Administrative Records Mask Racially Biased Policing

DEAN KNOX Princeton University
WILL LOWE Hertie School of Governance
JONATHAN MUMMOLO Princeton University

Researchers often lack the necessary data to credibly estimate racial discrimination in policing. In particular, police administrative records lack information on civilians police observe but do not investigate. In this article, we show that if police racially discriminate when choosing whom to investigate, analyses using administrative records to estimate racial discrimination in police behavior are statistically biased, and many quantities of interest are unidentified—even among investigated individuals—absent strong and untestable assumptions. Using principal stratification in a causal mediation framework, we derive the exact form of the statistical bias that results from traditional estimation. We develop a bias-correction procedure and nonparametric sharp bounds for race effects, replicate published findings, and show the traditional estimator can severely underestimate levels of racially biased policing or mask discrimination entirely. We conclude by outlining a general and feasible design for future studies that is robust to this inferential snare.

Concern over racial bias in policing, and the public availability of large administrative data sets documenting police–civilian interactions, have prompted a raft of studies attempting to quantify the effect of civilian race on law enforcement behavior. These studies consider a range of outcomes including ticketing, stop duration, searches, and the use of force (e.g., Antonovics and Knight 2009; Fryer 2019; Ridgeway 2006; Nix et al. 2017). Most research in this area attempts to adjust for omitted variables that may correlate with suspect race and the outcome of interest. In contrast, this study addresses a more fundamental problem that remains even if the vexing issue of omitted variable bias is solved: the inevitable statistical bias that results from studying racial discrimination using records that are themselves the product of racial discrimination (Angrist and Pischke 2008; Elwert and Winship 2014; Rosenbaum 1984). We show that when there is any racial discrimination in the decision to detain civilians—a decision that determines which encounters appear in police administrative data at all—then estimates of the effect of civilian race on subsequent police behavior are biased absent additional data and/or strong and untestable assumptions.

This study makes several contributions. We clarify the causal estimands of interest in the study of racially discriminatory policing—quantities that many studies appear to be targeting, but are rarely made explicit—and show that the conventional approach fails to recover any known causal quantity in reasonable settings. Next, we highlight implicit and highly implausible assumptions in prior work and derive the statistical bias when they are violated. We proceed to develop informative nonparametric sharp bounds for the range of possible race effects, apply these in a reanalysis and extension of a prominent article on police use of force (Fryer 2019), and present bias-corrected results that suggest this and similar studies drastically underestimate the level of racial bias in police–civilian interactions. Finally, we outline strategies for future data collection and research design that can mitigate these threats to inference. These are discussed in the context of a detailed and feasible proposed study of racial bias in traffic stops.

As we show in this article, the difficulty of estimating racial bias using police records stems from a thorny combination of mediation (Hernán, Hernández-Diáz, and Robins 2004; Imai et al. 2011; Pearl 2001; Robins, Hernán, and Brumbach 2000; VanderWeele 2009) and selection (Heckman 1979; Lee 2009): the effect of civilian race on the outcome of a police encounter is mediated by whether the civilian is stopped by police, but analysts only have data for one level of the mediator—that is, data on stopped individuals. Because of this, police records do not contain a representative sample of all individuals that police observe, but rather only those civilian encounters which escalated to the point of triggering a reporting requirement. If a civilian’s race affects whether officers choose to stop that civilian (Gelman, Fagan, and Kiss 2007; Glaser 2014), then analyzing administrative police records amounts to conditioning on a variable that is itself affected by suspect race, namely, whether a suspect appears in the data at all. This could occur if officers have a higher
threshold for stopping white civilians during the unseen first stage of police–civilian contact, meaning that white civilians observed in the data are incomparable because they tend to pose a greater threat to police than observed minorities. These unobserved differences can lead analysts to understate anti-minority racial bias—or even produce the appearance of anti-white bias—in the use of force. Despite claims to the contrary (Fryer 2018, 2), this statistical bias often cannot be eliminated with additional control variables, even if the goal is to estimate causal effects among the subset of police–civilian encounters that appear in police data. Moreover, the problem remains whether racial bias in detainment stems from so-called “taste-based” or “statistical” discrimination (Arrow 1972, see below for extended discussion on this point).

At the first glance, the problem of race-based selection into policing data may appear a classic case of sample selection bias (Elwert and Winship 2014; Heckman 1979) for which numerous remedies already exist. But policing data exhibit a constellation of features that render previous methodological approaches unsuitable or unusable in this setting, leading prominent scholars in this area to declare that “it is unclear how to estimate the extent of such bias or how to address it statistically,” (Fryer 2018, 5). For example, Heckman (1979) and more recent extensions like Lee (2009) provide methods for estimating or bounding average treatment effects in the population while accounting for sample selection. But with only data on stopped individuals, policing scholars rarely seek to estimate population treatment effects, instead targeting effects among individuals who actually interact with police. We show that even without attempting to generalize to the broader population, the issues we raise result in biased estimates of the effect of race on police behavior even among encounters in which civilians are detained.

A related large literature provides remedies for so-called “post-treatment bias”—statistical bias that results from conditioning on a variable that is affected by the causal variable of interest (Rosenbaum 1984). But implementation of these techniques requires either knowledge of the scale of the missing data (e.g., Nyhan, Skovron, and Titiunik 2017) or complete data on the posttreatment variable (e.g., Acharya, Blackwell, and Sen 2016). In the case of policing, administrative data sets only include observations with one level of the posttreatment variable (i.e., data on stopped individuals) and give no purchase on the number of individuals police observe but do not stop, meaning these techniques cannot be applied. This scenario also differs from situations of “truncation by death” (Frangakis and Rubin 2002) in which receipt of a treatment causes sample attrition and renders outcomes for some portion of units undefined. In the policing setting, individuals not detained by police are absent from the data, but many outcomes of interest are often still defined (e.g., the level of force applied to nonstopped individuals is zero, a realized outcome). This feature allows us to identify additional causal quantities that cannot be recovered in the “truncation by death” setting. In short, existing methods offer either unusable or suboptimal solutions to this pernicious threat to inference, absent strong assumptions about the unseen process mapping civilian race to officers’ decisions to detain individuals.

Our analysis indicates that existing empirical work in this area is producing a misleading portrait of evidence as to the severity of racial bias in police behavior. Replicating and extending the study of police behavior in New York in Fryer (2019), we show that the consequences of ignoring the selective process that generates police data are severe, leading analysts to dramatically underestimate or conceal entirely the differential police violence faced by civilians of color. For example, while a naïve analysis that assumes no race-based selection into the data suggests only 10,000 black and Hispanic civilians were handcuffed because of racial bias in New York City between 2003 and 2013, we estimate that the true number is approximately 56,000. And while analyses ignoring bias in stopping would conclude that 10% of uses of force against black and Hispanic civilians in these data were discriminatory, after bias-correction, we estimate that the true percentage is 39%.

While the techniques used to obtain our corrected results eliminate several facially implausible (and in some cases, empirically falsified) assumptions that are implicit in prior work, we caution that they nevertheless rely on weaker assumptions that in some cases are difficult to verify, as we discuss below. We seek to advance the study of racial bias in policing by explicitly stating these assumptions, discussing their plausibility in this context, and carefully grounding unobservable parameters—in particular, the proportion of racially discriminatory minority stops, which relates closely to the severity of the statistical bias—in prior research (Gelman, Fagan, and Kiss 2007; Goel, Rao, and Shroff 2016). We show that obtaining more precise bias-corrected estimates of racial discrimination in policing requires future research to be designed with this issue in mind. To that end, we outline a research design that alleviates these concerns. Our study also provides a general framework for analyzing the study of racial bias that can illuminate the causal interpretation of other longstanding tests for discrimination. For example, we show that under reasonable assumptions, so-called “outcome tests,” which compare the rates of finding evidence of criminal activity across detained suspects of different racial groups (Knowles, Perisco, and Todd 2001), imply a lower bound on the share of racial minorities who are discriminatorily detained. Outcome tests also appear elsewhere in criminal justice studies, for example, in capital sentencing (Alesina and...
Ferrara 2014) and bail decisions (Arnold, Dobbie, and Yang 2018). And as Ayres (2002) and Simoiu, Corbett-Davies, and Goel (2017) note, such tests have also been applied in a range of other social contexts, including financial lending and editorial decisions. By nesting the study of discrimination in a rigorous and general causal framework, our study can help synthesize results from a broad interdisciplinary literature on racial bias.

Our work also extends a growing literature in political science examining the political implications of law enforcement which, in recent decades, has largely studied policing indirectly, for example, as a means of explaining political participation (Burch 2013; Cohen et al. 2017; Lerman and Weaver 2014; White 2019) or as an instance of bureaucracy (Brehm and Gates 1999; Lipsky 1980; Ostrom and Whitaker 1973; Wilson 1989). This work is path breaking, but with some recent exceptions (Harvey and Mungan 2019; Magaloni, Franco, and Melo 2015; Mummolo 2018; Peyton et al. 2019; Soss and Weaver 2017), has tended to conceptualize policing as a cause of politics, rather than a political act in and of itself. The field’s relative inattention to policing was made evident by several recent officer-involved shootings of unarmed black men (Edwards, Lee, and Esposito 2019) and subsequent social unrest that caught many political scientists flatfooted, with little systematic evidence to offer as the demand for explanations of police behavior surged. As Soss and Weaver (2017) note, the field’s limited store of relevant knowledge in the aftermath of these events was especially glaring given law enforcement’s role as an everyday conduit of state power. According to one often-cited definition, politics is “who gets what, when, how” (Lasswell 1936). As a matter of routine, the dynamics of police-civilian interactions determine who gets protected, punished, or left to fend for themselves (Wilson 1968). Viewed in this way, the role of race in the state’s exercise of violence, as well as in the provision of safety more broadly, is inherently political (Alexander 2010; Gottschalk 2008; Key 1949). In addition to offering a rigorous analytic framework to help researchers contend with longstanding methodological hurdles, our study also underscores an often overlooked truth: policing is high-stakes politics.

CONCEPTUALIZING RACE AS A CAUSAL VARIABLE

We regard the investigation of racial bias in policing as an inherently causal inquiry, albeit a notoriously difficult one. That is, researchers seek to assess whether police behavior during police-civilian encounters would have differed if the civilian had belonged to another racial group, holding constant civilian behavior and circumstances. As noted in Fryer (2018), this “race effect”...is the proverbial ‘holy grail’—the parameter that we are all attempting to estimate but never quite do” (2). This task is distinct from the descriptive enterprise of merely documenting differential police behavior during encounters with various groups, as such disparities can arise via numerous processes that do not imply racial discrimination.3

The notion of a “causal effect of race” on an individual’s outcome is the subject of much contention in the literature on causal inference (Hernán 2016; Pearl 2018). Most notably, some have argued that this effect is undefined because race is an immutable, and hence nonmanipulable, characteristic (Holland 1986). Others argue that an individual’s race is a complex, multifaceted treatment—a “bundle of sticks,” in the words of Sen and Wasow (2016)—that affects outcomes through myriad channels, and therefore, researchers must be precise about the specific facets of race under consideration (Greiner and Rubin 2011).

Our analysis avoids this debate by focusing on police-civilian encounters—that is, sightings of civilians by police—as the unit of analysis, rather than individuals. The manipulation of race is conceptualized as the counterfactual substitution of an individual with a different racial identity into the encounter, while holding the encounter’s objective context—location, time of day, criminal activity, etc.—fixed. In other words, the “treatment” in this case is the entire “bundle of sticks” encapsulating the race of the civilian—including, for example, skin tone, dialect, and clothing. We note that the credibility of causal inferences and the exact interpretation of racial discrimination in this framework will depend crucially on how the analyst defines “race.”

We leave the specific operationalization in a given context to the analyst, and, in line with advice in Sen and Wasow (2016), encourage scholars to carefully convey their conceptualization of race when studying this and related questions.4

By conceptualizing the treatment in this way, we avoid consideration of the perhaps implausible counterfactual of holding all features of an individual constant but for their race. While various aspects of racial identity and its close correlates may not be separable in the observed world, there exists a subset of comparable situations in which minority and majority civilians are observed by police. If this subset can be identified, or approximated through covariate adjustment, we can estimate the counterfactual police behavior that would have occurred had the civilian in question been replaced with a member of another racial group.

While our approach considers a valid counterfactual and isolates racial discrimination that occurs during police-civilian encounters, it necessarily mutes the influence of pre-encounter macroinstitutional factors, such as decisions to deploy more officers to communities of color. In keeping with the goals of prior studies in this

3 For example, we may observe that members of one racial group are stopped more often by police than members of another racial group. While this result shows disparate police behavior, it does not conclusively demonstrate that the difference is due to civilian race. It could simply be the case that members of the first group participate in criminal activity more often in public.

4 Note that while the unit of analysis is the police-civilian encounter, for the sake of brevity, we occasionally refer to “minority civilians” as shorthand for “police-civilian encounters with minority civilians” in subsequent discussion. Readers are cautioned to keep this distinction in mind.
area, our approach holds such contextual features constant, allowing us to ask whether an encounter would have unfolded differently had it involved a civilian of differing race. But even if no such difference exists within encounters, law enforcement strategies adopted before encounters occur could still produce racially biased policing. We caution readers to keep this scope condition in mind.

**PRIOR RESEARCH ON RACIAL BIAS IN POLICING**

Race-based selection into policing data has been previously noted, and some scholars have devised research designs in an attempt to sidestep this issue. Groger and Ridgeway (2006), for example, leverage the so-called “veil of darkness” strategy, comparing patterns in traffic stops that occur before and after sunset under the logic that the race of the driver is plausibly hidden to police officers after dark. In this way, the study aims to identify a sample of police–civilian interactions that were initiated in a race-blind manner. Similarly, West (2018) examines data on police responses to traffic incidents, arguing that whether a co-racial officer responds to a motorist’s unanticipated accident is as-if random. If the assumptions in these studies hold, concerns over race-based sample selection are greatly alleviated.

These attempts to mitigate race-based selection remain rare, as most empirical studies in this literature focus nearly exclusively on mitigating the more familiar problem of omitted variable bias. For example, Fryer (2019) (detailed below), a study of racial bias in police violence, estimates discrimination using data on police–civilian encounters via multivariate regressions that control for a host of observables relating to civil-civilian interactions. In a related article, the author asserts that “regression can recover the ‘race effect’ if race is ‘as good as randomly assigned,’ conditional on the covariates” (Fryer 2018, 2). Fryer (2019) claims to find evidence of bias in sublethal force but none in lethal encounters.

A related study, Johnson et al. (2019), attempts to estimate racial bias in police shootings. Examining only positive cases in which fatal shootings occurred, they find that the majority of shooting victims are white and conclude from this that no antiminority bias exists. Knox and Mummolo (2020) show that this conclusion rests on the erroneous assumption that police encounter minority and white civilians in equal number.

Prior work has also examined racial bias in traffic enforcement, such as Ridgeway (2006) which employs propensity score weighting when estimating racial bias in traffic stops in Oakland, CA. The analysis examines outcomes including citations, stop duration, and the decision to search cars. The study claims this reweighting strategy can recover “the causal effect of race” (9) on poststop outcomes. In general, the analysis finds little evidence of racial bias on most outcomes, with the exception of stop duration. Antonovics and Knight (2009) use data on traffic citations from the Boston Police Department to estimate the probability that a ticketed driver was searched, controlling for driver attributes such as age, race, and gender as well as neighborhood traits. They interpret the coefficient on an indicator of whether the officer and ticketed driver are of different races as an estimate of “racial profiling based on prejudice,” as opposed to statistical discrimination (167). The claim is implicitly causal: some share of searches among racially mismatched driver–officer pairs would not have occurred had the driver belonged to another racial group.

The above examples represent a mere fraction of a decades-long, multidisciplinary effort to quantify the degree to which police discriminate against civilians of color [see Atiba Goff and Kahn (2012), Fridell (2017), and Ridgeway and MacDonald (2010) for more extensive reviews of this empirical literature]. We highlight these specific examples because they all contain several common features that are central to our critique.

For one, these studies analyze data that fail to capture the unseen selective process through which police come to engage civilians, a process that prior work shows is function of civilian race (Gelman, Fagan, and Kiss 2007). In this way, these studies all fail to account for the impact of race on the composition of the sample under study. As we show below, failing to account for this undocumented first stage of the police–civilian interaction will lead to statistical bias, even if the goal is to estimate the effect of suspect race within the sample of individuals who appear in police data and, in many cases, even with a “complete” set of control variables that render civilian race as-if randomly assigned to police encounters.

Second, the aforementioned studies, despite making at least implicitly causal claims, leave ambiguous the precise quantity of interest—whether it be the average treatment effect (ATE) of race in all encounters; the average treatment effect among the subset of encounters appearing in police data because a stop was made (ATE$_{M=1}$), which differs tremendously from the ATE; or the markedly more restrictive and difficult-to-interpret controlled direct effect among the same subset (CDE$_{M=1}$, defined below). While studies commonly discuss omitted variable bias and attendant assumptions, they rarely discuss the additional assumptions necessary to identify specific causal quantities of interest. As a result, readers are unable to assess the adequacy of research designs and estimators, rendering the interpretation and policy relevance of much prior work unclear.

**Taste-Based versus Statistical Discrimination**

A closely related literature attempts to parse “taste-based discrimination” (racial animus) from so-called “statistical discrimination” (Arrow 1972, 1998; Becker 1971; Eberhardt et al. 2004; Phelps 1972) as mechanisms for racially biased policing, and instead focuses on recovering the causal effect of civilian race on police behavior. In this study, we do not attempt to disentangle these mechanisms, and we note that taste-based and statistical discrimination both
pose serious normative concerns. While statistical discrimination is sometimes viewed as more innocuous, it nonetheless constitutes racial profiling because officers detain civilians due to the perceived actions of their racial group, not their observed individual behavior. Thus, quantifying the causal effect of civilian race on police behavior—our task here—is imperative regardless of the mechanism that produces such an effect.

CLARIFYING THE EFFECT OF CIVILIAN RACE: NOTATION, ESTIMANDS, ASSUMPTIONS, AND EXISTING APPROACHES

Researchers and policymakers examining the effects of racially biased policing are nominally interested in the relationship between two variables: the race of the civilian involved in encounter \(i\), which we operationalize through their minority status \(D_i \in \{0, 1\}\), and consequent police behavior \(Y_i \in \{0, 1\}\). However, analyses of administrative data on police-civilian encounters inherently involve a mediating variable that may be affected by race: whether an individual is stopped by police, which we denote \(M_i\). The causal ordering of these variables is depicted in the directed acyclic graph (DAG) in Figure 1. We note that analysts often possess rich contextual information about the objective context of the encounter, such as its location and time, which may relate to all of the above. We denote these covariates collectively as \(X_i\). However, administrative data invariably fail to capture unobservable subjective aspects of the encounter, \(U_i\), such as an officer’s suspicion or sense of threat.

As a motivating example, we consider the challenge of estimating racial bias in police violence as recently attempted in Fryer (2019). We ground our analysis in the potential outcomes framework (Rubin 1974) often used in the study of causal mediation (Imai et al. 2011; Pearl 2001). The potential mediator \(M_i(d)\) represents whether encounter \(i\) would have resulted in a stop if the civilian were of race \(d\). Similarly, the potential outcome \(Y_i(d, m)\) represents whether force would have been used in encounter \(i\) if the civilian were of race \(d\) and the mediating variable were \(m\). The observed mediator and outcome can be written in terms of these potential values as \(\hat{M}_i = M_i(D_i) = \sum_d M_i(d) \mathbb{1}\{D_i = d\}\) and \(\hat{Y}_i = Y_i(D_i, \hat{M}_i(D_i)) = \sum_d \sum_m Y_i(d, m) \mathbb{1}\{D_i = d, \hat{M}_i = m\}\), respectively. For any individual encounter, the (unobservable) causal effect of civilian race is the difference in potential force if the civilian were a minority and stopped as if they were a minority, versus if they were white and stopped accordingly, \(Y_i(1, M_i(1)) - Y_i(0, M_i(0))\).

This notation implicitly makes the stable unit treatment value assumption (SUTVA, Rubin 1990). “Stability” is of particular note: this stipulates that finer racial gradations must not affect the way that officers behave, above and beyond any differences between the broad binary categories \(D_i = 0\) and \(D_i = 1\). SUTVA also requires that each encounter is unaffected by a civilian’s race in other encounters; this might be violated if, for example, groups of individuals are stopped simultaneously.

Traditionally, analysts use data on stopped individuals to study bias by computing the difference in violence rates between stopped minority and white civilians, while controlling for observable differences between these two sets of encounters. We term this the “naive estimator,” \(\hat{\Delta}\), and it can be written as follows:

\[
\hat{\Delta} = \bar{Y}_1|D_i = 1, M_i = 1 - \bar{Y}_0|D_i = 0, M_i = 1,
\]

where conditioning on possible treatment-outcome founders, \(X_i\), is left implicit. Assuming the analyst has correctly measured and specified all such founders, \(\hat{\Delta}\) may appear entirely reasonable at the first glance. However, without further assumptions, this quantity will have no causal interpretation so long as the treatment affects the mediator (i.e., civilian race affects whether officers detain a civilian). As we show below, this is because treated encounters (with minority civilians) that result in a stop (\(M_i = 1\)) will not be comparable to those with stopped control (majority) civilians. As a simple example, suppose officers exhibited racial bias as follows: they detain white civilians if they observe them committing a serious crime (such as assault, potentially warranting the use of force) but detain nonwhite civilians regardless of observed behavior. When this is true, comparing stopped white and nonwhite civilians amounts to comparing fundamentally different groups. The analyst will observe force used against a greater proportion of stopped white civilians because of the differential physical threat they pose to officers. Under the traditional approach, the analyst would naively conclude that anti-white bias exists, yielding an erroneous portrait of racial discrimination in the use of force.

To formalize the limitations of the naive estimator, we begin by partitioning the population into principal

---

Notes:

1. While some police records indicate whether a suspect was engaged in violent behavior, allowing the analyst to control for this particular factor, a host of similar concerns (e.g., time-varying officer suspicion) are unmeasured and thus cannot be controlled away.
strata with respect to the mediator (Frangakis and Rubin 2002; VanderWeele 2011). That is, we conceptualize police-civilian encounters in terms of four latent classes within which \( M_i(1) \) and \( M_i(0) \) are constant. The general approach of principal stratification has proven useful for clarifying and bounding quantities of interest in areas ranging from instrumental variables (Angrist, Imbens, and Rubin 1996; Balke and Pearl 1997) to the closely related “truncation by death” problem (Rubin 2000; Zhang and Rubin 2003).

These principal strata include “always-stop” encounters in which \( M_i(0) = M_i(1) = 1 \), as well as stops that discriminate against racial minorities (“racial stops”) in which \( M_i(1) = 1 \) but \( M_i(0) = 0 \). Always-stop encounters may be conceptualized as relatively severe scenarios, such as violent crimes in progress, in which officers have no choice but to intervene regardless of civilian race. In contrast, previous work has identified certain behaviors, such as “furtive movements” (Gelman, Fagan, and Kiss 2007; Goel, Rao, and Shroff 2016), that appear to be acted on selectively by officers based on the race of suspects. “Never-stop” encounters, where \( M_i(0) = M_i(1) = 0 \), are situations in which civilians appear in conspicuous and would not be stopped, regardless of race. There also may be anti-white racial encounters, in which \( M_i(1) = 0 \) but \( M_i(0) = 1 \), though we believe these to be rare to nonexistent (discussed further below). Figure 2 shows encounters appearing in police records (principal strata for which \( M(D_i) = 1 \) are not comparable across civilian races. Minority police-civilian encounters that result in a stop are a mixture of “always-stop” and “antiminority racial stop” encounters, while encounters with white civilians that result in a stop are a combination of “always-stop” and “anti-white racial stop” encounters. These are fundamentally different groups, and without further assumptions, comparisons of rates of violence between them using the naïve estimator will be statistically biased.

To state this more formally, note that the naïve estimator recovers the weighted combination of violence rates in observed principal strata:

\[
E[Y|D_i = 1, M_i = 1] - E[Y|D_i = 0, M_i = 1]
\]

\[
= E[Y(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1] Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1)
+ E[Y(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 0] Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1)
- E[Y(0,1)|D_i = 0, M_i(1) = 1, M_i(0) = 1] Pr(M_i(1) = 1|D_i = 0, M_i(0) = 1)
- E[Y(0,1)|D_i = 0, M_i(1) = 0, M_i(0) = 1] Pr(M_i(1) = 0|D_i = 0, M_i(0) = 1).
\]

(2)

In equation (2), the first term is the average rate of force applied during encounters with racial minorities of the always-stop stratum, while the second term deals with minorities in the anti-minority racial-stop stratum. The third and fourth terms are the average violence rates among white civilian encounters in the always-stop and anti-white racial stop strata. Importantly, principal strata are not fully observable without further assumptions, and they exist even after conditioning on \( X_2 \) for any particular minority stop, it is fundamentally impossible to know with certainty whether a white civilian would have been stopped in identical circumstances. In sum, the naïve estimator compares groups with different potential outcomes, and because these groups are unobservable, the resulting bias is difficult to address.

A central quantity of interest in the study of policing bias is the average treatment effect of race, \( ATE = E[Y(1,M_i(1)) - Y(0,M_i(0))] \) — the extent to which civilians of color face greater risk of police violence than white civilians because of their race. The ATE considers both reported and unreported encounters, and it captures two related phenomena: first, whether members of the minority are differentially stopped; and second, if they are differentially subject to violence. However, police administrative records contain data only on reported encounters, meaning that this quantity cannot be estimated solely with police administrative data without untenable assumptions. The ATE can be restated as follows:

\[
ATE = \sum_d \sum_m \sum_f \left( E[Y_{1}(1,M_{i}(1))|D_{j} = 0, M_{i}(0) = m'] \right) 
= \sum_d \sum_m \sum_f \left( E[Y(1,M_{i}(1))|D_{j} = 0, M_{i}(0) = m'] \right) \times Pr(D_{j} = d, M_{i}(1) = m, M_{i}(0) = m'),
\]

(3)

where the second line illustrates how it sums over the principal strata depicted in Figure 2, taking into account the number of minority and white civilians in each stratum (the probabilities) and the local average treatment effects for each group (the expectations). In Online Appendices A.1–A.4, we use these quantities to derive bias and nonparametric sharp bounds.

No data are available for “never-stop” encounters, those with \( M_i(1) = M_i(0) = 0 \). Moreover, racial-stop encounters, with \( M_i(1) = 1 \) and \( M_i(0) = 0 \), are only recorded for minority civilians. However, consistent with Nyhan, Skovron, and Titiunik (2017), we show in Online Appendix A.6 that the ATE can be point identified if researchers collected two additional numbers: the count of total minority and white encounters, within levels of covariates \( X \) where applicable — a point we discuss further in our recommendations for future research.6

6 Nyhan, Skovron, and Titiunik (2017) examines the problem of studying the effect of party identification on turnout using voter registration files, given the fact that party ID likely affects who registers to vote. In an approach that is equivalent to our Proposition 2, the study uses registration rates of the treated and control voting-age populations to bound the ATE given the effect of party ID on registration. This option is not available in practice here since no data sets contain information on unreported encounter rates, or even their order of magnitude. As a result, analysts in this literature focus almost exclusively on the \( ATE_{M_{t-1}} \), which we examine with a different approach here. While some work in policing uses population figures as proxies for these encounter rates, we are skeptical of this approach, as police frequently stop civilians who reside in other jurisdictions.
Relatedly, analysts may seek to causally attribute the number of minority stops in which force would not have been used if the individual in question had been white (Yamamoto 2012). This value is proportional to the conditional average treatment effect among the treated (i.e., minority) stops, which can be written as follows:

\[
\operatorname{ATT}_{M=1} = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1].
\]

While the average treatment effects are of obvious policy importance, they are not the only quantity that researchers might seek to estimate. A closely related estimand is the controlled direct effect among the subset of reported encounters, \( \operatorname{CDE}_{M=1} = \mathbb{E}[Y_i(1, 1)|M_i = 1] - \mathbb{E}[Y_i(0, 1)|M_i = 1] \). This estimand differs from the \( \operatorname{ATE}_{M=1} \) in its conceptual approach to racially discriminatory stops. Where the \( \operatorname{ATE}_{M=1} \) asks whether a stop would have occurred at all if the individual were of differing race, the \( \operatorname{CDE}_{M=1} \) seeks to quantify what would have happened if the officer was forced to stop them anyway, perhaps against the officer’s will. In practice, the difference is one of interpretation—regardless of the target quantity, existing work in this domain is based on the naïve difference in reported outcomes, and the question lies in the interpretation of estimated results. We note that causal estimands in the literature are often left undefined, making it difficult to assess whether published results are intended to correspond to the \( \operatorname{ATE}_{M=1} \) or \( \operatorname{CDE}_{M=1} \) (e.g., Goel, Rao, and Shroff 2016; Simoiu, Corbett-Davies, and Goel 2017). In Online Appendix A.3, we discuss the \( \operatorname{CDE}_{M=1} \) at length. We show that it cannot be recovered in this setting unless analysts make the untenable assumption that no mediator-outcome confounding exists (Assumption 5, below). We refer readers to the Online Appendix for further details and focus on recovery of average treatment effects here.

**Necessary Assumptions**

In this subsection, we describe a number of statistical assumptions that the analyst must make for a causal study of racially biased policing when only administrative data on police-civilian interactions is available. Without these assumptions, causal quantities of interest in this substantive area cannot be identified in data.

**Assumption 1** (Mandatory Reporting). \( Y_i(d, 0) = 0 \) for all \( i \) and for \( d \in [0, 1] \).

We assume all encounters that escalate to the use of force also trigger a reporting requirement and are, therefore, observed in administrative data. Though there exist wide variability in data recording practices across jurisdictions, this assumption is plausible in the study of many major police departments. For example, New York Police Department (NYPD) officers are required to report a number of variables, including the specific type of force used, following each “stop, question, and frisk” encounter. Based on these and other reports, the NYPD releases detailed annual use-of-force reports (NYPD 2017). The completeness of

Because “never-stop” encounters are unobserved in current data sources, researchers seeking to understand the role of race in police behavior have, at least implicitly, focused on more narrowly defined estimands. Studies commonly restrict analysis to the subset of reported encounters, \( \operatorname{ATE}_{M-1} \). In contrast to the \( \operatorname{ATE} \), this estimand is by definition not concerned with unreported white encounters that would have escalated to a stop if the involved civilian was a minority. (The same is true for unreported black encounters that would have escalated if the involved civilian was white, to the extent that this group exists.) Formally, this quantity is given by the following equation:

\[
\operatorname{ATE}_{M=1} = \mathbb{E}[Y_i(1, M_i(1))|M_i = 1] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1].
\]

\(^7\) For example, Fryer (2018) notes that his analysis of police use of force is estimating the effect of suspect race “conditional on an interaction,” with police (4), rather than seeking its average treatment effect in the population.
these reports with respect to fatalities is informally enforced by standard journalistic practices which place high emphasis on documenting violent incidents (Iyengar 1994). Lesser forms of force are more likely to go unreported, to be sure, but the ubiquity of surveillance cameras, cell phone cameras, and media interest in police brutality makes unobserved uses of force increasingly unlikely (Fisher and Hermann 2015). We note that this assumption is implicit in all analyses of police use of force that rely on administrative data.

**Assumption 2 (Mediator Monotonicity).** $M_i(1) \geq M_i(0)$ for all $i$.

This assumption allows that there may be encounters in which minorities would be stopped ($M_i(1) = 1$) but whites would not ($M_i(0) = 0$), perhaps because officers racially discriminate in applying differential thresholds of “reasonable suspicion.” However, we assume that the reverse is never true: white civilians are never stopped in circumstances when their minority counterparts would be allowed to pass. This is clearly a stylized representation of a complex reality, and it would be violated if minority officers discriminate against white civilians. A violation could also occur if white civilians were more likely to be stopped by police because they appeared out of place in a predominantly black neighborhood, perhaps under the assumption that they were there to buy drugs (Gelman, Fagan, and Kiss 2007, 822).

These are rare occurrences, and a robustness check in Online Appendix B.3, our reanalysis of Fryer (2019) after dropping all stops based on suspicion of a drug transaction, shows substantively similar results.

**Assumption 3 (Relative Nonseverity of Racial Stops).** $E[Y_i(d, m)|D_i = d', M_i(1) = 1, M_i(0) = 1, X_i = x] \geq E[Y_i(d, m)|D_i = d', M_i(1) = 1, M_i(0) = 0, X_i = x]$.

We theorize that for encounters during criminal events severe enough to warrant stopping a police officer regardless of race (i.e., “severe” or “always-stop” encounters), the use of force is as or more likely to occur than during encounters in which police have more discretion over whether to stop an individual (i.e., those in which racial discrimination in stopping can occur) in expectation. We regard this assumption, which compares violence rates within encounters that hold civilian race fixed, as highly plausible. As one hypothetical example, this assumption would imply that police are as or more likely to use force against a white civilian observed committing assault than a white civilian observedjaywalking, on average.

**Assumption 4 (Treatment Ignorability).**

(a) With respect to potential mediator $M_i(d) \perp D_i|X_r$

(b) With respect to potential outcomes: $Y_i(d, m) \perp D_i|M_i(0) = m', M_i(1) = m^*, X_r$

This states that conditional on $X_r$, civilian race is “as good as randomly assigned” to encounters, and officers encounter minority civilians in circumstances that are objectively no different from white encounters. Part 4(a) stipulates that the observed covariates $X$ include the confounder $W$ in Figure 3(a). This assumption, while strong, has become more plausible in recent years as administrative data sets have come to include a host of encounter attributes that might largely capture features observable to police which correlate with suspect race and the potential for force. However, we note that this cannot be tested, even indirectly, without data on nonstopping individuals. This assumption would be violated if neighborhoods with high shares of minority residents were more heavily policed and the analyst failed to adjust for neighborhood, for example, using fixed effects. Part 4(b) implies that, for example, if police were more heavily armed during minority-neighborhood patrols and, hence, more likely to deploy force — represented by $V$ in Figure 3(b) — then $V$ must be included in $X$. Without Assumption 4, the range of possible racial effects is so wide as to be uninformative. We also note that every study claiming to estimate racial discrimination using similar data makes this assumption, often implicitly. Our aim in this study is not to assert the plausibility of treatment ignorability, but rather to clarify that deep problems remain even if this well-known issue is somehow solved.

**Strong Assumptions**

We now discuss further assumptions that are often left implicit in empirical studies of racially biased policing and that are implausible in many settings. We illustrate these scenarios graphically in Figure 3.

**Assumption 5 (Mediator ignorability).** $Y_i(d, m) \perp M_i(0)|D_i = d, M_i(1) = 1, X_r$

This is related to but dramatically stronger than Assumption 3, which merely requires that always-stop encounters are at least as severe in terms of observed criminal behavior. In contrast, for Assumption 5 to hold, violence rates in always-stop encounters must be identical to those in observationally equivalent racial stops. We find mediator ignorability to be highly implausible in the context of policing. Subjective factors such as an officer’s suspicion and sense of threat—depicted as $U$ in Figure 3(c)—can not only lead to investigation (stopping) but also a heightened willingness to use force. These mediator-outcome confounders must be captured in $X$ for this assumption to hold, but they are notoriously difficult to capture in officers’ self-reported accounts. Even when proxies based on qualitative officer narratives are available, strong legal incentives exist for distortion. Moreover, analysts must be sure to condition on all variables related to officer mindset that are causally upstream of stops, while taking care not to induce bias by conditioning on any that are downstream.

Below, we demonstrate that every analysis estimating a racial effect using only data on stopped individuals implicitly makes Assumption 5. We further note that Assumptions 4(a), 4(b), and 5 are jointly covered by the
slightly stronger assumption of sequential ignorability (Imai et al. 2011).

**Assumption 6** (No Racial Stops), $M_i(0) = M_i(1)|M_i = 1$.

In Figure 3, this amounts to assuming away the arrow between $D$ and $M$. Equivalently, this assumption states that all reported encounters were of the always-stop kind, or that there is no racial discrimination in stops. We show below that this assumption is implicitly made by all studies claiming to identify the average treatment effect of race, conditional on a reported interaction. Naturally, when there is no variation in $M_i(0)$, then this variable is ignorable and Assumption 5 is also satisfied.

However, in view of an overwhelming body of qualitative evidence and consistently massive quantitative differences in racial detainment rates across numerous policing domains, we find racial bias in police stops too plausible to dismiss by assumption (Alexander 2010; Baumgartner et al. 2017; Glaser 2014; Goel, Rao, and Shroff 2016; Lerman and Weaver 2014). A raft of studies have also found that racial disparities persist even after leading candidate omitted variables, such as differential criminal activity across racial groups, are accounted for (Gelman, Fagan, and Kiss 2007). While such patterns are not proof of a causal relationship, we consider the possibility that police exhibit anti-minority bias when engaging civilians strong enough to merit a careful consideration of the implications of that bias for the validity of studies of racially biased policing.

**Bias in the Naïve Estimator**

In this section, we clear up several misunderstandings about the conventional estimator, which compares reported minority stops to reported white stops (with or without covariates). First, we show that when there is any racial discrimination in detainment, selection on stops introduces unavoidable statistical bias in estimating the $\text{ATE}_{M=1}$, even when a perfect set of observed covariates renders race ignorable with respect to the potential mediator and outcomes. These results directly contradict prior assertions that “linear regression can recover the ‘race effect’ if race is ‘as good as randomly assigned,’ conditional on the covariates” (Fryer 2018, 2). The issue is not one of omitted variables, but rather posttreatment conditioning. Second, we clarify an important open question about the nature of this bias. Fryer (2018) comments in the context of selection into arrest data that, “It is unclear how to estimate the extent of such bias or how to address it statistically” (5). Here, we derive the exact form of this bias for the $\text{ATE}_{M=1}$ and the $\text{ATT}_{M=1}$; Online Appendix A.3 does the same for the $\text{CDE}_{M=1}$. We show that the bias is always negative, resulting in naive estimates that downplay the extent of racially discriminatory police violence. Below, we develop informative nonparametric sharp bounds that adjust the naïve estimates for the range of all possible selection bias.

Prior work on race and policing uses estimators that compare average reported outcomes in majority encounters to those in minority encounters. For simplicity of exposition, we present the special no-covariate case; Appendices A.1–A.3 derive the bias of the naïve estimator with covariate adjustment. We first refer readers to equation (1), which expresses the naïve estimator, $\Delta$, in terms of stratum mean potential outcomes. We demonstrate that this commonly used analytic approach fails to recover any quantity of interest under plausible assumptions. We first show that it is biased for the $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ unless Assumption 6 is true, and there are no racial stops. In Online Appendix A.3, we show it is also biased for the $\text{CDE}_{M=1}$ unless Assumption 5 holds—that is, always-stop encounters are identical in violence rates to racially discriminatory stops. As a result, the observed difference in means fails to recover any known causal quantity without additional, and highly implausible, assumptions.

In Online Appendix A.1, we derive the bias of $\Delta$ when it is used to estimate $\text{ATE}_{M=1}$ under the relatively plausible Assumptions 1–4. This bias can be written as follows:

\[
\Delta = \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1 \right] \\
- \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 0 \right] \\
\times \Pr(D_i = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
- \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i(1,1) | M_i(1) = 1, M_i(0) = 1 \right] \\
- \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i(1,1) | M_i(1) = 1, M_i(0) = 0 \right] \\
\times \Pr(D_i = 0 | D_i = 1, M_i = 1). 
\] (6)
We offer several comments on equation (6). The bias term is guaranteed to be negative, even with a perfect set of controls that render $D_i$ ignorable, as long as there exist any racially discriminatory stops of minority civilians (or in an empirically falsified edge case). The first term in the bias expression relates to heterogeneity in the average treatment effect, or the extent to which $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$ differs in expectation between always-stop and racial-stop encounters—respectively, those with $M_i(1) = M_i(0) = 1$ and $M_i(0) < M_i(1)$. Bias arises because in the latter type of encounter, a white civilian would never have been detained in the first place, and hence force would never have been used—that is, $E[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0] = 0$. Estimating the average potential outcomes of this group using stopped white civilians introduces unavoidable bias that the analyst cannot hope to eliminate simply by adding additional covariates to the estimating model. The second term is related to the difference in baseline violence rates between always-stop encounters and racially discriminatory stops; this term also vanishes if there are no racial stops.

Can the naive estimator be rehabilitated by simply redefining the quantity of interest? In Online Appendices A.2–A.3, we show that the answer is no. The structure of the bias when $\Delta$ is used to estimate the ATT$_{M=1}$ is simpler but leads to substantively identical conclusions: the naive estimator is biased unless there are no racial stops. We show that bias for the ATT$_{M=1}$ is given by $E[\Delta] - \text{ATT}_{M=1} = -E[Y_i(0, 0)|M_i(1) = 1, M_i(0) = 1] \Pr(M_i(0) = 0|M_i(1) = 1)$. While the identifying assumptions for the CDE$_{M=1}$ are slightly weaker, they are nonetheless wholly implausible. The sign of this bias for the ATT$_{M=1}$ and CDE$_{M=1}$ can also be shown to be negative under Assumption 1–4, except in the implausible edge cases described in the Online Appendix. Thus, regardless of the target quantity, the use of the observed difference in means will underestimate the rate of racially discriminatory police violence. In addition, we emphasize that these derivations show that statistical bias remains even after assuming a “complete” set of control variables that renders race ignorable. Posttreatment conditioning induces bias unless additional assumptions hold.

POTENTIAL SOLUTIONS

How should the analyst proceed in light of these results? We propose two approaches that eliminate the highly implausible assumptions outlined in the “Strong Assumptions” section, which are unstated but implicit in prior work. We caution that these solutions still rely on the weaker assumptions described in the “Necessary Assumptions” section, although we argue that these are often plausible in light of insights from extensive research on policing. Reasonable people can disagree on the plausibility of various assumptions, but by stating them explicitly, we seek to advance empirical work in an area which, at present, largely ignores such issues altogether.

In the first approach, we derive nonparametric sharp bounds representing the tightest possible range of causal effects that are consistent with the reported data (Manski 1995). Again, for simplicity, we begin by presenting bounds for the case in which treatment is unconditionally ignorable. To incorporate covariates, Online Appendix A.4 then describes a more general formulation in which bounds are computed within levels of $X$, without functional form assumptions, and reaggregated; this latter formulation is also applicable when a correctly specified regression is used. Both cases are demonstrated in a reanalysis of Fryer (2019) below.

A key limitation of the first proposed solution is that all quantities of interest remain only partially identified. This is fundamentally a consequence of selection into police administrative records; point identification simply cannot be achieved without either implausible assumptions or additional data. To this end, we outline an alternative approach that incorporates limited information about the missing encounters (those that do not result in a stop). We show that with additional data—which in some cases are already being collected by agencies—the prevalence of racially discriminatory stops and most racial effects of interest can be point identified. Following our applied example, we describe a feasible research design based on this approach in detail.

Bounds on Effect of Race

Here, we derive large-sample nonparametric sharp bounds on the ATE$_{M=1}$ and ATT$_{M=1}$, focusing first on the case in which Assumption 4 (treatment ignorability) holds without conditioning on further covariates. Proposition 1 quantifies and corrects for the range of possible bias induced by posttreatment conditioning, producing an informative interval of possible joint values for (1) the partially identified ATE$_{M=1}$ and (2) the proportion of racial stops among reported minority encounters, $\rho = \Pr(M_i(0) = 0|D_i = 1, M_i = 1)$. As equation (6) suggests, when there is no racial bias in police stops ($\rho = 0$), these bounds collapse on the observed difference in means. We further demonstrate in Figure 4 that these bounds are highly informative when $\rho$ is known or can be credibly estimated from supplemental data. When the prevalence of racially discriminatory detainment is unknown but a plausible range can be inferred from prior work, Figure 4 (discussed below) illustrates how this value can be used to assess the behavior of the bounds much like a sensitivity parameter.
**Proposition 1** (Nonparametric Sharp Bounds on $\text{ATE}_M$). When $D_i$ is ignorable, nonparametric sharp bounds on ($\text{ATE}_M$) under Assumptions 1–4 are jointly given by

$$
\frac{\Delta}{1} + \rho \mathbb{E}[Y_i | D_i = 1, M_i = 1] - \left( \mathbb{E}[Y_i | D_i = 0, M_i = 1] - \mathbb{E}[Y_i | D_i = 0, M_i = 1] \right)
\leq \text{ATE}_M \leq 
\frac{\Delta}{1} + \rho \mathbb{E}[Y_i | D_i = 1, M_i = 1] - \left( \mathbb{E}[Y_i | D_i = 0, M_i = 1] - \mathbb{E}[Y_i | D_i = 0, M_i = 1] \right)
\times \frac{1}{1} - (1 - \mathbb{E}[D_i | M_i = 1]) + \rho \mathbb{E}[Y_i | D_i = 0, M_i = 1] - \left( \mathbb{E}[Y_i | D_i = 0, M_i = 1] - \mathbb{E}[Y_i | D_i = 0, M_i = 1] \right)
$$

where $\Delta = Y_i | D_i = 1, M_i = 1 - Y_i | D_i = 0, M_i = 1$, and the (ATT$_M = \rho$) must similarly satisfy

$$
\text{ATT}_M = \frac{\Delta}{1} + \rho \mathbb{E}[Y_i | D_i = 1, M_i = 1]
$$

To derive Proposition 1, we reformulate the bias in terms of the unobserved joint distribution of (1) the use of force in minority encounters and (2) whether a minority stop was racially discriminatory. Following Knox et al. (2019), we then use Assumptions 1–4 and the Fréchet inequalities, in conjunction with the observed margins, to place sharp bounds on this joint distribution. These then imply sharp bounds on the $\text{ATE}_M$. A detailed proof is given in Online Appendix A.4 for the more general case in which $D_i$ is ignorable only after conditioning on prestop covariates. In this case, the local average treatment effect, $\text{ATE}_{M=1,x}$, is first bounded by applying Proposition 1 within levels of $X$ to obtain local bounds, $[\text{ATE}_{M=1,x} - \text{ATE}_{M=1,x}]$. These are then straightforwardly reaggregated to obtain bounds on the conditional treatment effect among stops, $\sum \text{ATE}_{M=1,x} \text{Pr}(X_i = x | M_i = 1)$, $\sum \text{ATE}_{M=1,x} \text{Pr}(X_i = x | M_i = 1)$. In Online Appendix A.5, we outline a Monte Carlo procedure for constructing confidence intervals that asymptotically contain both the true lower and upper bounds endpoints with probability $1 - \alpha$.

We note that the proportion of racially discriminatory stops may vary with $X$. However, when using these bounds as a sensitivity analysis, we suggest using the simplifying approximation of a constant $\rho$. This is because without additional data beyond civilian race, the use of force, or even prestop covariates, police administrative records alone are virtually uninformative about the range of $\rho$: any value in $[0, 1]$ could produce the observed data, although Proposition 1 shows that

---

**FIGURE 4. Bounds for Racially Discriminatory Use of Force, any Severity**

![Graph showing bounds for racially discriminatory use of force for black and Hispanic civilians.](https://example.com/figure4)

Notes: These plots present the $\text{ATE}_M$ for excess racial force, scaled by the number of stops (number of minority stops) to obtain the total number of civilians affected. The left panels consider the difference in the use of force if black civilians were substituted into each encounter of any race (each black encounter), versus white civilians; the right panels show the same quantities for Hispanic civilians. Blue points (error bars) denote the naive estimator (95% confidence intervals), which, conditional on the typical selection-on-observables assumption, is unbiased for the $\text{ATE}_M$. The horizontal line corresponds to an estimate of the proportion of discriminatory stops from Gelman, Fagan, and Kiss (2007). The top (bottom) panels present bounds based on a model with no controls (the main specification, adjusting for a wide range of covariates).
each possible $\rho$ value has differing implications for the set of possible racial effects.

**Point Identification of the ATE Given Additional Data**

The ATE is point identified with the collection of only two additional numbers—the count of total minority and white encounters, within levels of $X$ where applicable. Below, we propose an alternative design in which these data are collected from passive instruments such as traffic cameras or police body-worn cameras. Where such a design is infeasible (e.g., where traffic cameras cover only a subset of the jurisdiction under study), point identification can also be achieved by linking incomplete data on both reported and unreported encounters to police administrative records under mild assumptions.

**Proposition 2** (Point Identification of ATE). Under Assumptions 1–4, the ATE is identified by a weighted combination of the observed racial means,

$$E[Y|D_i = 1, M_i(D_i)] = \Pr(M_i = 1|D_i = 1)$$

$$-E[Y|D_i = 0, M_i(D_i)] = \Pr(M_i = 1|D_i = 0).$$

Intuitively, the proof breaks the ATE into the size-weighted sum of principal effects among always-stop and racial-stop encounters (the principal effect in never-stop encounters is known to be zero). Crucially, the additional data on nonstops allows the researcher to construct a contingency table representing the joint distribution of race and detainment. As part of the proof in Online Appendix A.6, we show that this can be used to straightforwardly recover the size of each principal stratum under Assumptions 2 and 4(a). However, it remains impossible to determine whether any individual stop was racially discriminatory.

When total encounter numbers are unknown, this joint distribution can nonetheless be estimated by attempting to link a representative sample of all encounters (e.g., using timestamps from traffic cameras) against administrative records (e.g., license plate databases); those that are unlinkable can be presumed unreported. After recovering principal strata sizes, we then proceed by noting that minority outcomes in reported administrative data are in fact a mixture of $Y(1, M_i(1))$ from both always-stop and racial-stop strata in precisely the required proportions; that reported white outcomes correspond to $Y(0, M_i(0))$ from the always-stop stratum; and that $Y(0, M_i(0))$ is known to be zero among the racial-stop stratum under Assumption 1. From this, the ATE can then be reconstructed.

**REANALYSIS OF FRYER (2019)**

We have shown that the standard approach to estimating racial bias in police data will always underestimate its degree, so long as police discriminate against minorities when choosing whom to investigate. To explore the magnitude of this statistical bias in an applied setting, we replicate and extend a section of Fryer (2019) which reports estimates of racial discrimination in the application of sublethal force using the NYPD’s “Stop, Question and Frisk” (SQF) database (2003–13). The NYPD data contain roughly 5 million records of pedestrian stops, the vast majority of which are of nonwhite suspects. The data record the use of varying levels of force, including laying hands on a suspect, handcuffing a suspect, pointing a weapon at a suspect, and pepper spraying a suspect, among others. The original analysis in Fryer (2019) utilized the simple naïve approach of equation (1) to predict the severity of force applied by police, as well as covariate-adjusted naïve models analogous to those we consider in Appendices A.1–A.3. Specifically, the study presented a logistic regression of police force on suspect race, along with additional specifications that added a host of control variables such as precinct fixed effects, to render the ignorability assumptions more plausible. We reproduce two of these models—the baseline specification including only racial group indicators, along with the richer “main specification” (21)—to estimate the conditional expectations in Proposition 1. For comparability to the original analysis, we take these models at face value, setting aside issues of potential model misspecification and the ignorability of civilian race.

One analysis in Fryer (2019) considered the use of any force against a suspect, while subsequent analyses examined force exceeding various severity thresholds, such as a binary outcome for “at least use of handcuffs.” Using the coding rules and estimation procedures in Fryer (2019), we were able to closely replicate the published results. However, in doing so, we discovered this procedure involved an unconventional and inadvisable step in which all observations with nonzero force below the threshold of interest were dropped—a severe case of selection on the dependent variable. In the most extreme case, in the analysis of police baton and pepper spray use, this resulted in the discarding of all encounters in which only lower levels of force were used, a set that comprised 21.5% of all observations and 99.8% of all uses of force. To present the most defensible results possible, for these outcomes, we depart from the analysis in Fryer (2019) and revise the procedure so that all encounters with a level of force at or above a given threshold are assigned an outcome of 1 (as before) and all other encounters are assigned a value of 0 (including those with lower levels of force, which are now retained). Section B.1 in the Online Appendix contains an extended discussion of the issue.

---

11 Because the replication material for Fryer (2019) was not posted at the time of analysis, these data were obtained directly from https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page.

12 The main specification in Fryer (2019) consists of a logistic regression of a force outcome on race dummies plus controls for gender, a quadratic in age, whether the stop was indoors or outdoors, whether the stop took place during the daytime, whether the stop took place in a high crime area, during a high crime time, or in a high crime area at a high crime time, whether the officer was in uniform, civilian ID type, whether others were stopped during the interaction, controls for civilian behavior, and precinct and year fixed effects. Fryer (2019) also notes that “missing indicators for all variables” are included as covariates. We omit these indicators as it was unclear how they were coded. See Figure 1 caption in Fryer (2019).
a comparison of the original, replicated, and corrected results; and a demonstration of the serious implications for statistical significance of the original estimates.

Based on the discussion in both Fryer (2018) and Fryer (2019), we interpret the published results as estimates of the $\text{ATE}_{M=1}$: “the difference in $Y$ that can be attributed to an individual’s race,” (Fryer 2018, 2), conditional on a recorded interaction with police (i.e., conditional on $M_i = 1$). We note that of the other quantities considered in this study, the unconditional $\text{ATE}$ cannot be estimated without information on unreported encounters, and the $\text{CDE}_{M=1}$ cannot be computed without strong assumptions about potential outcomes that can never be realized in observational settings. For these reasons, we focus on the $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ in this reanalysis.

Figure 4 depicts bounds on the $\text{ATE}_{M=1}$ when the binary outcome is any use of force, including the lowest recorded value of physically handling a civilian. Importantly, this specific outcome is unaffected by the outcome coding issue discussed above. (In Figures B.2 and B.3, we present additional bounds for varying force thresholds, up to whether a baton or pepper spray was used.) The plots also display estimates of the bias-corrected $\text{ATT}_{M=1}$ (dashed lines). As the plots show, the range of possible $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ values varies strongly with the severity of discrimination in stops.

In equation (6), we demonstrated that the use of the naïve estimator implied the substantively implausible assumption that police never discriminate in stops (i.e., $\rho = 0$). Similarly, contextual information also suggests that some depicted values of $\rho$ are implausibly large. To understand the range of empirically plausible values, we turn to two prior studies that use very different analytical approaches, gives us some empirical justification for narrowing the range of plausible racial effects in the use-of-force analysis. We note that the research design presented in the “Recommendations for Future Research” section below offers an alternative approach for obtaining better estimates of racially discriminatory stopping.

Figure 4 demonstrates that strong negative bias in the naïve estimator paints a wildly misleading portrait of police use of force. We turn first to estimates of the $\text{ATE}_{M=1}$ using the main specification, which adjusts for a battery of covariates. The naïve estimator (which assumes no racial bias in police stops) suggests that encounters with black (Hispanic) suspects are predicted to exhibit an additional 3.9 (0.4) instances of arrest rates by race, Gelman, Fagan, and Kiss (2007) as conservative.

Goel, Rao, and Shroff (2016) take an entirely different tack based on a comparison of “hit rates,” or the share of stops that produced evidence of the suspected crime for which the civilian was detained—a variant of an “outcome test” for discrimination (Anwar and Fang 2006; Knowles, Perisco, and Todt 2001). Using a flexible logistic regression to adjust for a vast array of indicators visible to officers prestop, the study shows that white hit rates exceeded those of “similarly situated” black civilians. We show in our Online Appendix A.7 that the difference in hit rates implies a minimum proportion of racial stops and, therefore, also implies a conservative estimate of $\rho$. The corresponding values of $\rho$ from these two studies are 0.32 and a lower bound of 0.34, respectively, when considering black civilians. While any estimate of this difficult-to-measure quantity from police data is sure to be imperfect, the fact that two independent estimates of racial bias in stopping so closely comport with one another, despite using wholly different analytical approaches, gives us some empirical justification for narrowing the range of plausible racial effects in the use-of-force analysis. We note that in Proposition 1 we consider binary minority status, whereas the specifications in Fryer (2019) take civilian race as a categorical variable. (However, only two races are considered for any particular $\text{ATE}_{M=1}$: estimate: black versus white, or Hispanic versus white. To accommodate this, in reported black $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ results, we use a slight generalization in which white civilian encounters are represented with $D_i = 0$, black encounters with $D_i = 1$, and subsequent minority groups with $D_i = 2, 3$, and so on. Proposition 1 and its covariate-adjusted counterpart in Online Appendix A can then be applied directly. The chief implication of this formulation is (1) a different average value for $Y(d, 1)$ is estimated for each minority group, and (2) that all minority groups are implicitly assumed to be racially stopped at the same rate, although this can easily be relaxed. (The same procedure is applied when the minority group of interest is Hispanic civilians, after setting the Hispanic indicator to $D_i = 1$.) To assess whether results were affected by this, in Online Appendix B.4, we conduct two additional analyses after first subsetting to black and white encounters, and Hispanic and white encounters, respectively. As the results makes clear, conclusions are virtually identical apart from differences that stem from the size of the subsampled data.

Note that we treat stops in which “other” was denoted as the use of force category as zero force, since the vast majority of these cases did not even not involve officers even laying hands on suspects.

13 We note that in Proposition 1 we consider binary minority status, whereas the specifications in Fryer (2019) take civilian race as a categorical variable. (However, only two races are considered for any particular $\text{ATE}_{M=1}$: estimate: black versus white, or Hispanic versus white. To accommodate this, in reported black $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ results, we use a slight generalization in which white civilian encounters are represented with $D_i = 0$, black encounters with $D_i = 1$, and subsequent minority groups with $D_i = 2, 3$, and so on. Proposition 1 and its covariate-adjusted counterpart in Online Appendix A can then be applied directly. The chief implication of this formulation is (1) a different average value for $Y(d, 1)$ is estimated for each minority group, and (2) that all minority groups are implicitly assumed to be racially stopped at the same rate, although this can easily be relaxed. (The same procedure is applied when the minority group of interest is Hispanic civilians, after setting the Hispanic indicator to $D_i = 1$.) To assess whether results were affected by this, in Online Appendix B.4, we conduct two additional analyses after first subsetting to black and white encounters, and Hispanic and white encounters, respectively. As the results makes clear, conclusions are virtually identical apart from differences that stem from the size of the subsampled data.

14 Note that we treat stops in which “other” was denoted as the use of force category as zero force, since the vast majority of these cases did not even not involve officers even laying hands on suspects.

15 Based on SQF data from 1998–99, Gelman, Fagan, and Kiss (2007) fit hierarchical Poisson models for the number of stops (by suspected crime, precinct, and race) per arrest in the previous year, which they model as $\text{e}^{\rho_{\text{race}}}$ within groups of stops defined by the suspected charges (violent crimes, weapons crimes, property crimes, and drug crimes) and precinct racial composition (<10%, 10–40%, and >40% black). Within each group, the excess black stopping rate is then given by $1 – e^{\rho_{\text{race}}}$, which we approximate the size of each group by multiplying the reported marginal probabilities of stop types (25%, 44%, 20%, and 11%, respectively) and composition groups (“each…” represents roughly 1/3 of the precincts”), since the joint distribution is not reported. The $\rho = 0.32$ estimate is then produced by taking the size-weighted average of subgroup excess black stopping rates. The corresponding estimate of $\rho$ for Hispanic civilians implied by Gelman, Fagan, and Kiss (2007) is slightly higher, at 0.35.
handcuffing per 1,000 encounters, compared with the same encounters had they involved white civilians. We then employ the most conservative racial stopping estimate, denoted by the vertical line in the figure, to generate bounds on the true race effect. Our bias-corrected results show the true effect is at least as high as 15.5 (13.0)—meaning that the conventional approach underestimates discriminatory force by a factor of at least 4 (32).

To characterize bias in estimates of the $\text{ATT}_{M=1}$, we again use the conservative racial stopping estimate from Gelman, Fagan, and Kiss (2007) to correct the naive estimate. Again, the naive approach substantially understates racially discriminatory police violence, suggesting that there were 75,000 instances in which police laid hands on black and Hispanic civilians, but would not have done so had those individuals been white. Our bias-corrected estimate shows the true number is approximately 307,000, meaning the naive approach masks 232,000 such incidents. Similarly, the naive approach indicates roughly 3,400 racially discriminatory instances in which officers pointed a weapon at a black or Hispanic civilian, whereas the bias-corrected $\text{ATT}_{M=1}$ shows the true number is almost five times as large.

To see how this statistical bias affects estimates for different levels of force, Table 1 presents naive estimates alongside $\text{ATE}_{M=1}$ bounds for excess force per 1,000 black and Hispanic encounters across the full spectrum of police actions—ranging from physical handling of a civilian to the use of pepper spray or a baton—again using the conservative racial-stop estimate from Gelman, Fagan, and Kiss (2007) to apply our bias correction. The results again show that the traditional approach substantially understates the degree of racial bias in police use of force. Our results also include numerous cases in which downward bias produces the illusion of no race effect. For example, while the approach in Fryer (2019) implies a statistically insignificant 2.4 instances per 1,000 encounters of pushing Hispanic suspects to a wall due to suspect race, our revised estimate shows the true number is at least 26—eleven times larger. We can also quantify the number of masked instances of racially discriminatory uses of force as a percentage of all uses of force against minorities, displayed in Figure 5. In the period we examine, black and Hispanic civilians experienced force at the hands of police 779,894 times. Using the approach in Fryer (2019), one would conclude that about 10% would not have occurred had those civilians been white. Using our bias-corrected approach, we find that in fact 39% were discriminatory. These underestimates persist across all force threshold analyses.17

**RECOMMENDATIONS FOR FUTURE RESEARCH**

The analysis above clarifies whether and when estimates of racial bias in police behavior identify causal quantities, shedding light on how traditional estimation approaches that fail to account for posttreatment conditioning can inadvertently mask racially biased policing. Our results suggest the body of evidence on this

17 These estimates were generated by computing the $\text{ATT}_{M=1}$ with covariate adjustment.
topic that relies on police administrative data may be largely uninformative or even misleading. While our bias-correction and bounding techniques are an improvement, they still rely on assumptions that many analysts may not be willing to entertain. Some of these assumptions, such as conditional treatment ignorability, are unavoidable. But others can be sidestepped or weakened through the use of research designs that preempt the problem of posttreatment conditioning. In what follows, we detail a feasible research design that addresses these concerns.

To estimate the effect of suspect race on poststop police behavior while avoiding the concerns outlined above, we describe a feasible study of police-civilian interactions during traffic stops. A key advantage of traffic studies is that much of the data needed to improve research are already collected passively by law enforcement agencies across the United States in an automated fashion via highway cameras. We note that before the advent of this technology, data on unreported police-civilian interactions had to be manually collected by researchers accompanying patrol officers on their shifts (Allen 1982; Smith, Visher, and Davidson 1984), a labor-intensive strategy highly vulnerable to researcher demand effects (Orne 1962).

Recall that a key problem in the typical study of police behavior while avoiding the concerns outlined above, we describe a feasible study of police-civilian interactions during traffic stops. A key advantage of traffic studies is that much of the data needed to improve research are already collected passively by law enforcement agencies across the United States in an automated fashion via highway cameras. We note that before the advent of this technology, data on unreported police-civilian interactions had to be manually collected by researchers accompanying patrol officers on their shifts (Allen 1982; Smith, Visher, and Davidson 1984), a labor-intensive strategy highly vulnerable to researcher demand effects (Orne 1962).

Recall that a key problem in the typical study of police behavior while avoiding the concerns outlined above, we describe a feasible study of police-civilian interactions during traffic stops. A key advantage of traffic studies is that much of the data needed to improve research are already collected passively by law enforcement agencies across the United States in an automated fashion via highway cameras. We note that before the advent of this technology, data on unreported police-civilian interactions had to be manually collected by researchers accompanying patrol officers on their shifts (Allen 1982; Smith, Visher, and Davidson 1984), a labor-intensive strategy highly vulnerable to researcher demand effects (Orne 1962).

Recall that a key problem in the typical study of police behavior while avoiding the concerns outlined above, we describe a feasible study of police-civilian interactions during traffic stops. A key advantage of traffic studies is that much of the data needed to improve research are already collected passively by law enforcement agencies across the United States in an automated fashion via highway cameras. We note that before the advent of this technology, data on unreported police-civilian interactions had to be manually collected by researchers accompanying patrol officers on their shifts (Allen 1982; Smith, Visher, and Davidson 1984), a labor-intensive strategy highly vulnerable to researcher demand effects (Orne 1962).

Recall that a key problem in the typical study of police behavior while avoiding the concerns outlined above, we describe a feasible study of police-civilian interactions during traffic stops. A key advantage of traffic studies is that much of the data needed to improve research are already collected passively by law enforcement agencies across the United States in an automated fashion via highway cameras. We note that before the advent of this technology, data on unreported police-civilian interactions had to be manually collected by researchers accompanying patrol officers on their shifts (Allen 1982; Smith, Visher, and Davidson 1984), a labor-intensive strategy highly vulnerable to researcher demand effects (Orne 1962).
Lange, Johnson, and Voas 2005), though in those studies, camera data on individual motorists were not linked to administrative data on police outcomes, as we propose below. Given a large random sample of passing cars captured by highway speed cameras, analysts could use video or photographic records to document license plate numbers that allow for a merge with other administrative data sets containing information on the registrant’s home neighborhood, whether each car went on to be stopped by nearby police at a proximate time, whether a summons was issued, and whether the encounter escalated to include a search or the use of force. As with all causal analyses of observational data, analysts must still make some version of Assumption 4(b)—no treatment-outcome confounding conditional on observable covariates—but in this case, the standard “treatment selection on observables” plausibly holds because virtually all prestop data available to an officer are in fact observable to the analyst. Using camera footage merged with administrative records, analysts could credibly measure this “complete” set of control variables. These factors would include not only the race, age, gender, and registered neighborhood of the driver but also the make, color, and condition of the car, along with weather and driving speed.

Given this set of covariates, researchers could credibly estimate the ATE for various outcomes, including searching, ticketing, and the use of force, by comparing the rates of outcomes between racial minority and majority motorists, regardless of whether they were stopped by police, conditional on $X$. The $\text{ATT}_{M=1}$ is similarly point identified because the proportion of racial stops can be calculated and used to correct estimates. However, the $\text{ATE}_{M=1}$ remains partially identified—the quantity can be bounded, as we show above, but not precisely estimated. And as Figure 6 makes clear, the $\text{CDE}_{M=1}$ remains fundamentally unidentifiable without covariates that make Assumption 5 plausible, such as controls for officer temperament that are specific to some stops but not others (i.e., time-varying), which likely influences both stopping decisions and subsequent treatment of civilians.

**CONCLUSION**

With the release of large and granular data on police-civilian interactions, many researchers have focused on estimating whether police exhibit racial bias in their treatment of civilians. Though some studies have acknowledged the threat of posttreatment bias in this setting (Fryer 2018), the issue has not been adequately addressed, and studies in this area have left ambiguous whether causal quantities are being approximated and the degree to which racial bias may be obscured by traditional estimation strategies. Given the policy relevance of this topic and the degree of selection bias inherent to these analyses, we believe social scientists need to devote substantial effort to develop research designs that can sidestep the threat of posttreatment conditioning rather than proceeding in the face of this threat and simply hoping for the best.

In this study, we clarify the statistical problems in the use of police administrative data in isolation to study racial bias. We offer bias-correction and bounding procedures for scholars analyzing these data, along with an improved research design that can avoid posttreatment conditioning altogether. Our results can inform the study of racial discrimination in a host of other settings beyond law enforcement. And though we focus on a case of racial bias in the United States, these results also speak to a rich literature on racial discrimination outside the U.S. context (e.g., Bruce-Jones 2015; Cano 2010). Our identifying assumptions may also be useful for researchers seeking to address biases stemming from posttreatment conditioning more generally, beyond studies of discrimination.

While we are optimistic about alternative designs and estimation strategies, we are under no illusions that eliminating this particular source of bias will remove others. Our research design suggestions may also limit the outcomes that are feasible to study. For example, rare events such as shootings may or may not occur during the observation periods proposed, meaning only lower level uses of force or sanctioning can be studied in
some cases. Our recommendations, therefore, place emphasis on bias reduction over latitude in the selection of research questions. But given the ease with which faulty conclusions can be reached as a result of the race-based selection we highlight, narrowing the scope of research to generate more reliable estimates may be preferable, especially because policy reforms could hinge on the results of studies in this area. Put differently, because of the pitfalls we highlight above, it is not clear that studies of rare phenomena that lack a sound design are generating usable knowledge anyway, so this trade-off in scope may be of only marginal concern (Samii 2016).

Regardless of which approach scholars pursue, this article highlights the need for further careful research into the first stage of police-civilian interactions—that is, the process by which officers decide whether or not to stop and investigate an individual for a crime. This effort is necessary not only to further our scholarly understanding of police-civilian interactions but also to craft effective policy reforms. If racial bias is concentrated in the initial stage of contact, reforms focused on reducing unnecessary police-civilian interactions may be most effective at curbing racially discriminatory police violence. On the other hand, if there exists more significant bias in the ultimate decision to use force, substantial improvements may require a wholly different reform strategy. Without serious consideration of the role of race in each stage of the complex police-civilian interactions under study, the benefits of data-driven reforms will be stunted, as will our collective understanding of the politics of policing.

SUPPLEMENTARY MATERIAL
To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055420000039.

Replication materials can be found on Dataverse at: https://doi.org/10.7910/DVN/KFOOCV.

REFERENCES
Bruce-Jones, Eddie. 2015. “German Policing at the Intersection: Race, Gender, Migrant Status and Mental Health.” Race & Class 56 (3): 36–49.


Knox et al. (2020) contained an error in the description of a numeric result. In describing a result in Table 1, the article stated: “For example, while the approach in Fryer (2019) implies a statistically insignificant 2.4 instances per 1,000 encounters of pushing Hispanic suspects to a wall due to suspect race, our revised estimate shows the true number is at least 22—nine times larger.” Thus, while estimates in Table 1 remain correct, the main-text discussion erroneously referred to the wrong entry in the table.

The authors regret the error.

REFERENCE
Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020
“Administrative Records Mask Racially Biased Policing.”
Administrative Records Mask Racially Biased Policing

Online Appendix

Contents

A Detailed proofs

A.1 Bias for $\text{ATE}_{M=1}$ ................................. 1
A.2 Bias for $\text{ATT}_{M=1}$ .............................. 4
A.3 Bias for $\text{CDE}_{M=1}$ ............................. 5
A.4 Nonparametric sharp bounds for $\text{ATE}_{M=1}$ .............................. 7
A.5 Uncertainty of bounds ............................... 10
A.6 Point identification of $\text{ATE}$ ....................... 11
A.7 Derivation of outcome test bounds on $\rho$ ............... 14

B Additional results

B.1 Coding schemes for dependent variables ............... 17
B.2 Varying levels of force ............................. 20
B.3 Excluding drug stops ............................... 23
B.4 Analysis of two races at a time ....................... 25
A Detailed proofs

A.1 Bias for ATE_{M=1}

We first derive the bias of the local difference in means (that is, among encounters with \(X_i = x\), \(\hat{\Delta}_x = Y_i|D_i = 1, M_i = 1, X_i = x - Y_i|D_i = 0, M_i = 1, X_i = x\)), in estimating the local average treatment effect among stops, \(ATE_{M=1,x} = \mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x]\). The overall bias is then given by \(\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - ATE_{M=1,x}\right) \Pr(X_i = x|M_i = 1)\).

\[\begin{align*}
\mathbb{E}[\hat{\Delta}_x] - ATE_{M=1,x} &= (\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x - Y_i|D_i = 0, M_i = 1, X_i = x]) \\
&\quad - (\mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x]) \\
&= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x] \\
&= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x] \\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)
\end{align*}\]

under mediator monotonicity, \(\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x) = 0\) and \(\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x) = 1\),

\[\begin{align*}
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
\end{align*}\]
\[
\begin{align*}
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 0 | M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(1) = 1 | D_i = 0, M_i(D_i) = 1, X_i = x) \Pr(D_i = 0 | M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
\Pr(M_i(1) = 0 | D_i = 0, M_i(D_i) = 1, X_i = x) \Pr(D_i = 0 | M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(1) = 1 | D_i = 0, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
\Pr(M_i(1) = 0 | D_i = 0, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 1 | D_i = 1, M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 1 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 1 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 1 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 1 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
\mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
\Pr(D_i = 1 | M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
\Pr(M_i(0) = 0 | D_i = 1, M_i(D_i) = 1, X_i = x) \Pr(D_i = 1 | M_i(D_i) = 1, X_i = x)
\end{align*}
\]
adding and subtracting \( \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \),

\[
\begin{align*}
&\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{align*}
\]

substituting potential mediators based on principal strata,

\[
\begin{align*}
&\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1) - Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{align*}
\]

under mandatory reporting, \( Y_i(d, 0) = 0 \),

\[
\begin{align*}
&\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad \Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\quad \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{align*}
\]
A.2 Bias for ATT

Next, we consider the bias that results when the local difference in means is used as an estimator for the local average racial effect among stopped minorities, $ATT_{M=1} = \mathbb{E}[Y_i(1,M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0,M_i(0))|D_i = 1, M_i = 1, X_i = x]$. Again, overall bias is found by the weighted average of local biases, $\sum_x \left( \mathbb{E}[\hat{\lambda}_x] - ATT_{M=1,x} \right) \text{Pr}(X_i = x|D_i = 1, M_i = 1)$.
by mediator monotonicity

\[
\begin{align*}
&= - \mathbb{E}[Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{align*}
\]

by mandatory reporting

\[
\begin{align*}
&= - \mathbb{E}[Y_i(0, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{align*}
\]

by treatment ignorability

\[
\begin{align*}
&= - \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|M_i(1) = 1, X_i = x)
\end{align*}
\]

A.3 Bias for CDE\textsubscript{M=1}

The CDE\textsubscript{M=1} is defined as

\[
\text{CDE}_{M=1} = \mathbb{E}[Y_i(1, 1)|M_i(D_i) = 1] - \mathbb{E}[Y_i(0, 1)|M_i(D_i) = 1] 
\]  

It is a somewhat contrived estimand because it necessarily involves counterfactuals that, for racial-stop encounters, could never realize even if researchers could somehow randomize civilian race in police-civilian encounters. For example, when a minority civilian is racially stopped for a “furtive movement” and reaches for their wallet, it makes little sense to consider an officer’s potential use of force if the civilian suddenly became white at that moment: had police observed a white civilian from the onset, a stop would never have occurred. Moreover, the assumptions required for such counterfactuals are fundamentally unverifiable (Robins and Greenland, 1992), unless the experimentalist can somehow also manipulate officer stopping decisions without distorting outcomes. (We note that always-stop encounters are not subject to this issue, but analysts cannot hope to estimate the conditional CDE in this group because it is impossible to identify which minority encounters belong to this group.)

In the previous bias expression for the ATE\textsubscript{M=1}, white individuals in the data—necessarily belonging to the always-stop group, \(M_i(1) = M_i(0) = 1\)—were used to estimate the \(Y_i(0, M_i(0))\) potential outcomes of minority encounters in the data. Unavoidable bias arose as long as any minority individuals in the data belonged to the racial-stop group—had these individuals been white,
under assumption 4(b),

\[ \hat{\Delta}_x - \text{CDE}_{M=1,x} = (\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]) \\
- (\mathbb{E}[Y_i(1, 1)|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, 1)|M_i = 1, X_i = x]) \\
= \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] \\
- \mathbb{E}[Y_i(1, 1)|M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, 1)|M_i(D_i) = 1, X_i = x] \\
= \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 0, M_i(0) = 1, M_i(0) = 1, X_i = x] \Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(1, 1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
+ \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
- \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)

under assumption 4(b),

\[ = (\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x] \\
- \mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x] \]
which reproduces Equation 7.
Finally, we demonstrate that this bias is weakly negative. Rearranging terms yields

\[
\begin{align*}
\mathbb{E}[Y_i(1, 1) - Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&= -\mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]
\end{align*}
\]

\[
\begin{align*}
&\times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, X_i = x] \\
&\times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x] \\
&\times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\times \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&- \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, X_i = x]
\end{align*}
\]

The sum of the first and second terms is weakly negative by Assumption 3, because the magnitude of the first term is greater than that of the second, and similarly the sum of the third and fourth terms is also weakly negative. Therefore, bias is weakly negative when the naïve estimator is used to estimate the CDE\(_{M=1}\).

### A.4 Nonparametric sharp bounds for ATE\(_{M=1}\)

In this section, we derive nonparametric sharp bounds for the ATE\(_{M=1}\). We begin with the case when the proportion of racially discriminatory stops among reported minority encounters, \(\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\), is known or can be assumed. Rearrangement of Equa-


We then construct bounds on $\mathbb{E}[Y_i|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]$ based on Fréchet inequalities for the joint distribution, $\Pr(Y_i(1, 1) = 1, M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x)$, which incorporate marginal information about $Y_i(1, 1)$ and $M_i(0)$.

\[
\text{max} \left\{ 0, \Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x) + \mathbb{E}[Y_i|D_i = 1, M_i(1) = 1, X_i = x] - 1 \right\}
\leq \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
\leq \min \left\{ \Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x), \mathbb{E}[Y_i|D_i = 1, M_i(1) = 1, X_i = x] \right\}
\]

These bounds are sharp given only marginal information, $\Pr(Y_i(1, 1) = 1|D_i = 1, M_i(1) = 1, X_i = x)$.
and \( \Pr(M_i(0) = 0|D_i = 1, M_i(1) = 1, X_i = x) \). However, the upper bound can be tightened further under Assumption 3, which implies \( \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \leq \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x]; \) this is at least as small as the upper Fréchet bound.

Finally, note that the reported data contain no information that can be used to constrain the proportion of racially discriminatory minority stops, \( \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \). If this proportion were zero, then the distribution of civilian race in police reports would reflect that of all police encounters (within levels of \( X \)). The reported data cannot distinguish between this possibility and an alternative population in which \( \rho_x = \Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x) \) is large, but white encounters are also larger by the proportion \( 1/(1 - \rho_x) \). Without side information about the total number of encounters, this proportion can take on any value in \( [0, 1) \). Therefore, sharp bounds on \( \text{ATE}_{M=1} \) alone are obtained by substituting Equation 5 into Equation 4 and setting the proportion of racial stops to unity. The bivariate bounds define the region in which \( \text{ATE}_{M=1}, \rho_x \) pairs are consistent with the observed data. When \( \rho_x \) is set to zero or one, these respectively recover the difference in reported means and the marginal upper bounds on \( \text{ATE}_{M=1} \). For \( \rho_x \in (0, 1) \),

\[
\mathbb{E}[\hat{\Delta}_x] + \rho_x \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x](1 - \Pr(D_i = 0|M_i = 1, X_i = x)) \\
\leq \text{ATE}_{M=1,x} \leq \\
\mathbb{E}[\hat{\Delta}_x] \\
+ \frac{\rho_x}{1 - \rho_x} \left( \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \max \left( 0, 1 + \frac{1}{\rho_x} \mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \frac{1}{\rho_x} \right) \right) \Pr(D_i = 0|M_i = 1, X_i = x) \\
+ \rho_x \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x](1 - \Pr(D_i = 0|M_i = 1, X_i = x)),
\]

which reduces to Proposition 1 in the no-covariate case. Otherwise, bounds on \( \text{ATE}_{M=1} \) are given by \( \sum_x \text{ATE}_{M=1,x} \Pr(X_i = x|M_i = 1) \leq \text{ATE}_{M=1} \leq \sum_x \text{ATE}_{M=1,x} \Pr(X_i = x|M_i = 1) \), where \( \text{ATE}_{M=1,x} \) (\( \overline{\text{ATE}}_{M=1,x} \)) denote the lower (upper) bounds on the local average treatment effect.

Finally, we note that per Equation 3, the \( \text{ATT}_{M=1,x} \) can be written

\[
\text{ATT}_{M=1,x} = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x) \\
= \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i = 1, X_i = x] \\
- \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x) \\
- \mathbb{E}[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x)
\]

under Assumption 1,

\[
= \mathbb{E}[Y_i(1, 1) - |D_i = 1, M_i = 1, X_i = x] \\
- \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x)
\]
and under Assumption 4,

\[ \begin{align*}
= \mathbb{E}[Y|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y|D_i = 0, M_i = 1, X_i = x] (1 - \text{Pr}(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x))
\end{align*} \]

which can be estimated from observed data if the proportion of racial stops is known. It then follows that

\[ \text{ATT}_{M=1} = \sum_x (\mathbb{E}[Y|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y|D_i = 0, M_i = 1, X_i = x]) + \rho_x \mathbb{E}[Y|D_i = 0, M_i = 1, X_i = x]) \]

\[ = \sum_x (\mathbb{E}[\Delta x] + \rho_x \mathbb{E}[Y|D_i = 0, M_i = 1, X_i = x]) \]

### A.5 Uncertainty of bounds

Here, we describe our approach for constructing confidence intervals for the bounds on these causal quantities. We take \(X_i, D_i\) and \(M_i\) as fixed, so that uncertainty in the bounds arises strictly from the estimation of the conditional expectations, \(\mathbb{E}[Y|D_i = d, M_i = 1, X_i = x]\). The asymptotic distribution of the estimated lower and upper bounds endpoints, \((\text{ATE}_{M=1}, \text{ATE}_{M=1}^\prime)\), then follows directly from the asymptotic joint distribution of \(\mathbb{E}[Y|D_i = d, M_i = 1, X_i = x]\) for all \(d\) and \(x\). We approximate this through a Monte Carlo simulation in which parameters of the logistic regression models described in Section 5 are sampled from a multivariate normal distribution centered on the parameter estimates and with the estimated covariance matrix. For each parameter sample \(\theta^\prime\), the corresponding bounds endpoint pair \((\text{ATE}_{M=1}, \text{ATE}_{M=1}^\prime)\) is computed deterministically; after drawing a sufficient number of such samples, we numerically obtain the shortest range that fully contains 95% of all simulated bounds intervals. Closely related alternatives to this approach are the bootstrap-based method of Horowitz and Manski (2000) and the fully Bayesian approach taken in Knox et al. (2019). For the analysis in Section 5, we follow Fryer (2019) in using a cluster-robust covariance estimator, clustering on precinct, and 5,000 samples were drawn for each force threshold and model specification.
A.6 Point identification of ATE

First, we note that strata sizes are identified with information on the total count of encounters by race (both reported and unreported).

\[
\begin{align*}
\Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) &= \Pr(M_i(1) = 1, M_i(0) = 1| D_i = 0, X_i = x) \\
&= \Pr(M_i = 1|D_i = 0, X_i = x) \\
\Pr(M_i(1) = 1, M_i(0) = 0, X_i = x) &= \Pr(M_i(1) = 1, X_i = x) \\
&= \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) \Pr(M_i = 1|D_i = 1, X_i = x) \\
&\quad - \Pr(M_i(1) = 1, M_i(0) = 0, X_i = x) \\
&= \Pr(M_i = 1|D_i = 0, X_i = x) \\
\Pr(M_i(1) = 0, M_i(0) = 1, X_i = x) &= 0 \\
\Pr(M_i(1) = 0, M_i(0) = 0, X_i = x) &= 1 - \Pr(M_i(1) = 0, M_i(0) = 1, X_i = x) \\
&\quad - \Pr(M_i(1) = 1, M_i(0) = 0, X_i = x) \\
&\quad - \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) \\
&= 1 - \Pr(M_i = 1|D_i = 1, X_i = x) \\
\end{align*}
\]

We then reexpress the ATE in terms of strata-specific mean potential outcomes and simplify.
\[ \text{ATE} = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \]
\[= \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \]
\[\Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x] \]
\[\Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 0, M_i(0) = 0, X_i = x] \]
\[\Pr(M_i(1) = 0, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \]
\[\Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \]
\[\Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 1, X_i = x] \]
\[\Pr(M_i(1) = 0, M_i(0) = 1|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
\[+ \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(1) = 0, M_i(0) = 0, X_i = x] \]
\[\Pr(M_i(1) = 0, M_i(0) = 0|D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x) \]
Pr(M_i(1) = 0, M_i(0) = 0 | D_i = 0, X_i = x) Pr(D_i = 0, X_i = x)

under mandatory reporting

= \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)

− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 0, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)
− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 0, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)

under mediator monotonicity

= \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)
+ \mathbb{E}[Y_i(1, M_i(1)) | D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)

− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 1, M_i(0) = 0 | D_i = 1, X_i = x) \Pr(D_i = 1, X_i = x)
− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]
   \Pr(M_i(1) = 0, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)
− \mathbb{E}[Y_i(0, M_i(0)) | D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x]
   \Pr(M_i(1) = 0, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)
Pr(M_i(1) = 1, M_i(0) = 1|D_i = 0, X_i = x) Pr(D_i = 0, X_i = x)

under treatment ignorability

\[ = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) \\
+ \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(1) = 1, M_i(0) = 0, X_i = x) \\
- \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(1) = 1, M_i(0) = 1, X_i = x) \]

which can be recovered from observed data

\[ = \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] \Pr(M_i = 1|D_i = 1, X_i = x) \\
- \mathbb{E}[Y_i|D_i = 0, M_i(D_i, X_i = x) = 1, X_i = x] \Pr(M_i = 1|D_i = 0, X_i = x) \]

which reduces to Proposition 2 in the no-covariate case.

A.7 Derivation of outcome test bounds on $\rho$

Our paper focuses on the difficulty of estimating a race effect on post-stop police behavior such as the use of force. However, another popular approach, the outcome test, focuses on establishing whether there exists any bias in the decision to stop a civilian (Becker, 1971; Goel, Rao and Shroff, 2016; Engel, 2008; Knowles, Perisco and Todd, 2001; Ridgeway and MacDonald, 2010). Because the degree of the statistical bias we explore is a function of racial discrimination in stopping decisions, it is useful to clarify the assumptions undergirding outcome tests. In the process, we demonstrate that the principal stratification framework sheds light on the precise interpretation of outcome tests, and we prove that the outcome test can be used to establish a lower bound on the share of police stops of racial minorities that are racially discriminatory.

Outcome tests compare the rates of finding evidence of a crime—conditional on a suspect being stopped by police—across racial groups. The logic behind the test is that if the decision to stop a civilian is unbiased, the rate of discovering evidence of a crime (“hit rates”) should be identical across groups. Proponents of outcome tests thus claim that differences in hit rates amount to evidence of racially biased policing. The empirical observation that hit rates are lower among minority stops can be written as $\mathbb{E}[Y_i|D_i = 0, M_i = 1] > \mathbb{E}[Y_i|D_i = 1, M_i = 1]$, where $Y_i$ is an indicator, say, for finding contraband on a suspect. However, interpreting the above inequality as evidence of racial discrimination in fact requires assumptions that closely mirror those we describe above.

To see this, first observe that the overall hit rate among minority stops can be decomposed into
the weighted average of the hit rate among always-stop encounters and the hit rate among the
(possibly nonexistent) set of racially discriminatory stops. In contrast, if we invoke Assumption 2
(which states that there are no white civilians stopped in circumstances where a minority civilian
would be allowed to pass), then stops involving white civilians belong exclusively to the always-
stop group. In this case, the empirical difference in hit rates can be rewritten in the potential
outcomes framework as

\[
\mathbb{E}[Y_i(0,1)|M_i(1) = 1, M_i(0) = 1, D_i = 0] - \mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 1, D_i = 1] \Pr(M_i(1) = 1, M_i(0) = 1|D_i = 1, M_i = 1)
\]

A major critique of the outcome test is that observed racial disparities in hit rates alone do not
constitute evidence of racially discriminatory stops because of the problem of “infra-marginality”
(Ayres, 2002; Simoiu, Corbett-Davies and Goel, 2017). This critique suggests that the above in-
equality may hold simply because white civilians in always-stop encounters engage in more crim-
inal conduct than minority suspects. In other words, the analyst might observe \( \mathbb{E}[Y_i|D_i = 1, M_i = 1] < \mathbb{E}[Y_i|D_i = 0, M_i = 1] \) even if \( \Pr(M_i(1) = 1, M_i(0) = 0) = 0 \)—that is, with no discrimination in stops—as long as \( \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 1] < \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 0] \). Some analysts employing the outcome test cast this scenario as unlikely, arguing that absent racial bias in stopping, “it would be difficult to explain why...whites for some reason had a sys-
tematically higher chance of possessing evidence of illegality” (Ayres, 2002) (137) and “there are
not compelling reasons to suspect” this to be the case (138). Indeed, the validity of the outcome
test hinges on the assumption that white and minority civilians in always-stop encounters com-
mit crimes at the same rates, or that \( \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 1] = \mathbb{E}[Y_i(0, 1)|M_i(1) = 1, M_i(0) = 1, D_i = 0] \). This assumption closely parallels Assumption 4, which requires that treat-
ment status is ignorable with respect to potential outcomes. (For simplicity, we suppose that this
holds without conditioning on covariates, but the result also holds within levels of \( X_i = x \).)
In this case, the observed racial difference in hit rates can be rewritten as

\[
\mathbb{E}[Y_i|D_i = 0, M_i = 1] - \mathbb{E}[Y_i|D_i = 1, M_i = 1] = (\mathbb{E}[Y_i(0,1)|M_i(1) = 1, M_i(0) = 1, D_i = 0] - \mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 0, D_i = 1]) \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, M_i = 1). \tag{7}
\]

This formulation makes clear that the observed evidence gap is due to the difference in hit

\[\text{1If we do not assume mediator monotonicity, and allow for the presence of stops of white suspects that would not have occurred if the suspect was a racial minority, then the inequality used to estimate the outcome test becomes uninformative with respect to racial discrimination.}\]
rates between always-stop minority encounters—in which officers would also have stopped a white civilian—and racially discriminatory minority stops. If the former is assumed to produce more evidence of criminal behavior (Assumption 3; this might hold if racially discriminatory stops are made under weaker standards of evidence), then it can be seen from Equation 7 that the empirical difference in hit rates implies that \( \Pr(M_i(1) = 1, M_i(0) = 0) > 0 \): that there must exist encounters in which minority civilians would be stopped but white civilians would not, precisely as proponents of the outcome test suggest.

Equation 7 also shows that outcome tests are unable to identify the exact prevalence of racial stops. Outcome tests allow the analyst to infer whether there is any racial bias in the decision to stop a suspect—but only if the analyst makes assumptions similar to those we outline above. However, we show that the outcome test can partially identify a range of possible proportions of racial stops. This clarification is useful, as it allows us later in this analysis to appeal to a published study of hit rates (Goel, Rao and Shroff, 2016) to help characterize the statistical bias in analyses of post-stop police behavior (e.g. Fryer, 2019).

By rearranging Equation 7 and substituting observed quantities, we arrive at

\[
\Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1, M_i = 1) = \frac{\mathbb{E}[Y_i|D_i = 0, M_i = 1] - \mathbb{E}[Y_i|D_i = 1, M_i = 1]}{\mathbb{E}[Y_i|D_i = 0, M_i = 1] - \mathbb{E}[Y_i(1, 1)|M_i(1) = 1, M_i(0) = 0, D_i = 1]}
\]

Although the second term in the denominator is unknown, the implied proportion of racially discriminatory stops is smallest when this value is zero—if, hypothetically, searches of racially stopped minorities never produce evidence. Thus, the outcome test suggests that at least \( \frac{\mathbb{E}[Y_i|D_i = 0, M_i = 1] - \mathbb{E}[Y_i|D_i = 1, M_i = 1]}{\mathbb{E}[Y_i|D_i = 0, M_i = 1]} \) of all minority stops are racially discriminatory, and to the extent that racially discriminatory searches result in any evidence of contraband, the proportion could potentially be much larger.
B  Additional results

B.1  Coding schemes for dependent variables

In this section, we reanalyze the NYPD SQF data using both the original and revised coding schemes for dependent variables in Fryer (2019). In an analysis of the use of force by level of severity, Fryer (2019) codes binary outcomes indicating whether force is used at or above some threshold. However, rather than coding all encounters with lower levels of force than a given threshold as a zero, the analysis coded only encounters with no force at all as a zero, while levels of force between no force and the threshold level were dropped from the data.\(^2\) This data dropping strategy, a form of selection on the dependent variable, is problematic. If the analyst suspects that civilian race affects which level of force is applied—the motivating hypothesis for this very analysis—then dropping data based on which level of force was applied is another form of post-treatment conditioning and will induce bias. Further, the amount of data lost under this coding scheme is substantial. In the case of the point-weapon threshold, for example, over one million encounters—over 20% of the data—appear to have been discarded despite containing sub-threshold force use, such as pushing a civilian to the ground. Table B1 displays the number of observations reported for various regressions in the original paper, our best attempts at replication, and the corrected procedure used in this paper.

We present results of our replication study using the original coding scheme and our corrected version side by side below. As the results show, a corrected analysis generally depresses the naïve treatment effects relative to the inadvisable coding scheme in Fryer (2019) and in most cases renders the original results statistically insignificant. However, these discrepancies in results across coding decisions do not alter the central point of our paper: post-treatment conditioning exerts a large downward bias on estimates of racially discriminatory uses of force.

\(^2\)Fryer (2019) acknowledges this data dropping strategy, writing "To be clear, an observation that records only hands would be in the hands regression but not the regression which restricts the sample to observations in which individuals were at least forced to the ground," (21, emphasis in original).
Table B1: **Comparison of SQF Data Dimensions Based on Outcome Coding.** The table displays the number of observations from bivariate analyses of the use of force by the NYPD using three coding procedures for force outcomes. The first column displays the number of observations as reported in results in Fryer (2019) (Appendix Tables 3A-3G in original). The second column reports the number of observations we recover when using the coding procedure in Fryer (2019) which drops observations where some level of force was used that was below a given threshold. The third column displays the number of observations we recover when using our corrected coding procedure, which codes outcomes as a 1 if a certain force threshold is reached and 0 otherwise.

<table>
<thead>
<tr>
<th>Outcome Description</th>
<th>N (published)</th>
<th>N (replicated)</th>
<th>N (corrected coding)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least use of hands</td>
<td>4,927,962</td>
<td>4,980,701</td>
<td>4,980,701</td>
</tr>
<tr>
<td>At least push to wall</td>
<td>4,152,918</td>
<td>4,245,091</td>
<td>4,980,701</td>
</tr>
<tr>
<td>At least use of handcuffs</td>
<td>4,017,783</td>
<td>4,122,329</td>
<td>4,980,701</td>
</tr>
<tr>
<td>At least draw weapon</td>
<td>3,957,687</td>
<td>3,965,721</td>
<td>4,980,701</td>
</tr>
<tr>
<td>At least push to ground</td>
<td>3,950,324</td>
<td>3,958,374</td>
<td>4,980,701</td>
</tr>
<tr>
<td>At least point weapon</td>
<td>3,918,741</td>
<td>3,926,805</td>
<td>4,980,701</td>
</tr>
<tr>
<td>Use baton or pepper spray</td>
<td>3,900,977</td>
<td>3,909,064</td>
<td>4,980,701</td>
</tr>
</tbody>
</table>
Figure B1: Replication of Fryer (2019) using various outcome coding rules. The figure displays odds ratios generated by OLS regressions that show the effect of suspect race without covariates on the use of force across all force types generated using three approaches: the published OLS results from the Appendix of Fryer (2019) (black points and bars), our best attempt at replication of these results (blue points and bars), and results using our corrected outcome coding scheme (red points and bars). Revising the coding scheme so as to retain data on sub-threshold uses of force generally deflates estimated treatment effects.
B.2 Varying levels of force
Figure B2: Corrected $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ for encounters with Black and white civilians, varying levels of force. This figure shows bounded effects comparing predicted levels of force when setting suspect race for all observations to black vs. white. These estimates use our corrected coding scheme for dependent variables (as described above). Results from regressions without covariates appear in the top panels and results from models with a full set of covariates appear in bottom panels.
Figure B3: Corrected $ATE_{M=1}$ and $ATT_{M=1}$ for encounters with Hispanic and white civilians, varying levels of force. This figure shows bounded effects comparing predicted levels of force when setting suspect race for all observations to Hispanic vs. white. These estimates use our corrected coding scheme for dependent variables (as described above). Results from regressions without covariates appear in the top panels and results from models with a full set of covariates appear in bottom panels.
B.3 Excluding drug stops
Figure B4: **Bounds on race effect excluding drug stops.** This analysis replicates the analysis in Figure 4 in the main text excluding stops that were motivated by suspicion of a drug transaction, as such instances may violate the mediator monotonicity assumption. The results remain substantively similar.
B.4 Analysis of two races at a time
Figure B5: **Bounds on race effect limiting analysis to two racial groups of suspects.** Plots in the main text estimated bounds using data on multiple racial groups of suspects by predicting counterfactual values for every observation, regardless of a suspect’s actual race, after model parameters were estimated. These figures reproduce the same analysis using only data on the two racial groups being compared, and exclude data on suspects who were not black, Hispanic or white entirely.
References


