andrew Gelman, Zhiyuan "Jerry" Lin, Julian Nyarko, Steven Raphael, and Amy Shoemaker for helpful conversations.

Online Discussion

This paper has generated considerable online discussion between the authors of this manuscript and Knox et al., summarized in the Twitter threads referenced below.

1. Mummolo: `https://twitter.com/jonmummolo/status/1275790509647241222`
2. Goel: `https://twitter.com/5harad/status/1275931524819386368`
3. Mummolo: `https://twitter.com/jonmummolo/status/1275938319432667141`
4. Gaebler: `https://twitter.com/jgaeb1/status/1276261257050353664`
5. Knox: `https://twitter.com/dean_c_knox/status/1276508868500164612`
6. Gaebler: `https://twitter.com/jgaeb1/status/1276551095347638273`

1 Introduction

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 [Greenberg et al., 2016], and criminal justice [Ayres, 2002, Fryer Jr, 2019, Gelman et al., 2007, MacDonald and Raphael, 2017, Rehavi and Starr, 2014].[1]

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 concerns that estimates are corrupted by post-treatment bias when one adjusts for covariates or restricts to samples of individuals determined downstream from the treatment variable. In the context of policing, Knox et al. [2020, p. 2] argue that post-treatment bias, particularly arising from sample selection, is so pervasive and pernicious that "existing empirical work in this area is producing a misleading portrait of evidence as to the severity of racial bias in police behavior."

To better understand these three challenges, 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, we assume prosecutors only have access to the incident report—and to no other information—when making their decisions. We allow, however, 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, we assume that the researcher has access to these incident reports but, importantly, not to any data on individuals that officers considered but ultimately decided against arresting.

The first challenge is to rigorously define the causal estimand 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, as well as how such an intervention might actually be performed [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

[1] One popular alternative to benchmark analysis is outcome analysis, in which one looks at the success rates of decisions rather than the decision rates themselves [Becker, 1957, 1993]. For example, in the lending context, one can look at race-specific default rates of applicants who ultimately received loans. If minority borrowers are found to default at lower rates than white borrowers, that suggests that lenders may discriminate by having a higher bar for minorities. Outcome tests can mitigate concerns about omitted-variable bias, but raise new concerns related to the problem of infra-marginality [Ayres, 2002, Simoiu et al., 2017]. For simplicity, we limit our primary discussion to benchmark analysis, though we note that our substantive conclusions largely extend to outcome tests as well.

those in which they were randomly described as "white." This conceptualization of discrimination is imperfect—partly due to implicit indications of race, as we discuss below—but it 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, an unlawful form of discrimination in which actions are motivated by animus or otherwise discriminatory intent [Goel et al., 2017].

In the absence of such an 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. That strategy, which mimics the key features of the randomized experiment described above, is formally justified by assuming 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 nearly all 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 [Cornfield et al., 1959, Rosenbaum and Rubin, 1983b], a classic statistical tack that remains underutilized in many applied settings [Cinelli and Hazlett, 2020].

Finally, there is the issue of post-treatment bias, especially due to sample selection. Knox et al. argue that researchers inadvertently introduce post-treatment bias in observational studies of discrimination by conditioning on intermediate outcomes, such as—in our charging example—being arrested, which themselves may be the product of discrimination. As a result, the authors assert 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 their mathematical framework applies more broadly; in particular, Knox et al.'s statistical claims—if true—would imply that one could not hope to measure discrimination in prosecutorial decisions using observational data.[2]

We re-examine these claims and show that the Knox et al. critique is flawed. The mathematical framework we adopt mirrors theirs, and, in particular, we consider identical causal estimands. Under their causal framework, we prove that one of the primary quantities of interest in discrimination studies is in fact nonparametrically identified under a standard ignorability condition. More specifically, we demonstrate that one can estimate discrimination in charging decisions in a principled manner, even when arrest decisions are discriminatory in unknown ways. We trace our qualitatively different conclusion to a logical error in the Knox et al. analysis: the authors conflate sufficiency with necessity. Knox et al. derive a complex set of conditions—including no discrimination in earlier stages of the decision making process—that collectively are sufficient to guarantee identifiability, but then mistakenly conclude that those conditions are also necessary. In fact, the standard ignorability assumption is also sufficient.

At a conceptual level, the authors fail to account for the precise timing of events. Indeed, in our charging example, there are effectively two treatments, one affecting an officer's arrest decision and the other affecting a 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

[2] The authors state that their "results can inform the study of racial discrimination in a host of other settings beyond law enforcement ... Our identifying assumptions may also be useful for researchers seeking to address biases stemming from posttreatment conditioning more generally, beyond studies of discrimination" [p. 16].

but pre-treatment relative to the prosecutor's perception of race. In such a two-decider situation, as Greiner and Rubin [2011] suggest, one can recover estimates of discrimination by the second decider (e.g., in the charging decision) under the standard ignorability assumption, even if there is discrimination by the first decider (e.g., in the arrest decision).[3] The motivating example of Knox et al. involves a single decider making two decisions—an officer first deciding whether to make a stop and then deciding whether to use force—but their formal mathematical statements apply equally to both the one-decider and the two-decider settings.

We illustrate these ideas by analyzing synthetic data, as well as by analyzing a detailed dataset of actual charging decisions for approximately 20,000 felony arrests in a major U.S. city. On the synthetic data, we construct simple examples that satisfy standard ignorability but which violate the Knox et al. conditions, and then show that traditional benchmark analysis recovers the true causal effect. On the real-world data, we show that Black and white individuals with similar observable characteristics were charged at comparable rates, but that men were charged about 2.5% more often than otherwise similar women. Under an ignorability assumption, one could interpret this gap causally and say that prosecutors charged men more often than similarly situated women *because of* their gender. A standard sensitivity analysis [Cinelli and Hazlett, 2020] shows that this result is robust to moderate violations of the ignorability assumption.

To engage with past work, we prove our mathematical results and conduct our primary empirical analysis within a common causal framework of discrimination, where one measures the effects of protected traits, like race and gender, on decisions. More generally, though, we argue that this narrow focus on such *disparate treatment* fails to capture *disparate impact*, an alternative form of discrimination central to many legal judgments and policy prescriptions. In spirit, disparate impact enquiries flip the usual causal question of disparate treatment studies: instead of measuring the effect of race on decisions, one measures the effect of decisions on racial disparities. Suppose, hypothetically, that prosecutors adopted a policy of charging wealthy individuals less often than poorer ones who had otherwise similar cases, and that this policy led to racial disparities in charging rates. Under a traditional disparate treatment framework—in which racial disparities would be estimated after adjusting for income and other factors that are pre-treatment relative to the prosecutor's perception of race—one would conclude there was no discrimination in prosecutorial decisions. However, that analysis would overlook the policy's unjustified disparate impact [Ayres, 2005, Jung et al., 2018].

The criminal justice system involves a large number of actors making an even larger number of decisions, ranging from an officer's decision to patrol a certain block to a judge's decision to impose a certain sentence. As such, we share Knox et al.'s concerns with viewing discrimination narrowly through the lens of a single decision point. Even if—hypothetically—officers did not discriminate in post-stop use of force, communities of color would still likely be subject to disproportionate police violence stemming from discrimination in initial stop decisions, as has been found in previous research [Gelman et al., 2007, Goel et al., 2016, Pierson et al., 2020]. Nevertheless, careful examination of specific decision points can often help one target interventions and inform policy reforms. Further, regardless of purpose, the application of faulty mathematical and theoretical reasoning jeopardizes the entire enterprise of quantitative discrimination studies. By clarifying the statistical foundations of discrimination analysis, we hope our contributions help to advance this important area of work.

[3] Greiner and Rubin [2011, p. 779] write: "In complicated transactions . . . a researcher may have more than one choice of decider to study. . . . By focusing on a decider who perceives the unit's immutable characteristic 'late' in the interaction, the researcher implicitly chooses a later timing of treatment assignment, rendering more measured variables pretreatment and thus properly characterized as covariates. That in turn can make an ignorability assumption more plausible. But by treating such variables as covariates (and thus conditioning on them in the analysis), the researcher forgoes the detection of any prior discrimination that may have affected the values of these covariates."

2 Defining and Estimating a Measure of Discrimination

To formally state our results, we first introduce a simple two-step model of discrimination and define the main causal quantity of interest within this general framework. The initial setup mirrors that of Knox et al., though we more explicitly account for the precise timing of events. We then state an ignorability condition and prove that it is sufficient to guarantee that the *controlled direct effect among the observed* (CDE_{Ob}), the primary measure of discrimination we consider here, is nonparametrically identified, in contrast to the claims of Knox et al. Finally, we compare the CDE_{Ob} to the *total effect* (TE), a second causal estimand considered by Knox et al., and show that although these estimands appear formally distinct, in many scenarios they are two sides of the same coin. Indeed, this correspondence illustrates that post-treatment bias is an elusive concept in observational studies of discrimination. Our notation differs slightly from that of Knox et al., but our key mathematical definitions are identical to theirs.

2.1 A Model of Discrimination

We start by describing a two-step model to characterize discriminatory decision making in a variety of real-world situations. In the first stage, each individual in some population is subject to a binary decision, such as an offer of employment, admission to college, or law enforcement action. The main illustrative example we consider in this paper concerns arrest decisions in the U.S. criminal justice system, where a police officer may decide to arrest an individual if sufficient evidence exists to suggest participation in criminal activity. Those who receive a "positive" first-stage decision (e.g., those who are arrested) proceed to a second stage, where another binary decision 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. Importantly, 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. The central aim of this paper is to clarify the technical assumptions needed to statistically identify discrimination—more precisely, disparate treatment—in the second stage (e.g., the charging decision) when data are only available for those who made it past the first stage (e.g., those who were arrested), particularly when first-stage decisions may themselves be discriminatory.

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 other trait—can be independently altered prior to their decisions. There are, however, examples in which one can plausibly intervene twice even for a single decider making 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. However, regardless of whether one imagines there are two deciders or a single one, our formal statistical results hold unaltered. Similarly, the Knox et al. statistical framework does not distinguish between these two possibilities.

We now formally describe our model of decision making, beginning with the first stage. We denote by $D_i \in \{w, b\}$ the race of the i-th individual as perceived by the first decider at the first stage, 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 $M_i(w)$ and $M_i(b)$ denote the potential first-stage decisions for the i-th individual (e.g., whether they were arrested), and write $M_i = M_i(D_i)$ for the observed first-stage decision.

Next, we let $Z_i \in \{w, b\}$ denote the race of the i-th individual as perceived by the second decider, at the second stage. In our running example, Z_i denotes the race of the i-th individual as perceived by the prosecutor reviewing that person's file, while D_i denotes race as perceived by the police officer during the encounter. Finally, we define the second-stage potential outcome as a function of both the first-stage outcome (e.g., the arrest decision) and the second decider's perception of race, Z_i (e.g., race as perceived by the prosecutor). Thus, assuming once again that there is no interference, the observed second-stage outcome for individual i can be denoted $Y_i(Z_i, M_i(D_i))$, and we count four potential outcomes for each individual: $Y_i(z, m)$, where $z \in \{w, b\}$ and $m \in \{0, 1\}$. In our example, charging decisions are made only for individuals who were arrested, and so $Y_i(b, 0) = Y_i(w, 0) = 0$.

We further allow each individual to have an associated vector of (non-race) covariates, X_i, representing, for example, an individual's behavior during a police encounter, 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 below is unlikely to hold. For simplicity of exposition, we assume X_i is discrete. In practice, X_i 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_i and Z_i, as we believe this helps to clarify the timing of events and the meaning of causal quantities. However, our focus is observational settings, in which disagreement between Z_i and D_i may be realized only rarely, if at all. For this reason, we make the further assumption that $D_i = Z_i$, where, for example, the officer's perception of race is the same as the prosecutor's, and we henceforth refer only to Z_i. This setting is the one implicitly considered in Knox et al. Though we restrict our attention to this special case, we note that Proposition 3, one of our main technical results, applies directly to the more general setting.

With this framing, we now describe the primary causal estimand we consider. This quantity, which we call the *controlled direct effect among the observed* (CDE_{Ob})—and which Knox et al. denote by $\text{CDE}_{M=1}$—reflects discrimination in the second stage of the decision-making process outlined above, such as discrimination in the prosecutor's charging decision.

Definition 1 (CDE_{Ob})**.** The *controlled direct effect among the observed*, denoted CDE_{Ob}, is given by:

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

In defining the CDE_{Ob}, 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 estimand compares outcomes under the two race perception scenarios. Moreover, this estimand restricts to the subset of individuals who had a "positive" first-stage decision (e.g., those who were in reality arrested). The CDE_{Ob} thus captures discrimination in the second-stage decision among those who made it past the first stage, such as discrimination in prosecutorial charging decisions among those who were arrested.

Knox et al. argue that the CDE_{Ob} is a contrived estimand, as it involves second-stage counterfactual outcomes for individuals who may never have made it to the second stage if their race had been counterfactually altered before the first-stage decision [Knox et al., 2020, Online Appendix, p. 5]. We note, however, that the CDE_{Ob} precisely maps to common understandings of disparate treatment in second-stage decisions, including in our charging example. Further, the causal quantity primarily considered by Knox et al.—the TE, defined below—requires one to similarly consider counterfactual outcomes for individuals who might not have made it to the first-stage decision point had they been a different race. For example, a counterfactually white individual might not have been in the vicinity of officers, and thus not subject to a first-stage stop or arrest decision. We explore these estimands in more detail in Section 2.3, and discuss the central role of the CDE_{Ob} in discrimination studies. First, though, we show that a standard ignorability assumption together with an overlap condition is sufficient to estimate the CDE_{Ob}.

2.2 Estimating the CDE_{Ob}

Our goal is to estimate the CDE_{Ob} using data that only describe those who received a "positive" decision in the first stage, a decision that itself may have been the product of discrimination. For example, we seek to estimate discrimination in charging decisions based only on data describing those who were arrested. Under this statistical setup, Knox et al. claim that the CDE_{Ob} cannot be identified absent a specific set of "highly implausible" ignorability assumptions.[4] However, as we show now, a standard ignorability assumption, together with an overlap condition, is sufficient to guarantee the CDE_{Ob} can be 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

$$Y(z,1) \perp\!\!\!\perp Z \mid X, M = 1 \tag{2}$$

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. Subset ignorability is thus just a restatement of 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. Indeed, almost all causal analyses implicitly rely on a version of subset ignorability, since researchers rarely make inferences about their full sample; for instance, it is standard in propensity score matching to subset to the common support of the treated and untreated units' propensity scores.

In the traditional, non-staged setting, ignorability is sufficient to guard against omitted-variable bias and obtain consistent estimates of the average treatment effect. Likewise in a staged model of discrimination, subset ignorability is sufficient to guarantee consistent estimates of the CDE_{Ob}. Importantly, the first-stage decision, M, and the covariates, X, are pre-treatment relative to the second-stage intervention. As a result, it may be possible in practice to (approximately) satisfy subset ignorability and, in particular, to consider estimands for which post-treatment bias is not a concern.

Proposition 3. *Suppose $Y(z,1)$, Z, M, and X satisfy subset ignorability, and that overlap holds, meaning that* $\Pr(Z = z \mid X = x, M = 1) > 0$ *for all x and z. Then,*

$$\begin{aligned}\text{CDE}_{\text{Ob}} = &\sum_x \mathbb{E}[Y \mid Z = b, X = x, M = 1] \cdot \Pr(X = x \mid M = 1) \\ &- \sum_x \mathbb{E}[Y \mid Z = w, X = x, M = 1] \cdot \Pr(X = x \mid M = 1).\end{aligned} \tag{3}$$

[4] Specifically, Knox et al. claim that treatment ignorability, mediator ignorability, and mediator monotonicity, as defined in Section 5 below, are necessary assumptions.

Proof. Conditioning on X, we have

$$\begin{aligned}\textsc{cde}_{\text{Ob}} = \sum_x \mathbb{E}[Y(b,1) \mid X = x, M = 1] \cdot \Pr(X = x \mid M = 1) \\ - \sum_x \mathbb{E}[Y(w,1) \mid X = x, M = 1] \cdot \Pr(X = x \mid M = 1).\end{aligned} \tag{4}$$

By the subset ignorability assumption in Eq. (2), and our assumption of overlap, we can condition the summands in Eq. (4) on $Z = b$ and $Z = w$, respectively, without changing their values, yielding

$$\begin{aligned}\textsc{cde}_{\text{Ob}} = \sum_x \mathbb{E}[Y(b,1) \mid Z = b, X = x, M = 1] \cdot \Pr(X = x \mid M = 1) \\ - \sum_x \mathbb{E}[Y(w,1) \mid Z = w, X = x, M = 1] \cdot \Pr(X = x \mid M = 1).\end{aligned} \tag{5}$$

Finally, the statement of the proposition follows by replacing the potential outcomes with their realized values. $\square$

Proposition 3 immediately implies that when subset ignorability holds—and there is overlap—the $\textsc{cde}_{\text{Ob}}$ is identified by the observed data. Suppose we have n i.i.d. observations $(X_i, Z_i, Y_i)_{i=1}^n$ with $M_i = 1$ (e.g., data on the subset of arrested individuals). Let $S_{zx} = \{i : Z_i = z \wedge X_i = x\}$ represent the set of observations with $Z = z$ and $X = x$, and let n_x denote the total number of observations with $X = x$. Then the standard difference-in-means estimator,

$$\Delta_n = \sum_x \left[\frac{1}{|S_{bx}|} \sum_{i \in S_{bx}} Y_i \right] \cdot \frac{n_x}{n} - \sum_x \left[\frac{1}{|S_{wx}|} \sum_{i \in S_{wx}} Y_i \right] \cdot \frac{n_x}{n}, \tag{6}$$

yields a consistent estimate of the $\textsc{cde}_{\text{Ob}}$. To see this, note that

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

Consequently, each of the terms in Δ_n converges to the corresponding terms in the definition of the $\textsc{cde}_{\text{Ob}}$, in Eq. (1). Further, a straightforward calculation shows that the following expression yields a consistent estimate of the standard error of Δ_n:

$$\widehat{\text{SE}}(\Delta_n) = \sqrt{\sum_x \left(\frac{n_x}{n}\right)^2 \left[\frac{c_{bx}(1 - c_{bx})}{|S_{bx}|} + \frac{c_{wx}(1 - c_{wx})}{|S_{wx}|} \right]} \tag{7}$$

where

$$c_{bx} = \frac{1}{|S_{bx}|} \sum_{i \in S_{bx}} Y_i \quad \text{and} \quad c_{wx} = \frac{1}{|S_{wx}|} \sum_{i \in S_{wx}} Y_i.$$

Eq. (7) accordingly allows us to form confidence intervals for Δ_n.

The stratified difference-in-means estimator is the basis for nearly all applications of benchmark analysis in discrimination studies. In practice, as we discuss in Section 3, it is common to approximate Δ_n via a regression model, but the theoretical underpinnings of that approximation are the same. In that sense, our analysis above simply grounds the traditional statistical approach within a specific causal framework, and demonstrates that the standard ignorability assumption, together with overlap, is sufficient to yield valid estimates. In Section 5, we directly compare this ignorability condition to the criteria proposed by Knox et al. However, we note that our results above stand in opposition to the Knox et al. claim that the "difference in means fails to recover any known causal quantity" in reasonable settings [p. 1].

2.3 An alternative measure of discrimination

To better understand the CDE_{Ob}, we now contrast it with the *total effect* (TE) [Imai et al., 2010a], a second estimand considered by Knox et al. The total effect and the CDE_{Ob} differ in this context in two ways: (1) the population of individuals about which we are making inferences; and (2) the timing of the counterfactual comparison. The total effect is not restricted to individuals who had a "positive" first-stage decision (e.g., those who were arrested). Additionally, in the total effect, one imagines that perception of race is counterfactually determined before the first-stage decision (instead of after the first-stage decision, as with the CDE_{Ob}), and is the same at both stages.

Definition 4 (TE)**.** The *total effect*, denoted TE, is given by:

$$\text{TE} = \mathbb{E}[Y(b, M(b)) - Y(w, M(w))]. \tag{8}$$

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, 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 controlled direct effect.

In experimental settings, where researchers actively randomize treatments, the TE captures the total effect of treatment on downstream outcomes. In such experimental settings, the CDE_{Ob} is perhaps an unusual quantity to consider, as it conditions on an intermediate, post-treatment outcome.[5] But in studies of discrimination—particularly racial discrimination—there is no clear intervention point, and the difference between the TE and the CDE_{Ob} is largely 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 CDE_{Ob} in another, equally valid description of the same events, as we explain next.

In our running example, the implicit population of interest consists of those individuals stopped by the police, and the TE reflects a process in which the decision-making process starts—and perception of race is counterfactually determined—after the stop has occurred but before the arrest decision has been made. We can, however, imagine moving back the clock and starting the process after an officer first spots an individual but before a stop decision has been made, with the population of interest now comprising those individuals spotted by an officer. In this case, the original TE is equivalent to the CDE_{Ob} on this newly defined population, where the first-stage decision indicates whether an individual was stopped. Both the original TE and the new CDE_{Ob} capture combined discrimination in the arrest and charging decisions, among the subset of individuals who were stopped. One can similarly measure cumulative discrimination that includes the stop decision itself, either in terms

[5] However, even in randomized trials, researchers may be interested in restricting to intermediate outcomes. For instance, there is a large literature studying the effects of job training programs on wages through random assignment to program [LaLonde, 1986]. In these settings the outcome of interest (wages) is defined only for units who are employed—but the employment status of the experimental units is itself affected by the intervention. A popular framework for studying these problems is principal stratification [Frangakis and Rubin, 2002], which in effect restricts the estimand to the subset of units who would have had a job regardless of the intervention (sometimes referred to as the always-employed). This estimand is related to but distinct from the CDE_{Ob}.

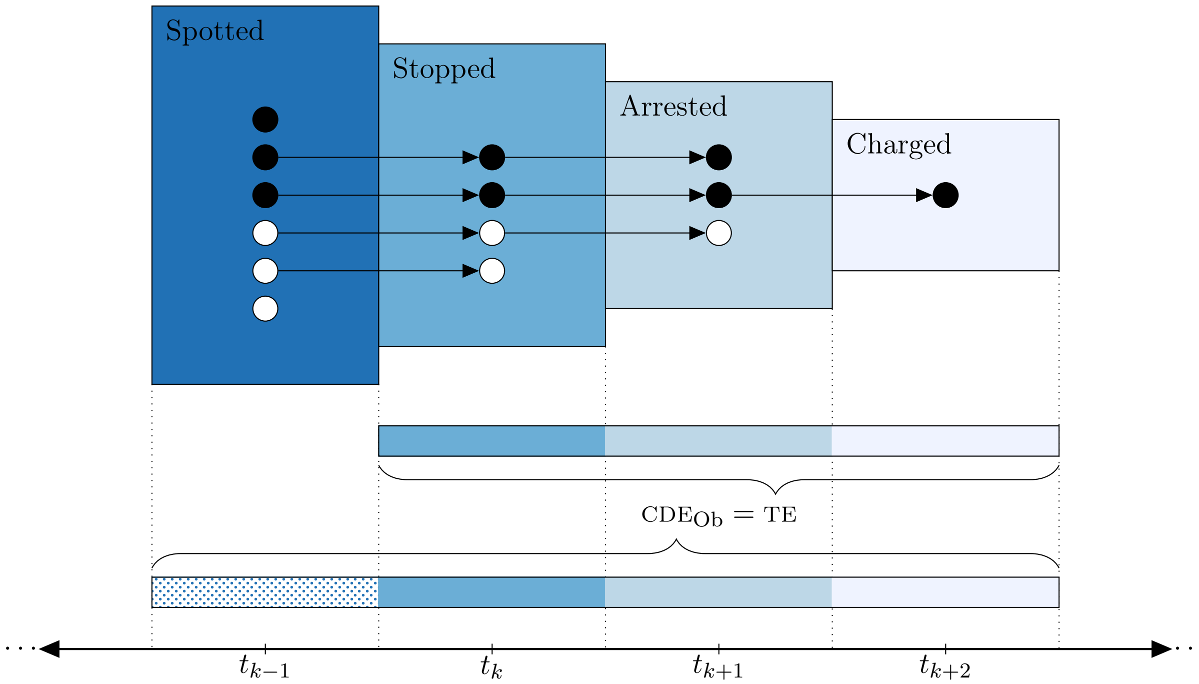


Figure 1: *One can measure combined discrimination in arrest and charging decisions either via the* TE *or the* CDE_{Ob}. *In studies of discrimination, there is no clear point at which race is "assigned" and so both the* TE *and the* CDE_{Ob} *can be used interchangeably to express the same underlying causal effect. 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* CDE_{Ob}*, 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.

of the TE or the CDE_{Ob}. For the former, 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 latter, we back up the clock once again and imagine the first-stage indicating whether an individual was spotted by an officer, among an even larger population of people who perhaps reside in the city under consideration. Figure 1 provides a graphical depiction of this interchangeability.

The above discussion illustrates that the very idea of "intermediate outcomes"—a concept central to the Knox et al. critique—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.[6] Indeed, such generational counterfactuals may be critical for understanding systemic, institutional discrimination. Although the TE appears 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.

[6]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.

One interpretation of this reality is an even more extreme version of the Knox et al. claim: that post-treatment bias is fundamentally unavoidable in the study of discrimination. We propose two alternatives that we believe are more productive. First, one can measure disparate treatment in the actions of specific deciders at specific decision points. In this vein, the CDE_{Ob} aims to narrowly quantify the effects of protected attributes on specific decisions, net of any discrimination that may have preceded it. Such assessment is particularly useful for informing legal liability—especially given constitutional prohibitions against intentional discrimination—and, consequently, for reforming behavior. For instance, to audit and improve prosecutorial practices, it is useful to understand the degree to which charging decisions may be motivated by animus or implicit bias [Eberhardt, 2019, O'Flaherty and Sethi, 2019], even if those decisions are just one part of a complex criminal justice system. Second, one can measure the broader disparate impacts of decisions on groups defined by race, gender, or other characteristics, regardless of whether those characteristics themselves affected the decisions—a point briefly mentioned in the Introduction and to which we return in the Discussion. Like disparate treatment, disparate impact plays a prominent role in legal and social understandings of discrimination. This alternative also circumvents many of the statistical challenges with defining race to be the treatment in a causal analysis, including concerns with post-treatment bias. However, given our focus on the causal effects of race, we do not discuss disparate impact in detail here, despite its broader importance.

To deconstruct concerns over post-treatment bias, we concentrate our efforts on elucidating the statistical properties of the CDE_{Ob}. Nonetheless it is worth highlighting some of the conceptual challenges with framing discrimination in this manner. As alluded to above, many in the causal inference community have expressed skepticism towards the use of so-called immutable characteristics, such as gender or race, as valid "causes" because appropriate counterfactuals are hard to define [Holland, 1986, 2003, VanderWeele and Robinson, 2014].[7] What does it mean, for example, to imagine an individual were a different race? The use of *perceived* traits [Greiner and Rubin, 2011] partially mitigates the concern, since it is easier to imagine interventions altering perceptions of traits rather than the traits themselves. Here, though, the exact form of the intervention likely impacts outcomes. Altering perceptions of race by changing the name on an employment application [Bertrand and Mullainathan, 2004] can have different effects than by selectively listing membership in a racial affinity group [Kang et al., 2016]. As such, any particular operationalization of race (or perception of race) may only be loosely attached to the underlying theoretical construct. Much has been written elsewhere on the subtleties of these issues [Jung et al., 2018], and it is not our aim to break new conceptual ground on the subject. We instead proceed as in Knox et al. and presume the mathematical objects we study have meaningful interpretations for questions of science and policy.

3 A Stylized Example

Our results above show that subset ignorability, in theory, is sufficient to ensure nonparametrically identified estimates of the CDE_{Ob}, even when the first-stage decisions are discriminatory. We illustrate that idea by investigating in detail a hypothetical scenario involving arrest decisions in the first stage and charging decisions in the second stage where subset ignorability holds by construction. This example includes discrimination in the first-stage decisions and, in particular, violates the conditions that Knox et al. claim are necessary for identifiability to hold. We then explore the

[7] VanderWeele and Robinson question the validity of race counterfactuals specifically, though they propose an alternative framework for interpreting causal effects of race. Relatedly, Sen and Wasow [2016] argue that race, instead of being "immutable", is better conceptualized as a "bundle of sticks" that can be disaggregated into elements whose causal effects can indeed be measured.

properties of simple estimators in this setting through a simulation study that demonstrates how one can accurately estimate the true CDE_{Ob} using only prosecutorial records.

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 are brought to a prosecutor who decides whether or not to press charges. We assume the prosecutor observes only the documented race and criminal record of the arrestee, and thus, by construction, the charging decision depends only on these two factors. For example, the prosecutor may choose only to charge individuals who have several previous drug convictions. Importantly, and in contrast to the officer, the prosecutor does not observe the individual's behavior leading up to the arrest.

Our goal is to estimate the CDE_{Ob} given the prosecutor's records on arrested individuals, namely their criminal history and race. Intuitively, subset ignorability holds in this simple scenario because the prosecutor's dataset contains all factors used in the charging decision, even though the prosecutor does not know all the factors that led to an arrest, a decision that may itself have been discriminatory. 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. Further, while criminal histories themselves may be the product of past discrimination, they reflect a form of complex, long-term discrimination that we do not aim to measure here. Such alternative 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 time, 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 this generative process, we can both compute the true CDE_{Ob} and compute estimates based only on information available to the prosecutor. Our data generation proceeds in three steps.

Step 1. First, for each person in our (infinite) population of individuals encountered by the police, we stochastically define their race Z_i, criminal history X_i, and behavior A_i. For simplicity, all three variables are binary—for example, one can imagine that individuals are either white or Black, that X_i indicates whether an individual had at least one previous drug conviction, and that A_i indicates whether they were seen actively engaging in a drug transaction. Officers observe Z_i and A_i for all individuals, and prosecutors observe Z_i and X_i for the subset of individuals who were arrested. For appropriately defined constants μ_Z, μ_X, μ_A, δ, and γ, these three covariates are generated as follows:

$$\begin{aligned} Z_i &\sim \textsc{Bern}(\mu_Z), \\ X_i &\sim \textsc{Bern}(\mu_X + \mathbb{1}_{\{Z_i=b\}} \cdot \delta), \\ A_i &\sim \textsc{Bern}(\mu_A + \mathbb{1}_{\{Z_i=b\}} \cdot \gamma). \end{aligned}$$

For ease of exposition, we assume the Bernoulli race variable Z_i takes values in $\{w, b\}$ rather than $\{0, 1\}$. The specification above allows for the distributions of criminal history (X) and behavior (A) to vary by race (Z).

Step 2. We next define each individual's potential outcomes for the arrest decision. Let $U_i \sim \textsc{Unif}(0, 1)$ be a uniform random variable on the unit interval. Then for constants α_0, α_A, and α_{black},

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.*

Z_i	A_i	X_i	$M_i(b)$	$M_i(w)$	$Y_i(b,1)$	$Y_i(w,1)$	M_i	Y_i
b	0	0	0	0	0	0	0	0
b	0	1	0	0	1	0	0	0
b	1	1	1	0	1	1	1	1
w	0	0	1	1	0	0	1	0
w	0	1	0	0	0	0	0	0

set

$$M_i(b) = \begin{cases} 1 & U_i \leq \alpha_0 + \alpha_A \cdot A_i + \alpha_{\text{black}} \\ 0 & \text{otherwise} \end{cases}$$

and

$$M_i(w) = \begin{cases} 1 & U_i \leq \alpha_0 + \alpha_A \cdot A_i \\ 0 & \text{otherwise.} \end{cases}$$

When $\alpha_{\text{black}} \geq 0$, anyone who would be arrested if white would also be arrested if Black (i.e., $M_i(b) \geq M_i(w)$). While not necessary for our results, this monotonicity property ensures our construction comports with intuitive theories of discrimination; this property is also one of the Knox et al. assumptions, as we discuss in Section 5. When $\alpha_{\text{black}} > 0$ (and $\alpha_0 + \alpha_A < 1$), 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.

Step 3. Finally, we define the charging potential outcomes in an analogous manner to the arrest potential outcomes. Let $V_i \sim \textsc{Unif}(0, 1)$ be a uniform random variable on the unit interval. Then for constants β_0, β_X, and β_{black}, set

$$Y_i(b,1) = \begin{cases} 1 & V_i \leq \beta_0 + \beta_X \cdot X_i + \beta_{\text{black}} \\ 0 & \text{otherwise} \end{cases}$$

and

$$Y_i(w,1) = \begin{cases} 1 & V_i \leq \beta_0 + \beta_X \cdot X_i \\ 0 & \text{otherwise.} \end{cases}$$

As above, $Y_i(b, 1) \geq Y_i(w, 1)$ 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. This property aids interpretation but is not necessary for our main results. Also as above, we say the charging decision is discriminatory when $\beta_{\text{black}} > 0$ (and $\beta_0 + \beta_X < 1$). To complete the definition of the charging potential outcomes, recall that $Y_i(b, 0) = 0$ and $Y_i(w, 0) = 0$, since we assume one cannot be charged without first being arrested.

Features of our Data Generating Process. Table 1 displays a sample of five rows of data generated from our three-step process. From the full set of potential outcomes, we can compute the true CDE_{Ob} by directly applying Definition 1 to the generated data, taking the average difference between $Y_i(b, 1)$ and $Y_i(w, 1)$ among arrested individuals.

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 do not observe A_i). Second, our set-up allows for situations where the arrest decisions are themselves discriminatory—those where $\alpha_{\text{black}} > 0$. As we discuss in more detail in Section 5, this property violates a condition that Knox et al. assert is necessary to identify the CDE_{Ob}. Third, the prosecutor's records include the full set of information on which charging decisions are based (i.e., Z_i and X_i). Moreover, the charging potential outcomes (generated in Step 3) depend only on one's criminal history, X_i, not on one's realized race, Z_i, and, consequently, $Y(z, 1) \perp\!\!\!\perp Z \mid X, M = 1$. Thus by construction, our generative process satisfies subset ignorability.

3.2 Estimating the CDE_{Ob}

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. We explore the performance of two methods for estimating the CDE_{Ob} based on data observed by the prosecutor: the stratified difference-in-means estimator described in Eq. (6), and a regression-based estimator.

The prosecutor can implement the stratified difference-in-means estimator in three steps. First, partition arrested individuals into those with $X_i = 0$ and $X_i = 1$. 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, the prosecutor can apply Eq. (7) 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 CDE_{Ob}, at least 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 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 CDE_{Ob} via a parametric model that estimates observed charging decisions as a function of criminal history and race.

To demonstrate this approach, we use a linear probability model fit on the prosecutor's data:

$$\mathbb{E}[Y_i \mid X_i, Z_i] = \beta_0 + \beta_1 X_i + \beta_2 Z_i. \tag{9}$$

Under such a linear model, the CDE_{Ob} is approximated by β_2, 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. (9) is in fact correctly specified, and so we are guaranteed to obtain statistically consistent estimates. However, even when the model is not perfectly specified, this general approach often gives reasonable approximations in practice, especially when one applies a suitably flexible statistical procedure (e.g., a random forest or regularized regression [Hastie et al., 2009]) to model observed outcomes.

3.3 Simulation Results

We perform a small simulation study to understand the properties of these two estimators across various assumptions about levels of discrimination. We simulate 1,000 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., $\mu_Z = 0.3$); that 30% of white individuals and 40% of Black individuals have a past drug conviction, indicated by X_i; and that 30% of white individuals and 20% of Black individuals are seen engaging in a drug transaction, indicated by A_i.[8] 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 CDE_{Ob} using both the stratified difference-in-means estimator and the regression-based estimator, and compare the results to the true population-level CDE_{Ob}. Figure 2 demonstrates that both estimators produce results that are unbiased and that are typically close to the true estimand across all parameter settings. For each combination of α_{black} and β_{black}, the estimates on the 1,000 synthetic datasets yield the approximate sampling distributions for the difference-in-means estimator and the regression-based estimator. 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, and the hollow points to the regression-based estimator. The horizontal lines indicate the true population-level CDE_{Ob}. In all cases, the points lie on the horizontal lines, 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.

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. (7); and for the regression-based estimator, we used the standard OLS estimate of standard error. For each parameter setting, we computed the proportion of confidence intervals for the 1,000 datasets that contained the true value of the CDE_{Ob}. We found the true coverage was in line with the nominal coverage, ranging from 94% to 96% across parameter specifications.

4 An Analysis of Real-World Charging Decisions

We now assess possible race and gender discrimination in real-world charging decisions. We start with the set of individuals in a major U.S. city who were arrested for a felony offense between 2013 and 2019.[9] 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).[10] As in our stylized example in Section 3, case information is sent to the prosecutor's office after an officer makes an arrest. In the jurisdiction we consider, an arrested individual is then charged if the prosecutor handling the case believes there is sufficient evidence to establish the individual committed the alleged offense beyond a reasonable doubt, and that pressing charges would be in line with the city's justice goals.[11]

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

[9] For privacy reasons, we are unable to release individual-level data.

[10] 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].

[11] In contrast to these charging criteria, police officers are legally allowed to make arrests when they believe there is probable cause of criminal activity, a lower evidentiary standard.

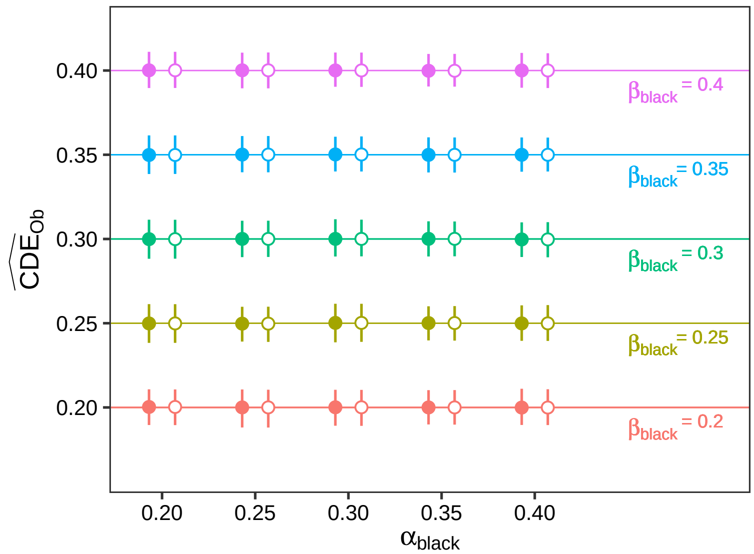


Figure 2: *In our hypothetical example of officer and prosecutor behavior, one can obtain accurate estimates of discrimination in charging decisions based only on information available to prosecutors on the subset of individuals who were arrested. The plot shows results of 1,000 simulations for each of 25 different combinations of discrimination in officer and prosecutor decisions, given by* α_{black} *and* β_{black}*, respectively. The true value of the* $\mathrm{CDE}_{\mathrm{Ob}}$*, indicated by the horizontal colored lines, is computed based on the full set of potential outcomes for each individual. For each parameter choice, we summarize the sampling distribution for our two estimators: the stratified difference-in-means estimator (solid circle) and the regression-based estimator (hollow circle). Specifically, for each combination of* α_{black} *and* β_{black}*, we display the mean of the sampling distribution (indicated by the points), as well as the interval spanned by the 2.5th and 97.5th percentiles of the sampling distribution. These results show that both estimators can recover the true value of the* $\mathrm{CDE}_{\mathrm{Ob}}$*, and also confirm that the true value of the* $\mathrm{CDE}_{\mathrm{Ob}}$ *does not depend on the degree of discrimination in the first stage, as seen by the constant value of the* $\mathrm{CDE}_{\mathrm{Ob}}$ *across different values of* α_{black}.

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, slightly higher than the 50% charging rate for 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 $\mathrm{CDE}_{\mathrm{Ob}}$. But before doing so, we first check that overlap is satisfied—meaning that $\Pr(Z = z \mid X = x, M = 1) > 0$—for both our race-based and our gender-based analyses. Here, $Z_i = 1$ 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_i is a vector of observed case features, and $M = 1$ means we restrict to those individuals who were arrested. In contrast to ignorability, overlap can be assessed directly by examining the data. To do so, we estimate propensity

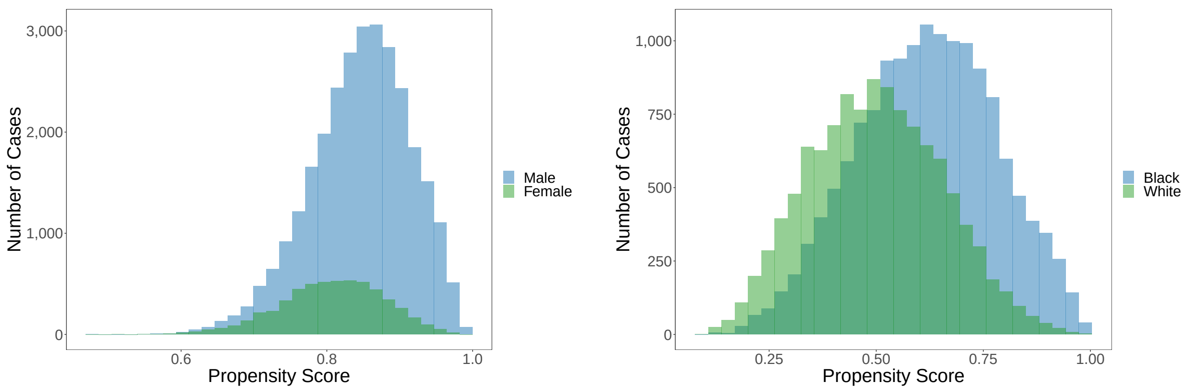


Figure 3: *We plot, for both our gender-based (left) and race-based (right) analyses, the distribution of propensity scores, disaggregated by observed treatment status. We find that the propensity scores are concentrated away from the interval endpoints, satisfying overlap.*

scores [Rosenbaum and Rubin, 1983a], $\Pr(Z = z \mid X = x, M = 1)$, via an L^1-regularized (lasso) logistic regression model.

In Figure 3, 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.

We now estimate the CDE_{Ob} via a linear probability model fit to the data, as in our synthetic examples above:

$$\mathbb{E}[Y_i \mid X_i, Z_i] = \beta_0 + X_i\beta_1 + Z_i\beta_2, \tag{10}$$

where $Y_i = 1$ if individual i was charged, and X_i denotes the vector of covariates described above. In the gender model, we find that the $\widehat{\text{CDE}}_{\text{Ob}}$—as given by $\hat{\beta}_2$—is 0.025 (95% CI: [0.014, 0.037]); and in the race model, we have $\widehat{\text{CDE}}_{\text{Ob}}$ 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 discrimination 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 differences in the initial prosecutorial charging decision between Black and white individuals—specifically for charges with statutory mandatory minimum sentences—was a primary driver for sentencing disparities. 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 analysis of prosecutors at the San Francisco District Attorney's Office, MacDonald and Raphael [2017] 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.

The AUC of our outcome model 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 is likely violated. 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 proceed by positing relationships between the unmeasured confounder and both the treatment variable (e.g., race or gender) and the outcome (e.g., the charge decision), and then examine the sensitivity of estimates under the model of confounding.

Here 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 $\hat{\beta}_2$ in Eq. (10)—might change if one were to refit the model including an unmeasured confounder U. More specifically, under the extended model

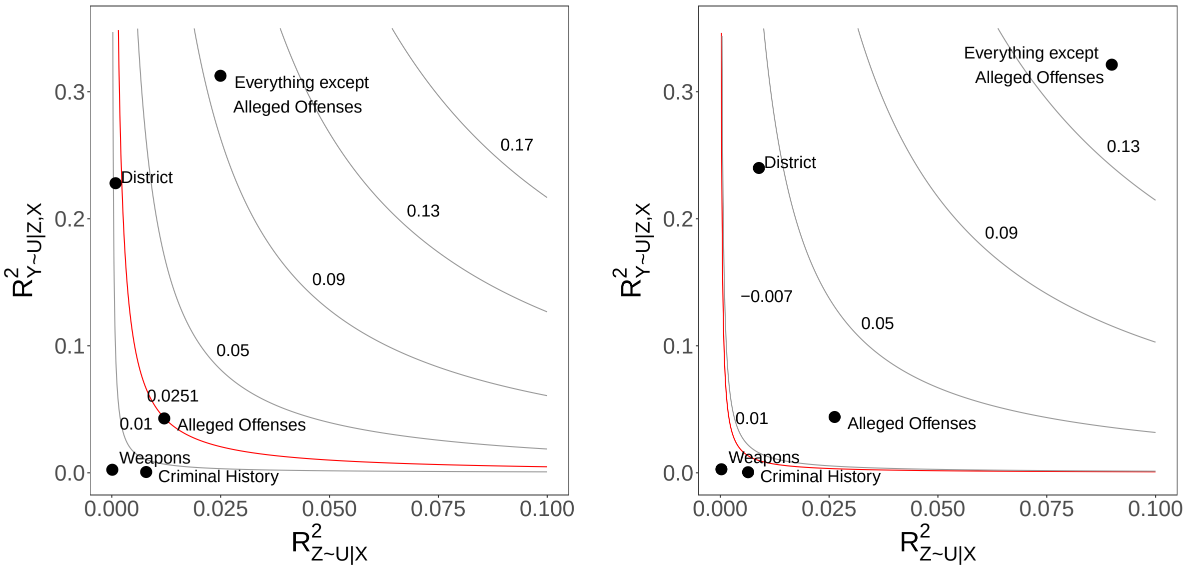
$$\mathbb{E}[Y_i \mid X_i, Z_i, U_i] = \beta_0 + X_i\beta_1 + Z_i\beta_2 + U_i\gamma,$$

Cinelli and Hazlett bound the change in $\hat{\beta}_2$ 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, $R^2_{Y\sim U|Z,X} = (R^2_{\text{full}} - R^2_{\text{red}})/(1 - R^2_{\text{red}})$. $R^2_{Z\sim U|X}$ is defined analogously. As these partial R^2 values increase, so does the amount by which $\hat{\beta}_2$ may change.

The contour plots in Figure 4 show the maximum amount by which the $\widehat{\text{CDE}}_{\text{Ob}}$ 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 $\widehat{\text{CDE}}_{\text{Ob}}$. 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. The method of Cinelli and Hazlett partially addresses this issue by parameterizing bounds in terms of partial R^2 values, which may be easier to directly interpret than some alternative methods. To further aid interpretability, 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. (10) 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 4 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 far 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 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



(a) Gender-based analysis

(b) Race-based analysis

Figure 4: *Contour plots describing the sensitivity of the $\widehat{\mathrm{CDE}}_{\mathrm{Ob}}$ to unmeasured confounding, for our analysis of gender (left) and race (right). The plots indicate the maximum amount the $\widehat{\mathrm{CDE}}_{\mathrm{Ob}}$ 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 $\widehat{\mathrm{CDE}}_{\mathrm{Ob}}$ 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.*

red curve in both cases, suggesting that an unobserved confounder of similar strength could again substantially alter our results.

By definition, it is not possible to know the exact nature and impact of unmeasured confounding. Thus, as in many applied statistical problems, one must ultimately rely in large part on domain expertise and intuition to form reasonable conclusions. In this case, based on conversations with attorneys in our partner jurisdiction, and on our own experience working with criminal justice data, we interpret these results as providing moderately robust evidence that gender and race likely have limited effects on prosecutorial charging decisions in this jurisdiction.

5 A Comparison to Alternative Ignorability Conditions

We conclude by stating and investigating the specific mathematical claims of Knox et al. Their main assertion is that a set of assumptions—which they call treatment ignorability, mediator ignorability, and mediator monotonicity—are, together, necessary and sufficient conditions for the $\mathrm{CDE}_{\mathrm{Ob}}$ to be nonparametrically identified by the observed (second-stage) data. As we show below, these conditions are sufficient but they are not in fact necessary. In particular, 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 $\mathrm{CDE}_{\mathrm{Ob}}$ can accordingly 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.

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

Definition 5 (Sequential ignorability)**.** We say that *sequential ignorability* is satisfied when the following two conditional independence criteria hold:

$$\{Y(z,m), M(z')\} \perp\!\!\!\perp Z \mid X, \tag{11}$$

$$Y(z,m) \perp\!\!\!\perp M \mid Z, X, \tag{12}$$

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

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 9, 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, X may encode an individual's criminal history but not that individual's behavior at the time of the arrest. In this case, one would in general expect the first condition of sequential ignorability—in Eq. (11)—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 introduce the definitions of treatment ignorability, mediator ignorability, and mediator monotonicity proposed by Knox et al., starting with treatment ignorability.

Definition 6 (Treatment ignorability)**.** *Treatment ignorability* is the combination of the following two conditional independence criteria: for $z \in \{w, b\}$ and $m \in \{0,1\}$,

$$M(z) \perp\!\!\!\perp Z \mid X, \tag{13}$$

$$Y(z,m) \perp\!\!\!\perp Z \mid M(w), M(b), X. \tag{14}$$

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—this condition is indeed violated in the simple example discussed in Section 3.1, as $M(b)$ depends on A, which in turn depends on Z. 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 CDE_{Ob}, which only quantifies discrimination in the second-stage decision. Proposition 3 makes that statement precise, showing that subset ignorability—which does not consider first-stage potential outcomes—is sufficient to ensure the CDE_{Ob} 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; for example, both subset ignorability and the second condition of treatment ignorability are satisfied in the stylized model of Section 3.1. In theory, however, the difference between the two is large. As we show in Theorem 9 below, treatment ignorability alone—even with its strong assumption on the first stage—is not sufficient to ensure the CDE_{Ob} is identified by the second-stage data.

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

Definition 7 (Mediator ignorability)**.** For $z \in \{w, b\}$ and $m \in \{0, 1\}$, *mediator ignorability* is satisfied when

$$Y(z, m) \perp\!\!\!\perp M(w) \mid Z = z, M(b) = 1, X. \tag{15}$$

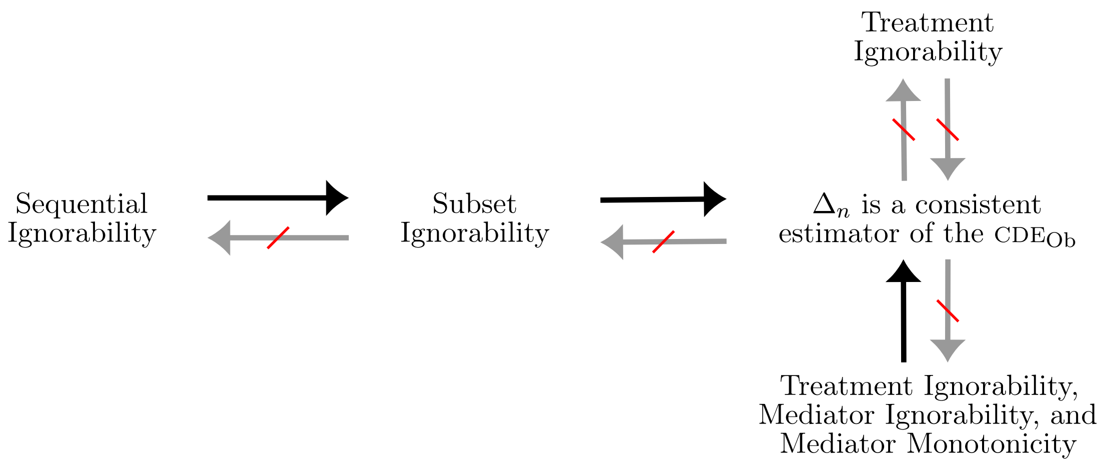
Definition 8 (Mediator monotonicity)**.** *Mediator monotonicity* is satisfied when

$$M_i(b) \geq M_i(w). \tag{16}$$

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 that Knox et al. argue is also necessary for valid estimation: $M_i(b) \geq M_i(w)$, meaning that an individual who would be arrested if white would also be arrested if Black. The monotonicity requirement is perhaps intuitively plausible given our understanding of racial discrimination, but the conditional independence assumption of mediator ignorability seems harder to interpret. Nevertheless, Knox et al. write that "every analysis estimating a racial effect using only data on stopped individuals [i.e., when data are censored except for the stratum $M = 1$] implicitly [assumes Mediator Ignorability]" [Knox et al., 2020, p. 8].

Having introduced the key definitions, we can now present our main analytic result, Theorem 9, 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 Proposition 3) that subset ignorability is a sufficient condition for the difference-in-means estimator Δ_n to yield consistent estimates of the CDE_{Ob}. Further, we show that treatment ignorability is not a necessary condition for Δ_n to yield consistent estimates, in contrast to the Knox et al. claim. We show this by explicitly constructing examples for which $\Delta_n \overset{\text{a.s.}}{\to} \text{CDE}_{\text{Ob}}$, 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 CDE_{Ob}. Accordingly, no estimator, including Δ_n, can yield a consistent estimate of the CDE_{Ob} for every instance in the family.

Theorem 9. *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:*



Theorem 9, which we prove in Appendix A, refutes the key technical claim of Knox et al. Namely, we show that the conditions they assert are required for valid inference—treatment ignorability, mediator ignorabilty, and mediator monotonicity—are not in fact necessary. Importantly, the more conventional assumption of subset ignorability is sufficient to ensure the CDE_{Ob} can be identified from data on the second-stage decisions. The conclusions in Knox et al. stem from mistaking a proof of sufficiency for one of necessity. Using treatment ignorability and mediator monotonicity, the authors first show that $\text{CDE}_{\text{Ob}} - \mathbb{E}[\Delta_n] = \alpha + \beta$ for certain complex expressions α and β. They then show that $\alpha = 0$ and $\beta = 0$ if mediator ignorability holds. This argument shows that treatment ignorability, mediator monotonicity, and mediator ignorability are jointly *sufficient* to ensure Δ_n provides an unbiased estimate of the CDE_{Ob}, but it does not show that these three conditions are also necessary.

6 Discussion

In response to recent statistical concerns raised about discrimination studies, we have worked to clarify the theoretical underpinnings of such research. We specifically showed that subset ignorability, together with overlap, is sufficient to guarantee that a standard causal measure of discrimination (the CDE_{Ob}) is nonparametrically identified in many common settings. In particular, 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. Our conclusions stand in contrast to those of Knox et al., whose critique of discrimination studies rests on an erroneous mathematical claim. In carrying out this work, we have aimed to make explicit the often unstated assumptions in studies of discrimination, and to put that research on more solid theoretical footing.

Stepping back, there are at least two general frameworks for conceptualizing discrimination, which approximately map to the legal notions 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. Both disparate treatment and disparate impact play important roles in legal and policy discussions. But it is critical to distinguish between the two, as the perspective one adopts in any given situation affects the choice of statistical estimation strategy and the interpretation of results.

Throughout this paper, we have focused on the statistical foundations and measurement of disparate treatment. In our primary example—prosecutorial charging decisions—we found the effect of race and gender on decisions was relatively small. But 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, suggesting little disparate treatment. However, among the 748 charged cases, 464 (62%) involved a Black individual. Charging decisions for these cases thus impose a heavy burden on Black individuals, even if those decisions were not tainted by animus or implicit bias. To the extent that prosecution of such drug crimes is misaligned with community goals, these decisions create an unjustified, and discriminatory, disparate impact.

In observational studies of discrimination, concerns about post-treatment bias may be misplaced, but rigorously measuring discrimination is still a daunting task. At an empirical level, it is often difficult to gain access to detailed data on individual decisions, in which case benchmark analysis—even when coupled with sensitivity analysis—may be an inadequate approach. At a theoretical level, we still 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, much of the literature has framed discrimination in terms of causal effects 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., charging decisions) does not reflect specific or cumulative discrimination at other stages—for example, arrests. Looking forward, we hope our work offers a path toward quantifying disparities, 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.

I. Ayres. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of "included variable" bias. *Perspectives in Biology and Medicine*, 48(1):68–S87, 2005.

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.

G. S. Becker. *The economics of discrimination.* University of Chicago Press, 1957.

G. S. Becker. Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, pages 385–409, 1993.

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.

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.

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.

J. L. Eberhardt. *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. Viking, 2019.

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

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.

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.

S. Goel, J. M. Rao, R. Shroff, et al. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016.

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.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media, 2009.

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.

P. W. Holland. Causation and race. *ETS Research Report Series*, 2003(1):i–21, 2003.

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.

S. K. Kang, K. A. DeCelles, A. Tilcsik, and S. Jun. Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly*, 61(3):469–502, 2016.

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

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

J. MacDonald and S. Raphael. An analysis of racial and ethnic disparities in case dispositions and sentencing outcomes for criminal cases presented to and processed by the Office of the San Francisco District Attorney. Technical report, 2017.

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.

B. O'Flaherty and R. Sethi. *Shadows of Doubt: Stereotypes, Crime, and the Pursuit of Justice*. Harvard University Press, 2019.

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.

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.

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.

M. Sen and O. Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522, 2016.

C. Simoiu, S. Corbett-Davies, and S. Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

T. J. VanderWeele and W. R. Robinson. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473, 2014.

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`.

A Proof of Theorem 9

Proposition 3 shows that subset ignorability implies that Δ_n is a consistent estimator of the CDE_{Ob}. We show the remaining seven implications and non-implications in turn, starting with the claim that sequential ignorability implies subset ignorability. For completeness, we prove that the conjunction of treatment ignorability, mediator ignorability, and mediator monotonicity implies that Δ_n is a consistent estimator of the CDE_{Ob}—a fact initially stated and proved by Knox et al.

Case 1 (Sequential ignorability implies subset ignorability)**.** The first condition of sequential ignorability, in Eq. (11), 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. (12)—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,

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

Subset ignorability now follows, as it is the special case in which $M = 1$ in Eq. (17).

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

$$\begin{aligned} &\Pr(Z = z, M(b) = m_b, M(w) = m_w) \\ &\quad = \Pr(Z = z) \cdot \Pr(M(b) = m_b \mid Z = z) \cdot \Pr(M(w) = m_w \mid Z = z), \end{aligned} \tag{18}$$

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:

$$\Pr(Y(z,1) = 1 \mid Z) = \begin{cases} 1 & z = b \wedge Z = b, \\ 0 & z = w \wedge Z = b, \\ \frac{1}{2} & Z = w. \end{cases} \tag{19}$$

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. (19), $Y(z, 1) \not\perp\!\!\!\perp Z$. (Because X and M are constant, we need not condition on them when considering the subset ignorability criterion.) We further have,

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

Finally,

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

Thus, even though subset ignorability is violated in this example, Δ_n yields a consistent estimate of the CDE_{Ob}.

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 Proposition 3, Δ_n is a consistent estimator of the CDE_{Ob}. However, that example does not satisfy treatment ignorability, as $M(z) \not\perp\!\!\!\perp Z$, contrary to Eq. (13). (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 CDE_{Ob})**.** We show, more generally, that the CDE_{Ob} 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 CDE_{Ob}. As a result, no estimator—including Δ_n—can consistently estimate the CDE_{Ob} 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

$$\begin{aligned}
&\Pr(Y(z,1) = y_{zm}, M(w) = m_w, Z = z) \\
&\quad = \Pr(Y(z,1) = y_{zm} \mid M(w) = m_w) \cdot \Pr(M(w) = m_w) \cdot \Pr(Z = z),
\end{aligned} \tag{20}$$

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

$$\Pr(Y(z,1) = 1 \mid M(w)) = \begin{cases} \alpha & M(w) = 0 \land z = w, \\ 1 & \text{otherwise.} \end{cases} \tag{21}$$

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. (20) and $M(b) \perp\!\!\!\perp Z$ since $M(b)$ is constant. Consequently, the first condition of treatment ignorability is satisfied. Eq. (20) 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 $M(b) = 1$,

$$\Pr(Y = y, Z = b, M = 1) = \begin{cases} 0 & y = 0,\\ \frac{1}{2} & y = 1.\end{cases} \tag{22}$$

Now consider the case $z = w$. By Eq. (20),

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

By Eq. (21), $\Pr(Y(w,1) = 1 \mid M(w) = 1) = 1$, and so,

$$\Pr(Y = y, Z = w, M = 1) = \begin{cases} 0 & y = 0,\\ \frac{1}{4} & y = 1.\end{cases} \tag{23}$$

Finally, combining Eqs. (22) and (23) 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 $\textsc{cde}_{\text{Ob}}$ differs across these examples. Specifically, we have

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

where second to last equality follows from Eq. (21) 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. (21), 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 $\textsc{cde}_{\text{Ob}}$, implying that the $\textsc{cde}_{\text{Ob}}$ 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 $\textsc{cde}_{\text{Ob}}$)**.** The proof is in two pieces. First, we derive an expression for the $\textsc{cde}_{\text{Ob}}$ holding X constant, and then prove the general claim.

Supposing $X=x$ is constant, recall that by definition $M_i=1$ if and only if $M_i(z)=1$ where $Z_i=z$. 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 $\textsc{cde}_{\text{Ob}}$ in Eq. (1). It follows from the law of total expectation that:

$$\begin{aligned}
\textsc{cde}_{\text{Ob}} &= \mathbb{E}[Y(b,1) - Y(z,1) \mid M=1] \\
&= \mathbb{E}[Y(b,1) - Y(z,1) \mid E_1] \cdot \Pr(E_1 \mid M=1) \\
&\qquad + \mathbb{E}[Y(b,1) - Y(z,1) \mid E_2] \cdot \Pr(E_2 \mid M=1)
\end{aligned} \tag{24}$$

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. (14), we are free to condition both terms on the right hand side by levels of Z, yielding

$$\mathbb{E}[Y(b,1) \mid Z=b, E_1] - \mathbb{E}[Y(w,1) \mid Z=w, E_1] = \mathbb{E}[Y \mid Z=b, E_1] - \mathbb{E}[Y \mid Z=w, E_1], \tag{25}$$

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. (15), 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. (14). Replacing potential outcomes with their realizations, it follows that

$$\mathbb{E}[Y(b,1) - Y(w,1) \mid E_2] = \mathbb{E}[Y \mid Z=b, E_2] - \mathbb{E}[Y \mid Z=w, E_1]. \tag{26}$$

Now, we substitute Eqs. (25) and (26) into Eq. (24).

$$\begin{aligned}\text{CDE}_{\text{Ob}} &= (\mathbb{E}[Y \mid Z=b, E_1] - \mathbb{E}[Y \mid Z=w, E_1]) \cdot \Pr(E_1 \mid M=1) \\ &\qquad + (\mathbb{E}[Y \mid Z=b, E_2] - \mathbb{E}[Y \mid Z=w, E_1]) \cdot \Pr(E_2 \mid M=1) \\ &= (\mathbb{E}[Y \mid Z=b, E_1] - \mathbb{E}[Y \mid Z=w, M=1]) \cdot \Pr(E_1 \mid M=1) \\ &\qquad + (\mathbb{E}[Y \mid Z=b, E_2] - \mathbb{E}[Y \mid Z=w, M=1]) \cdot \Pr(E_2 \mid M=1) \\ &= \big(\mathbb{E}[Y \mid Z=b, E_1] \cdot \Pr(E_1 \mid M=1) + \mathbb{E}[Y \mid Z=b, E_2] \cdot \Pr(E_2 \mid M=1)\big) \\ &\qquad - \big(\mathbb{E}[Y \mid Z=w, M=1] \cdot (\Pr(E_1 \mid M=1) + \Pr(E_2 \mid M=1))\big) \\ &= \mathbb{E}[Y \mid Z=b, M=1] - \mathbb{E}[Y \mid Z=w, M=1],\end{aligned} \tag{27}$$

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

$$\begin{aligned}\mathbb{E}[Y(b,1) - Y(w,1) \mid M=1] &= \sum_x \mathbb{E}[Y(b,1) - Y(w,1) \mid M=1, X=x] \cdot \Pr(X=x \mid M=1) \\ &= \sum_x \mathbb{E}[Y \mid Z=b, M=1, X=x] \cdot \Pr(X=x \mid M=1) \\ &\qquad - \mathbb{E}[Y \mid Z=w, M=1, X=x] \cdot \Pr(X=x \mid M=1),\end{aligned} \tag{28}$$

where the second equality follows from Eq. (27), using the fact that X is constant on each of the events $\{X=x\}$. Eq. (28) is identical to Eq. (3) in Proposition 3, and so the estimator Δ_n converges almost surely to the quantity on the right-hand side of Eq. (28) by precisely the same argument as there.

□