The Bias Is Built In: How Administrative Records Mask Racially Biased Policing

Dean Knox
dcknox@princeton.edu

Will Lowe
wlowe@princeton.edu

Jonathan Mummolo*
jmummolo@princeton.edu

Abstract

Researchers often lack the necessary data to credibly estimate racial bias in policing. In particular, police administrative records lack information on civilians police observe but do not investigate. In this paper, 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, rendering many quantities of interest 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 approaches. We develop a bias-correction procedure and nonparametric sharp bounds for race effects, replicate published findings, and show traditional estimation techniques 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.

*Dean Knox is an Assistant Professor of Politics at Princeton University. Will Lowe is a Senior Research Specialist and Lecturer in Politics at Princeton University. Jonathan Mummolo is an Assistant Professor of Politics and Public Affairs at Princeton University. We thank Michael Pomirchy for research assistance.

Electronic copy available at: https://ssrn.com/abstract=3336338

Concern over racial bias in policing combined with the public availability of large administrative data sets documenting police-citizen interactions have prompted a raft of studies attempting to quantify the effect of citizen race on law enforcement behavior. These studies consider a range of outcomes including citations, stop duration, searches and the use of force (e.g. Antonovics and Knight, 2009; Fryer, 2019; Nix et al., 2017; Ridgeway, 2006). Most research in this area attempts to adjust for omitted variables that may correlate with suspect race and the outcome of interest. This paper 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 (Rosenbaum, 1984; Angrist and Pischke, 2008; Elwert and Winship, 2014). We show that when there is any racial discrimination in the decision to detain civilians—a decision which determines whether civilians appear in police administrative data at all—then estimates of the effect of citizen race on subsequent police behavior are biased absent additional data and/or strong and untestable assumptions.

This paper 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 that employs this approach, then derive the statistical bias when they are violated. We proceed to develop informative nonparametric sharp bounds for the range of possible racial effects, apply these to a reanalysis and extension of a prominent paper 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-citizen 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 paper, 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; Robins, Hernán and Brumback, 2000; Pearl, 2001; VanderWeele, 2009) and selection (Heckman, 1977; Lee, 2009): the effect of civilian race on the outcome of a police encounter is mediated by whether a suspect is stopped by police, but the analyst only has data for one level of the mediator—i.e., data on stopped individuals. Because of this, police records do not contain a representative sample of all individuals that police

Electronic copy available at: https://ssrn.com/abstract=3336338

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), 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 this unseen first stage of police-citizen contact, which would render white civilians in these data sets incomparable to racial minorities in the data, and thereby bias estimates of racial discrimination.[1] 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-citizen encounters that appear in police data, and 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 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, 1977) 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).[2] For example, path-breaking work including Heckman (1977), as well as 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.[3] 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 *in sample*.

A related large literature provides remedies for so-called "post-treatment bias"—statistical

[1] For example, if police are more likely to stop minority citizens regardless of their behavior, but tend to stop white citizens only when criminal behavior is observed, the set of white individuals in police data may pose a greater threat to police than the set of minority individuals, on average.

[2] This comment was made in reference to an analysis of arrest data in (Fryer, 2019). Further, (Fryer, 2019) includes an analysis aimed at characterizing selection into police data sets, and finds mixed results depending on the outcome examined. The study states: "Taken together, this evidence demonstrates how difficult it is to understand whether there is potential selection into police datasets ...Solving this is outside the scope of this paper," (19).

[3] In other words, policing scholars are targeting the sample average treatment effect (*SATE*), not the population average treatment effect (*PATE*).

Electronic copy available at: https://ssrn.com/abstract=3336338

bias which 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 post-treatment variable (e.g. Acharya, Blackwell and Sen, 2016).[4] In the case of policing, administrative data sets only include observations with one level of the post-treatment 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 non-stopped individuals is zero, i.e. a realized outcome). This feature allows us to identify additional causal quantities that cannot be recovered in the "truncation by death" setting. In short, absent strong assumptions about the unseen process mapping civilian race to officers' decisions to detain individuals, existing methods offer either unusable or sub-optimal solutions to this pernicious threat to inference.

Following a series of controversial police shootings of unarmed black men and subsequent violent clashes between police and protesters, racial bias in policing has once again become a central fixture of legal, political and scholarly debate (Alexander, 2010; Lerman and Weaver, 2014*a*,*b*). Academic research in this area has and will be relied upon by lawmakers and courts (Gelman, Fagan and Kiss, 2007) and may serve as the basis for policy reforms, making causal validity paramount. But 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 citizens of color. For example, while a naïve analysis that assumes no race-based selection into the data suggests that only 2,900 black and Hispanic civilians were handcuffed due to racial bias in New York City between 2003 and 2013, we estimate that the true number is approximately 60,000.

While the techniques used to obtain these these corrected results eliminate several fa-

[4]In addition, the remedy proposed in Blackwell (2013), which requires re-weighting across all strata of the post-treatment variable, cannot be implemented in the situation we describe. However, the alternative designs we propose in Section 4.2 are amenable to this approach.

Electronic copy available at: https://ssrn.com/abstract=3336338

cially 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 in Section 3.1. 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 (Goel, Rao and Shroff, 2016; Gelman, Fagan and Kiss, 2007). We show that obtaining more precise bias-corrected estimates of racial discrimination in policing requires that future research be designed with this pernicious variant of sample selection bias in mind. To that end, we outline a research design that alleviates these concerns.

In what follows, we outline scope conditions for our analysis, and discuss causal estimands of interest in the study of racially biased policing and their identifying assumptions. We quantify the statistical bias for these estimands resulting from conventional analytic approaches, then derive bias-free nonparametric sharp bounds for the effect of race on police behavior. We apply these bounds in a reanalysis of Fryer (2019), showing that the study's estimates of racially discriminatory police violence are likely substantially understated. We then present a research design robust to these concerns and conclude.

1 Conceptualizing race as a causal variable

We regard the investigation of racial bias in policing as an inherently causal endeavor, albeit a notoriously difficult one. That is, researchers seek to assess whether police behavior during police-citizen encounters would have unfolded differently if the civilian had belonged to another racial group, holding constant criminal behavior and circumstance. 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 treatment across racial groups, as such disparities can arise via numerous processes that do not imply racial discrimination.

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 non-manipulable, characteristic (Gelman and Hill, 2007; Holland, 1986). Others argue that an individual's race is a complex, multifaceted treatment—a "bundle of sticks," in the words

Electronic copy available at: https://ssrn.com/abstract=3336338

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-citizen *encounters*—i.e., 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—e.g. skin tone, dialect, clothing, or some combination thereof. 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.[5]

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 citizens 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.

2 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. Grogger 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

[5]Note that while the unit of analysis is the police-citizen encounter, for the sake of brevity, we occasionally refer to "minority citizens" as shorthand for "police-citizen encounters with minority citizens" in subsequent discussion. Readers are cautioned to keep this distinction in mind.

Electronic copy available at: https://ssrn.com/abstract=3336338

sample of police-citizen interactions that were initiated in a race-blind manner. Similarly West (2018) examines data on police responses to traffic accidents, arguing that the dyadic relationships between the race of motorists and responding officers in these unanticipated events 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. Several recent studies attempt to estimate the effect of civilian race on police use of force. For example, Fryer (2019) (detailed below), a study of racial bias in police violence in various settings, estimates racial bias using data on police-citizen encounters via multivariate regressions that control for a host of observables relating to civilians, officers and circumstance. 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) finds some evidence of bias in sub-lethal force but none in lethal encounters. Nix et al. (2017) analyzes a recently assembled database of police-involved shootings by *The Washington Post* to study whether suspect race affected various attributes of shootings. Though the data contain no information on non-shootings, (i.e., the study selects on the dependent variable), the authors make explicitly causal claims. For example, the study reports that, "Black citizens were no more or less likely than White citizens to have been attacking the officer(s) or other citizens when they were fatally shot by police. These results provide support for an implicit bias effect with respect to non-black minority groups. That is, citizens of other races/ethnicities were significantly more likely than Whites to have been fatally shot because of an apparent threat perception failure," (325).

Prior work has also examined racial bias during traffic stops. For example, Ridgeway (2006) 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 re-weighting strategy can recover "the causal effect of race" (9) on post-stop outcomes. In general the analysis finds little evidence of racial bias on most outcomes, with the exception of stop duration. Antonovics and Knight (2009) uses 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

Electronic copy available at: https://ssrn.com/abstract=3336338

of "racial profiling based upon 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, multi-disciplinary effort to quantify the degree to which police discriminate against citizens of color (see Fridell (2017) and Ridgeway and MacDonald (2010) for more extensive reviews of this empirical literature). We highlight these examples because they all contain several common features that are central to our critique. For one, all of these studies analyze data that fail to capture the unseen selective process through which police come to engage civilians, a process that prior work strongly suggests may be a function of citizen 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-citizen 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, all of the aforementioned studies, despite making at least implicitly causal claims, leave ambiguous the precise quantity of interest—whether it be the total effect (TE) of race in all encounters; the total effect among the subset of encounters appearing in police data because a stop was made (TE_S), which differs tremendously from the TE; or the markedly more restrictive and difficult-to-interpret controlled direct effect among the same subset (CDE_S, 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 ambiguous.

2.1 Taste-based vs. statistical discrimination

The aforementioned studies differ from a closely related literature that 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 focus on recovering the causal effect of civilian race on police behavior. In this paper, we do not attempt to decipher the mechanism for racially

Electronic copy available at: https://ssrn.com/abstract=3336338

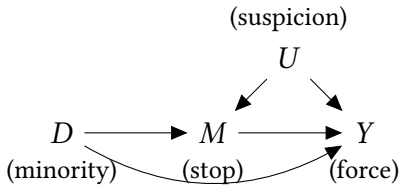
biased policing, and note that taste-based and statistical discrimination both pose serious normative concerns. Even statistical discrimination, sometimes viewed as more innocuous, constitutes racial profiling since it implies officers are detaining civilians not due to their observed actions but to the actions of the racial groups to which civilians belong. Federal courts have outlawed such justifications for detaining civilians (see discussions in Goel, Rao and Shroff (2016) and Mummolo (2018)). As such, quantifying the causal effect of civilian race on police behavior—our task here—is imperative regardless of the mechanism that motivates such an effect.

3 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. However, analyses of administrative data on police-citizen 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

Figure 1: Directed acyclic graph of racial discrimination in the use of force by police. Observed X is left implicit; these covariates may be causally prior to any subset of D, M, and Y.



Electronic copy available at: https://ssrn.com/abstract=3336338

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 (Pearl, 2001; Imai et al., 2011). 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 $M_i = M_i(D_i) = \sum_d M_i(d)\mathbf{1}\{D_i = d\}$ and $Y_i = Y_i(D_i, M_i(D_i)) = \sum_d \sum_m Y_i(d, m)\mathbf{1}\{D_i = d, M_i = m\}$, respectively. 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.

Our analysis begins by partitioning the population into principal strata with respect to the mediator (Frangakis and Rubin, 2002; VanderWeele, 2011). That is, we conceptualize police-citizen 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 racially discriminatory stops ("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 upon selectively by officers based on the race of suspects. Importantly, principal strata are not fully observable without further assumptions, and they exist even after conditioning on X_i: 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 our analysis, we subdivide principal strata further according to the realized race of the civilian involved, for a total of eight disjoint and collectively exhaustive groups corresponding to unique combinations of D_i, $M_i(1)$, and $M_i(0)$.

Electronic copy available at: https://ssrn.com/abstract=3336338

A central quantity of interest in the study of policing bias is the average total effect of race, $\text{TE} = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))]$—the extent to which citizens of color face greater risk of police violence than white citizens *because of their race*. The total effect 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. This quantity can be written as:

$$\begin{aligned}\text{TE} &= \mathbb{E}[Y_i(1, M_i(1))] - \mathbb{E}[Y_i(0, M_i(0))] \\ &= \sum_d \sum_m \sum_{m'} \Big(\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0)) | D_i = d, M_i(1) = m, M_i(0) = m'] \\ &\qquad \times \Pr(D_i = d, M_i(1) = m, M_i(0) = m') \Big) \end{aligned} \tag{1}$$

However, police administrative records contain data only on reported encounters, meaning that this quantity cannot be estimated solely with police administrative data. No data is available for "never-stop" encounters, those with $M_i(1) = M_i(0) = 0$, such as instances when civilians ask for directions. Moreover, racial-stop encounters, with $M_i(1) = 1$ and $M_i(0) = 0$, are only recorded for minority civilians. As a consequence, researchers seeking to understand the role of race in police behavior have, at least implicitly, focused on more narrowly defined estimands.[6] Studies commonly restrict analysis to the subset of reported encounters, i.e., they seek to estimate effects among those stopped by police. We denote this conditional average total effect—the "total effect among stops"—as $\text{TE}_S = \mathbb{E}[Y_i(1, M_i(1)) | M_i = 1] - \mathbb{E}[Y_i(0, M_i(0)) | M_i = 1]$. In contrast with the TE, 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. However, the TE_S does include all reported minority stops, some of which would not have occurred if the civilian were

[6] 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 total effect.

Electronic copy available at: https://ssrn.com/abstract=3336338

white. Formally, the TE_S is given by

$$\begin{aligned}
\text{TE}_S &= \mathbb{E}[Y_i(1, M_i(1))|M_i = 1] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1] \\
&= \sum_m \Big(\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = m] \\
&\qquad \times \Pr(M_i(0) = m|D_i = 1, M_i = 1) \Pr(D_i = 1|M_i = 1) \Big) \\
&\quad + \sum_m \Big(\mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 0, M_i(1) = m, M_i(0) = 1] \\
&\qquad \times \Pr(M_i(1) = m|D_i = 0, M_i = 1) \Pr(D_i = 0|M_i = 1) \Big) \qquad (2)
\end{aligned}$$

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 total effect among treated (i.e., minority) stops, which can be written as:

$$\begin{aligned}
\text{TE}_{ST} &= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1] \\
&= \sum_m \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = m] \\
&\qquad \Pr(M_i(0) = m|D_i = 1, M_i = 1). \qquad (3)
\end{aligned}$$

While the total 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, $\text{CDE}_S = \mathbb{E}[Y_i(1, 1)|M_i = 1] - \mathbb{E}[Y_i(0, 1)|M_i = 1]$. This estimand differs from the TE_S in its conceptual approach to racially discriminatory stops. Where the TE_S asks whether a minority stop would have occurred at all if the individual were white, the CDE_S seeks to quantify what would have happened *if the officer was forced to stop them anyways*, 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 difference in reported outcomes, and the question lies in the interpretation of estimated results. We note that estimands in the literature are often left undefined, making it difficult to assess whether published results are intended to correspond to the TE_S or CDE_S.

The CDE_S is a somewhat contrived estimand because it is based on a counterfactual for encounter i that in many cases could never realize, even in experiments where civilians of differing races could somehow be randomized into police-citizen encounters. For example,

Electronic copy available at: https://ssrn.com/abstract=3336338

when a minority civilian is racially stopped (e.g. for a "furtive movement") and reaches for their wallet, it makes little sense to consider the potential for an officer to deploy 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 "cross-world" counterfactuals are fundamentally unverifiable (Robins and Greenland, 1992). The CDE_S is less problematic in situations where civilians are as-if randomly detained by police, thus negating the issue of race-based selection. This might occur, for example, if police agencies institute a procedure for drunk driving checkpoints whereby every fifth car is stopped, a race-blind procedure. In this case, it is simply equivalent to the TE_S divided by the probability of a stop.

The CDE_S can be expressed as:

$$\begin{aligned}
\text{CDE}_S &= \mathbb{E}[Y_i(1,1)|M_i(D_i) = 1] - \mathbb{E}[Y_i(0,1)|M_i(D_i) = 1] \\
&= \sum_m \mathbb{E}[Y_i(1,1) - Y_i(0,1)|D_i = 1, M_i(1) = 1, M_i(0) = m] \\
&\qquad \Pr(M_i(0) = m|D_i = 1, M_i(D_i) = 1) \Pr(D_i = 1|M_i(D_i) = 1) \\
&\quad + \sum_m \mathbb{E}[Y_i(1,1 - Y_i(0,1)|D_i = 0, M_i(1) = m, M_i(0) = 1] \\
&\qquad \Pr(M_i(1) = m|D_i = 0, M_i(D_i) = 1) \Pr(D_i = 0|M_i(D_i) = 1) \qquad (4)
\end{aligned}$$

To visualize the components needed to recover these estimands, Figure 2 displays potential outcomes across various values of realized treatment, D_i, and principal strata, as defined by the mediator $M_i(d)$. The TE, TE_S, TE_{ST} and CDE_S are all defined as differences between average potential outcomes under some form of treatment (left side of the figure) and control (right side of figure). To illustrate, consider the top two cells of the left half of the figure. These two strata collectively represent all minority encounters ($D_i = 1$) that resulted in a recorded stop by police, i.e., $M_i(D_i) = M_i(1) = 1$. These encounters fall into one of two categories—those for which $M_i(0) = M_i(1) = 1$ (the always-stop encounters), and those for which $M_i(0) = 0$ (encounters in which a white civilian would not have been stopped, i.e., encounters in which the minority civilian was stopped due to racial bias), though this distinction is not observable to the analyst. To estimate any of the four causal quantities described above, analysts must first estimate the average potential outcome $Y_i(1,1)$ across these strata. This is straightforward because, as the red outline in these two cells indicates, the $Y_i(1,1)$ potential outcome is in fact recorded for these two strata in police administrative data. For most other strata, however, encounters are either en-

Electronic copy available at: https://ssrn.com/abstract=3336338

tirely unreported or the relevant potential outcome is counterfactual and hence remains unobserved. Absent data on these unobserved strata-specific mean potential outcomes, the analyst must make additional assumptions (outlined below) in order to estimate these causal quantities without statistical bias.

3.1 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-citizen 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* $d \in \{0, 1\}$.

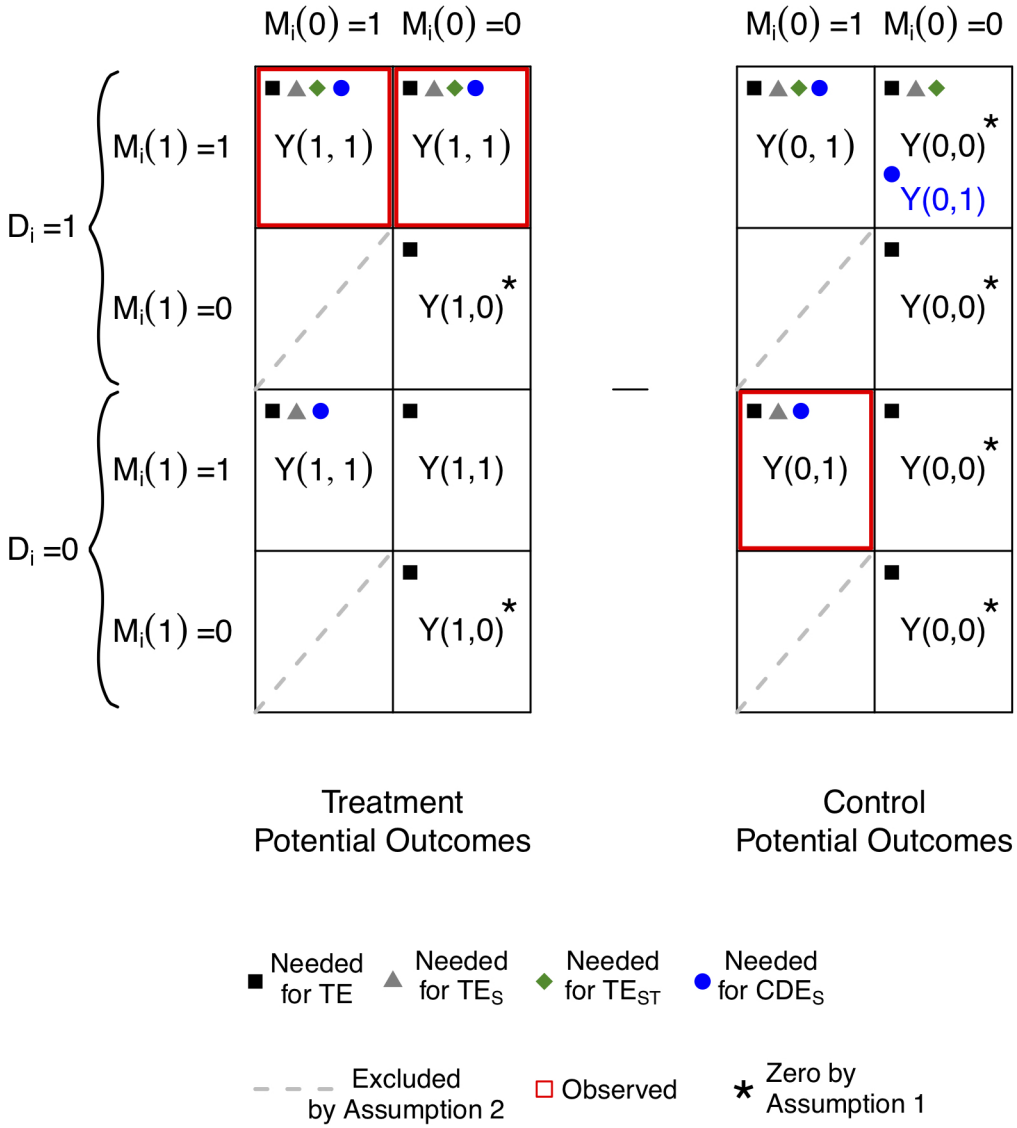
We assume all encounters that escalate to the use of force also trigger a reporting requirement and are therefore observed in administrative data. Though there exists wide variability in data recording practices across jurisdictions, this assumption is plausible in the study of many major police departments. 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, 2014). The completeness of these reports with respect to fatalities is informally enforced by standard journalistic reporting practices which place high emphasis on documenting incidents of violent crime (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 the 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.

There undoubtedly exist many encounters in which civilians of both groups would be treated equally, including egregious crimes that would result in arrest regardless of suspect race, as well as mundane encounters in which no civilian would be detained. We allow 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

Electronic copy available at: https://ssrn.com/abstract=3336338

Figure 2: **Potential Outcomes by Strata.** Encounters fall into one of eight strata based on three factors: the encountered civilian's actual race, D_i; whether a minority would be stopped in those circumstances, $M_i(1)$; and whether white civilians would be stopped, $M_i(0)$. Below, each stratum is represented with a square, and each group of eight squares represents the full dataset. This categorization of the dataset is conceptual: an encounter's stratum membership is only partially observable, and the share of cases in each stratum is unknown. The figure displays the stratum-specific potential outcomes that, when averaged, comprise various causal estimands. The TE, TE_S and TE_{ST} are based on $Y_i(1, M_i(1))$ and $Y_i(0, M_i(0))$; in contrast, the CDE_S is based on $Y_i(1, 1)$ and $Y_i(0, 1)$. The TE is the size-weighted average of all stratum-specific effects, whereas the TE_S, TE_{ST} and CDE_S reflect only those strata in which $M_i = 1$. Only the stratum potential outcomes boxed in red are observed in police data; as a result, causal quantities cannot be estimated without assumptions detailed in Section 3.1 and additional data on unobserved cases.



Electronic copy available at: https://ssrn.com/abstract=3336338

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—for example, it would be violated if minority officers discriminate against white civilians. However, previous studies on the racial composition of police forces suggest this is an unlikely scenario—prior results show that departments with more racial diversity behave similarly to those with higher shares of white officers (see Sklansky (2005) for a review of this issue).[7]

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

We theorize that for encounters during criminal events severe enough to warrant stopping a civilian 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.

Assumption 4 (Treatment Ignorability)**.**

(a) With respect to potential mediator $M_i(d) \perp\!\!\!\perp D_i|X_i$.

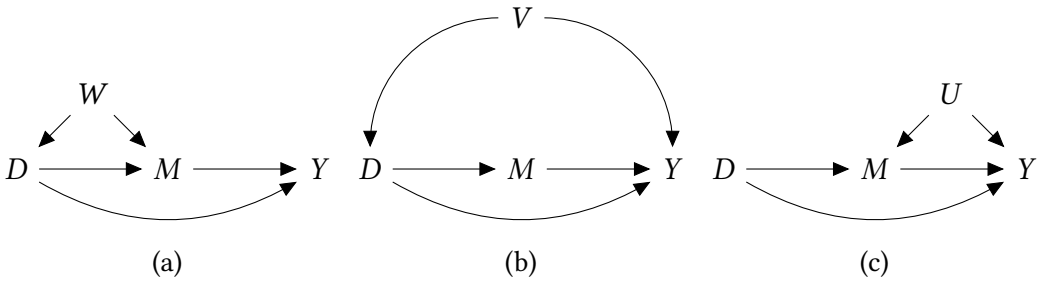
(b) With respect to potential outcomes: $Y_i(d, m) \perp\!\!\!\perp D_i|M_i(0) = m', M_i(1) = m'', X_i$.

This states that conditional on X_i, 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 non-stopped 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 fixed effects. Part 4(b) implies that, for

[7] Another violation would 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). As a robustness check to probe the validity of this assumption, Figure B4 shows our reanalysis of Fryer (2019) after dropping all stops based on suspicion of a drug transaction and shows substantively similar results.

Electronic copy available at: https://ssrn.com/abstract=3336338

Figure 3: **Violations of assumptions.** DAGs (a), (b), and (c) respectively illustrate the violation of Assumptions 4(a), 4(b), and 5. Note that the variable U depicted in DAG (c) is almost certain to exist in the policing context, and we do not advocate the use of Assumption 5.



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 paper 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.

3.2 Strong assumptions

We now discuss further assumptions necessary 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)**.** $Y_i(d, m) \perp\!\!\!\perp M_i(0) | D_i = d, M_i(1) = 1, X_i$

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

Electronic copy available at: https://ssrn.com/abstract=3336338

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 mindset-related variables that are causally upstream of stops, while taking care not to induce bias by conditioning on any that are downstream.

In Section 3.3, 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 in Section 3.3 that this assumption is implicitly made by all studies claiming to identify the average total 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; Goel, Rao and Shroff, 2016; Glaser, 2014; Lerman and Weaver, 2014*a*).[8] 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.

3.3 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 covari-

[8]To cite one of many striking examples, in 2011, the height of "stop, question and frisk" in New York City, there were more stops by the New York Police Department (NYPD) of African American males age 14–24 (close to 200,000 stops) than there were members of that group living in the city at the time (Gardiner, 2012). The comparable figure for white males in the same age group that year was about 25,000 stops.

Electronic copy available at: https://ssrn.com/abstract=3336338

ates). We also refer to this as the naïve estimator. First, we show that when there is any racial discrimination in detainment, selection on stops introduces unavoidable statistical bias in the TE_S even when a perfect set of observed covariates renders race ignorable. 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 post-treatment 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 TE_S, the TE_{ST}, and the CDE_S. We show that it is always negative, resulting in naïve estimates that downplay the extent of racially discriminatory police violence. In Section 5, 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 A1.1–A1.3 derive the bias of the naïve estimator with covariate adjustment. We begin by reexpressing the expectation of the naïve difference-in-means estimator, $\mathbb{E}[\hat{\Delta}]$, in terms of stratum mean potential outcomes. This estimator can be written as:

$$
\begin{aligned}
\mathbb{E}[\hat{\Delta}] &= \mathbb{E}[Y_i | D_i = 1, M_i = 1] - \mathbb{E}[Y_i | D_i = 0, M_i = 1] \\
&= \mathbb{E}[Y_i(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 1] \Pr(M_i(0) = 1 | D_i = 1, M_i(1) = 1) \\
&\quad + \mathbb{E}[Y_i(1,1) | D_i = 1, M_i(1) = 1, M_i(0) = 0] \Pr(M_i(0) = 0 | D_i = 1, M_i(1) = 1) \\
&\quad - \mathbb{E}[Y_i(0,1) | D_i = 0, M_i(1) = 1, M_i(0) = 1] \Pr(M_i(1) = 1 | D_i = 0, M_i(0) = 1) \\
&\quad - \mathbb{E}[Y_i(0,1) | D_i = 0, M_i(1) = 0, M_i(0) = 1] \Pr(M_i(1) = 0 | D_i = 0, M_i(0) = 1). \qquad (5)
\end{aligned}
$$

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 TE_S and TE_{ST} unless Assumption 6 is true and there are no racial stops. Next, we show that it is also biased for the CDE_S 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 Supplementary Materials A1.1, we derive the bias of $\hat{\Delta}$ when used to estimate TE_S

Electronic copy available at: https://ssrn.com/abstract=3336338

under the relatively plausible Assumptions 1–4. This bias can be written as:

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - \mathrm{TE}_S \\
&\quad = \big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad - \mathbb{E}[Y_i(1,1) - Y_i(0,0) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\quad - \big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1). \qquad (6)
\end{aligned}
$$

We offer several comments on Equation 6. The first term in the bias expression relates to heterogeneity in the total 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 $M_i(0) < M_i(1)$.[9] This bias term is guaranteed to be nonzero, even with a perfect set of controls that render D_i ignorable, as long as there exist any racially discriminatory stops of minority civilians. This is because in those encounters, a white civilian would never have been detained in the first place, and hence force would never have been used—that is, $\mathbb{E}[Y_i(0,0) | D_i = 1, M_i(1) = 1, M_i(0) = 0] = 0$. 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.[10] Therefore, bias cannot be eliminated unless assumption 6 (no racial stops) holds. 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. The structure of the bias when $\hat{\Delta}$ is used to estimate the TE_{ST} is simpler, but leads to substantively identical conclusions: the naïve estimator is biased unless there are no racial stops. In Supplementary Materials A1.2, we show that bias for the TE_{ST} is given by $\mathbb{E}[\hat{\Delta}] - \mathrm{TE}_{ST} = -\mathbb{E}[Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \Pr(M_i(0) = 0 | M_i(1) = 1)$. An important exception to these results arises when there is zero use of force against white civilians—but as we show in Section 5, this possibility is empirically falsified.

Can the naïve estimator be rehabilitated by simply redefining the quantity of interest?

[9] Note that $M_i(d)$ simplifies in Equation 6, because it is constant within principal strata.

[10] The sole exception is if both of the following conditions hold: (1) majority civilians are never subject to force, which can be tested using reported administrative data; and (2) the weaker but nonetheless implausible Assumption 5 holds.

Electronic copy available at: https://ssrn.com/abstract=3336338

In Supplementary Materials A1.3, we show that the answer is no. The bias of $\hat{\Delta}$ when used to estimate CDE_S is almost identical in structure and can be expressed as:

$$\begin{aligned}
&\mathbb{E}[\hat{\Delta}] - \text{CDE}_S \\
&\quad = \big(\mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad - \mathbb{E}[Y_i(1,1) - Y_i(0,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1) \Pr(D_i = 1 | M_i = 1) \\
&\quad - \big(\mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 1] \\
&\qquad - \mathbb{E}[Y_i(1,1) | M_i(1) = 1, M_i(0) = 0] \\
&\qquad \big) \Pr(M_i(0) = 0 | D_i = 1, M_i = 1). \qquad (7)
\end{aligned}$$

Bias for the CDE_S differs merely in that all individuals are held at $Y(d, 1)$ rather than $Y(d, M(d))$. Among the subset of reported encounters, this only affects the racially discriminatory stops, or those in which $M_i(0) < M_i(1)$. In other words, Equation 7 simply substitutes $Y_i(0, 1)$ for Equation 6's $Y_i(0, M_i(0)) = Y_i(0, 0)$ in this stratum. The chief implication is that the identifying assumptions for the CDE_S are weaker, but nonetheless implausible.

In the previous bias expression for the TE_S, 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, they would never have been stopped, and hence would not be subject to force. Changing the target estimand to the CDE_S conceptually sidesteps this specific issue by considering a different counterfactual, $Y_i(0, 1)$ instead of $Y_i(0, M_i(0))$, but does not eliminate bias. For encounters in which only minority civilians would be stopped—that is, encounters with $M_i(1) = 1$ and $M_i(0) = 0$—this new counterfactual represents an impossible cross-world scenario. The CDE_S asks whether force would have been used if officers were forced, against nature, to stop a white individual in this encounter as if they were a minority. Even so, bias remains unless officers are as violent toward minorities in always-stop encounters (where they are forced to intervene) as they are in racially discriminatory stops (where they are free to exercise discretion). In other words, for the naïve estimator to recover the CDE_S, Assumption 5 must hold.

For all quantities considered here, we further note that the bias can be rewritten as a series of comparisons between potential violence rates in always-stop and racial-stop

Electronic copy available at: https://ssrn.com/abstract=3336338

encounters. Under assumption 3, the sign of this bias can be shown to be strictly negative unless the relevant assumptions are met. Thus, regardless of whether the TE_S, the TE_{ST}, or the CDE_S is the target quantity, use of the observed difference in means will understate 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. Post-treatment conditioning induces bias unless additional assumptions hold, beyond the typical ones relating to treatment-outcome confounding.

4 Potential solutions

How should the analyst proceed in light of these results? We propose two approaches that eliminate the highly implausible assumptions outlined in Section 3.2, which are unstated but implicit in prior work. We caution that these solutions still rely on the weaker assumptions described in Section 3.1, although we argue that these are often reasonable in light of insights gained 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, Supplementary Materials A1.4 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) in Section 5.

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 by collecting additional data—which in some cases is already available—the prevalence of racially discriminatory stops and most racial effects of inter-

Electronic copy available at: https://ssrn.com/abstract=3336338

est can be point identified. In Section 6, we describe a feasible research design based on this approach in detail.

4.1 Bounds on the racial effect

Here, we derive large-sample nonparametric sharp bounds on the TE_S and TE_{ST}, 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 post-treatment conditioning, producing an informative interval of possible joint values for (1) the partially identified TE_S 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 ρ 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 TE_S)**.** *When D_i is ignorable, bounds on (TE_S, ρ) under Assumptions 1–4 are jointly given by*

$$\mathbb{E}[\hat{\Delta}] + \rho\, \mathbb{E}[Y_i | D_i = 0, M_i = 1]\, (1 - Pr(D_i = 0 | M_i = 1))$$

$$\leq \; TE_S \; \leq$$

$$\mathbb{E}[\hat{\Delta}] + \frac{\rho}{1-\rho} \left(\mathbb{E}[Y_i | D_i = 1, M_i = 1] - \max \left\{ 0, 1 + \frac{1}{\rho} \mathbb{E}[Y_i | D_i = 1, M_i = 1] - \frac{1}{\rho} \right\} \right) Pr(D_i = 0 | M_i = 1)$$

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

where $\hat{\Delta} = \overline{Y_i | D_i = 1, M_i = 1} - \overline{Y_i | D_i = 0, M_i = 1}$, and the (TE_{ST}, ρ) must similarly satisfy

$$TE_{ST} = \mathbb{E}[\hat{\Delta}] + \rho\, \mathbb{E}[Y_i | D_i = 0, M_i = 1]$$

We reformulate the bias in terms of the unobserved joint distribution of (1) 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, which then imply sharp bounds on the TE_S. A detailed proof is given in

Electronic copy available at: https://ssrn.com/abstract=3336338

Supplementary Materials A1.4 for the more general case in which D_i is ignorable only after conditioning on pre-stop covariates. In this case, the local total effect, TE_{Sx}, is first bounded by applying Proposition 1 within levels of X to obtain local bounds, $[\underline{\text{TE}}_{Sx}, \overline{\text{TE}}_{Sx}]$. These are then straightforwardly reaggregated to obtain bounds on the average total effect, $\left[\sum_x \underline{\text{TE}}_{Sx} \Pr(X_i = x|M_i = 1),\ \sum_x \overline{\text{TE}}_{Sx} \Pr(X_i = x|M_i = 1)\right]$. In Supplementary Materials A1.5, 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 ρ. (Section 4.2 describes an alternative approach for estimating the local proportion of racial stops directly.) This is because without additional data beyond civilian race, use of force, or even pre-stop covariates, police administrative records alone are virtually uninformative about the range of ρ: any value in $[0, 1)$ could produce the observed data,[11] although Proposition 1 shows that each possible ρ value has differing implications for the set of possible racial effects.

4.2 Point identification of the TE given additional data

The TE 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. Section 6 proposes an alternative design in which this data is collected from passive instruments such as traffic cameras or police body-worn cameras. Where such a design is infeasible (for example, 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. Where record linkage is impossible, surveys will also work under stronger assumptions, discussed below.

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

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

[11] If all stops were racially discriminatory, then we would observe no white stops.

Electronic copy available at: https://ssrn.com/abstract=3336338

Intuitively, the proof breaks the TE 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 non-stops allows the researcher to construct a contingency table representing the joint distribution of race and detainment. As part of the proof in Supplementary Materials A1.6, 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 plates databases); those that are unlinkable can be presumed unreported. If neither of these techniques are feasible, alternatives such as the Police-Public Contact Survey offer yet another avenue for estimating stop rates by group. However, one challenge in using surveys that sample on individuals, rather than encounters, is that some civilians may have a higher frequency of police interaction. This should in principle be incorporated as an additional respondent weight, but in practice, encounter frequency is difficult to accurately proxy. Reliance on surveys also requires the usual assumptions about survey representativeness and the truthfulness of self-reports.

After recovering principal strata sizes, we then proceed by noting that minority outcomes in reported administrative data are in fact a mixture of $Y_i(1, M_i(1))$ from both always-stop and racial-stop strata in precisely the required proportions; that reported white outcomes correspond to $Y_i(0, M_i(0))$ from the always-stop stratum; and that $Y_i(0, M_i(0))$ is known to be zero among the racial-stop stratum under Assumption 1. From this, the TE can then be reconstructed.

5 Reanalysis of Fryer (2019)

The results above show 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 which civilians 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 sub-lethal force using the NYPD's

Electronic copy available at: https://ssrn.com/abstract=3336338

"Stop, Question and Frisk" (SQF) database (2003-2013).[12] 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 5 to predict the severity of force applied by police, as well as covariate-adjusted naïve models analogous to those we consider in Appendices A1.1–A1.3. 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)[13]—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 non-zero force below the threshold of interest were dropped—a severe case of researcher-induced 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. In order to present the most defensible results possible, for these outcomes, we therefore 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 en-

[12]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`.

[13]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 year fixed effects. See Figure 1 caption in Fryer (2019).

Electronic copy available at: https://ssrn.com/abstract=3336338

counters are assigned a value of 0 (including those with lower levels of force, which are now retained). Section B.1 in the Supplementary Materials contains an extended discussion of the issue; a comparison of the original, replicated, and corrected results; and a demonstration of the implications for statistical significance of the original estimates.

Based on the discussion in both Fryer (2019) and Fryer (2018), we interpret the published results as estimates of the TE_S: "the difference in Y that can be attributed to an individual's race," (Fryer, 2018, 2), conditional on a recorded interaction with police (that is, conditional on $M_i = 1$). We note that of the other quantities considered in this paper, the TE cannot be estimated without information on unreported encounters, and the CDE_S cannot be computed without strong assumptions about cross-world potential outcomes that are unverifiable, even in experimental settings. For these reasons, we focus on the TE_S and TE_{ST} in this reanalysis.[14]

Figure 4 depicts bounds on the TE_S when the binary outcome is any use of force, including the lowest recorded value of physically handling a civilian.[15] Importantly, this specific outcome is unaffected by the coding issue discussed above. (In Figures B2 and B3, 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 TE_{ST} (dashed lines). As the plots show, the range of possible TE_S and TE_{ST} values varies strongly with the severity of discrimination in stops.

In Equation 6, we demonstrated that 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 ρ are implausibly large. To understand the range of empirically plausible values, we turn to two prior studies that

[14]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 TE_S estimate: black versus white, or Hispanic versus white). To accommodate this, in reported black TE_S and TE_{ST} 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 $D_i = 2, 3, \ldots$. Proposition 1 and its covariate-adjusted counterpart in Supplementary Materials A1.4 can then be applied directly. The chief implication of this formulation is (1) a different average value for $Y_i(d, 1)$ 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 Supplementary Materials 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 subsetted data.

[15]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 involve officers laying hands on suspects.

Electronic copy available at: https://ssrn.com/abstract=3336338

use very different analytic approaches to shed light on the degree of racial bias in the decision to detain civilians. Using the SQF data and controlling for precinct, suspected crime, and prior local arrest rates by race, Gelman, Fagan and Kiss (2007) produce estimates that—by our calculations—imply 32% of black-civilian stops made by the NYPD could not be explained even by differential criminality between racial groups of suspects, as proxied by prior arrest rates.[16] Their analyses are run separately by precinct and crime type; for simplicity, we take the weighted average of racial-stop proportions. This analytic approach most likely underestimates the proportion of racially discriminatory stops—the number of prior arrests in a precinct and racial group is not a direct measure of criminality, but is itself contaminated by discrimination in previous detainments and arrests. We therefore regard the value of ρ implied by 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 Todd, 2001). Using a flexible logistic regression to adjust for a vast array of indicators visible to officers pre-stop, the study shows that white hit rates exceeded those of "similarly situated" black civilians. We show in our Supplementary Materials (section A1.7) that the difference in hit rates implies a minimum proportion of racial stops, and therefore also implies a conservative estimate of ρ.[17] The corresponding values of ρ 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 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 the research

[16] Based on SQF data from 1998-1999, 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 $e^{\mu+\alpha_{\text{race}}}$ 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 $1 - e^{\alpha_{\text{white}} - \alpha_{\text{black}}}$. 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 ρ for Hispanic civilians implied by Gelman, Fagan and Kiss (2007) is slightly higher, at 0.35.

[17] Using SQF data from 2008–2012, Goel, Rao and Shroff (2016) estimate a hit rate of 3.8% for white suspects and 2.5% for black suspects (379), which implies that ρ is at least 0.34.

Electronic copy available at: https://ssrn.com/abstract=3336338

design presented in Section 6 offers an alternative approach for obtaining better estimates of racially discriminatory stopping.

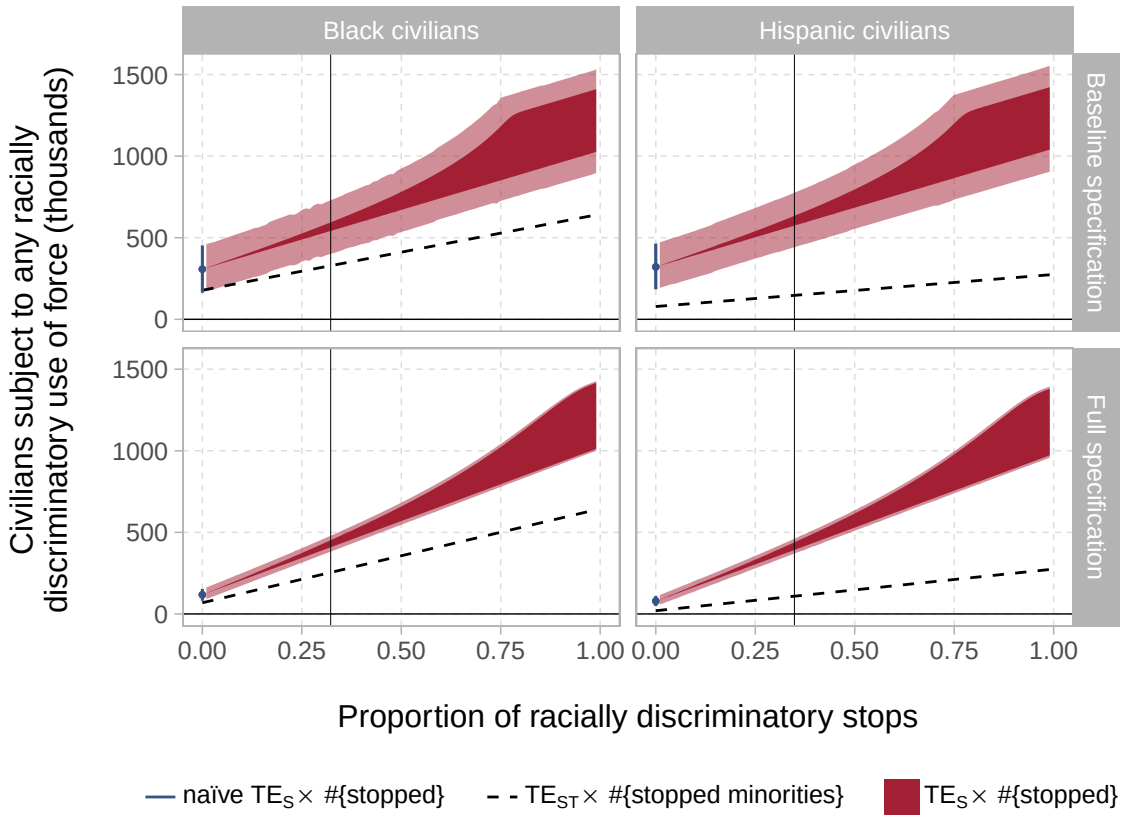
Figure 4 demonstrates that strong negative bias in the naïve estimator paints a wildly misleading portrait of police use of force. We turn first to estimates of the TE_S using the main specification, which adjusts for a battery of covariates. The naïve estimator (which assumes no racial bias in police stops) suggests that encounters with black (Hispanic) suspects are predicted to exhibit an additional 3.9 (0.33) instances of handcuffing per 1,000 encounters, versus if white civilians had appeared in those same encounters. 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 (40).

To characterize bias in estimates of the TE_{ST}, we again use the conservative racial stopping estimate from Gelman, Fagan and Kiss (2007) to correct the naïve estimate. Again, the naïve approach substantially understates racially discriminatory police violence, suggesting that there were 88,000 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 362,000, meaning the naïve approach masks 274,000 such incidents. Similarly, the naïve approach indicates roughly 1,220 racially discriminatory instances in which officers pointed a weapon at a black or Hispanic civilian, whereas the bias-corrected TE_{ST} shows the true number is almost five times as large.

To see how this statistical bias affects estimates relating to different levels of force, Table 1 presents naïve estimates alongside TE_S 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 22—nine times as large.

Electronic copy available at: https://ssrn.com/abstract=3336338

Figure 4: **Bounds for racially discriminatory use of force, any severity.** These plots present the TE_S (TE_{ST}) 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 use of force if black civilians appeared in each encounter (each black encounter), versus white civilians; the right panels show the same quantities for Hispanic civilians. Blue points (error bars) denote the naïve estimator (95% confidence intervals), which, conditional on the typical selection on observables assumption, is unbiased for the TE_S if there are no discriminatory stops of minority civilians (corresponding to zero on the x-axis). The dark (light) regions represent the range of possible values (95% CI) for the TE_S and proportion of discriminatory stops in reported data jointly, per Proposition 1. The vertical line corresponds to an estimate of the proportion of discriminatory stops from Gelman, Fagan and Kiss (2007), suggesting a plausible value for this unobservable parameter. The top (bottom) panels present bounds based on a model with no controls (the main specification, adjusting for a wide range of covariates).



Electronic copy available at: https://ssrn.com/abstract=3336338

Table 1: **TE_S, by severity of force and minority group.** Excess use of force used against minority civilians (versus white civilians) per 1,000 encounters. Bounds intervals indicate the range of possible TE_S values when the unknown proportion of discriminatory stops is approximated with the conservative estimate from Gelman, Fagan and Kiss (2007). Estimates in bold. 95% confidence intervals in italics.

	TE_S for encounters with black civilians (vs. white)			
	No covariates		Full specification	
Minimum force	bounds	naïve	bounds	naïve
Use of hands	**(112.66, 124.59)**	**61.69**	**(86.99, 96.74)**	**23.53**
	(84.6, 151.84)	*(32.89, 90.63)*	*(81.7, 102.15)*	*(16.41, 30.61)*
Push to wall	**(24.15, 27.75)**	**4.2**	**(26.48, 30.21)**	**6.67**
	(15.5, 37.35)	*(-5.29, 14.02)*	*(24.29, 32.38)*	*(3.73, 9.52)*
Use of handcuffs	**(14.6, 16.92)**	**1.32**	**(16.56, 19.02)**	**3.9**
	(9.45, 22.61)	*(-4.83, 7.53)*	*(15.05, 20.55)*	*(1.87, 5.88)*
Draw weapon	**(4.52, 5.14)**	**1.26**	**(4.71, 5.35)**	**1.46**
	(3.13, 6.67)	*(-0.33, 2.83)*	*(4.22, 5.86)*	*(0.79, 2.13)*
Push to ground	**(4.04, 4.58)**	**1.22**	**(4.11, 4.66)**	**1.26**
	(2.79, 5.97)	*(-0.21, 2.66)*	*(3.68, 5.09)*	*(0.68, 1.82)*
Point weapon	**(1.49, 1.7)**	**0.36**	**(1.64, 1.86)**	**0.55**
	(0.96, 2.29)	*(-0.29, 1)*	*(1.37, 2.13)*	*(0.18, 0.91)*
Baton or pepper spray	**(0.17, 0.19)**	**0.08**	**(0.17, 0.19)**	**0.07**
	(0.1, 0.26)	*(-0.01, 0.15)*	*(0.12, 0.24)*	*(-0.01, 0.14)*

	TE_S for encounters with Hispanic civilians (vs. white)			
	No covariates		Full specification	
	bounds	naïve	bounds	naïve
Use of hands	**(115.44, 127.53)**	**64.48**	**(79, 88.37)**	**15.53**
	(88.94, 155.96)	*(37.06, 92.91)*	*(74.53, 92.81)*	*(9.61, 21.35)*
Push to wall	**(26.41, 30.14)**	**6.46**	**(22.19, 25.7)**	**2.39**
	(19.54, 37.79)	*(-1.12, 14.26)*	*(20.19, 27.66)*	*(-0.26, 4.91)*
Use of handcuffs	**(12.54, 14.76)**	**-0.74**	**(13, 15.25)**	**0.33**
	(9.1, 18.24)	*(-5.27, 3.57)*	*(11.72, 16.56)*	*(-1.4, 2.03)*
Draw weapon	**(3.42, 3.98)**	**0.16**	**(3.11, 3.66)**	**-0.14**
	(2.41, 5.08)	*(-1.04, 1.33)*	*(2.63, 4.16)*	*(-0.77, 0.48)*
Push to ground	**(3.11, 3.6)**	**0.29**	**(2.72, 3.19)**	**-0.13**
	(2.18, 4.61)	*(-0.83, 1.37)*	*(2.3, 3.61)*	*(-0.68, 0.39)*
Point weapon	**(0.73, 0.9)**	**-0.41**	**(0.8, 0.98)**	**-0.28**
	(0.32, 1.29)	*(-0.94, 0.08)*	*(0.53, 1.25)*	*(-0.64, 0.06)*
Baton or pepper spray	**(0.05, 0.06)**	**-0.05**	**(0.05, 0.07)**	**-0.05**
	(-0.01, 0.12)	*(-0.13, 0.02)*	*(0, 0.12)*	*(-0.12, 0.02)*

Electronic copy available at: https://ssrn.com/abstract=3336338

6 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 post-treatment conditioning can inadvertently mask racially biased policing. Our results suggest the body of evidence on this topic that relies on police administrative data may be largely uninformative or even misleading. While an improvement, our bias correction and bounding techniques 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 post-treatment conditioning. In what follows, we detail a feasible research design that addresses these concerns.

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

Recall that a key problem in the typical study of police administrative data is the unobservability of those encounters that do not generate police reports. However, given the prevalence of highway speed cameras across police jurisdictions, it is entirely feasible to collect data on every passing car (or, a random sample of passing cars) whether or not police pulled the car over and recorded the stop. This mode of data collection has already been utilized in prior work (Kocieniewski, 2002; Lange, Johnson and Voas, 2005), though in those studies camera data on individual motorists were not linked to data on policing outcomes, as we propose below.

Given a large random sample of passing cars generated by highway speed cameras, analysts could use these 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 obser-

Electronic copy available at: https://ssrn.com/abstract=3336338

vational data, analysts must still make some version of Assumption 4(b), (no treatment-outcome confounding conditional on observable covariates X)—but in this case, the standard "treatment selection on observables" plausibly holds, because virtually all pre-stop data available to an officer are in fact also observable using camera footage. With these merged administrative records, analysts could credibly measure this "complete" set of control variables.[18] 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 TE 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 TE_{ST} is similarly point identified, because the proportion of racial stops can be calculated and corrected for. However, the TE_{S} remains partially identified—the quantity can be bounded, as we show above, but not precisely estimated. And as Figure 5 makes clear, the CDE_{S} 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., temporal variation in officer temperament), which likely influences both stopping decisions and subsequent treatment of civilians. For example, a driver's offensive bumper sticker in a given encounter (unobservable to the analyst) could anger an officer, thereby leading to both a stop and the use of violence. Because such elements would not appear in administrative data sets, and would not be captured by officer fixed effects (in the event they were available), mediator-outcome confounding would persist.

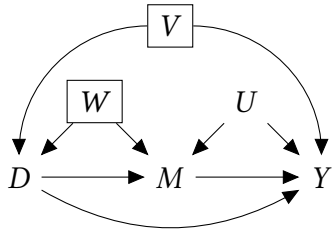
7 Conclusion

With the release of large and granular data on police-citizen 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 post-treatment bias in this setting (Fryer, 2018), the issue has not been adequately addressed, and studies in this area

[18]This approach is akin to the design of Hainmueller and Hangartner (2013), another rare instance in which the analyst could claim to measure all relevant covariates in an observational setting. In that study, citizens made judgments about individuals applying for citizenship in Switzerland. Because all information on potential citizens was contained on a flier distributed by the government, the authors could credibly account for all possible factors that contributed to the average citizen's judgment of applicants.

Electronic copy available at: https://ssrn.com/abstract=3336338

Figure 5: **Traffic stop design.** The DAG illustrates potential back-door paths for stops (through W, e.g., heavily policed neighborhoods) and for the use of force (through V, e.g., car registrant has warrant for arrest) that may correlate with the presence of minority drivers. These are blocked (boxed) by conditioning on pre-stop variables, including license plates as well as administrative records that can be linked through them. Many mediator-outcome confounders (U) cannot be blocked but do not pose a threat to inference for the TE or TE_S.



have left ambiguous which 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 that policing scholars need to devote substantial effort to develop research designs that can sidestep the threat of post-treatment conditioning rather than proceeding in the face of this threat and simply hoping for the best.

While we are optimistic that alternative designs can be pursued, we are under no illusions that strategies aimed at eliminating this one source of bias will remove concerns over others (discussed in the previous section). Our research design suggestions may also limit the outcomes that are feasible to study. For example, rare events like 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 in order to generate more reliable estimates may be preferable, especially since policy reforms will often 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 tradeoff in scope may be of only marginal concern (Samii, 2016).

Regardless of which approach scholars pursue, this paper highlights the need for further careful research into the first stage of police-citizen interactions—that is, the process

Electronic copy available at: https://ssrn.com/abstract=3336338

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-citizen interactions, but to craft effective policy reforms. If racial bias is concentrated in the initial stage of contact, reforms focused on reducing unnecessary police-citizen 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, the greatest improvements may emanate from a wholly different reform strategy. Without serious attention to the threat of post-treatment bias during the study design phase, the potential for data-driven reforms to promote equitable policing is unlikely to be realized.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects." *Biometrics* 110(3):512–529.

Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.

Allen, David. 1982. "Police Supervision on the Street: An Analysis of Supervisor/Officer Interaction During the Shift." *Journal of Criminal Justice* 10(2):91–109.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434):444–455.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Antonovics, Kate and Brian G. Knight. 2009. "A new look at racial profiling: Evidence from the Boston Police Department." *The Review of Economics and Statistics* 91(1).

Anwar, Shamena and Hanming Fang. 2006. "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence." *The Review of Economic Studies* 96(1):127–151.

Arrow, Kenneth J. 1972. Models of Job Discrimination. In *Racial Discrimination in Economic Life*, ed. A.H. Pascal. Lexington, Mass.: D.C. Heath pp. 83–102.

Electronic copy available at: https://ssrn.com/abstract=3336338

Arrow, Kenneth J. 1998. "What has economics to say about racial discrimination?" *Journal of economic perspectives* 12(2):91–100.

Balke, Alexander and Judea Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92(439):1171–1176.

Baumgartner, Frank R., Derek A. Epp, Kelsey Shoub and Bayard Love. 2017. "Targeting young men of color for search and arrest during traffic stops: evidence from North Carolina, 2002-2013." *Politics, Groups, and Identities* 5(1):107–131.

Becker, Gary. 1971. *The economics of discrimination.* University of Chicago Press.

Blackwell, Matthew. 2013. "A framework for dynamic causal inference in political science." *American Journal of Political Science* 57:504–520.

Eberhardt, Jennifer, Phillip Atiba Goff, Valerie J. Purdie and Paul G. Davies. 2004. "Seeing Black: Race, Crime, and Visual Processing." *Journal of Personality and Social Psychology* 87(6):876–893.

Elwert, Felix and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *The Annual Review of Sociology* 40:31–53.

Fisher, Marc and Peter Hermann. 2015. "Did the McKinney, Texas, Police Officer Know He Was Being Recorded?" *The Washington Post* .

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

Fridell, Lorie A. 2017. "Explaining the Disparity in Results Across Studies Assessing Racial Disparity in Police Use of Force: A Research Note." *Journal of economic perspectives* 42(3):502–513.

Fryer, Roland G. 2018. "Reconciling Results on Racial Differences in Police Shootings." *American Economic Review (Papers and Proceedings)* .

Fryer, Roland G. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* .

Electronic copy available at: https://ssrn.com/abstract=3336338

Gardiner, Sean. 2012. "Report Finds Stop-and-Frisk Focused on Black Youth." *The Wall Street Journal* . https://blogs.wsj.com/metropolis/2012/05/09/report-finds-stop-and-frisk-focused-on-black-youth/.

Gelman, Andrew, Jeffrey Fagan and Alex Kiss. 2007. "An Analysis of the New York City Police Department's "Stop-and-Frisk'" Policy in the Context of Claims of Racial Bias." *Journal of the American Statistical Association* 102(429):813–823.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Glaser, Jack. 2014. *Suspect Race: Causes and Consequences of Racial Profiling*. Oxford University Press.

Goel, Sharad, Justin M. Rao and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-And-Frisk Policy." *Annals of Applied Statistics* 10(1):365–394.

Greiner, James D. and Donald B. Rubin. 2011. "Causal effects of perceived immutable characteristics." *Review of Economics and Statistics* 93:775–785.

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

Hainmueller, Jens and Dominik Hangartner. 2013. "Who gets a Swiss passport? A natural experiment in immigrant discrimination." *American Political Science Review* 107(1):159–187.

Heckman, James J. 1977. "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)." *NBER Working Paper* (No. 172).

Hernán, Miguel A. 2016. "Does water kill? A call for less casual causal inferences." *Annals of Epidemiology* 26(10):674–680.

Hernán, Miguel, Sonia Hernández-Diáz and James Robins. 2004. "A Structural Approach to Selection Bias." *Epidemiology* 15:615–625.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 81(396):945–60.

Electronic copy available at: https://ssrn.com/abstract=3336338

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765–789.

Iyengar, Shanto. 1994. *Is anyone responsible?: How television frames political issues*. Chicago: University of Chicago Press.

Knowles, J., N. Perisco and P. Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* pp. 203–229.

Knox, Dean, Teppei Yamamoto, Matthew A. Baum and Adam J. Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *Journal of the American Statistical Association* .

Kocieniewski, David. 2002. "Study Suggests Racial Gap In Speeding In New Jersey." *The New York Times* . https://www.nytimes.com/2002/03/21/nyregion/study-suggests-racial-gap-in-speeding-in-new-jersey.html.

Lange, James E., Mark B. Johnson and Robert B. Voas. 2005. "Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike." *Justice Quarterly* 22(2):193–223.

Lee, David S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76(3):1071–1102.

Lerman, Amy and Vesla Weaver. 2014*a*. *Arresting Citizenship: The Democratic Consequences of American Crime Control*. University of Chicago Press.

Lerman, Amy and Vesla Weaver. 2014*b*. "Staying Out of Sight? Concentrated Policing and Local Political Action." *The Annals of the American Academy of Political and Social Science* 651(1):202–219.

Manski, C. F. 1995. *Identification Problems in the Social Sciences*. Harvard University Press.

Mummolo, Jonathan. 2018. "Modern Police Tactics, Police-Citizen Interactions and the Prospects for Reform." *Journal Of Politics* 80(1):1–15.

Nix, Justin, Bradley A. Campbell, Edward H. Byers and Geoffrey P. Alpert. 2017. "A Bird's Eye View of Civilians Killed by Police in 2015 Further Evidence of Implicit Bias." *Criminology & Public Policy* 16(1):309–340.

Electronic copy available at: https://ssrn.com/abstract=3336338

Nyhan, Brendan, Christopher Skovron and Rocío Titiunik. 2017. "Differential Registration Bias in Voter File Data: A Sensitivity Analysis Approach." *American Journal of Political Science* 61:744–760.

NYPD. 2014. Use of Force Report, 2017. Technical report. `https://www1.nyc.gov/assets/nypd/downloads/pdf/use-of-force/use-of-force-2017.pdf`.

Orne, Martin T. 1962. "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American Psychologist* 17(1):776–783.

Pearl, Judea. 2001. "Direct and indirect effects." *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* pp. 411–420.

Pearl, Judea. 2018. "Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes." *Journal of Causal Inference* 6(2):1–7.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62(1):659–661.

Ridgeway, Greg. 2006. "Assessing the Effect of Race Bias in Post-traffic Stop Outcomes Using Propnesity Scores." *Journal of Quantitative Criminology* pp. 1–29.

Ridgeway, Greg and John MacDonald. 2010. *Race, Ethnicity, and Policing: New and Essential Readings*. NYU Press chapter Methods for Assessing Racially Biased Policing.

Robins, James M., Miguel A. Hernán and Babette Brumback. 2000. "Marginal structural models and causal inference in Epidemiology." *Epidemiology* 11(5):550–60.

Robins, J.M. and S. Greenland. 1992. "Identifiability and exchangeability for direct and indirect effects." *Epidemiology* 3(2):143–155.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society* 147(5):656–666.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66:688–701.

Electronic copy available at: https://ssrn.com/abstract=3336338

Rubin, Donald B. 1990. "Formal mode of statistical inference for causal effects." *Journal of statistical planning and inference* 25(3):279–292.

Rubin, Donald B. 2000. "Causal Inference Without Counterfactuals: Comment." *Journal of the American Statistical Association* 95(450):435–438.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3):941–955.

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

Sklansky, David Alan. 2005. "Not your father's police department: Making sense of the new demographics of law enforcement." *J. Crim. L. & Criminology* 96:1209–1244.

Smith, Douglas A., Christy A. Visher, and Laura A. Davidson. 1984. "Equity and discretionary justice: The influence of race on police arrest decisions." *J. Crim. L. & Criminology* 75.

VanderWeele, Tyler J. 2009. "Marginal structural models for the estimation of direct and indirect effects." *Epidemiology* pp. 18–26.

VanderWeele, Tyler J. 2011. "Principal stratification–uses and limitations." *The International Journal of Biostatistics* 7(1):1–14.

West, Jeremy. 2018. "Racial Bias in Police Investigations.". Working Paper `https://people.ucsc.edu/~jwest1/articles/West_RacialBiasPolice.pdf`.

Yamamoto, Teppei. 2012. "Understanding the Past: Statistical Analysis of Causal Attribution." *American Journal of Political Science* 56(1):237–256.

Zhang, Junni L. and Donald B. Rubin. 2003. "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death"." *Journal of Educational and Behavioral Statistics* 28(4):353–368.

Electronic copy available at: https://ssrn.com/abstract=3336338

The Bias is Built In: How Administrative Records Mask Racially Biased Policing

Supplementary Materials

Electronic copy available at: https://ssrn.com/abstract=3336338

A1 Detailed proofs

A1.1 Bias for TE_S

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

$$
\begin{aligned}
&\mathbb{E}[\hat{\Delta}_x] - \text{TE}_{Sx} \\
&\quad = (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}]) \\
&\qquad - (\mathbb{E}[Y_i(1, M_i(1))|M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|M_i = 1, X_i = x]) \\
&\quad = \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x] \\
&\quad = \mathbb{E}[Y_i|D_i = 1, M_i(D_i) = 1, X_i = x] - \mathbb{E}[Y_i(1, M_i(1))|M_i(D_i) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i|D_i = 0, M_i(D_i) = 1, X_i = x] + \mathbb{E}[Y_i(0, M_i(0))|M_i(D_i) = 1, X_i = x] \\
&\quad = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\qquad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\qquad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\qquad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&\qquad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x]\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

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

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

Electronic copy available at: https://ssrn.com/abstract=3336338

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)$$

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

$$\cancel{\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)}\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)$$

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

$$\cancel{\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x)}$$

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

$$\cancel{\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x)}\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\cancel{\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)}$$

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

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

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

$$\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

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

$$\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x)$$

Electronic copy available at: https://ssrn.com/abstract=3336338

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

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

substituting potential mediators based on principal strata,

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

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

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

Electronic copy available at: https://ssrn.com/abstract=3336338

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

invoking assumption 4(b) (treatment ignorability),

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

which is Equation 6.

A1.2 Bias for TE_{ST}

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, $\text{TE}_{STx} = \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1, X_i = x]$. Again, overall bias is found by the weighted average of local biases, $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{TE}_{STx}\right) \Pr(X_i = x|D_i = 1, M_i = 1)$.

$$
\begin{aligned}
\mathbb{E}[\hat{\Delta}_x] - \text{TE}_{STx} &= (\mathbb{E}[\overline{Y_i|D_i = 1, M_i = 1, X_i = x} - \overline{Y_i|D_i = 0, M_i = 1, X_i = x}]) \\
&\quad - (\mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i = 1, X_i = x]) \\
&= \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(1, M_i(1))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&= - \mathbb{E}[Y_i(0, M_i(0))|D_i = 0, M_i(D_i) = 1, X_i = x] \\
&\quad + \mathbb{E}[Y_i(0, M_i(0))|D_i = 1, M_i(D_i) = 1, X_i = x] \\
&= - \mathbb{E}[Y_i(0,1)|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(1) = 1|D_i = 0, M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0,1)|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x]\Pr(M_i(1) = 0|D_i = 0, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0,1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 1|D_i = 1, M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0,0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)
\end{aligned}
$$

Electronic copy available at: https://ssrn.com/abstract=3336338

by mediator monotonicity

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

by mandatory reporting

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

by treatment ignorability

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

A1.3 Bias for CDE_S

We now turn to the question of whether the local difference in means is biased for the local controlled direct effect, $\text{CDE}_{Sx} = \mathbb{E}[Y_i(1,1)|M_i=1, X_i=x] - \mathbb{E}[Y_i(0,1)|M_i=1, X_i=x]$. The derivation is almost identical to that of the TE_{Sx}, differing only in that all individuals are held at $M_i = 1$ instead of being allowed to vary with civilian race, $M_i(D_i)$. Bias for CDE_S is then given by the weighted average of local biases, $\sum_x \left(\mathbb{E}[\hat{\Delta}_x] - \text{CDE}_{Sx}\right)\Pr(X_i=x|M_i=1)$.

$$
\begin{aligned}
\mathbb{E}[\hat{\Delta}_x] - \text{CDE}_{Sx} = & (\mathbb{E}[\overline{Y_i|D_i=1, M_i=1, X_i=x} - \overline{Y_i|D_i=0, M_i=1, X_i=x}]) \\
& -(\mathbb{E}[Y_i(1,1)|M_i=1, X_i=x] - \mathbb{E}[Y_i(0,1)|M_i=1, X_i=x]) \\
= & \mathbb{E}[Y_i|D_i=1, M_i(D_i)=1, X_i=x] - \mathbb{E}[Y_i|D_i=0, M_i(D_i)=1, X_i=x] \\
& -\mathbb{E}[Y_i(1,1)|M_i(D_i)=1, X_i=x] + \mathbb{E}[Y_i(0,1)|M_i(D_i)=1, X_i=x] \\
= & \mathbb{E}[Y_i(1,1)-Y_i(0,1)|D_i=0, M_i(1)=1, M_i(0)=1, X_i=x]\Pr(D_i=1|M_i(D_i)=1, X_i=x) \\
& -\mathbb{E}[Y_i(1,1)-Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=1, X_i=x] \\
& \qquad \Pr(M_i(0)=1|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=1|M_i(D_i)=1, X_i=x) \\
& -\mathbb{E}[Y_i(1,1)-Y_i(0,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x] \\
& \qquad \Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=1|M_i(D_i)=1, X_i=x) \\
& +\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=1, X_i=x] \\
& -\mathbb{E}[Y_i(1,1)|D_i=0, M_i(1)=1, M_i(0)=1, X_i=x]
\end{aligned}
$$

Electronic copy available at: https://ssrn.com/abstract=3336338

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

under assumption 4(b),

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

which reproduces Equation 7.

A1.4 Nonparametric sharp bounds for TE_S

In this section, we derive nonparametric sharp bounds for the TE_{Sx}. 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 Equation 6 (within levels of X) yields

$$\begin{aligned}
\text{TE}_{Sx} &= \mathbb{E}[\hat{\Delta}_x] \\
&\quad + \mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1|M_i(D_i) = 1, X_i = x)) \\
&\quad - \mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)(1 - \Pr(D_i = 1|M_i(D_i) = 1, X_i = x)) \\
&\quad + \mathbb{E}[Y_i(0,1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&= \mathbb{E}[\hat{\Delta}_x] \\
&\quad + \frac{\mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad - \frac{\mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)^2 \qquad (1) \\
&\qquad \Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 0, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i(D_i) = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(0,1)|M_i(1) = 1, M_i(0) = 1, X_i = x]\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 1|M_i(D_i) = 1, X_i = x) \\
&= \mathbb{E}[\hat{\Delta}_x] \\
&\quad + \frac{\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)\Pr(D_i = 0|M_i = 1, X_i = x) \\
&\quad - \frac{\mathbb{E}[Y_i(1,1)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x]}{\Pr(M_i(0) = 1|D_i = 1, M_i(1) = 1, X_i = x)}\Pr(M_i(0) = 0|D_i = 1, M_i(D_i) = 1, X_i = x)^2 \qquad (2) \\
&\qquad \Pr(D_i = 0|M_i = 1, X_i = x)
\end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=3336338

$$- \mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=0|M_i=1, X_i=x)$$
$$+ \mathbb{E}[Y_i|D_i=0, M_i=1, X_i=x]\Pr(M_i(0)=0|D_i=1, M_i(D_i)=1, X_i=x)\Pr(D_i=1|M_i=1, X_i=x) \quad (3)$$

We then construct bounds on $\mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x]$ 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 $M_i(0)$.

$$\begin{gathered}\frac{\max\{0, \Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x) + \mathbb{E}[Y_i|D_i=1, M_i=1, X_i=x]-1\}}{\Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x)} \\ \leq \ \mathbb{E}[Y_i(1,1)|D_i=1, M_i(1)=1, M_i(0)=0, X_i=x] \ \leq \\ \frac{\min\{\Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x), \mathbb{E}[Y_i|D_i=1, M_i=1, X_i=x]\}}{\Pr(M_i(0)=0|D_i=1, M_i(1)=1, X_i=x)}\end{gathered} \quad (4)$$

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 TE_S alone are obtained by substituting Equation 4 into Equation 3 and setting the proportion of racial stops to unity. The bivariate bounds define the region in which (TE_S, ρ_x) pairs are consistent with the observed data. When ρ_x is set to zero or one, these respectively recover the difference in reported means and the marginal upper bounds on TE_S. For $\rho_x \in (0,1)$,

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

which reduces to Proposition 1 in the no-covariate case. Otherwise, bounds on TE_S are given by $\sum_x \underline{\text{TE}}_{Sx}\Pr(X_i=x|M_i=1) \leq \text{TE}_S \leq \sum_x \overline{\text{TE}}_{Sx}\Pr(X_i=x|M_i=1)$, where $\underline{\text{TE}}_{Sx}$ ($\overline{\text{TE}}_{Sx}$) denote the

Electronic copy available at: https://ssrn.com/abstract=3336338

lower (upper) bounds on the local average total effect.

Finally, we note that per Equation 3, the TE_{STx} can be written

$$\begin{aligned}
\text{TE}_{STx} &= \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x) \\
&\quad + \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x) \\
&= \mathbb{E}[Y_i(1, 1)|D_i = 1, M_i = 1, X_i = x] \\
&\quad - \mathbb{E}[Y_i(0, 1)|D_i = 1, M_i(1) = 1, M_i(0) = 1, X_i = x] \Pr(M_i(0) = 1|D_i = 1, M_i = 1, X_i = x) \\
&\quad - \mathbb{E}[Y_i(0, 0)|D_i = 1, M_i(1) = 1, M_i(0) = 0, X_i = x] \Pr(M_i(0) = 0|D_i = 1, M_i = 1, X_i = x)
\end{aligned}$$

under Assumption 1,

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

and under Assumption 4,

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

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

$$\begin{aligned}
\text{TE}_{ST} &= \sum_x \left(\mathbb{E}[Y_i|D_i = 1, M_i = 1, X_i = x] - \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x] + \rho_x \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]\right) \\
&= \sum_x \left(\mathbb{E}[\Delta_x] + \rho_x \mathbb{E}[Y_i|D_i = 0, M_i = 1, X_i = x]\right).
\end{aligned}$$

A1.5 Uncertainty of TE 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, $\mathbb{E}[Y_i|D_i = d, M_i = 1, X_i = x]$. The asymptotic distribution of the estimated lower and upper bounds endpoints, $(\hat{\underline{\text{TE}}}_S, \hat{\overline{\text{TE}}}_S)$, then follows directly from the asymptotic joint distribution of $\hat{\mathbb{E}}[Y_i|D_i = d, M_i = 1, X_i = x]$ 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 θ^*, the corresponding bounds endpoint pair $(\underline{\text{TE}}^*_S, \overline{\text{TE}}^*_S)$ is computed deterministically; after drawing a sufficient number of such samples, we numerically obtain the shortest range that fully con-

Electronic copy available at: https://ssrn.com/abstract=3336338

tains 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.

A1.6 Point Identification of TE

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

$$
\begin{aligned}
\Pr(M_i(1)=1, M_i(0)=1, X_i=x) &= \Pr(M_i(1)=1, M_i(0)=1 | D_i=0, X_i=x) \\
&= \Pr(M_i=1 | D_i=0, X_i=x) \\
\Pr(M_i(1)=1, M_i(0)=0, X_i=x) &= \Pr(M_i(1)=1, X_i=x) \\
&\quad - \Pr(M_i(1)=1, M_i(0)=1, X_i=x)\Pr(M_i=1 | D_i=1, X_i=x) \\
&\quad - \Pr(M_i=1 | D_i=0, X_i=x) \\
\Pr(M_i(1)=0, M_i(0)=1, X_i=x) & \\
\Pr(M_i(1)=0, M_i(0)=0, X_i=x) &= 1 - \Pr(M_i(1)=0, M_i(0)=1, X_i=x) \\
&\quad - \Pr(M_i(1)=1, M_i(0)=0, X_i=x) \\
&\quad - \Pr(M_i(1)=1, M_i(0)=1, X_i=x) \\
&= 1 - \Pr(M_i=1 | D_i=1, X_i=x)
\end{aligned}
$$

We then reexpress the TE in terms of strata-specific mean potential outcomes and simplify.

Electronic copy available at: https://ssrn.com/abstract=3336338

$$
\begin{aligned}
\text{TE} = {} & \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(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] \\
& \quad \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] \\
& \quad \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] \\
& \quad \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(1, M_i(1))|D_i = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
& \quad \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] \\
& \quad \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(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
& \quad \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(1, M_i(1))|D_i = 0, M_i(1) = 0, M_i(0) = 0, X_i = x] \\
& \quad \Pr(M_i(1) = 0, 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] \\
& \quad \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] \\
& \quad \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] \\
& \quad \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] \\
& \quad \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 = 0, M_i(1) = 1, M_i(0) = 1, X_i = x] \\
& \quad \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(0, M_i(0))|D_i = 0, M_i(1) = 1, M_i(0) = 0, X_i = x] \\
& \quad \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 = 0, M_i(1) = 0, M_i(0) = 1, X_i = x] \\
& \quad \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) = 0, M_i(0) = 0, X_i = x]
\end{aligned}
$$

Electronic copy available at: https://ssrn.com/abstract=3336338

$$\Pr(M_i(1) = 0, M_i(0) = 0 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)$$

under mandatory reporting

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

under mediator monotonicity

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

Electronic copy available at: https://ssrn.com/abstract=3336338

$$\Pr(M_i(1) = 1, M_i(0) = 1 | D_i = 0, X_i = x) \Pr(D_i = 0, X_i = x)$$

under treatment ignorability

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

which can be recovered from observed data

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

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

A1.7 Outcome tests can identify the lower bound of the share of racially discriminatory stops

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 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, even if the rate of criminal activity differs 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 $\mathbb{E}[Y_i | M_i = 1, D_i = 0] > \mathbb{E}[Y_i | M_i = 1, D_i = 1]$, 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.

Electronic copy available at: https://ssrn.com/abstract=3336338

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.[1] In this case, the empirical difference in hit rates can be rewritten in the potential outcomes framework as

$$\begin{aligned}
&\mathbb{E}[Y_i(0,1)|M_i(1)=1, M_i(0)=1, D_i=0] \\
&\quad -\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1, D_i=1]\Pr(M_i(1)=1, M_i(0)=1|D_i=1) \\
&\quad -\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, D_i=1]\Pr(M_i(1)=1, M_i(0)=0|D_i=1) \;>\; 0
\end{aligned} \tag{5}$$

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 inequality may hold simply because white civilians in always-stop encounters engage in more criminal conduct than minority suspects. In other words, the analyst might observe $\mathbb{E}[Y_i|M_i=1, D_i=1] < \mathbb{E}[Y_i|M_i=1, D_i=0]$ 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 systematically 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 are equally criminal, 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 treatment 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

$$\begin{aligned}
&\mathbb{E}[Y_i|M_i=1, D_i=0] - \mathbb{E}[Y_i|M_i=1, D_i=1] \\
&\quad = \Big(\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=1, D_i=1] \\
&\qquad -\mathbb{E}[Y_i(1,1)|M_i(1)=1, M_i(0)=0, D_i=1]\Big)\Pr(M_i(1)=1, M_i(0)=0|D_i=1).
\end{aligned} \tag{6}$$

[1] If 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.

Electronic copy available at: https://ssrn.com/abstract=3336338

This formulation makes clear that the observed evidence gap is due to the difference in hit 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 6 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 6 also suggests 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. Specifically, it can be shown that $Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1)$ is *at least* as large as $(\mathbb{E}[Y_i|M_i = 1, D_i = 0] - \mathbb{E}[Y_i|M_i = 1, D_i = 1]) \,/\, \mathbb{E}[Y_i|M_i = 1, D_i = 1]$ (see Supplementary Materials A1.7 for proof). 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).

We begin with Equation 6:

$$\begin{aligned}
&\mathbb{E}[Y_i|M_i = 1, D_i = 0] - \mathbb{E}[Y_i|M_i = 1, D_i = 1] \\
&\quad = \big(\mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 1, D_i = 1] \\
&\qquad -\mathbb{E}[Y_i(1,1)|M_i(1) = 1, M_i(0) = 0, D_i = 1]\big) \Pr(M_i(1) = 1, M_i(0) = 0|D_i = 1)
\end{aligned}$$

Given the aforementioned assumptions, this implies:

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

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* $(\mathbb{E}[Y_i|M_i = 1, D_i = 0] - \mathbb{E}[Y_i|M_i = 1, D_i = 1])/\mathbb{E}[Y_i|M_i = 1, D_i = 1]$ 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.

Electronic copy available at: https://ssrn.com/abstract=3336338

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).

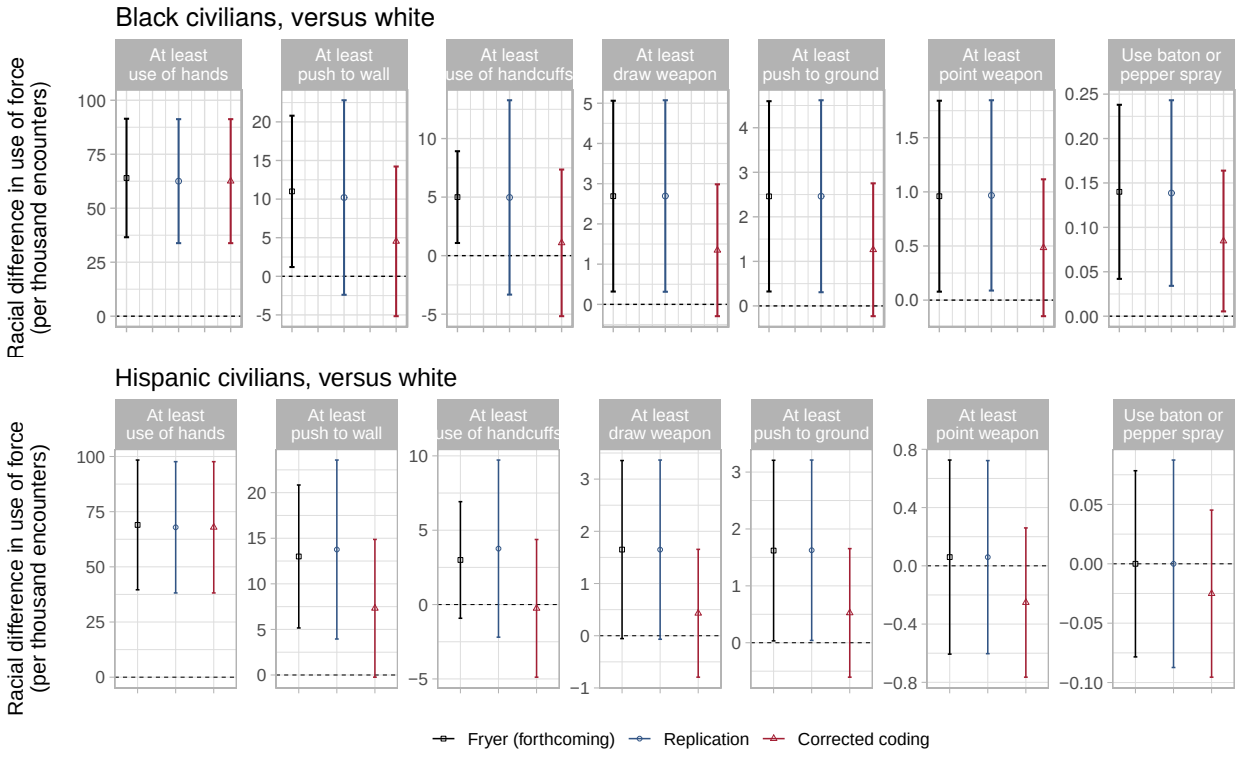
Electronic copy available at: https://ssrn.com/abstract=3336338

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.

	N (published)	*N* (replicated)	*N* (corrected coding)
any force	4,927,962	4,982,090	4,982,090
wall	4,152,918	4,246,329	4,982,090
handcuffs	4,017,783	4,123,523	4,982,090
draw weapon	3,957,687	3,966,879	4,982,090
push to ground	3,950,324	3,959,530	4,982,090
point weapon	3,918,741	3,927,956	4,982,090
pepper spray/baton	3,900,977	3,910,210	4,982,090

Electronic copy available at: https://ssrn.com/abstract=3336338

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.

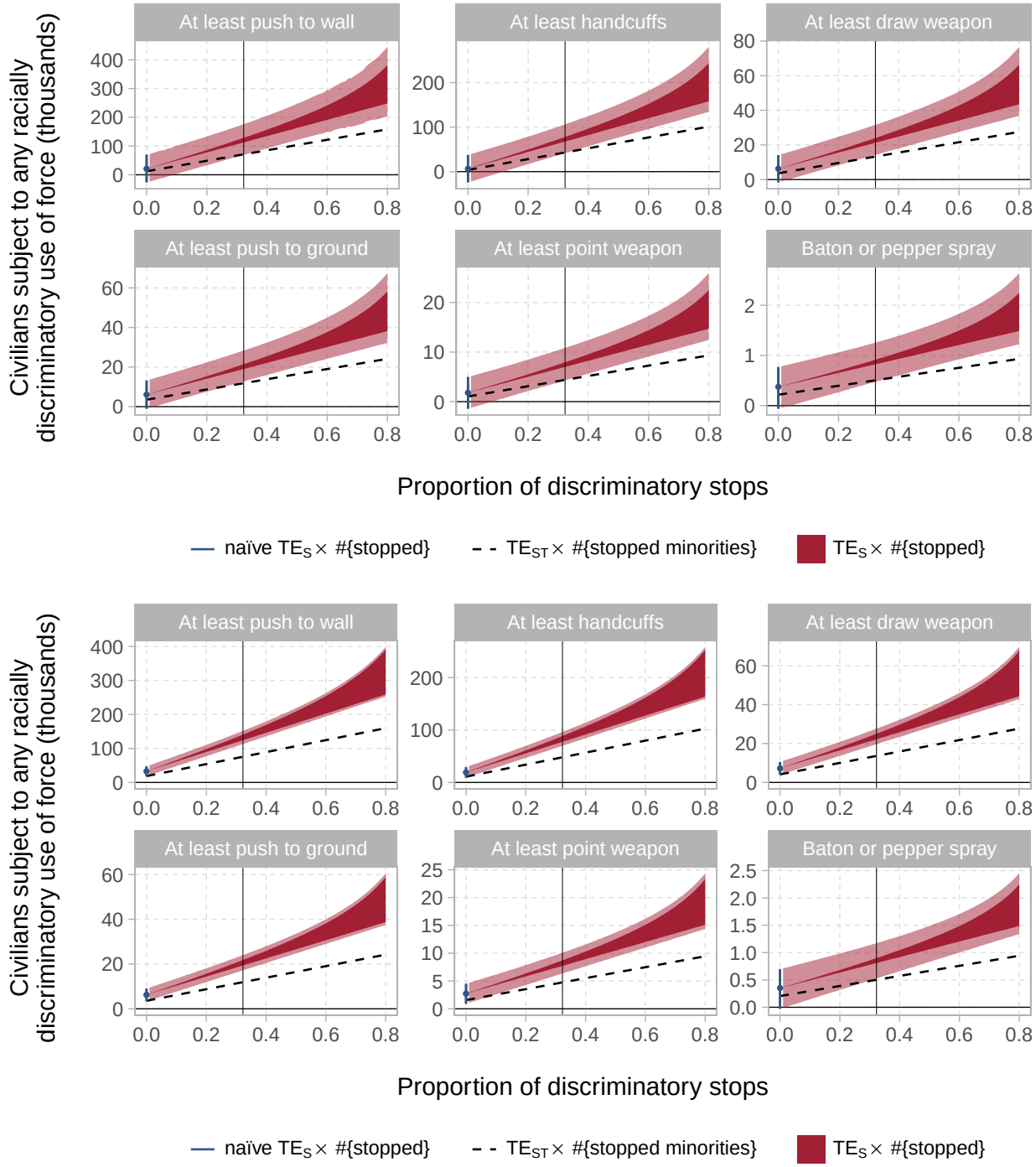


Electronic copy available at: https://ssrn.com/abstract=3336338

B.2 Varying levels of force

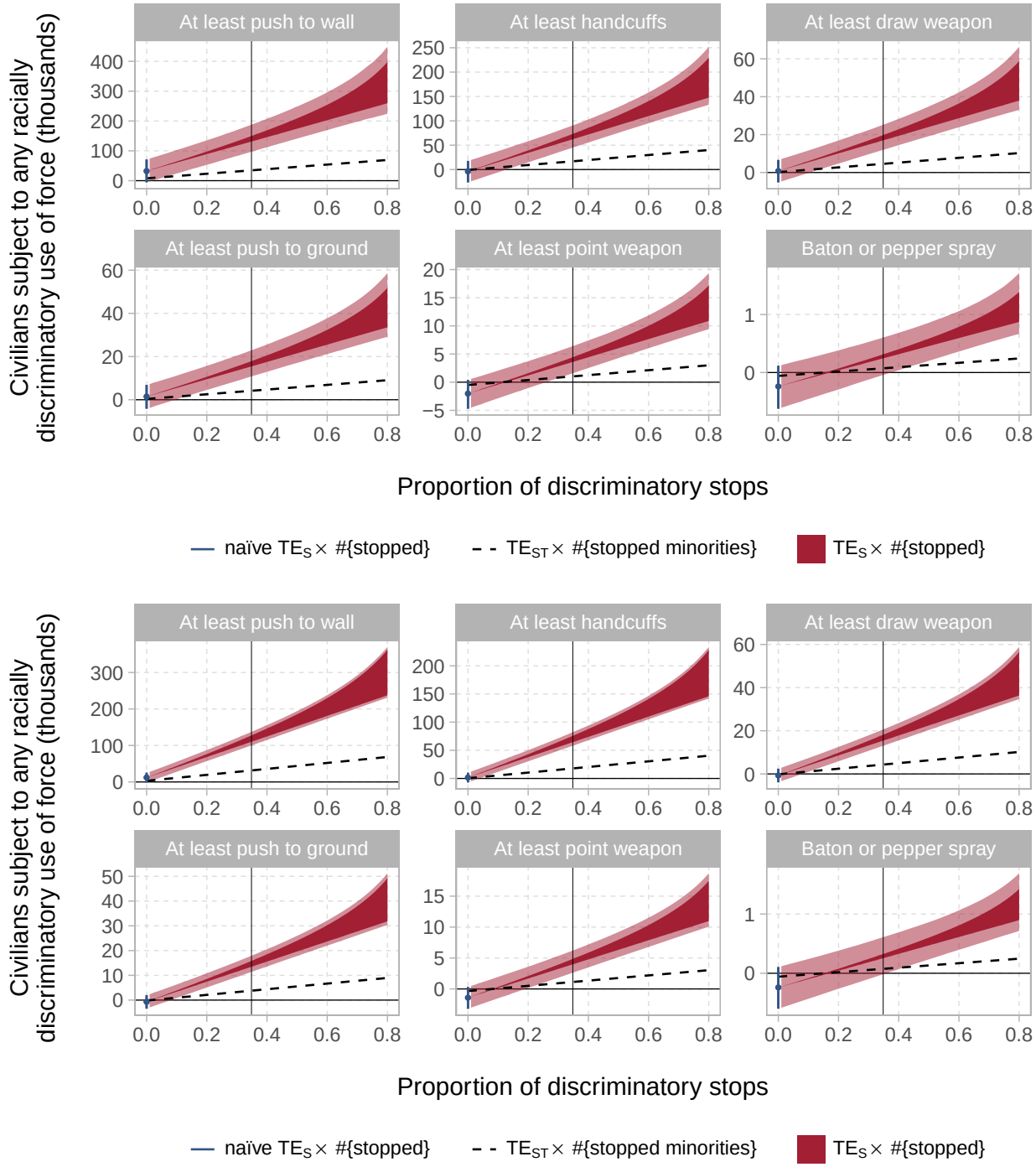
Electronic copy available at: https://ssrn.com/abstract=3336338

Figure B2: **Corrected TE_S and TE_{ST} 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.



Electronic copy available at: https://ssrn.com/abstract=3336338

Figure B3: **Corrected TE_S and TE_{ST} 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.

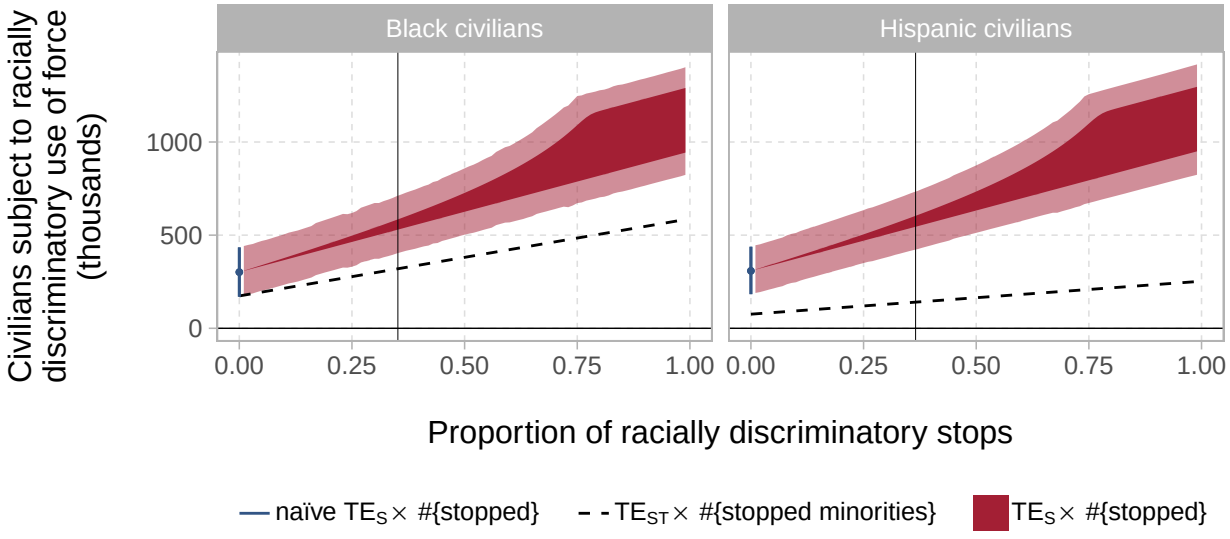


Electronic copy available at: https://ssrn.com/abstract=3336338

B.3 Excluding drug stops

Electronic copy available at: https://ssrn.com/abstract=3336338

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.

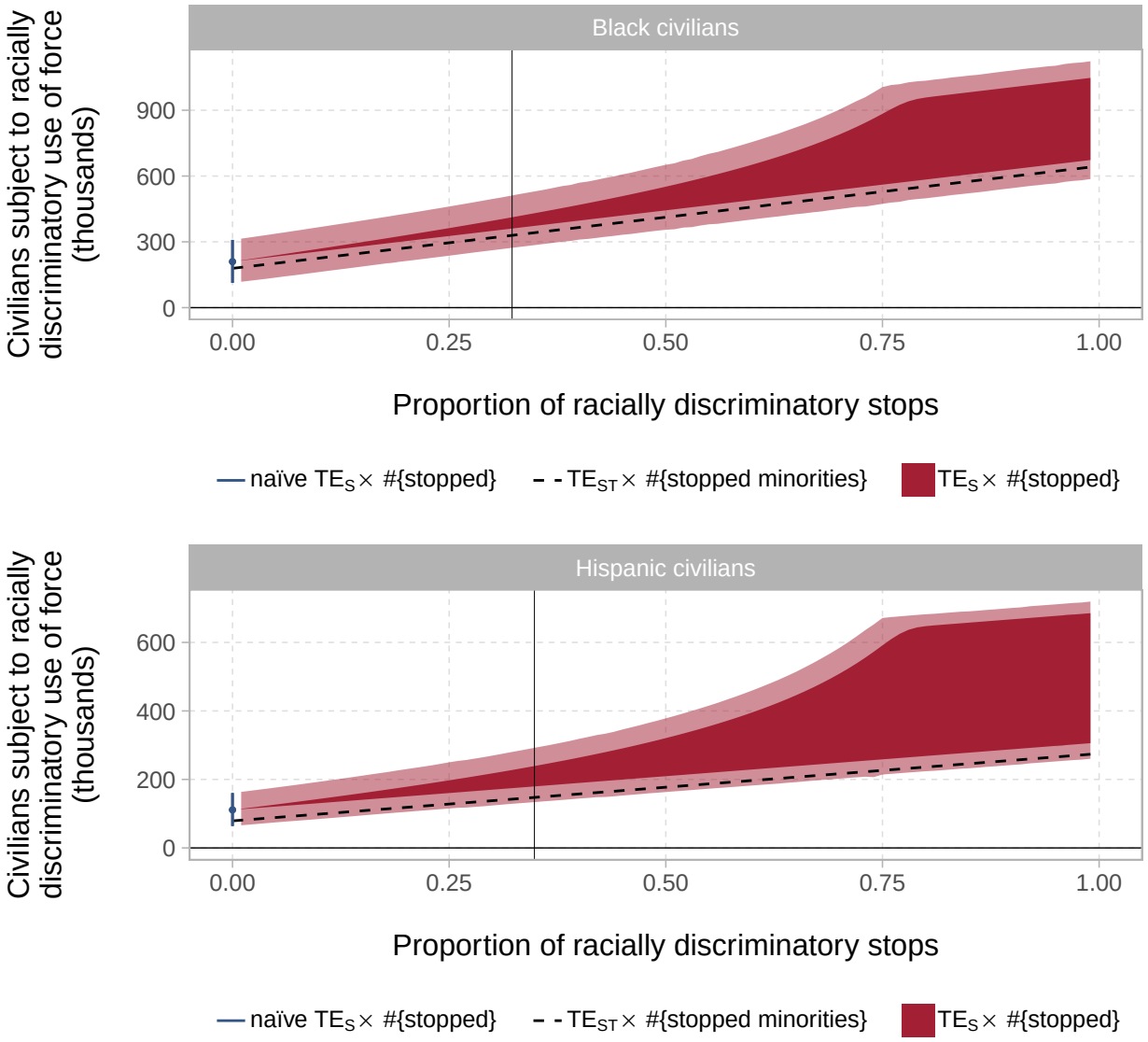


Electronic copy available at: https://ssrn.com/abstract=3336338

B.4 Analysis of two races at a time

Electronic copy available at: https://ssrn.com/abstract=3336338

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.



Electronic copy available at: https://ssrn.com/abstract=3336338

References

Ayres, Ian. 2002. "Outcome Tests of Racial Disparities in Police Practices." *Justice Research and Policy* 4(1-2):131–142.

Becker, Gary. 1971. *The economics of discrimination.* University of Chicago Press.

Engel, Robin. 2008. "A Critique of the "Outcome Test" in Racial Profiling Research." *Justice Quarterly* 25(1):1–36.

Fryer, Roland G. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* .

Goel, Sharad, Justin M. Rao and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-And-Frisk Policy." *Annals of Applied Statistics* 10(1):365–394.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data." *Journal of the Americal Statistical Association* 95(449):77–84.

Knowles, J., N. Perisco and P. Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* pp. 203–229.

Knox, Dean, Teppei Yamamoto, Matthew A. Baum and Adam J. Berinsky. 2019. "Design, Identification, and Sensitivity Analysis for Patient Preference Trials." *Journal of the American Statistical Association* .

Ridgeway, Greg and John MacDonald. 2010. *Race, Ethnicity, and Policing: New and Essential Readings.* NYU Press chapter Methods for Assessing Racially Biased Policing.

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

Electronic copy available at: https://ssrn.com/abstract=3336338